МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ
«ХАРКІВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ»

О. С. Мельников

ІНТЕЛЕКТУАЛЬНИЙ АНАЛІЗ ДАНИХ

Навчально-методичний посібник
для студентів другого (магістерського) рівня підготовки спеціальностей
122 – Комп'ютерні науки та 124 – Системний аналіз

Затверджено
редакційно-видавничою
радою університету,
протокол № 3 від 12.10.2023 р.

Харків
НТУ «ХПІ»
2023

УДК 519.23:681.3
М 48

Рецензенти:
О. С. Мазманішвілі, д-р фіз.-мат. наук,
с.н.с. ННЦ «Харківський фізико-технічний інститут»
Л. М. Любчик, д-р техн. наук,
професор кафедри комп'ютерної математики та аналізу даних НТУ «ХПІ»

Мельников О. С.
М 48 Інтелектуальний аналіз даних : навчально-методичний посібник для студентів другого (магістерського) рівня підготовки спеціальностей 122 – Комп'ютерні науки, 124 – Системний аналіз / О. С. Мельников. – Харків : НТУ «ХПІ», 2023. – 196 с.

ISBN 978-617-05-0434-0

Розглянуто методичні основи та математичний апарат інтелектуального аналізу даних. Наведені основні алгоритми вирішення найбільш поширених задач інтелектуального аналізу даних – класифікації, регресії, кластеризації, пошуку асоціативних правил. Всі теми забезпечені прикладами для ілюстрації основних концепцій і алгоритмів, супроводжуються контрольними запитаннями та завданнями для самостійної роботи.

Призначено для студентів магістерського рівня підготовки спеціальностей галузі знань «Інформаційні технології» та інших технічних спеціальностей.

Іл. 63. Табл. 38. Бібліогр. 70 назв.

УДК 519.23:681.3

ISBN 978-617-05-0434-0

ЗМІСТ

ПЕРЕДМОВА

Одним із найбільш помітних тенденцій розвитку сучасного світу протягом останніх десятиліть є експоненційне зростання обсягів інформації. За оцінками, наведеними в [54], в 2018 р. в світі щоденно створювалось 2,5 квінтильйона ($2{,}5\times10^{18}$) байта даних. Очікується, що до 2025 року це число зросте до 463 квінтильйонів. Тільки на Фейсбуці щоденно публікується 350 млн. фотографій і біля 100 млн. годин відеоінформації. Google кожний день виконує 3,5 мільярда пошукових запитів. Можна навести і багато інших приголомшливих цифр.

Зрозуміло, що при такій кількості інформації тільки мізерна її частина має шанс бути побаченою людським оком. Єдина надія знайти щось цікаве і корисне в океані інформації полягає у широкому застосуванні методів автоматизованого пошуку закономірностей, спільно відомих в англомовній літературі під назвою Data Mining (видобуток даних).

Data Mining – це процес виявлення в даних раніше невідомих, нетривіальних і корисних з практичної точки зору закономірностей. Це міждисциплінарна галузь на перетині інформатики та статистики із загальною метою перетворення інформації, що міститься в базах даних, в зрозумілу структуру для подальшого використання. Для визначення цієї галузі у вітчизняній літературі використовується термін «інтелектуальний аналіз даних». На думку автора, цей термін є не дуже вдалим, оскільки із назви випливає протиставлення методів Data Mining традиційним методам аналізу даних. Насправді, ці методи не виключають застосування класичного апарату математичної статистики, а розвивають і доповнюють його. Втім, англійський термін Data Mining теж є неточним, оскільки мова йдеться не стільки про видобуток даних, скільки про їх збагачення – вилучення «корисних речовин» (закономірностей) із «сирої руди» (необроблених даних).

Мета цього посібника полягає в тому, щоб не тільки ознайомити читача з найбільш поширеними методами Data Mining, але й окреслити місце цієї галузі знань серед інших дисциплін, пов’язаних з аналізом даних. Значна увага приділяється аналізу переваг і недоліків окремих методів, виділенню сфери їх застосування при вирішенні практичних задач. Посібник містить багато прикладів для ілюстрації основних концепцій і алгоритмів інтелектуального аналізу даних. Всі алгоритми реалізовані програмно. Наводяться основні кроки їх виконання і обговорюються «підводні камені», які можуть зустрітися при практичній імплементації алгоритмів.

Посібник супроводжується файлами даних, які дають можливість читачеві самостійно розібрати наведені приклади. Матеріал не прив’язаний до використання якихось конкретних мов програмування чи пакетів програм. Насправді, за рідким виключенням, всі алгоритми можуть бути реалізовані навіть у Microsoft Excel і наочність процесу виконання у цьому середовищі має

певні методологічні переваги. Іншими хорошими альтернативами можуть бути Python з пакетом NumPy, система R чи Matlab.

Зміст посібника можна поділити на три частини. В перших трьох главах викладаються методичні основи та математичний апарат інтелектуального аналізу даних. Ці глави призначені для того, щоб нагадати читачеві найважливіші концепції теорії ймовірностей та математичної статистики. Також висвітлюються теми, які зазвичай не входять в стандартну програму цих дисциплін, але є важливими з точки зору викладення подальшого матеріалу (логістичний розподіл, порядкові статистики тощо). У другій частині (глави 4–5) розглядається виявлення залежностей в кількісних даних на базі багатовимірного статистичного аналізу. Третя частина посібника (глави 6–10) присвячена методам дискретного аналізу даних – класифікації, кластеризації та асоціативним правилам.

Глави посібника мають наскрізну нумерацію, від першої до десятої. Розділи мають подвійну нумерацію, яка складається з номера глави та номеру розділу всередині глави. Так само нумеруються таблиці, рисунки та приклади. Кінець кожного прикладу позначається символом «■». Термінологія виділяється курсивом. Для всіх термінів наводиться їх англійський еквівалент. Кінець кожної глави супроводжується контрольними запитаннями для перевірки засвоєння студентами матеріалу та завданнями для самостійної роботи. Задачі підвищеної складності відзначені зірочкою.

Файли даних, використані в прикладах, можна завантажити за посиланням https://tinyurl.com/pv7srnc6.

Цей посібник є результатом багаторічного досвіду викладання дисципліни «Інтелектуальний аналіз даних» для магістрів спеціальностей «Комп'ютерні науки», «Системний аналіз» та «Економічна кібернетика» в НТУ «Харківський політехнічний інститут». Вона орієнтована на студентів магістерського рівня підготовки галузі знань 12 – Інформаційні технології, але також може бути корисною студентам спеціальностей «Статистика», «Прикладна математика», «Економіка» та всім зацікавленим у сучасних методах аналізу даних.

Автор висловлює глибоку вдячність співробітникам кафедри системного аналізу та інформаційно–аналітичних технологій за цінні коментарі і створення сприятливих умов для роботи над посібником.

1. ОСНОВИ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ ДАНИХ

1.1 Поняття інтелектуального аналізу даних

Інтелектуальний аналіз даних (надалі ІАД) – це мультидисциплінарна область, пов'язана з пошуком в даних прихованих закономірностей (*шаблонів даних* – англ. *data patterns*). Поняття інтелектуального аналізу даних приблизно відповідає поширеному в англомовній літературі терміну *Data Mining*, який перекладається також як видобуток даних, глибинний аналіз даних, вилучення знань, розкопка знань в базах даних [5, 7, 10].

Поняття ІАД в сучасному трактуванні набуло популярності в 1990–х роках у зв'язку з розповсюдженням систем автоматизованого обліку інформації. До цього часу задачі обробки і аналізу даних вважалися сферою розгляду прикладної статистики.

Традиційна методологія статистичних досліджень передбачає використання даних як засобу для перевірки гіпотез. Типове статистичне дослідження містить наступні кроки:

1) висувається гіпотеза – припущення про наявність певного зв'язку між досліджуваними явищами (процесами, показниками);

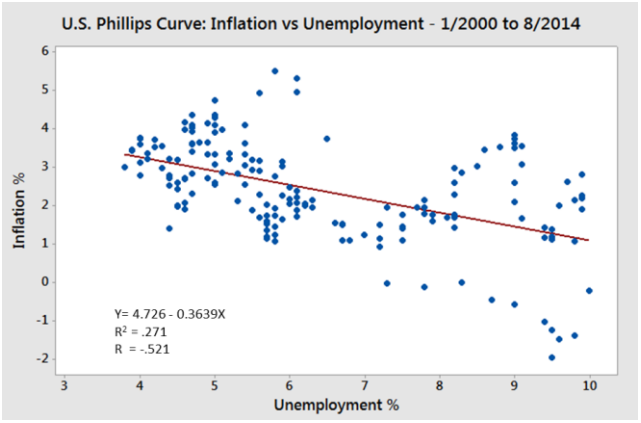
2) збираються дані, які могли б підкріпити або спростувати цю гіпотезу;

3) гіпотеза перевіряється на зібраних даних із використанням статистичних методів, вибір яких принаймні частково зумовлений структурою гіпотези.

Така структура є типовою для наукового підходу взагалі, хоча статистичні дослідження не обов'язково пов'язані з перевіркою наукових теорій, а можуть мати інші цілі (наприклад, оцінку шансів різних кандидатів на перемогу у виборах). У будь-якому разі, статистичні дослідження є зумовленими і керованими чітко сформульованою теорією (theory driven).

В якості прикладу можна привести так звану криву Філліпса – гіпотезу про наявність негативного зв'язку між рівнем безробіття та рівнем інфляції, висунуту новозеландським економістом Вільямом Філліпсом в 1958 р. [41]. Філліпс аргументував, що низький рівень безробіття змушує підприємців підвищувати заробітну плату, щоб залучити працівників. Відповідно, зростають доходи працівників, що підвищує попит на більшість товарів. Обидва фактори сприяють зростанню цін, тобто інфляції. Для перевірки цих міркувань Філліпс використав дані про рівень безробіття і середню заробітну плату в Великобританії за період 1861–1957 рр. і надійшов до висновку, що емпіричні дані непогано узгоджуються з його теорією (рис. 1.1*а*). У подальшому ця теорія неодноразово перевірялась на інших даних з неоднозначними висновками ([Д5], рис. 1.1*б*), і, згідно з сучасними уявленнями, зв'язок між безробіттям і інфляцією є менш тісним, ніж вважалося Філліпсом.

Зауважимо також, що бази даних, які використовувались в наведених вище дослідженнях, мали порівняно невеликий розмір.



а *б*

Рис. 1.1. Оцінка кривої Філліпса на даних статистики праці:
а – Великобританії, 1861–1913 [41]; *б* – США, 2000–2014 [Д5]

Наприкінці минулого століття відбулось поступове переосмислення цієї класичної методології внаслідок повсюдного запровадження автоматизованих систем обліку інформації і радикальним зниженням вартості її зберігання. Внаслідок цього підприємства та організації почали накопичувати величезні обсяги інформації, і закономірно постало питання про їх використання для пошуку цікавих та корисних з практичної точки зору закономірностей.

Наприклад, типова мережа супермаркетів зберігає всі касові чеки, які містять інформацію про куплені товари, їх ціни, дату та час купівлі, розташування магазину, форму оплати тощо. Якщо при оплаті покупець користується дисконтною карткою мережі, то ці дані можуть бути автоматично пов’язані з особистістю покупця та його демографічними характеристиками, вказаними ним при отриманні картки. Це надає можливість вивчати особливості поведінки покупців:

– на загальномережевому рівні – які товари купляються одночасно, як залежить попит від часу доби;

– на рівні групи споживачів – які товари купляють споживачі певної вікової категорії, як відрізняється попит на товари в містах та у сільській місцевості;

– на індивідуальному рівні – які товари купляє конкретний споживач.

Виявлені закономірності можуть бути використані для покращення діяльності мережі – наприклад, при розташуванні товарів на полицях, для персоналізації реклами тощо. При цьому теоретичне обґрунтування виявлених закономірностей є питанням другорядної важливості.

Отже, на відміну від класичної методології статистичних досліджень, ІАД відштовхується від наявних даних (data driven) і використає їх для формування гіпотез (як правило, у автоматичному або напівавтоматичному режимі). ІАД можна охарактеризувати як технологію, яка призначена для пошуку у масивах

даних таких закономірностей, які відповідають наступним критеріям [10, 27]:

– *неочевидність*, тобто знайдені закономірності не випливають автоматично із специфіки досліджуваних процесів і не виявляються елементарними методами обробки інформації;

– *об'єктивність*, тобто виявлені закономірності відповідають реальній ситуації, на відміну від експертної думки, яке завжди є суб'єктивною;

– *практична корисність*, тобто висновки мають конкретне значення, якому можна знайти практичне застосування.

Хрестоматійним прикладом такої закономірності, який наводиться майже в усіх посібниках з ІАД, є історія про пиво та пелюшки. Згідно з цією історію, дослідження в однієї з американських мереж супермаркетів виявило, що чоловіки віком 25–35 років, які купували пелюшки, майже завжди одночасно купували пиво. Напевно, це пояснюється тим, що молоді батьки у такий спосіб борються із стресом, але асоціація між пивом та пелюшками точно не є очевидною. Виявлена залежність навела менеджерів мережі до ідеї розташовувати поруч стелажі з цими товарами, що нібито призвело до зростання продажів обох категорій товарів майже на третину[1].

Процедури ІАД виконуються на постійній основі, по мірі накопичення даних. Це вимагає тісної інтеграції алгоритмів ІАД з системами управління базами даних (СУБД), а певна «самостійність» процедур ІАД у генерації гіпотез про наявність зв’язків у досліджуваних даних споріднює ІАД зі штучним інтелектом та машинним навчанням.

Місце ІАД серед інших дисциплін, пов’язаних з обробкою даних, можна зобразити за допомогою діаграми, наведеної на рис. 1.2.

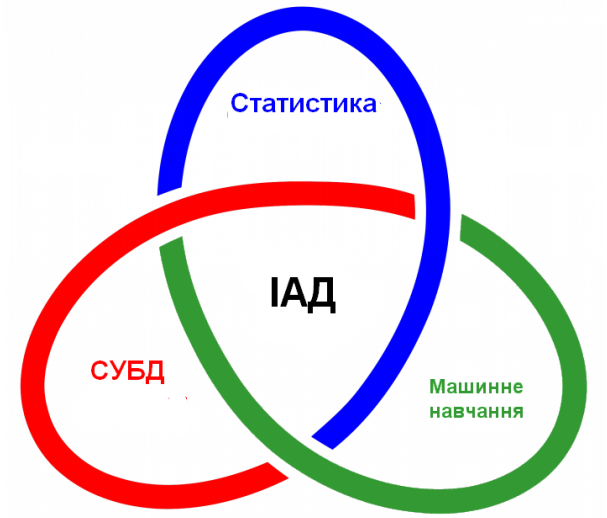


Рис. 1.2. Місце ІАД серед інших дисциплін

[1] Згідно з [24], це дослідження було виконано консалтинговою компанією NCR для мережі Osco ще в 1992 році. Проте, невідомо, чи насправді менеджери Osco відреагували на цю інформацію. Основним результатом дослідження було виявлення біля 5000 товарів, що повільно продаються. Після видалення цих товарів з полиць магазинів споживачі вирішили, що асортимент мережі покращився, бо їм стало простіше знаходити потрібні товари. Цей результат не набув широкої популярності, але теж є досить неочікуваним.

Термін «інтелектуальний аналіз даних» міцно укоренився у вітчизняній літературі, але, на думку автора, є дуже невдалим. Він створює хибне враження, начебто процедури ІАД є більш «інтелектуальними», ніж класичні методи аналізу даних. Насправді, методи ІАД не виключають застосування традиційних методів аналізу даних (кореляційного, дисперсійного, регресійного, факторного аналізу), а розвивають і доповнюють їх. Англійській термін “Data Mining” наводить аналогії з видобутком корисних копалин (mining), коли для отримання кінцевого продукту (знань) треба «перелопатити» тони сировини (даних). Саме в цьому і полягає призначення методів ІАД.

1.2 Дані та їх атрибути

В широкому розумінні дані – це інформація (найчастіше цифрова), подана у формалізованому вигляді, прийнятному для обробки людиною (як правило, із використанням автоматичних засобів). Дані можуть бути представлені у вигляді тексту, таблиць, графіків, фотографій, звуків, карт, відеозаписів, тощо. Вони можуть бути взяті з опублікованих праць та інтернет ресурсів, отримані в результаті вимірювань, експериментів, математичних операцій. При цьому мають бути точно описані умови отримання даних та способи їх розрахунків.

Дані повинні бути представлені у формі, придатній для зберігання, передачі і обробки. Іншими словами, дані – це необроблений матеріал, що надається постачальниками даних і використовується споживачами для аналізу, планування та прийняття рішень.

Дані можуть бути структурованими та неструктурованими. *Структуровані дані* (англ. *structured data*) найчастіше організовані у вигляді двомірної таблиці, де строки відповідають досліджуваним об’єктам, а стовпці – їх атрибутам. Приклад такого набору даних наведено у табл. 1.1. *Атрибут* (англ. *attribute*) – це певна властивість, що характеризує об'єкт, наприклад, вік людини, її стать, колір очей, тощо. Атрибут також називають полем таблиці, виміром, змінною.

Таблиця 1.1 – Приклад набору даних

Атрибути				
Код об’єкту	**Вік**	**Сімейний стан**	**Місячний дохід, грн.**	**Стать**
1	18	Неодружений	5000	Ч
2	32	Одружений	9000	Ч
3	24	Неодружений	3600	Ч
4	22	Неодружений	5400	Ч
5	46	Розведений	25000	Ж
6	39	Одружений	12000	Ж
7	54	Розведений	6400	Ч
8	40	Одружений	18000	Ж

Неструктуровані дані (англ. *unstructured data*) не мають чітко визначеної організації. До цієї категорії належать статті в ЗМІ, листи електронної пошти, повідомлення в соціальних мережах, фотографії в Instagram та відео на Youtube.

Як правило, значно простіше працювати із структурованими даними. Проте, переважна більшість доступної для аналізу інформації є неструктурованою. Це зумовлює розвиток спеціальних методів ІАД (text mining, web mining тощо).

Вимірювання (англ. *measurement*) – це процес присвоєння кодів або числових значень характеристикам досліджуваних об'єктів згідно із заздалегідь визначеними правилами. *Шкала вимірювань* (англ. *measurement scale*) – це правило, згідно з яким атрибутам присвоюються коди або числові значення. Більшість програмних інструментів ІАД при імпорті даних з інших джерел пропонують обрати тип шкали і/або тип даних для кожної змінної.

Змінні можуть бути якісними та кількісними. *Якісні*, або *категоріальні* дані (англ. *qualitative* або *categorical data*) – це інформація про наявність у досліджуваного об’єкта певних ознак, які неможливо виміряти, виразити в цифрах або порахувати. Прикладами якісних даних можуть бути колір, смак, стать, місце проживання, карткова масть тощо. Якісні дані можуть задаватися числами або символами. В більшості мов програмування для представлення такої інформації зручно використовувати перечислювальний тип даних (enumeration, enum). Наприклад, в C для представлення географічних напрямів можна ввести змінну `enum direction {North, East, South, West}`.

Також якісні дані можуть виступати як ідентифікатор об’єкту (номер паспорту, поштовий індекс, адрес електронної пошти).

Кількісні дані (англ. *quantitative data*), в свою чергу, можуть бути дискретними і безперервними.

Дискретні дані (англ. *discrete data*) є значеннями ознаки, загальна кількість яких може бути підрахована за допомогою натуральних чисел. Прикладом дискретних даних може бути кількість товарів на складі або номер курсу студента ВНЗ. В мовах програмування таким ознакам відповідає цілочисельний тип даних (integer, int) зі знаком +/– чи без (signed, unsigned).

Безперервні дані (англ. *continuous data*) – це дані, значення яких можуть брати яке завгодно значення в деякому інтервалі. Вимірювання безперервних даних передбачає велику точність. Прикладами безперервних даних можуть бути температура, висота, вага. В мовах програмування для представлення такої інформації використовуються типи даних з рухомою комою (real, float).

Стенлі Стівенс в 1946 р. запропонував класифікацію шкал вимірювання, яка згодом стала загальноприйнятою [51]. Згідно з нею виділяється чотири типи шкал, а саме: номінальна, порядкова, інтервалів, та відношень. Відмінності між ними зведені у таблицю 1.2.

Таблиця 1.2 – Типи шкал

Тип шкали	Характеристики	Приклади	Операції
Номінальна (найменувань)	Об'єкти класифіковані, класам присвоєні словесні найменування або умовні номери–коди. Те, що номер одного класу більше або менше іншого, свідчить лише про те, що вони розрізняються.	Національність, колір очей, номер гравця футбольної команди, стать, діагноз, автомобільний номер	=, ≠
Порядкова	Об'єкти класифіковані, а класи позначені номерами (закодовані). Значення чисел, що надаються класам, відображають ступінь прояву певних властивостей об’єктів, що належать до цих класів. На цій підставі класи можна порівнювати між собою.	Місце у рейтингу академічної успішності, місце на змаганнях, сорт товару, кількість «зірок» у готелю	=, ≠, >, <
Інтервалів	Існує одиниця виміру, за допомогою якої класи можна не тільки впорядкувати, але й обчислювати різниці між ними. Рівні різниці відповідають рівним відмінностям у вимірюваних властивостях. Нульова точка є довільною (умовною) і не вказує на відсутність властивості.	Календарний час, шкали температур за Цельсієм і Фаренгейтом.	=, ≠, >, <, +, –
Відношень	Числа, які надані класам, мають всі властивості шкали інтервалів. Крім цього, на шкалі існує абсолютний нуль, який відповідає повній відсутності вимірюваної властивості. Відношення чисел відбивають кількісні відносини між вимірюваними властивостями.	Зріст, вага, час, ціна, температура за Кельвіном (існує абсолютний нуль).	=, ≠, >, <, +, –, /

Номінальна шкала (англ. *nominal scale*) – це шкала, яка містить тільки категорії; дані в ній не мають природного порядку, з ними не можуть бути зроблені ніякі арифметичні дії. Номінальна шкала складається з категорій для класифікації та сортування об'єктів або спостережень за певною ознакою. Приклади такої шкали: професія, місто проживання, сімейний стан. Для цієї шкали можна застосовувати лише операції дорівнює/не дорівнює.

Важливим окремим випадком номінальної шкали є *дихотомічна* або *бінарна* шкала (англ. *binary scale*), яка містить тільки дві категорії. Приклад такої шкали: стать (чоловіча/жіноча), вид підсумкового контролю дисципліни (залік/іспит).

Номінальні змінні також зазвичай використовуються в якості ідентифікатору об’єкта (номер паспорту, телефонний номер тощо).

Порядкова шкала (англ. *ordinal scale*) – це така шкала, в якій об’єктам можна присвоїти числа для позначення їх відносної позиції, але не величини відмінностей між ними.

Порядкова шкала дає можливість ранжувати значення змінних. Вимірювання в порядковій шкалі містять інформацію про послідовність

величин, але не дозволяють сказати, наскільки одна з величин більше або менше іншої. Приклади такої шкали: місце, яке команда отримала на змаганнях; номер країни в світовому рейтингу за певною ознакою, тощо. При цьому невідомо, наскільки одна країна краще іншої; відомий лише її номер в рейтингу. Для порядкової шкали можуть бути застосовні операції порівняння: рівність (=), нерівність (≠), більше (>), менше (<).

Шкала інтервалів (англ. *interval scale*) – це шкала, різниці між значеннями якої можуть бути обчислені, проте їхні відношення не мають сенсу. Ця шкала дозволяє знаходити різницю між двома величинами, має властивості номінальної і порядкової шкал, а також дозволяє визначити зміну ознаки кількісно. Прикладом такої шкали може бути шкала температури за Цельсієм. Якщо температура повітря вранці становить –2ºС, а ввечері +2ºС, то можна сказати, що протягом дня температура зросла на 4 градуси, але не можна сказати, у скільки разів або на скільки відсотків підвищилась температура.

Номінальна і порядкова шкали є дискретними, а шкала інтервалів – безперервною. Шкала інтервалів дозволяє здійснювати точні вимірювання ознаки і виконувати арифметичні операції додавання (+) та віднімання (–).

Шкала відношень (англ. *ratio scale*) – це шкала, в якій є певна точка відліку і можливо знайти відношення між будь-якими двома значеннями. Прикладами такої шкали можуть бути вага в кілограмах, об’єм в літрах або ціна в гривнях. Так, банка об’ємом 0,5л містить на 50% більше пива, ніж банка обсягом 0,33л; внаслідок інфляції ціна кави може протягом року зрости на 20%, тощо.

Для шкали відношень можуть застосовуватись ті ж самі операції, що і для шкали інтервалів, і, крім того, операція ділення (/).

Приклад використання різних шкал для вимірювання властивостей різних об’єктів наведено в табл.1.3.

Таблиця 1.3 – Приклад використання шкал вимірювання

№ студента	Рік народження	Код спеціальності	Результати тестування, у балах	Оцінка ECTS
1	2004	122	87	B
2	2002	124	92	A
3	2005	186	74	D
Шкала:				
номінальна	інтервалів	номінальна	відношень	порядкова

Як свідчить приклад в таблиці 1.3, різниця між шкалами вимірювання не є абсолютною. Так, підсумкова оцінка за дисципліну часто виводиться на базі результатів тестування шляхом перетворення кількісних величин в якісні за певними правилами.

Іноді використання категоріальних змінних замість кількісних надає більш

об’єктивну інформацію про досліджувані явища та процеси. Наприклад, добре відомо, що під час соціологічних опитувань люди схильні занижувати свій рівень доходів. Це пояснюється як невпевненістю респондентів щодо цілей опитування, так і тим, що люди часто зневажують вторинні джерела доходів – оренду майна, відсотки за депозитами, матеріальну допомогу від батьків тощо. В той же час більшість людей адекватно оцінюють свій матеріальний стан у широко визначених грошових рамках. Отже, використання категоріальної змінної для оцінки рівня доходів може надавати точнішу, хоча й спрощену, картину розподілу доходів населення.

Аналогічно, категоріальні змінні іноді можуть бути змістовно перетворені в кількісні. Наприклад, для більшості професій (номінальна шкала) в Україні встановлюється тарифний розряд (порядкова шкала), на базі якого в сукупності з мінімальним розміром оплати праці, який визначається Державним бюджетом на відповідний рік, можна визначити базовий рівень заробітної плати працівника (шкала відношень).

1.3 Задачі інтелектуального аналізу даних

Немає єдиної думки щодо того, які задачі слід відносити до ІАД. Проте, зазвичай до сфери ІАД відносять такі задачі [5, 7, 10, 27]:

- класифікація та регресія;
- кластеризація;
- пошук асоціативних правил;
- виявлення аномалій;
- підведення підсумків.

Задачі класифікації та регресії – це сукупність методів дослідження впливу однієї чи декількох незалежних змінних $X_1, X_2, ..., X_m$ на залежну змінну *Y*. Ціль аналізу полягає у побудові функції або алгоритму, які б дозволяли за заданими значеннями $X_1 = x_1, X_2 = x_2, ..., X_m = x_m$ встановити значення залежної змінної $Y = y$. Якщо при цьому залежна змінна може приймати тільки дискретні значення, то говорять про задачі *класифікації* (англ. *classification*); якщо ж залежна змінна є безперервною, то говорять про задачі *регресії* (англ. *regression*). Можливі значення залежної змінної в задачах класифікації зазвичай інтерпретуються як мітки класів, до одного з яких слід віднести досліджуваний об’єкт за його ознаками.

Задача *кластеризації* (англ. *clustering*) є логічним розвитком задач класифікації. На відміну від них, при кластеризації класи об’єктів не визначені заздалегідь, тому ця задача є більш складною. Результатом кластеризації є розбиття об’єктів на декілька груп, які мають між собою суттєві відмінності.

В задачах пошуку *асоціативних правил* (англ. *association rules*) встановлюються закономірності між пов’язаними подіями, зареєстрованими в

наборах даних. Відмінність асоціації від вищезгаданих задач ІАД полягає в тому, що пошук закономірностей здійснюється на основі властивостей не окремого об’єкту, а декількох пов’язаних подій, які відбуваються одночасно.

Задача *виявлення аномалій* (англ. *outlier detection*) полягає у пошуку незвичайних записів у базі даних, які можуть бути цікавими, або виникнути в результаті помилок при введенні даних. В будь-якому випадку, такі записи потребують окремої уваги і подальшого дослідження.

Підведення підсумків (англ. *summarization*) полягає у забезпеченні більш компактного представлення набору даних, включаючи візуалізацію та формування звітів. В результаті *візуалізації* (англ. *visualization*) створюється графічний образ досліджуваних даних, який наочно демонструє виявлені закономірності. Поширені в ІАД методи візуалізації включають часові ряди, діаграми розкиду, гістограми, розфарбування карт тощо. На рис. 1.3 в якості прикладу візуалізації наводиться *хмара слів* (англ. *word cloud*) для ключового слова “Data Mining”. Так називається зображення, що складається зі слів, які використовуються в певному тексті чи темі. Розмір кожного слова вказує на його частоту чи важливість.

Рис. 1.3. Хмара слів для ілюстрації найбільш важливих концепцій Data Mining

Задачі ІАД можуть бути описовими (дескриптивними) і прогностичними.

В результаті рішення *описових* задач (англ. *descriptive analysis*) аналітик отримує шаблони, які компактно описують дані і легко піддаються інтерпретації. Знайдені закономірності можуть бути корисними для виявлення недоліків і вузьких місць досліджуваних процесів, вдосконалення управління ними. Вирішення описових задач потрібно також для порівняння двох або більшої кількості наборів даних з метою пошуку спільних рис і відмінностей. *Прогностичні задачі* (англ. *predictive analytics, forecasting*) ґрунтуються на аналізі даних з метою створення моделі для передбачення тенденцій розвитку досліджуваних процесів. На основі виявлених закономірностей в історичних даних оцінюються відсутні або ж майбутні значення цільових показників.

Прогностичний аналіз є майже синонімом для машинного навчання, бо метою обох галузей є вивчення закономірностей для генерації прогнозів щодо майбутніх або інших подій з невідомими результатами. Розрізняють задачі безумовного та умовного прогнозування. В задачах *безумовного* прогнозування (англ. *unconditional forecasting*) слід оцінити найбільш вірогідний розвиток подій. В задачах *умовного*, або *сценарного* прогнозування (англ. *conditional* або *scenario forecasting*) визначаються наслідки того, що відбудеться певна подія.

Контрольні запитання

1. В чому полягають основні задачі ІАД?
2. В чому пролягають відмінності між ІАД та прикладною статистикою?
3. Яким критеріям мають відповідати знайдені за допомогою ІАД закономірності?
4. Наведіть приклади структурованих та неструктурованих даних.
5. Наведіть приклади якісних та кількісних даних.
6. Як відрізняються кількісні дані, які використовуються в ІАД?
7. Що таке шкала вимірювання?
8. Охарактеризуйте чотири шкали вимірювання за Стівенсом.
9. В чому полягає різниця між шкалою інтервалів та шкалою відношень?
10. В чому полягають переваги і недоліки використання категоріальних та кількісних змінних?
11. Які задачі зазвичай відносять до сфери ІАД?
12. Чим відрізняються задачі класифікації від задач регресії?
13. Чим відрізняються задачі класифікації від задач кластеризації?
14. В чому полягають описові та прогностичні задачі ІАД?
15. В чому полягає відмінність між задачами безумовного та умовного прогнозування?

2. ІМОВІРНІСНИЙ ПІДХІД ДО АНАЛІЗУ ДАНИХ

Теорія ймовірностей – це розділ математики, який вивчає випадкові події, явища, процеси, тощо. Теорія ймовірностей є математичним підґрунтям всіх дисциплін, пов'язаних з аналізом даних. Тому для подальшого вивчення дисципліни доцільно нагадати деякі концепції з теорії ймовірностей.

2.1 Випадкові події

Відправною точкою теорії ймовірностей є поняття випадкового експерименту, результати якого неможливо передбачити заздалегідь. Можливі наслідки такого експерименту називаються *випадковими подіями* (англ. *random events*). Вважається, що є можливість повторювати експеримент велику кількість разів. Приклади випадкових експериментів наведені в Табл. 2.1.

Випадкові події зазвичай позначають великими латинськими або грецькими буквами. Подія Ω, яка настає при кожній реалізації експерименту, називається *достовірною* (англ. *sure event*). Подія $\varnothing$, яка не може настати ні при одній реалізації експерименту, називається *неможливою* (англ. *impossible event*).

Таблиця 2.1 – Приклади випадкових експериментів

Випадковий експеримент	**Пов'язані події**
Кидання монети	Випадає цифра, випадає герб
Кидання грального кубика	Випадає 4 очка Випадає парна кількість очок Випадає число очок менше 4
Очікування потягу в метро	Час очікування не перевищує двох хвилин Час очікування складає від однієї до трьох хвилин

Із кожної подією A можна пов'язати подію, яка полягає у тому, що A не настає. Цю подію називають *протилежною* до A і позначають $\bar{A}$ (англ. *complementary event*).

Сумою двох подій A і B (позначається $A+B$ або $A\cup B$) називається така подія, яка полягає в настанні принаймні однієї з цих подій (англ. *union of events*).

Добутком двох подій A і B (позначається AB або $A\cap B$) називається така подія, яка полягає в тому, що обидві події відбуваються одночасно (англ. *intersection of events*).

Дві події A і B називаються *несумісними* (англ. *mutually exclusive*), якщо їх сумісне настання неможливе: $AB=\varnothing$.

Подія A *спричиняє* подію B (B є *наслідком* A), якщо із настання події A випливає настання події B, тобто $A\subset B$ (англ. *A causes B*).

Сукупність подій $A_1,...,A_n$ утворюють *повну групу*, якщо одна і тільки одна із цих подій в результаті експерименту обов'язково настає: $A_1 \cup ... \cup A_n = \Omega,\ A_i \cap A_j = \varnothing,\ i \neq j$ (англ. *jointly exhaustive events*).

Подія ω називається *елементарною* (англ. *elementary*), якщо для довільної події A вона спричиняє або подію A, або $\bar{A}$. Тобто елементарні події є найпростішими наслідками випадкового експерименту.

Множина $\Omega=\{\omega\}$ всіх елементарних подій називається *простором елементарних подій* (англ. *sample space*). Випадкові події розглядаються як підмножини простору елементарних подій.

2.2 Ймовірності випадкових подій

Існує багато інтерпретацій поняття ймовірності [26]. З точки зору інтелектуального аналізу даних найбільш корисною з них є частотна інтерпретація.

Нехай в однакових умовах проводиться серія із n випадкових експериментів, у кожному із яких може настати деяка подія A. Якщо $n(A)$ – число експериментів, у яких подія A настала, то відношення $\nu(A) = n(A)/n$ називається *відносною частотою* настання події A (англ. *relative frequency*).

В багатьох випадках при проведенні різних серій із великої кількості експериментів відносні частоти події для цих серій наближаються до певного числа. Ця закономірність називається властивістю статистичної стійкості відносних частот. Таким чином, з кожною випадковою подією можна пов'язати деяке стале число, яке і вважається ймовірністю випадкової події.

Означення. Число, до якого прямує відносна частота події A при зростанні числа експериментів, називається *ймовірністю* події A і позначається $P(A)$ (англ. *probability*).

На практиці, при великій кількості експериментів, за ймовірність наближено приймають відносну частоту.

Із даного означення випливають такі властивості ймовірності:

1) $0 \le P(A) \le 1$;

2) $P(\Omega) = 1$;

3) $P(\varnothing) = 0$;

4) якщо $A \cap B = \varnothing$, то $P(A+B)=P(A)+P(B)$.

Розрахунок ймовірностей подій зручно проілюструвати за допомогою діаграм Венна (рис. 2.1).

Зокрема, з діаграми негайно випливає формула для визначення ймовірності суми двох подій:

$$P(A+B) = P(A) + P(B) - P(AB). \qquad (2.1)$$

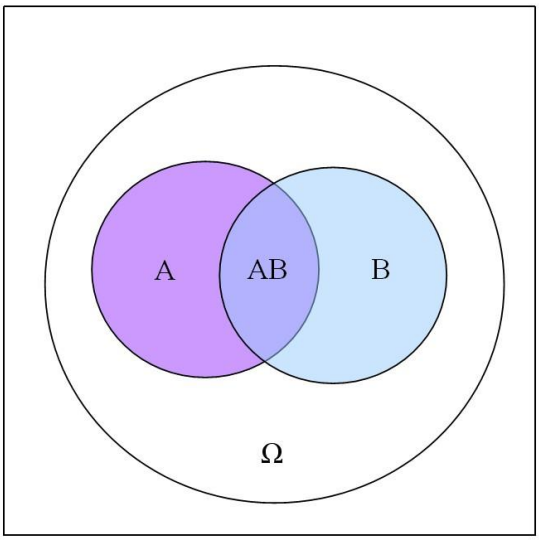


Рис. 2.1. Діаграма Венна для подій *A* і *B*.

2.3 Умовні ймовірності

Якщо при обчисленні ймовірності події *A* не накладається ніяких умов, крім тих, якими визначається випадковий експеримент, то ймовірність *P*(*A*) називають безумовною. Але часто необхідно обчислити ймовірності подій при додатковій умові, що настала деяка інша подія *B*.

Ймовірність події *A*, обчислена за припущенням, що подія *B* уже настала, називається *умовною ймовірністю* події *A* за умови *B* і позначається $P(A|B)$ (англ. *conditional probability of A given B*)

Розглянемо, як знаходяться умовні ймовірності в класичній моделі. Позначимо через n_A, n_B, n_{AB} кількість елементарних подій, що спричиняють відповідно події *A, B, AB*. Тоді:

$$P(A)=\frac{n_A}{n};\ P(B)=\frac{n_B}{n};\ P(AB)=\frac{n_{AB}}{n}. \tag{2.2}$$

Якщо подія *B* вже настала, то змінюються умови експерименту і у новому (умовному) експерименті число можливих наслідків буде рівне n_B – числу елементарних подій, що спричиняють подію *B*, а подію *A* будуть спричиняти тільки ті елементарні події, які спричиняють *AB*. Тому

$$P(A|B)=\frac{n_{AB}}{n_B}=\frac{n_{AB}/n}{n_B/n}=\frac{P(AB)}{P(B)}. \tag{2.3}$$

Аналогічно отримаємо

$$P(B|A)=\frac{n_{AB}}{n_A}=\frac{n_{AB}/n}{n_A/n}=\frac{P(AB)}{P(A)}. \tag{2.4}$$

З останніх двох формул випливають два способи обчислення добутку подій *A* і *B*:

$$P(AB)=P(A|B)P(B)=P(B|A)P(A), \tag{2.5}$$

а також формула для зв’язку між двома умовними ймовірностями:

$$P(A|B)=\frac{P(B|A)P(A)}{P(B)}. \tag{2.6}$$

Формула (2.6) відома як *правило Байєса* (англ. *Bayes rule*).

Приклад 2.1. Нехай при киданні грального кістки стало відомо, що випало більше 2 очок. Знайти ймовірність того, що випало 6 очок.

Нехай x – кількість очок, що випали. Визначимо подію A як «$x=6$», а подію B – як «$x>2$». Тоді за формулою (2.3) шукана ймовірність дорівнює

$$P(x=6|x>2)=\frac{P(x=6,x>2)}{P(x>2)}=\frac{P(x=6)}{P(x>2)}=\frac{1/6}{4/6}=\frac{1}{4}. \ \blacksquare$$

Часто формулу (2.6) надають в дещо іншій формі. Нехай подія A може настати із однією із подій $H_1,...,H_n$, що утворюють повну групу подій. Із попарної несумісності подій $H_1,...,H_n$ випливає, що події $AH_1,...,AH_n$ також несумісні. Тому

$$P(A)=\sum_{i=1}^{n}P(AH_i)=\sum_{i=1}^{n}P(A|H_i)P(H_i). \tag{2.7}$$

Формула (2.7) називається *формулою повної ймовірності* (англ. *law of total probability*). Поєднавши цю формулу з формулою (2.6), отримаємо:

$$P(H_i|A)=\frac{P(A|H_i)P(H_i)}{P(A)}=\frac{P(A|H_i)P(H_i)}{\sum_{i=1}^{n}P(A|H_i)P(H_i)}. \tag{2.8}$$

В аналізі даних події $H_1,...,H_n$ називаються *гіпотезами* (англ. *hypotheses*), а $P(H_i)$ – *апріорною ймовірністю* гіпотези H_i (англ. *prior probability*). Умовна ймовірність $P(H_i|A)$ називається *апостеріорною ймовірністю* гіпотези H_i (англ. *posterior probability*).

Приклад 2.2. Припустімо, що тест на COVID–19 має *чутливість* 99% та *специфічність* 99% (англ. *sensitivity* та *specificity*, відповідно). Тобто, цей тест даватиме 99% правильних позитивних результатів для тих, хто хворий на COVID, і 99% правильних негативних результатів для тих, хто ні. Припустімо, що в даний конкретний час 0.5% від загальної кількості людей хворіють на COVID. Якщо для випадково вибраної особи перевірка виявляється позитивною, то якою є ймовірність, що вона дійсно хвора на COVID?

Позначимо наявність хвороби через *yes*/*no*, а результати тесту через +/–. За формулою (2.8)

$$P(yes|+)=\frac{P(+|yes)P(yes)}{P(+)}=\frac{P(+|yes)P(yes)}{P(+|yes)P(yes)+P(+|no)P(no)};$$

$$P(yes|+)=\frac{0,99\times0,005}{0,99\times0,005+0,01\times0,995}\approx33.2\%.$$

Отже, навіть якщо індивідуальна перевірка дає позитивний результат, то

ймовірніше, що людина не хворіє на COVID.

Цей несподіваний результат виникає тому, що кількість «здорових» є дуже великою у порівнянні з кількістю хворих. Таким чином, кількість хибних позитивних результатів (0,995%) переважує кількість правильних позитивних результатів (0,495%).

На конкретних цифрах, якщо перевірено 1000 осіб, то слід очікувати 995 здорових і 5 хворих на COVID. Для 995 здорових очікується $0{,}01 \times 995 \approx 10$ хибних позитивних результатів. Для п'яти хворих очікується $0{,}99 \times 5 \approx 5$ правильних позитивних результатів. Отже, із 15 позитивних результатів лише 5, тобто близько 33%, є істинними. ■

На базі формули (2.5) вводиться одне із найважливіших понять теорії ймовірностей – поняття незалежності подій.

Події A і B називаються *незалежними* (англ. *independent*), якщо

$$P(AB) = P(A)P(B). \tag{2.9}$$

З формули (2.9) безпосередньо випливає, що для незалежності подій A і B необхідно і достатньо, щоб виконувалась одна із наступних умов:

$$P(A \mid B) = P(A);\ P(B \mid A) = P(B). \tag{2.10}$$

2.4 Випадкові величини та їх характеристики

Випадковою величиною X називається величина, значення якої залежить від певної випадкової події (англ. *random variable*). Прикладом може бути кількість очок, що випала під час кидання гральної кістки, або час роботи електричної лампочки до перегорання. В першому з цих прикладів випадкова величина може прийняти одне із скінченної (або зліченної) множини можливих значень; такі випадкові величини називають *дискретними* (англ. *discrete*). В другому прикладі випадкова величина може прийняти будь-яке значення з певного діапазону числової осі, тобто множина можливих значень не є скінченною. Такі випадкові величини називають *неперервними* (англ. *continuous*).

У подальшому будемо позначати випадкові величини великими літерами, а невипадкові – малими.

Законом розподілу випадкової величини X називається співвідношення, яке встановлює зв'язки між її можливими значеннями та ймовірностями отримання таких значень (англ. *distribution*).

Найбільш повну інформацію про поведінку випадкової величини надає її *функція розподілу* (також використовують терміни інтегральна або кумулятивна функція розподілу; англ. *cumulative distribution function* або скорочено *cdf*). Вона визначається як ймовірність того, що випадкова величина X прийме значення, яке не перевищатиме заданого рівня $x \in (-\infty, +\infty)$:

$$F(x) = P\{X \le x\}. \tag{2.11}$$

Із визначення функції розподілу безпосередньо випливають її властивості:

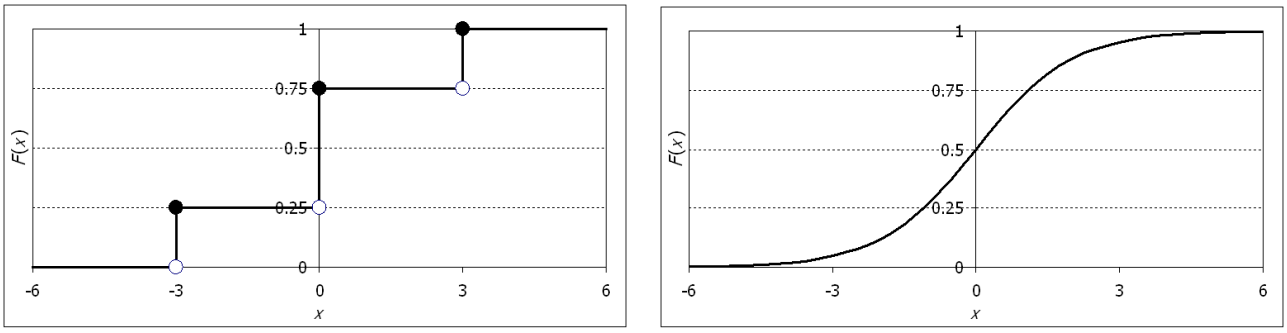
1. $0 \le F(x) \le 1 \; \forall x \in (-\infty, \infty)$.

2. $F(-\infty) = 0$.

3. $F(\infty) = 1$.

4. $F(x)$ монотонно не зменшується за своїм аргументом.

Приклади функцій розподілу для дискретних і неперервних випадкових величин наведені на рис. 2.2.



а *б*

Рис. 2.2. Приклади функцій розподілу випадкових величин:
а – дискретних; *б* – неперервних

Якщо відома функція розподілу, то ймовірність влучання випадкової величини в інтервал $(a,b]$ визначається як

$$P\{a < X \le b\} = F(b) - F(a). \qquad (2.12)$$

Для дискретних випадкових величин більш зручною формою опису часто виявляється її *ряд розподілу* (англ. *probability mass function*). Це таблиця, яка задає перелік можливих значень випадкової величини та ймовірності отримання таких значень:

X	x_1	x_2	…	x_i	…
P	p_1	p_2	…	p_i	…

,

де

$$p_i = f(x_i) = P\{X = x_i\}, \; i = 1,2,... \qquad (2.13)$$

За наявністю ряду розподілу, інтегральну функцію розподілу для дискретних випадкових величин можна отримати як

$$F(x) = \sum_{y \le x} f(y). \qquad (2.14)$$

Для неперервних випадкових величин ряд розподілу задати неможливо, адже вони можуть приймати будь-яке з незліченної множини значень. З цього також випливає, що ймовірність кожного окремого значення неперервної випадкової величини дорівнює нулю:

$$P\{X = x\} = 0. \qquad (2.15)$$

Замість цього, можна розрахувати за формулою (2.12) ймовірність влучання неперервної випадкової величини в маленький інтервал біля x:

$$P\{x \le X \le x+dx\} = F(x+dx) - F(x) = \frac{F(x+dx)-F(x)}{dx}dx. \qquad (2.16)$$

(Зверніть увагу, що внаслідок формули (2.15) $P\{X \le x\} = P\{X < x\}$, отже для неперервних випадкових величин із знаками нерівності можна поводитись досить вільно).

Здійснивши в останній формулі граничний перехід при $dx \to 0$, отримаємо

$$f(x) = \lim_{dx\to 0} \frac{F(x+dx)-F(x)}{dx} = \frac{dF(x)}{dx}. \qquad (2.17)$$

Функція $f(x)$ називається *щільністю розподілу* або *густиною ймовірності* випадкової величини X (англ. *probability density function* або скорочено *pdf*). Величина $f(x)dx$ називається *елементом ймовірності* і приблизно дорівнює ймовірності влучання неперервної випадкової величини X в інтервал $[x, x+dx]$.

Згідно з визначенням, щільність розподілу будь-якої випадкової величини є невід'ємною, $f(x) \ge 0$, і має властивість

$$\int_{-\infty}^{\infty} f(x)dx = 1. \qquad (2.18)$$

Графік щільності розподілу називається *кривою розподілу*.

Ймовірність влучання неперервної випадкової величини X в деякий інтервал $[a,b]$ дорівнює

$$P\{a \le X \le b\} = \int_{a}^{b} f(x)dx. \qquad (2.19)$$

Функція розподілу неперервної випадкової величини X виражається через її щільність розподілу як

$$F(x) = P\{-\infty < X \le x\} = \int_{-\infty}^{x} f(u)du, \qquad (2.20)$$

що, власне, й зумовлює назву «інтегральна функція розподілу».

Функція розподілу надає повну інформацію про поведінку випадкової величини, але для визначення такої функції в реальних умовах потрібен великий обсяг статистичної інформації, яку може бути складно отримати. Більш стисло можна охарактеризувати випадкову величину за допомогою чисельних характеристик, найбільш розповсюдженими з яких є математичне сподівання та дисперсія.

Математичне сподівання випадкової величини X характеризує її середнє значення. Воно визначається як

$$M[X] = \mu_x = \sum_{x:\, f(x)>0} xf(x) \qquad (2.21)$$

для дискретних випадкових величин та

$$M[X] = \mu_x = \int_{-\infty}^{\infty} xf(x)dx \qquad (2.22)$$

для неперервних випадкових величин (англ. *expected value*, позначається $E[x]$).

Аналогічно визначається математичне сподівання деякої функції від випадкової величини $h(X)$:

$$M[h(X)] = \sum_{x:\, f(x)>0} h(x) f(x) \qquad (2.23)$$

для дискретних випадкових величин та

$$M[h(X)] = \int_{-\infty}^{\infty} h(x) f(x)dx \qquad (2.24)$$

для неперервних випадкових величин.

Для лінійної функції $h(X) = a + bX$, де a,b – довільні константи, з формул (2.23)–(2.24) випливає, що

$$M[a + bX] = a + bM[X]. \qquad (2.25)$$

Дисперсія випадкової величини X характеризує її розкид навколо середнього значення. Вона визначається як

$$D[X] = \sigma_x^2 = M[(X - M[X])^2] = M[X^2] - (M[X])^2 \qquad (2.26)$$

(англ. *variance*, позначається $Var[x]$

Якщо підставити $Y = a + bX$ в формулу (2.26), то після спрощення отримаємо наступну важливу властивість дисперсії:

$$D[a + bX] = b^2 D[X]; \qquad (2.27)$$

зокрема, $D[a] = 0$.

Квадратний корінь від дисперсії називається *середньоквадратичним відхиленням* (англ. *standard deviation*) випадкової величини X:

$$\sigma_x = \sqrt{D[X]}. \qquad (2.28)$$

Середньоквадратичне відхилення часто використовується для грубої оцінки діапазону можливих значень випадкової величини. Підстави для такого підходу надає нерівність Чебишева (наводиться у формі з [9]):

$$P\{\mu_x - k\sigma_x \le X \le \mu_x + k\sigma_x\} \ge 1 - \frac{1}{k^2}, \qquad (2.29)$$

де k – довільна позитивна константа. Наприклад, для k=2 отримуємо, що з ймовірністю принаймні 0,75 будь-яка випадкова величина буде знаходитись в межах $\pm 2\sigma$ від свого математичного сподівання. Для деяких окремих розподілів можна отримати значно точніші оцінки.

2.5 Деякі важливі розподіли випадкових величин

Надамо далі стислу характеристику деяких поширених в теорії ймовірностей розподілів , які є важливими для ІАД.

2.5.1 Дискретні розподіли

Розподіл Бернуллі (англ. *Bernoulli distribution*). Це розподіл дискретної випадкової величини X, яка приймає значення 1 з ймовірністю p і значення 0 – з ймовірністю $q = 1 - p$. Часто значення «1» інтерпретується як «успіх» у деякому випробовуванні. Чисельні характеристики випадкової величини із розподілом Бернуллі задаються формулами

$$M[X] = 1 \times p + 0 \times q = p;$$
$$D[X] = M[X^2] - (M[X])^2 = p - p^2 = pq. \qquad (2.30)$$

Біноміальний розподіл (англ. *binomial distribution*) характеризується двома параметрами n та p. Це розподіл кількості успіхів в серії із n незалежних випробувань, результат кожного з яких описується розподілом Бернуллі з ймовірністю успіху p. Скорочено позначається як $X \sim B(n, p)$. Розподіл Бернуллі є окремим випадком біноміального розподілу при $n = 1$.

Ряд розподілу для випадкової величини X задається формулами:

$$p_k = P\{X = k\} = C_n^k p^k q^{n-k}, \qquad (2.31)$$

де $C_n^k = \dfrac{n!}{k!(n-k)!}$ – кількість комбінацій, які наводять до k успіхів в серії з n спроб, $0 \le k \le n$. Приклади кривих біноміального розподілу наведені на рис. 2.3.

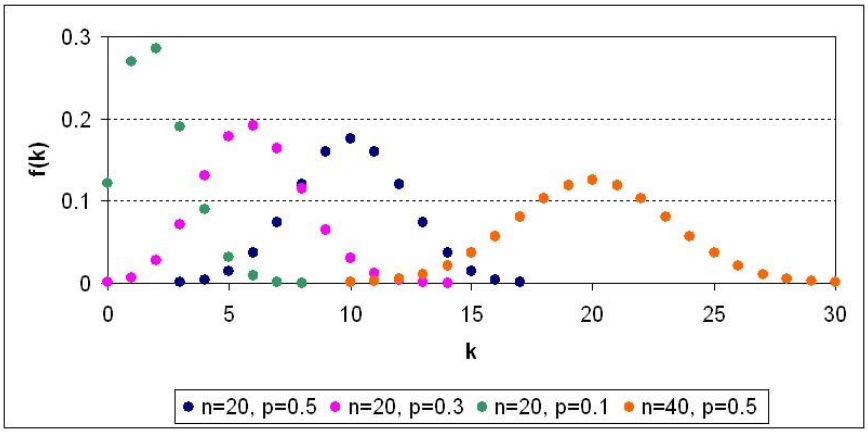


Рис. 2.3. Приклади кривих біноміального розподілу для різних значень n та p

Чисельні характеристики випадкової величини $X \sim B(n, p)$ задаються формулами:

$$M[X] = np;\ D[X] = npq. \qquad (2.32)$$

2.5.2 Неперервні розподіли

Рівномірний розподіл (англ. *uniform distribution*) характеризується верхньою та нижньою межами $[a,b]$. Всі значення у середині цього інтервалу є рівноймовірними. Прикладом може бути час очікування потягу в метро. Скорочено позначається як $X \sim U[a,b]$.

Рівномірний розподіл має постійну щільність:

$$f(x)=\begin{cases}1/(b-a), & x\in[a,b]\\ 0, & x\notin[a,b]\end{cases}. \tag{2.33}$$

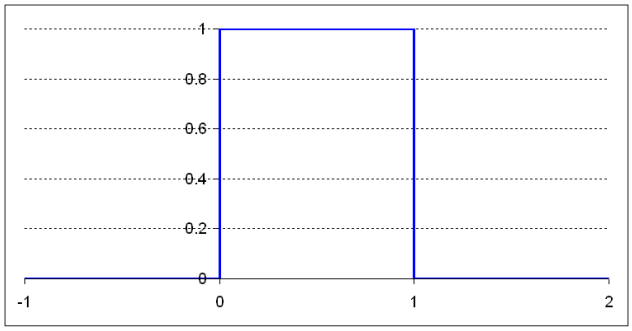
Інтегральна функція рівномірного розподілу є кусково–лінійною:

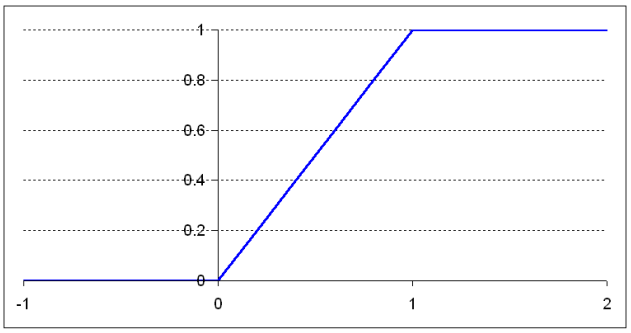
$$F(x)=\begin{cases}0, & x<a\\ \dfrac{x-a}{b-a}, & a\le x\le b.\\ 1, & x>b\end{cases} \tag{2.34}$$

Якщо $a=0$, а $b=1$, тобто $X\sim U[0,1]$, то такий рівномірний розподіл називають стандартним. Має місце твердження: якщо випадкова величина $X\sim U[0,1]$, то випадкова величина $Y=a+(b-a)X\sim U[a,b]$.

Графіки щільності та інтегральної функції стандартного рівномірного розподілу наведені на рис. 2.5. Чисельні характеристики випадкової величини $X\sim U[a,b]$ задаються формулами:

$$M[X]=\frac{a+b}{2};\ D[X]=\frac{(b-a)^2}{12}. \tag{2.35}$$





а *б*

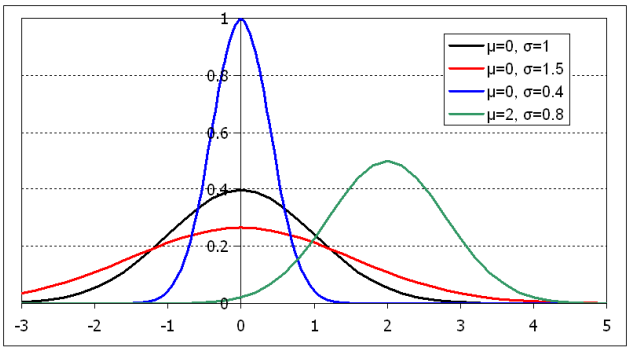
Рис. 2.5. Стандартний рівномірний розподіл:
а – щільність розподілу; *б* – функція розподілу

Рівномірний розподіл грає провідну роль в імітаційному моделюванні, бо є основою для побудови більш складних розподілів.

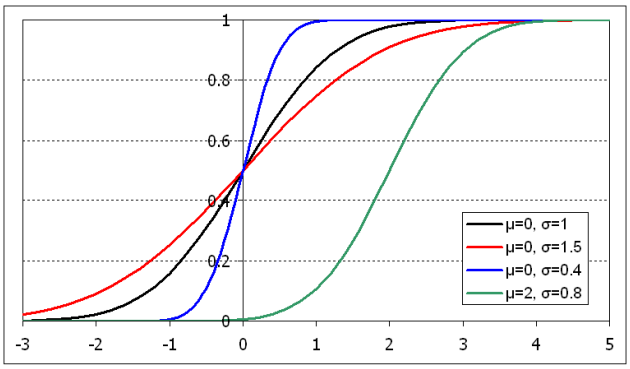
Нормальний розподіл (англ. *normal distribution*). Нормальний розподіл характеризується середнім значенням μ та дисперсією σ^2 і скорочено позначається як $X\sim N(\mu,\sigma^2)$. Щільність нормального розподілу з такими параметрами задається функцією

$$f(x)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right). \tag{2.36}$$

У відповідності із назвами параметрів $M[X]=\mu$, $D[X]=\sigma^2$. Приклади кривих нормального розподілу для різних значень μ та σ^2 наведені на рис. 2.6.



а



б

Рис. 2.6. Криві нормального розподілу для різних значень параметрів μ та σ^2: *а* – щільність розподілу; *б* – функція розподілу

Якщо $\mu = 0$, а $\sigma = 1$, тобто $X \sim N(0,1)$, то такий нормальний розподіл називають стандартним. Має місце твердження: якщо випадкова величина $Z \sim N(0,1)$, то випадкова величина $X = \mu + \sigma Z \sim N(\mu, \sigma^2)$ (і, відповідно, якщо $X \sim N(\mu, \sigma^2)$, то $Z = \dfrac{X - \mu}{\sigma}$ має стандартний нормальний розподіл).

Інтегральна функція стандартного нормального розподілу, яку зазвичай позначають як $\Phi(x)$, не може бути отримана в аналітичній формі. За визначенням вона дорівнює:

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} dt = \frac{1}{2}\left(1 + \operatorname{erf}\left(\frac{x}{\sqrt{2}}\right)\right), \qquad (2.37)$$

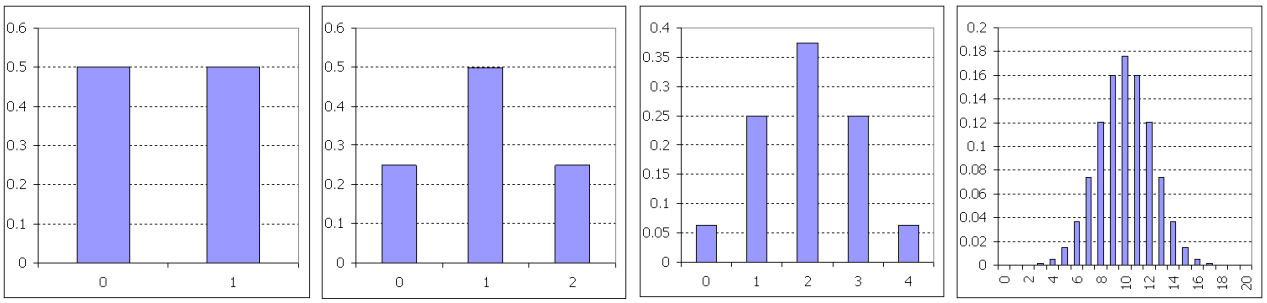
де $\operatorname{erf}(x)$ – так звана функція похибок:

$$\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_{0}^{x} e^{-t^2} dt\,. \qquad (2.38)$$

Реалізація цієї функції є в переважній більшості мов програмування.

Нормальний розподіл займає винятково важливе місце в теорії ймовірностей. Його значення зумовлено *центральною граничною теоремою* (англ. *central limit theorem*), яка (у класичній формі) стверджує, що сума незалежних однаково розподілених випадкових величин при достатньо великої їх кількості збігається до нормального розподілу. В якості прикладу на рис. 2.7 ілюструється збіжність біноміального розподілу до нормального із збільшенням кількості експериментів.

Інші варіанти центральної граничної теореми встановлюють, що (за певних умов) збіжність відбуватиметься, навіть якщо доданки матимуть різний розподіл та/або є залежними [8, 25]. Це створює підґрунтя для використання нормального розподілу в тих випадках, коли досліджуваний об’єкт зазнає впливу великої кількості окремих випадкових факторів.



а *б* *в* *г*

Рис. 2.7. Розподіл кількості гербів в серії із n підкидань монети:
а – $n = 1$; *б* – $n = 2$; *в* – $n = 4$; *г* – $n = 20$

Нормальний розподіл має низку важливих властивостей.

1. Він симетричний відносно точки $x = \mu$, яка одночасно є модою, медіаною і математичним сподіванням розподілу.

2. Суми та різниці нормально розподілених випадкових величин також матимуть нормальний розподіл.

3. Правило трьох сигм та його варіанти:

$$P\{\mu - 3\sigma \le X \le \mu + 3\sigma\} \approx 0{,}997, \qquad (2.39)$$

тобто в 997 випадках із 1000 нормально розподілена випадкова величина буде знаходитись в межах $\pm 3\sigma$ від свого математичного сподівання.

В математичній статистиці зазвичай використовується рівень достовірності 0,95. Для такого рівня впевненості можна встановити більш точні межі:

$$P\{\mu - 1{,}96\sigma \le X \le \mu + 1{,}96\sigma\} \approx 0{,}95. \qquad (2.40)$$

Ці властивості нормального розподілу ілюструються рис. 2.8.

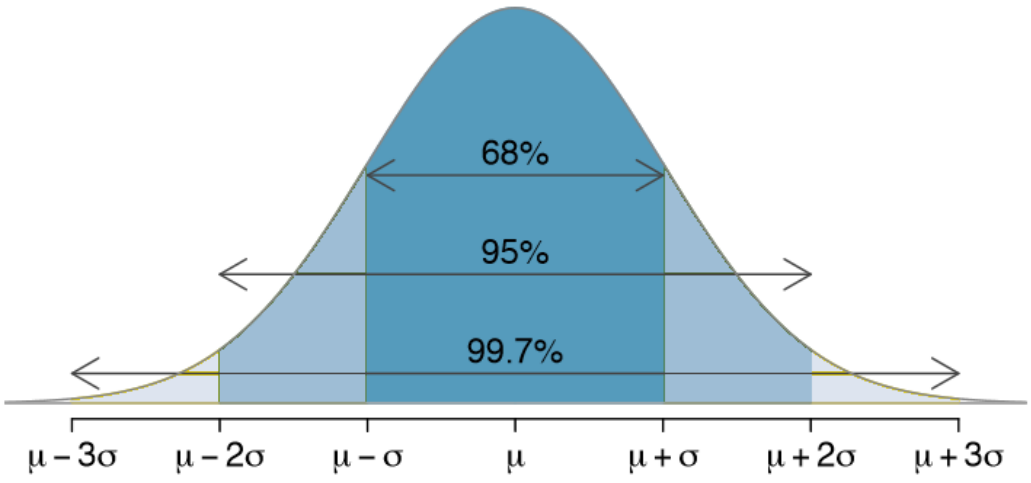
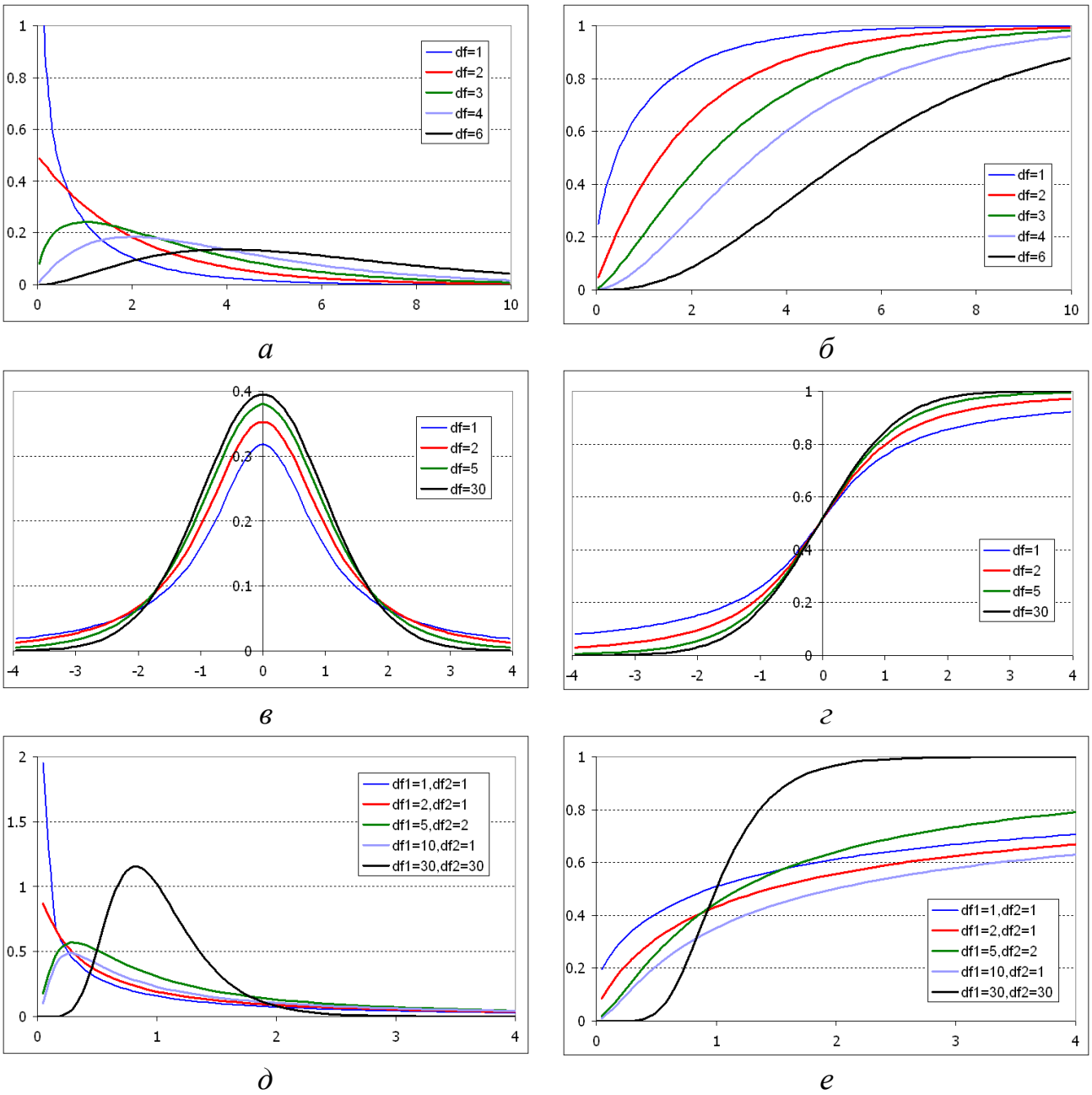


Рис. 2.8. Ймовірності влучання нормально розподіленої випадкової величини в інтервали $\pm 1, 2, 3\sigma$ від математичного сподівання μ

Розподіли математичної статистики. Так узагальнено називають три розподіли, які є спорідненими до нормального і дуже часто зустрічаються в задачах математичної статистики. Це розподіл хі–квадрат, t–розподіл

Стьюдента та *F*–розподіл Фішера. Параметрами всіх цих розподілів є кількість ступенів свободи (англ. *degrees of freedom*, скорочено *df*). Зміст цього поняття буде розкритий у наступному розділі.

Якщо випадкова величина $z \sim N(0,1)$, то $x = z^2$ буде мати розподіл *хі–квадрат* (англ. *chi–squared*) з одним ступенем свободи, що скорочено позначається як $x \sim \chi^2[1]$. Якщо $x_1,...,x_n$ є незалежними випадковими величинами із розподілом $x_i \sim \chi^2[1]$, то їх сума матиме розподіл хі-квадрат із n ступенями свободи: $\sum_{i=1}^{n} x_i \sim \chi^2[n]$. Форма розподілу хі-квадрат при різній кількості ступенів свободи ілюструється на рис. 2.9*а,б*.



а *б*

в *г*

д *е*

Рис. 2.9. Щільності та інтегральні функції розподілів математичної статистики: *а,б* – хі–квадрат; *в,г* – *t*–розподіл Стьюдента; *д,е* – *F*–розподіл Фішера

Математичне сподівання та дисперсія випадкової величини $X \sim \chi^2[n]$ дорівнюють n та $2n$, відповідно.

Якщо $z \sim N(0,1)$, $x \sim \chi^2[n]$ і не залежить від z, то відношення

$$t[n] = \frac{z}{\sqrt{x/n}} \tag{2.41}$$

матиме *t–розподіл Стьюдента* із n ступенями свободи (англ. *t–distribution*). Цей розподіл має таку ж форму, що й нормальний, але товстіші «хвости» (рис. 2.9*в,г*). При збільшенні n розподіл Стьюдента збігається до нормального.

Якщо дві незалежні змінні $x_1 \sim \chi^2[n_1]$, $x_2 \sim \chi^2[n_2]$, то відношення

$$F[n_1, n_2] = \frac{x_1/n_1}{x_2/n_2} \tag{2.42}$$

буде мати *F–розподіл Фішера* (англ. *F–distribution*). Цей розподіл є двопараметричним. Його форма ілюструється на рис. 2.9*д,е*.

Іноді є корисним наступний факт: якщо $t \sim t[n]$, то $t^2 \sim F[1,n]$.

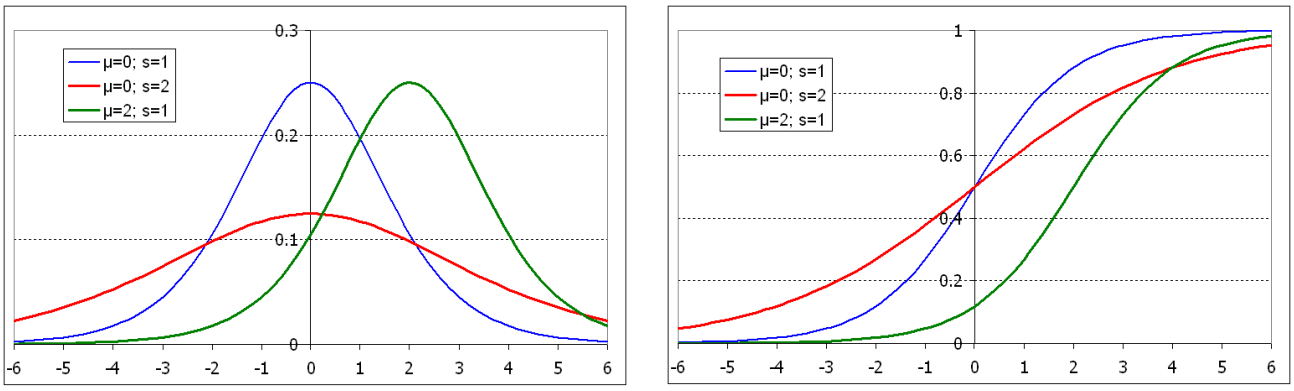
Логістичний розподіл (англ. *logistic distribution*) має багато застосувань, найважливішим серед яких з точки зору ІАД є логістична регресія (п. 8.2). Логістичний розподіл характеризується двома параметрами, μ та s, і скорочено позначається як $X \sim L(\mu, s)$. Функція логістичного розподілу задається як

$$F(x) = \frac{1}{1+e^{-(x-\mu)/s}}, \tag{2.43}$$

а щільність розподілу становить:

$$f(x) = \frac{e^{-(x-\mu)/s}}{s(1+e^{-(x-\mu)/s})^2}. \tag{2.44}$$

Приклади кривих логістичного розподілу для різних значень μ та s наведені на рис. 2.10. Як можна бачити, за формою логістичний розподіл є схожим на нормальний, але він має значно важчі «хвости». В [25] наводиться оцінка, що логістичний розподіл найближче до розподілу Стьюдента з $df = 7$.



а *б*

Рис. 2.10. Криві логістичного розподілу для різних значень параметрів μ та s: *а* – щільність розподілу; *б* – функція розподілу

Логістичний розподіл з параметрами $\mu = 0$ та $s = 1$ вважається стандартним; його інтегральна функція розподілу позначається як

$$\Lambda(x) = \frac{1}{1+e^{-x}}. \tag{2.45}$$

Математичне сподівання логістичного розподілу дорівнює μ, а дисперсія становить $\frac{\pi^2}{3}s^2$. Щільність логістичного розподілу факторизується як:

$$f(x) = \frac{1}{s} \times \frac{1}{1+e^{-(x-\mu)/s}} \times \frac{e^{-(x-\mu)/s}}{1+e^{-(x-\mu)/s}} = \frac{1}{s}F(x)(1-F(x)). \tag{2.46}$$

Зокрема, для стандартного логістичного розподілу

$$f(x) = \Lambda(x)(1-\Lambda(x)). \tag{2.47}$$

Розподіл екстремальних значень або *розподіл Гумбеля* (англ. *extreme value distribution, Gumbel distribution*) використовується, відповідно до своєї назви, для моделювання розподілу максимумів або мінімумів множин випадкових величин із деяких інших розподілів. Розподіл Гумбеля тісно пов'язаний із логістичним розподілом. На ньому базується мультиноміальна логістична модель, яка буде описана далі в п. 8.3.

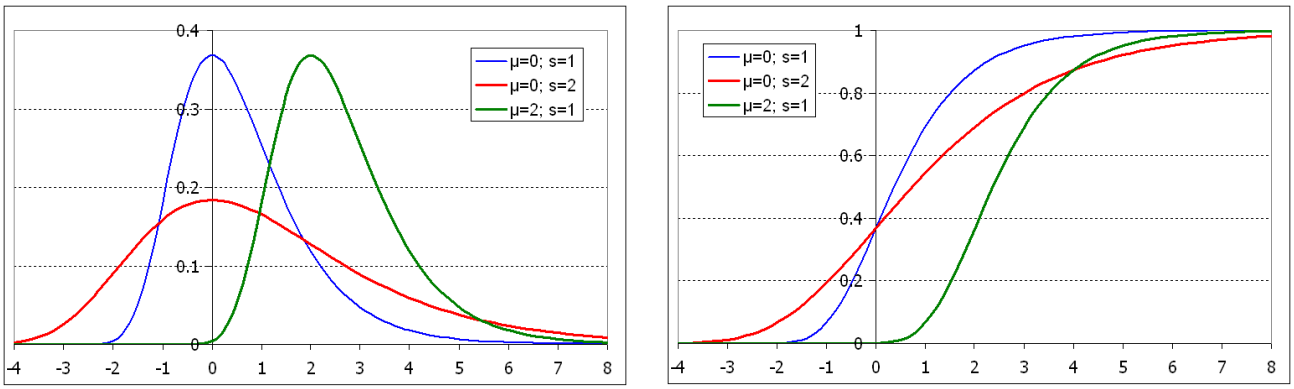
Розподіл Гумбеля характеризується двома параметрами, μ та s, позначається $X \sim EV(\mu, s)$, і задається інтегральною функцією

$$F(x) = e^{-e^{-(x-\mu)/s}}. \tag{2.48}$$

Щільність розподілу становить:

$$f(x) = \frac{1}{s}\exp\left(-\left(\frac{x-\mu}{s} + \exp\left(-\frac{x-\mu}{s}\right)\right)\right). \tag{2.49}$$

Форма кривих розподілу Гумбеля для різних значень μ та s показана на рис. 2.11. Стандартними значеннями параметрів вважаються $\mu = 0$ та $s = 1$. Розподіл є асиметричним; його математичне сподівання дорівнює $\mu + s\gamma$, де $\gamma \approx 0{,}5772$ – константа Ейлера. Дисперсія становить $\pi^2 s^2 / 6$.



а *б*

Рис. 2.11. Криві розподілу Гумбеля для різних значень параметрів μ та s:
а – щільність розподілу; *б* – функція розподілу

Розподіл має декілька важливих властивостей:

1) якщо $G_1,...,G_m$ – незалежні випадкові величини, що мають розподіл Гумбеля із параметрами μ та s, то $\max\{G_1,...,G_m\} \sim EV(\mu + s\ln m, s)$;

2) якщо випадкові величини $X \sim EV(\mu_x, s)$, $Y \sim EV(\mu_y, s)$ незалежні, то $X - Y \sim L(\mu_x - \mu_y, s)$.

Саме ці властивості роблять розподіл Гумбеля привабливим для моделювання раціонального вибору серед декількох альтернатив.

2.6 Системи випадкових величин

Сумісний розподіл двох випадкових величин X та Y (англ. *joint distribution*) може бути заданий через інтегральну функцію як

$$F(x,y) = P\{X \le x, Y \le y\}. \tag{2.50}$$

Для дискретних випадкових величин сумісний розподіл може також бути надано таблицею значень

$$f(x,y) = P\{X = x, Y = y\}, \tag{2.51}$$

а для неперервних випадкових величин може бути задана сумісна щільність розподілу $f(x,y)$ так, що

$$P\{a \le X \le b, c \le X \le d\} = \int_a^b \int_c^d f(x,y)dxdy. \tag{2.52}$$

Граничний розподіл (англ. *marginal distribution*) випадкових величин X та Y визначається як

$$F_X(x) = \lim_{y\to\infty} F(x,y\}; \; F_Y(y) = \lim_{x\to\infty} F(x,y\}. \tag{2.53}$$

Граничний розподіл випадкової змінної отримується шляхом підсумовування або інтегрування за значеннями іншої змінної:

$$f_X(x) = \sum\nolimits_y f(x,y); \; f_Y(y) = \sum\nolimits_x f(x,y) \tag{2.54}$$

для дискретних випадкових величин або

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)dy; \; f_Y(y) = \int_{-\infty}^{\infty} f(x,y)dx \; - \tag{2.55}$$

для неперервних випадкових величин.

Умовний розподіл (англ. *conditional distribution*) випадкової величини Y при відомому значенні $X = x$ визначається як

$$f(y \mid x) = \frac{f(x,y)}{f_X(x)}. \tag{2.56}$$

Умовний розподіл випадкової величини є звичайним розподілом ймовірностей і для нього можна розрахувати будь-які чисельні характеристики. Зокрема, *умовне математичне сподівання* (англ. *conditional expectation*)

випадкової величини Y при відомому значенні випадкової величини $X = x$ визначається як

$$M[Y \mid X = x] = \sum_{i=1}^{n} y_i P\{Y = y_i \mid X = x\} \tag{2.57}$$

для дискретних випадкових величин та

$$M[Y \mid X = x] = \int_{-\infty}^{\infty} y f(y \mid x) dy \tag{2.58}$$

для неперервних випадкових величин.

Всі ці поняття природним чином узагальнюються на випадок сумісного розподілу трьох та більше випадкових величин.

2.7 Визначення залежностей між випадковими величинами

Випадкові величини X та Y є *незалежними* (англ. *independent*), якщо знання однієї із змінних ніяк не впливає на розподіл іншої, тобто $f(y \mid x) = f_Y(y)$; $f(x \mid y) = f_X(x)$. Застосувавши ці рівняння до формули (2.56), отримаємо наступне правило:

$$f(x, y) = f_X(x) f_Y(y) \Leftrightarrow X, Y \text{ є незалежними.} \tag{2.59}$$

Це правило виконується також і для інтегральної функції розподілу:

$$F(x, y) = F_X(x) F_Y(y) \Leftrightarrow X, Y \text{ є незалежними.} \tag{2.60}$$

Приклад 2.3. Нехай X, Y – результати двох кидань монети; X,Y = 1 якщо випав герб і 0 в іншому випадку. Нехай $Z = X + Y$. Можливі 4 комбінації значень X та Y: 00,01,10,11, ймовірність кожної з яких складає 0,25.

Сумісний розподіл X та Y може бути заданий таблицею

X \ Y	Y=0	Y=1	$f_X(x)$
X=0	0,25	0,25	0,5
X=1	0,25	0,25	0,5
$f_Y(y)$	0,5	0,5	1

Для всіх значень x та y $f(x, y) = f_X(x) f_Y(y)$. Отже, X та Y є незалежними.

Розглянемо тепер сумісний розподіл X та Z.

X \ Z	Z=0	Z=1	Z=2	$f_X(x)$
X=0	0,25	0,25	0	0,5
X=1	0	0,25	0,25	0,5
$f_Y(y)$	0,25	0,5	0,25	1

Для $x = 1, z = 0$ маємо $f(1,0) = 0 \neq f_X(1) f_Z(0) = 0,5 \times 0,25$. Отже, X та Z є залежними. ■

Умовне математичне сподівання $M[Y \mid X = x]$ називається також *регресією* Y на X (англ. *regression*) . Якщо $M[Y \mid X = x] = M[Y]$ при будь-якому значенні x, то змінні Y та X називають незалежними за математичним очікуванням.

Коваріацією (англ. *covariance*) змінних X та Y називається величина

$$Cov(X,Y) = M[(x - M[x])(y - M[y])] = M[xy] - M[x]M[y]. \qquad (2.61)$$

Зауважимо, що за попереднім визначенням $Cov(X,X) = D[x]$.

Якщо випадкові величини X та Y незалежні, то:

$$Cov(X,Y) = M[(x - \mu_x)(y - \mu_y)] = M[x - \mu_x]M[y - \mu_y] = 0. \qquad (2.62)$$

Зворотне твердження невірно: якщо $Cov(X,Y) = 0$ (в цьому випадку величини X та Y називаються некорельованими), то вони можуть бути залежними.

Сенс коваріації полягає у наступному. Якщо між X та Y є позитивний зв'язок, то великим значенням змінної X найчастіше будуть відповідати великі значення змінної Y. Тоді обидва співмножники в формулі (2.62) будуть мати однаковий знак і коваріація буде позитивною. Аналогічно, від'ємне значення коваріації свідчить про наявність зворотного зв'язку між змінними.

На жаль, абсолютне значення коваріації нічого не говорить про силу такого зв'язку. Для цього використовується поняття кореляції.

Кореляція (англ. *correlation*) випадкових величин X та Y визначається як

$$Corr(X,Y) = \rho_{XY} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}. \qquad (2.63)$$

Можна довести, що $-1 \le \rho_{XY} \le 1$. Чим ближче до одиниці абсолютне значення кореляції, тим тіснішим є зв'язок між змінними X та Y.

Якщо $Y = a + bX$, тобто Y є лінійною функцією від X, то:

$$\rho_{XY} = \frac{cov(a + bX, X)}{\sqrt{D[X]D[Y]}} = \frac{bD[X]}{\sqrt{D[X]D[Y]}} = b\frac{\sigma_X}{\sigma_Y}. \qquad (2.64)$$

Якщо до того ж змінні X та Y будуть стандартизовані так, що $\sigma_X = \sigma_Y = 1$ (цього можна досягти шляхом ділення на σ_X, σ_Y), то коефіцієнт кореляції буде співпадати з коефіцієнтом b. Отже, кореляція є показником лінійного зв'язку між змінними.

Таким чином, можна виділити три ступеня незалежності між випадковими величинами X та Y [45].

1. Незалежність: $f(x,y) = f_X(x) f_Y(y)$.

2. Незалежність за математичним сподіванням: $M[Y \mid X] = M[Y]$.

3. Відсутність кореляції: $\rho_{XY} = 0$.

Перше визначення є «найсильнішим», а третє – «найслабшим». Тобто, з визначення 1 випливає визначення 2, а з нього, у свою чергу – визначення 3.

Зворотні твердження в загальному випадку не є вірними.

Відсутність незалежності між випадковими змінними означає наявність

зв’язку між ними. Пошук таких взаємозв’язків і є центральним питанням інтелектуального аналізу даних.

Контрольні запитання

1. Наведіть приклади випадкових експериментів.
2. В чому полягає частотна інтерпретація ймовірності?
3. Що таке простір елементарних подій?
4. Як визначаються сума та добуток подій?
5. Як визначається умовна ймовірність?
6. В чому полягає теорема Байєса?
7. Як визначається незалежність випадкових подій?
8. Розкрийте зміст понять апріорної та апостеріорної ймовірності.
9. В чому полягає формула повної ймовірності?
10. Дайте визначення поняття «випадкова величина».
11. Що називається розподілом випадкової величини?
12. Як визначається функція розподілу випадкової величини?
13. Чим розрізняються дискретні та неперервні випадкові величини?
14. Назвіть основні властивості інтегральної функції розподілу.
15. Як знайти ймовірність влучання випадкової величини у визначений інтервал числової осі?
16. За якою формулою розраховується математичне сподівання випадкової величини? Що воно характеризує?
17. Як знайти математичне сподівання функції від випадкової величини?
18. Як визначається і що характеризує дисперсія випадкової величини?
19. Що характеризує середньоквадратичне відхилення випадкової величини? Як воно пов’язано з дисперсією?
20. Чим пояснюється особливе місце нормального розподілу в теорії ймовірностей?
21. Як визначається сумісний розподіл двох випадкових величин?
22. Що таке граничний розподіл ймовірностей в системі двох випадкових величин?
23. Як знайти розподіл ймовірностей випадкової величини за деякої відомої умови?
24. Коли дві випадкові величини можна вважати незалежними?
25. Як визначається умовне математичне сподівання?
26. Як визначається коваріація між двома випадковими величинами?
27. Чим відрізняється кореляція від коваріації?
28. Назвіть три ступеня незалежності між випадковими величинами.
29. Чи можуть бути корельованими незалежні випадкові величини?
30. Чи можуть бути залежними некорельовані випадкові величини?

Завдання для самостійної роботи

2.1. Із колоди в 36 карт вибирають три карти. Знайти ймовірність того, що:

а) вони виявляться однакової масті;

б) різної масті;

в) рівно одна карта буде тузом;

г) принаймні одна карта буде тузом.

2.2. Двомоторний літак може успішно приземлитися на одному двигуні. Чотирьохмоторний літак може успішно приземлитися на двох двигунах. Відмова будь-якого з двигунів не залежить від виходу з ладу інших двигунів і складає p. Який з двох літаків є безпечнішим?

2.3. По каналу зв'язку передається одна з двох кодових комбінацій 11111 або 00000. Апріорні ймовірності передачі цих команд дорівнюють 0,6 і 0,4, відповідно. Внаслідок перешкод ймовірність правильного прийому окремого символу дорівнює 0,8 незалежно від попередніх помилок.

Прийнято комбінацію 10110. Визначити, яку комбінацію було передано.

2.4. Монета кинута 4 рази. Випадкова величина X дорівнює кількості появ герба. Знайти ряд розподілу випадкової величини X, кумулятивну функцію розподілу, математичне сподівання і дисперсію.

2.5. У темряві людина намагається відчинити двері в свою квартиру, випадково перебираючи n ключів у своїй кишені. Позначимо через N кількість спроб, які знадобились, щоб нарешті відкрити двері. Знайти $M[N]$ та $D[N]$, якщо:

а) ключі, які не підійшли, далі не перевіряються;

б) ключі, які не підійшли, повертаються назад до кишені і можуть бути випробувані наново.

2.6. Вивести формули (2.35).

2.7. Випадкова величина $X \sim N(1,3^2)$. Знайти:

а) $P(|X| > 2)$;

б) $P(X > -1 | X \le 1)$.

2.8. Випадкова величина X має експоненційний розподіл, якщо її щільність розподілу задається функцією:

$$f(x) = \lambda e^{-\lambda x},\ \lambda > 0, x \ge 0.$$

Знайдіть $M[X]$ та $D[X]$.

2.9*. Нехай $u_1 = d_1 + \varepsilon_1$, $u_2 = d_2 + \varepsilon_2$, де d_1, d_2 – константи, а $\varepsilon_1, \varepsilon_2$ – незалежні випадкові величини із розподілом Гумбеля $EV(0,1)$. Доведіть, що

$$P(u_1 > u_2) = \frac{e^{d_1}}{e^{d_1} + e^{d_2}}.$$

2.10. Сумісний розподіл випадкових величин X та Y задається таблицею

		X		
		0	1	2
Y	0	0,05	0,10	0,03
	1	0,20	0,10	0,20
	2	0,08	0,16	0,08

а) Знайти наступні ймовірності:

$P(Y < 2)$;

$P(Y < 2, X > 0)$;

$P(Y > 0 \mid X > 0)$.

б) Знайти граничні розподіли X та Y.

в) Знайти $M[X], M[Y], D[X], D[Y]$.

г) Знайти $\mathrm{cov}(X,Y), \rho_{XY}, M[X^2Y^3], \mathrm{cov}(X,Y^2)$.

д) Знайти умовні розподіли $f(Y \mid X = 2); f(X \mid Y > 0)$.

е) Знайти $M[Y \mid X]$ та $D[Y \mid X]$.

2.11. Доведіть, що $D[X+Y] = D[X] + D[Y] + 2\,\mathrm{cov}(X,Y)$.

2.12*. Доведіть, що $D[Y] = M\left[D[Y \mid X]\right] + D\left[M[Y \mid X]\right]$.

3. СТАТИСТИЧНИЙ АНАЛІЗ ТА ВІЗУАЛІЗАЦІЯ ДАНИХ

3.1 Предмет, задачі та основні концепції математичної статистики

Математична статистика – це наука, що займається методами обробки результатів експериментів або спостережень над випадковими явищами.

Задачі, які вирішує математична статистика, є зворотними до тих, які вивчає теорія ймовірностей. В теорії ймовірностей на базі певних теоретичних припущень про характер випадкових явищ можна отримати імовірнісні характеристики результатів серії спостережень над цими явищами. В статистиці вихідною точкою є серія таких спостережень, на базі яких ми намагаємось встановити імовірнісні характеристики відповідних явищ.

Наприклад, якщо інтервал руху потягів метрополітену складає 5 хвилин, а пасажир прибуває до станції метро у довільний момент часу, то час очікування потягу (у хвилинах) буде випадковою величиною із рівномірним розподілом $U[0,5]$. Згідно з властивостями цього розподілу (див. п. 2.5) математичне сподівання часу очікування потягу складатиме 2,5 хвилини; ймовірність того, що потяг доведеться чекати більше 4 хвилин дорівнюватиме 0,2 і так далі.

Часто в нас немає точної інформації про справжню циклічність руху потягів. Проте, можна записувати власні спостереження над часом очікування потягів (або витягти відповідну інформацію із Google Maps). В результаті отримаємо дані, подібні до наведених в табл. 3.1.

Таблиця 3.1 – Результати спостережень за часом очікування потягу метро

№	Дата	Час доби	Лінія	Час очікування, хв. (x)	Ранг за зростанням
1	1.06.23	8:30	Червона	3	7
2	3.06.23	9:00	Синя	5	10
3	4.06.23	16:00	Зелена	2,5	5
4	7.06.23	10:00	Зелена	2,75	6
5	9.06.23	14:30	Синя	0,25	1
6	10.06.23	19:15	Зелена	1	2
7	12.06.23	7:30	Червона	4,5	9
8	12.06.23	18:00	Червона	1,25	3
9	15.06.23	11:15	Зелена	3,25	8
10	17.06.23	15:45	Червона	1,5	4

На базі аналізу цієї інформації ми можемо спробувати зробити власні висновки щодо імовірнісних характеристик руху потягів. При цьому виникають наступні задачі, які є типовими для математичної статистики в цілому:

– як компактно охарактеризувати найважливішу інформацію, що міститься в табл. 3.1? (описова статистика);

– яка найкраща оцінка інтервалу руху між потягами випливає із зібраних даних? (точкові оцінки параметрів);

– наскільки можна бути впевненим у побудованій оцінці? чому дорівнює її похибка? (інтервальні оцінки параметрів);

– скільки потрібно спостережень, щоб оцінити інтервал руху потягів в межах 5% похибки із визначеним рівнем впевненості у результатах? (планування експериментів);

– чи можна стверджувати, що час очікування потягу насправді має рівномірний розподіл? (оцінка закону розподілу);

– чи співпадає інтервал руху потягів на різних лініях метрополітену? (перевірка гіпотез);

– чи існує залежність між часом доби та інтервалом руху потягів і якщо так, то який вона має вигляд? (кореляційний та регресійний аналіз).

Для відповіді на ці та інші запитання формується *статистика* – певна функція від наявних спостережень, встановлюються її властивості і на цій основі робляться висновки.

При аналізі даних, як правило, немає можливості розглянути всю сукупність об'єктів, що нас цікавлять (*генеральну сукупність*, англ. *population*). Збір та вивчення дуже великих обсягів даних є дорогим процесом, що вимагає великих витрат коштів та часу. В наведеному вище прикладі вивчення генеральної сукупності вимагало би детальних спостережень за рухом потягів на всіх трьох лініях метрополітену протягом декількох діб.

Набагато частіше для дослідження доступна тільки деяка частина всієї сукупності, або *вибірка* (англ. *sample*), подібна до наведеної в табл. 3.1. На основі вивчення вибірки можна зробити певні висновки щодо властивостей генеральної сукупності. Але для цього вибірка мусить бути *репрезентативною*, тобто такою, властивості якої повністю відбивають властивості генеральної сукупності. Таким чином, у вибірці повинні бути представлені різні комбінації і елементи генеральної сукупності. Різноманітність об’єктів, представлених в генеральної сукупності, впливає, зокрема, на розмір вибірки, достатній для цілей дослідження.

З точки зору організації даних, розповсюджені наступні типи вибірок.

1. *Часовий ряд* (англ. *time series*) є серією спостережень за одним і тим же об’єктом протягом певного періоду часу. Найчастіше спостереження відповідають послідовними рівновіддаленим моментам часу. Прикладом може бути середньомісячний рівень заробітної плати в Україні починаючи з 1992 р. Елементи часового ряду прийнято індексувати літерою *t* (від англ. time – час).

2. *Просторові дані* (англ. *cross section*) є множиною спостережень за різними об’єктами в один і той же момент часу. Прикладом може бути середній рівень заробітної плати в різних областях Україні в червні 2023 року. Елементи просторових даних індексуються літерами *i, j, k*.

3. *Панельні дані* (або просто панель; англ. *panel data*) є набором

спостережень за різними об’єктами в різні моменти часу і представляють собою двовимірний масив, у якого один із вимірів – просторовий, а другий – часовий. Отже, панельні дані мають два індекси (i, t). Прикладом може бути середній рівень заробітної плати в різних областях Україні за період 1992–2022 рр. Панель називається *збалансованою*, якщо множина спостережуваних об’єктів не змінюється з протягом часу. В іншому випадку панель називається *незбалансованою*. Так, державна служба статистики України з очевидних причин не може збирати дані, які стосуються окупованих регіонів України.

4. *Транзакційні дані* (англ. *transaction data*) представляють собою особливий тип даних, де кожен запис відбиває результат виконання транзакції (закінченої бізнес-операції). Прикладом транзакції може бути касовий чек або переказ коштів з банківського рахунку.

3.2 Описова статистика та візуалізація одновимірних даних

Одновимірними (англ. *univariate*) в статистиці називаються дані спостережень за однією характеристикою чи атрибутом. Наявність в таких даних певних закономірностей часто можна помітити навіть неозброєним оком.

Для часових рядів простим та ефективним засобом візуалізації є побудова простого *лінійного графіку* (англ. *line graph*). Це тип діаграми, який відображає інформацію як серію точок даних («маркерів»), з’єднаних відрізками. Приклади лінійних графіків наводяться на рис. 3.1 та 3.2.

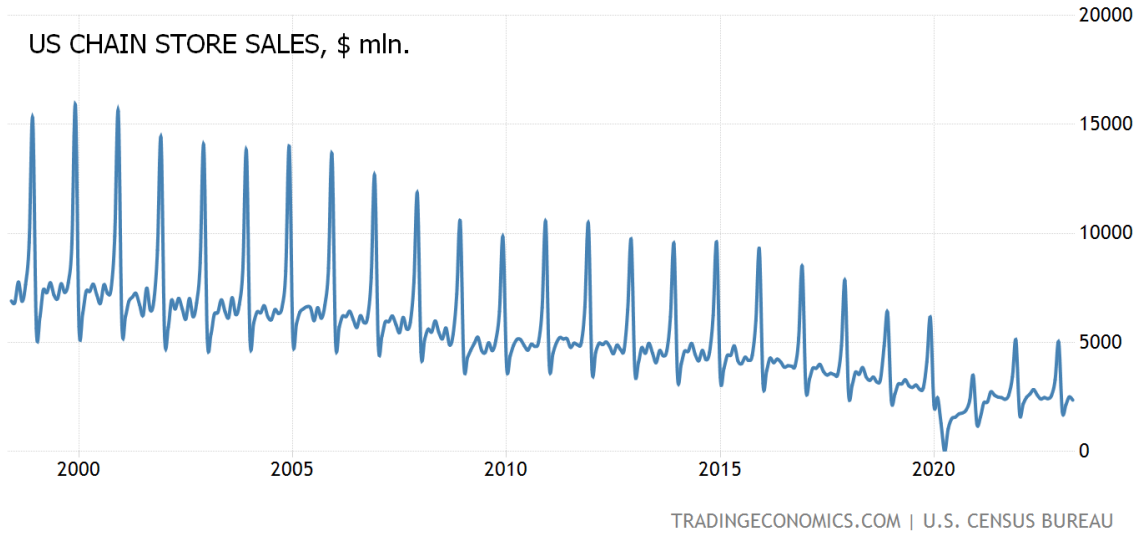


Рис. 3.1. Обсяг продажів у торгівельних мережах США, 1998–2023 [Д11]

На рис. 3.1 наведені щомісячні обсяги продажів у торгівельних мережах США. Одразу видно, що продажі мають яскраво виражений сезонний характер. Пік продажів щороку припадає на різдвяні та новорічні свята. Також помітна загальна тенденція до зниження обсягу продажів, що вірогідно пояснюється поступовою втратою покупців традиційними мережевими магазинами на користь торгівельних інтернет–платформ.

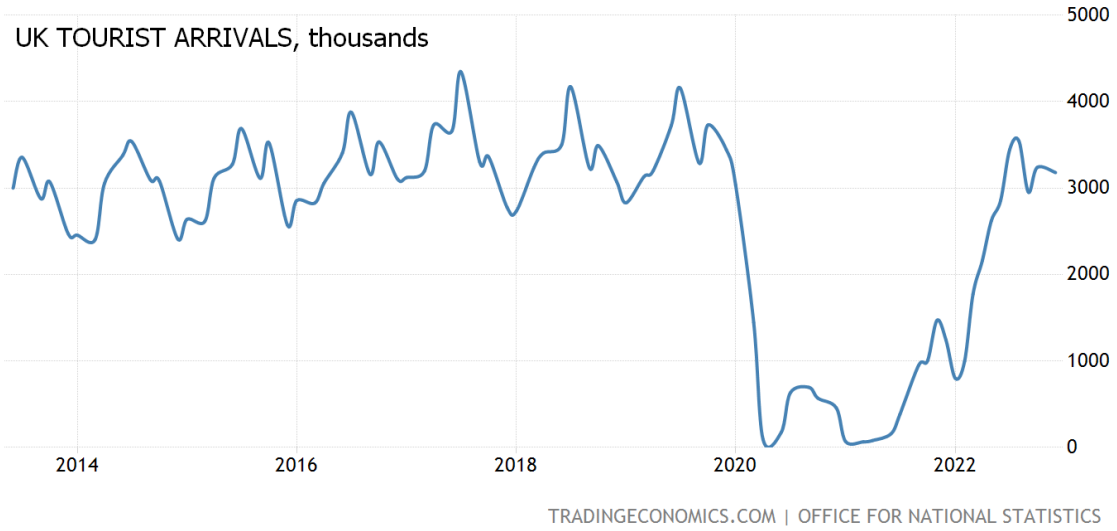
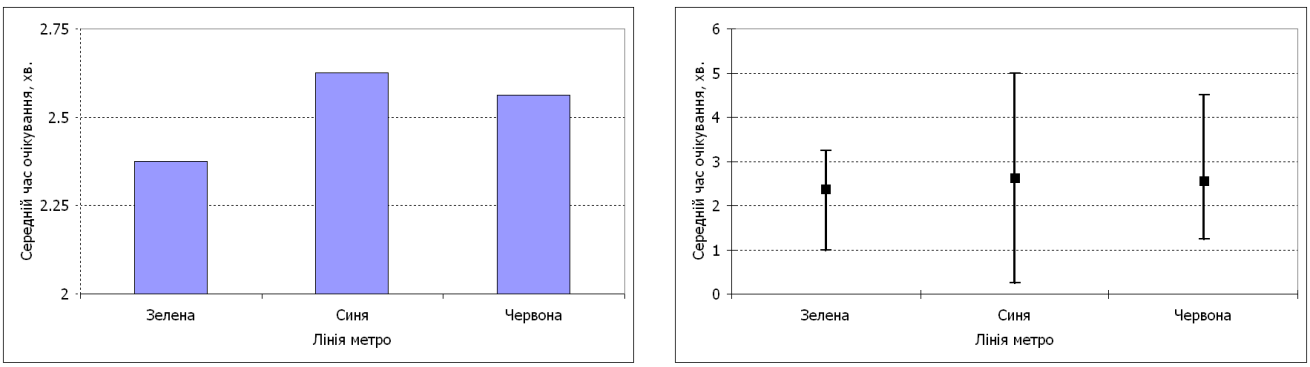


Рис. 3.2. Кількість туристів, які щомісяця відвідують Великобританію, 2013–2023 [Д11]

Рис. 3.2 зображує щомісячну кількість прибуттів пасажирів до Великобританії з туристичною метою. Привертає увагу практично повне припинення туристичних подорожей, пов'язане з початком всесвітньої епідемії COVID та їх повільне відновлення після завершення пандемії.

У сукупності рис. 3.1 та 3.2 ілюструють три особливості, які привертають увагу при аналізі часових рядів: циклічність, довгострокові тенденції зміни даних (*тренд*) та структурні зміни протягом досліджуваного періоду.

Для зображення просторових даних з порівняно невеликою кількістю об'єктів краще підходить *стовпчикова діаграма* (англ. *bar chart*). На таких діаграмах стовпчики відповідають окремим категоріям, а довжина стовпчика є індикатором їх досліджуваних характеристик. Різновиди подання інформації за допомогою стовпчикових діаграм наведені на рис. 3.3, який зображує дані табл. 3.1 у графічній формі. Рис. 3.3 (*а*) надає середній час очікування потягів на різних лініях. Рис. 3.3 (*б*) містить більш детальну інформацію: відрізки вказують інтервал зміни часу очікування від мінімального до максимального значення, а маркер посередині відрізка відповідає середньому значенню.



а *б*

Рис. 3.3. Характеристики часу очікування потягу для різних ліній метро: *а* – стовпчикова діаграма; *б* – діаграма з маркерами

Рис. 3.3 також ілюструє той факт, що вибір діапазону на осі значень (вертикальній осі) впливає на сприйняття інформації тими, кому вона адресована. З рис. 3.3 (*а*) складається враження, що між часом очікування потягів на різних лініях існують суттєві відмінності, в той час як на рис. 3.3 (*б*) різниця виглядає незначною. Це пояснюється тим, що шкала на першому графіку відповідає (відносно малому) інтервалу зміни середніх значень, і це візуально підсилює різницю між ними. На жаль, це створює можливості для маніпуляції при поданні інформації.

В тих випадках, коли категорії даних відповідають географічним регіонам, ефективним способом подання інформації є її зображення на карті. На рис. 3.4 у такій спосіб представлені дані про річне споживання алкоголю (у літрах чистого спирту) на душу населення у різних країнах світу. Для візуального відображення числової інформації використовується інтенсивність кольору.

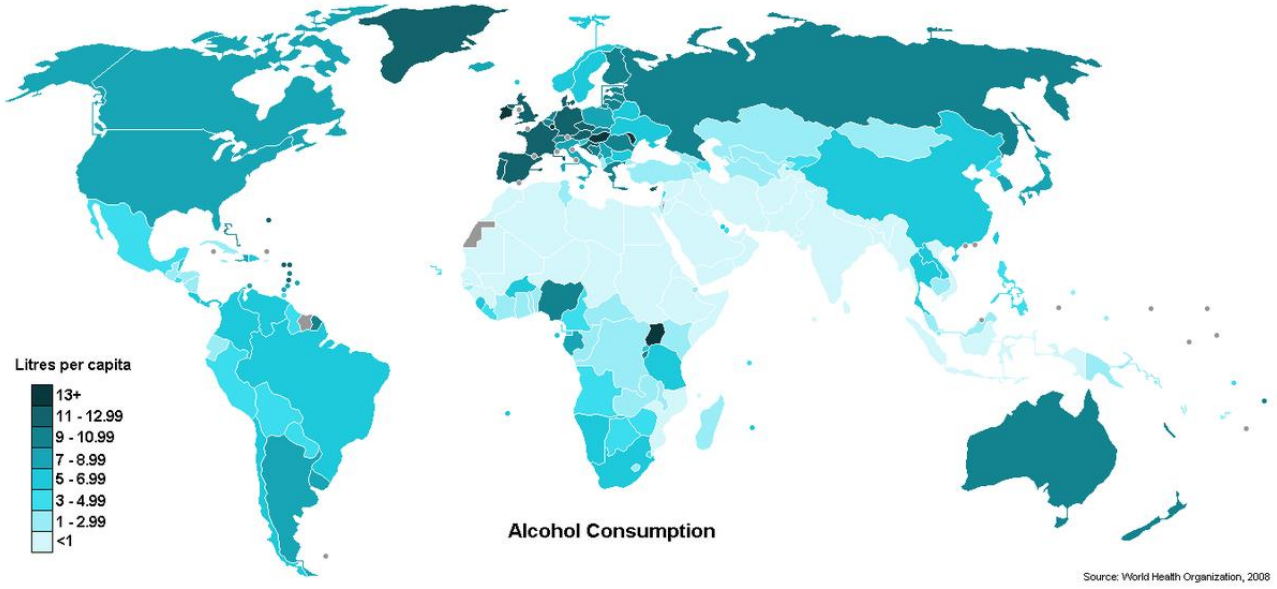


Рис. 3.4. Споживання алкоголю на душу населення в різних країнах світу [Д12]

Якщо впорядкувати досліджувані дані x_i, $i = 1,...,n$ за зростанням, отримуємо *варіаційний ряд* (англ. *variational series*): $x_{(1)} < x_{(2)} < ... < x_{(n)}$. В табл. 3.1 останній стовпчик наводить позицію елемента у варіаційному ряді.

Якщо елемент x_i зустрічається у вибірці n_i разів тоді число n_i називається *частотою* (англ. *frequency*) або абсолютною частотою елемента x_i, а відношення $w_i = n_i / n$ – його *відносною частотою* (англ. *relative frequency*). Дані, згруповані у вигляді впорядкованої послідовності пар $\{(x_i, n_i)\}$ називаються *статистичним рядом* (англ. *discrete series*).

Якщо кількість можливих значень x_i є великою (або взагалі нескінченною у випадку неперервних даних), то доцільно розбити діапазон варіації даних $[\min\limits_{1 \le i \le n} x_i, \min\limits_{1 \le i \le n} x_i]$ на декілька інтервалів. Як правило, обирають інтервали рівної

довжини. Загальноприйнятної методики визначення кількості інтервалів не існує, однак на практиці для цього часто використовують правило Стерджеса:

$$m = 1 + \log_2 n. \qquad (3.1)$$

Для кожного з побудованих у такий спосіб інтервалів $I_j = [\underline{x}_j, \overline{x}_j], j = 1,...,m$ визначають кількість спостережень $n_j \in I_j$, які йому належать. Послідовність пар $\{(I_j, n_j)\}$ або $\{(I_j, w_j = n_j / n)\}$ називають *інтервальним статистичним рядом* (англ. *continuous series*).

Графічно статистичні ряди зображають за допомогою *гістограми* (англ. *histogram*). На горизонтальній осі гістограми вказуються значення або інтервали значень, а на вертикальній – абсолютні або відносні частоти. На рис. 3.5 наводиться приклад гістограми розподілу часу очікування потягу, побудований за даними табл. 3.1.

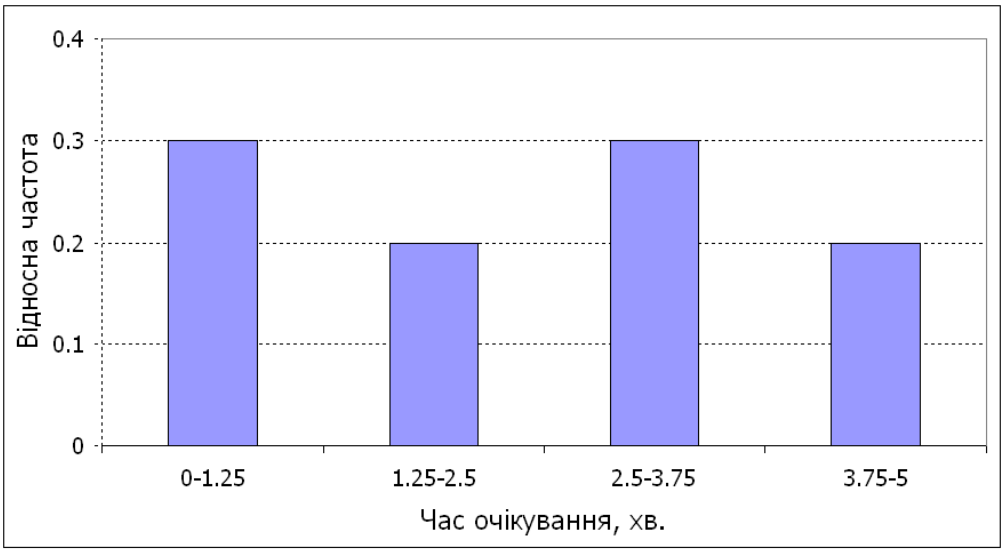


Рис. 3.5. Гістограма розподілу часу очікування потягу за даними табл. 3.1

Часто форма гістограми надає добре уявлення про закон розподілу значень у генеральній сукупності. На рис. 3.6 зображено гістограму розподілу довжини крил кімнатної мухи на базі 100 спостережень [Д9]. Видно, що дані добре узгоджуються з припущенням про їх нормальний розподіл. Для наочності на гістограму накладено криву цього розподілу.

Емпіричною функцією розподілу або *кумулятою* (англ. *empirical cumulative distribution function* або скорочено *ecdf*) називається функція $F^*(x)$ аргументу x, $x \in \mathbf{R}$, що визначає відносну частоту спостережень, які за значенням не перевищують x. Значення $F^*(x)$ можна отримати двома способами:

1) як кількість елементів множини $\{x_i : x_i \le x\}$, поділену на обсяг вибірки;

2) як суму відносних частот елементів варіаційного ряду, які передують x.

Емпірична функція розподілу є кусково–постійною функцією, яка змінює свої значення у послідовних точках варіаційного ряду. Графічно вона зображується ступінчастою функцією.

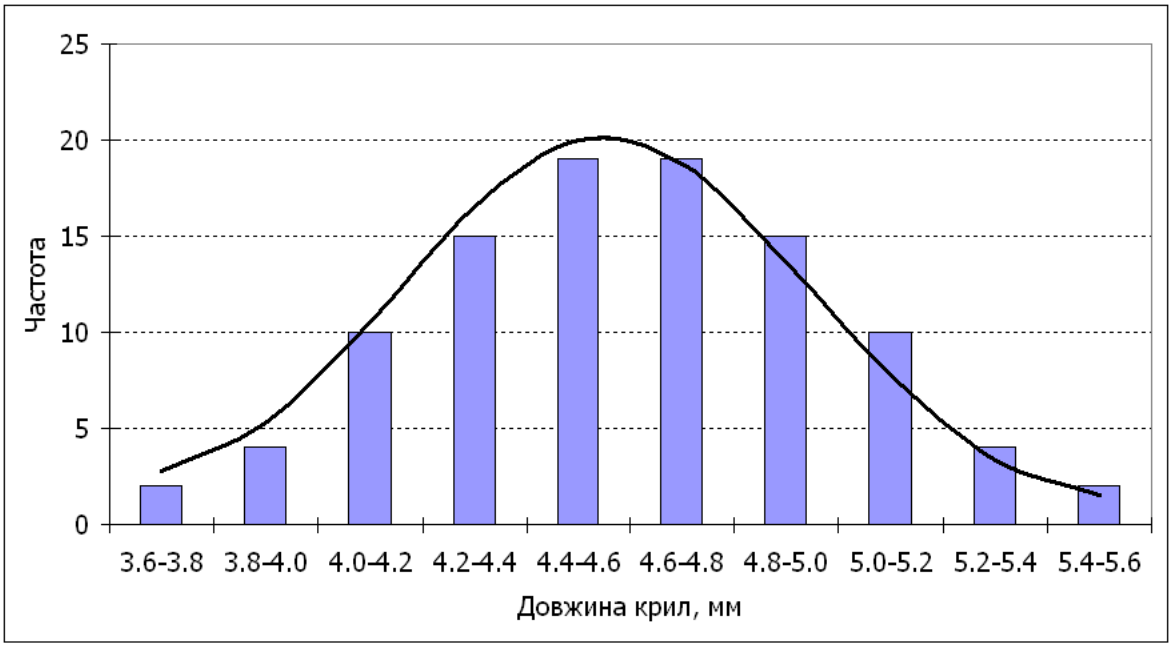


Рис. 3.6. Гістограма розподілу довжини крил кімнатної мухи
(дані з [Д9], доступні в електронній формі на сайті [Д8])

Емпірична функція розподілу є статистичним аналогом функції розподілу в теорії ймовірностей і має аналогічні властивості (див. п. 2.4). Приклад емпіричної функції розподілу часу очікування потягу для даних з табл. 3.1 наведений на рис. 3.7.

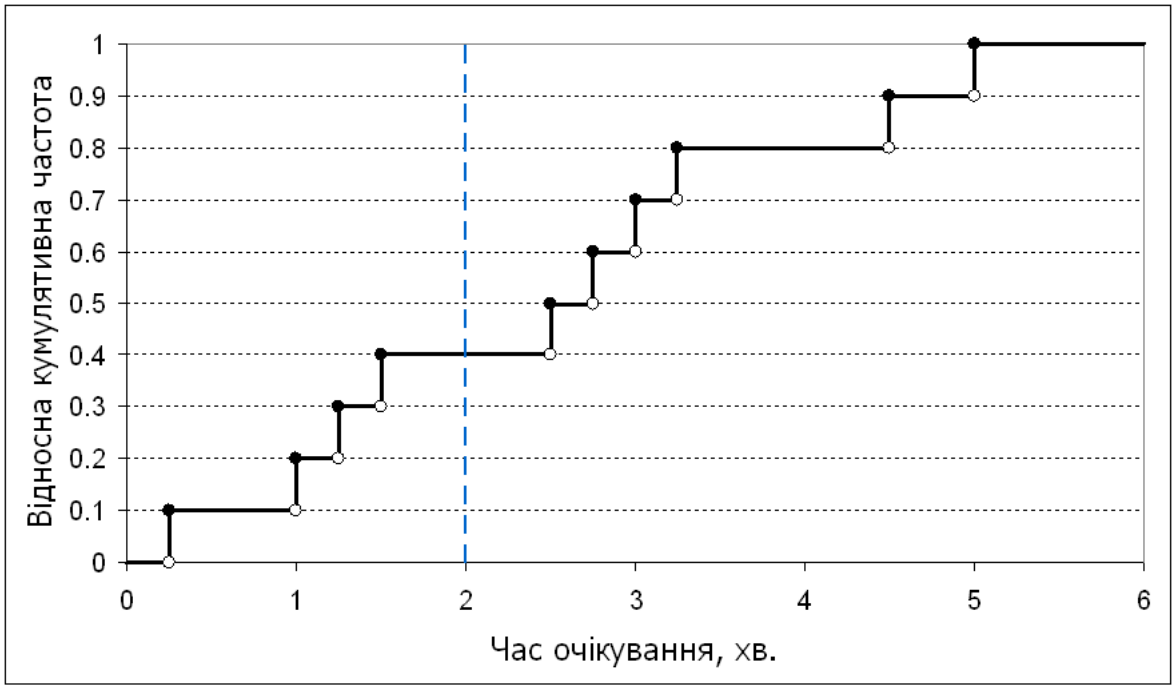


Рис. 3.7. Емпірична функція розподілу часу очікування потягу метро
за даними табл. 3.1. Наприклад, $F^*(2) = 0.4$

Систематизація та візуалізація вибірки є важливою частиною розвідувального аналізу даних. Однак у багатьох випадках, особливо коли обсяг досліджуваних даних є значним, зручніше користуватися кількісними

показниками, які б подавали інформацію про дані у стислому вигляді. Такі показники називаються числовими характеристиками вибірки, а їх сукупність – *описовою статистикою* вибірки (англ. *descriptive statistics*). Найбільш важливими є характеристики вибіркового розподілу, які кількісно описують його центр та розсіювання навколо центру.

Для характеристики центра розподілу використовуються такі показники.

Вибірковим середнім (англ. *sample mean*) називається середнє арифметичне значень досліджуваної вибірки:

$$\overline{x} = \frac{x_1 + x_2 + ... + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i. \qquad (3.2)$$

Так, для даних з табл. 3.1 середній час очікування складає

$$\overline{x} = \frac{3+5+2{,}5+2{,}75+0{,}25+1+4{,}5+1{,}25+3{,}25+1{,}5}{10} = 2{,}5.$$

Якщо дані подані у вигляді статистичного ряду, то вибіркове середнє обчислюється за формулою:

$$\overline{x} = \frac{x_1 n_1 + x_2 n_2 + ... + x_m n_m}{n_1 + n_2 + ... + n_m} = \sum_{j=1}^{m} x_j w_j \,. \qquad (3.3)$$

Якщо порівняти останню формулу з формулою (2.21) і нагадати частотну інтерпретацію ймовірностей, то стає зрозумілим, що вибіркове середнє є статистичним аналогом математичного сподівання. Цей зв’язок буде розкрито повніше у наступних розділах.

Вибіркове середнє не підходить для характеристики даних, представлених у номінальній або порядковій шкалах вимірювання. Для таких даних в якості характеристики центра розподілу часто застосовується мода. *Модою* (англ. *mode*) називається значення, яке зустрічається в даних найчастіше. За допомогою моди можна визначити, наприклад, найбільш популярну модель ноутбука, фаворита майбутніх виборів чи спортивних змагань тощо. При знаходженні моди може виникнути ситуація, коли існує кілька значень з максимальною частотою. Якщо існує лише одне значення з максимальною частотою, то розподіл називається унімодальним; якщо таких значень два – бімодальним, якщо більше – мультимодальним. Якщо ж усі значення зустрічаються у виборці однаково часто, то кажуть, що ряд не має моди. Саме такий випадок має місце для даних про час очікування з табл. 3.1.

Медіаною (англ. *median*) називається таке значення, яке міститься посередині варіаційного ряду, тобто ділить його на дві рівновеликі частини за кількістю спостережень. Якщо обсяг вибірки є парним числом, то медіана визначається як арифметичне середнє елементів ряду з індексами $\frac{n}{2}$ та $\frac{n}{2}+1$.

Для даних з табл. 3.1 медіана складає $\frac{x_{(5)}+x_{(6)}}{2}=\frac{2,5+2,75}{2}=2,625$. Медіану можна обчислити для даних любої шкали вимірювання, крім номінальної.

Медіана не залежить від значень, що містяться по обидва боки від неї. Завдяки цій властивості медіана має певні переваги над вибірковою середньою, бо на останню можуть досить суттєво впливати статистичні викиди або загальна нерівномірність вибірки. Наприклад, якщо на підприємстві працює 99 працівників з заробітною платою 10 тис. грн. на місяць і менеджер із зарплатою 5 млн., то середня заробітна плата на підприємстві складатиме приблизно 60 тис. грн. Медіана ж складатиме 10 тис. грн., що, напевно, краще характеризує «середній» стан справ на підприємстві.

Узагальненням поняття медіани є квантилі. *Квантиль* (англ. *quantile*) – це таке число, яке ділить варіаційний ряд у заданій пропорції. Квантилем порядку p статистичного розподілу є число x_p, нижче якого у варіаційному ряді міститься p-та частина всіх спостережуваних значень, тобто $F^*(x_p)=p$. В залежності від кількості частин, на які розбивається сукупність, виділяють квартилі (4 рівних частини), децилі (10 частин) і перцентилі (100 частин).

В якості прикладу використання квантилей можна навести децильний коефіцієнт, який характеризує нерівномірність розподілу доходів населення. Він визначається як відношення дев'ятого та першого децилей, тобто показує у скільки разів нижня планка доходів 10% найбільш забезпечених верств населення перевищує верхню планку доходів 10% найменш забезпечених.

Показники розсіювання, або варіації, дають змогу оцінити, наскільки дані розкидані навколо центрального значення.

Найпростішим показником розсіювання є *розмах варіації* (англ. *range*), який дорівнює різниці між найменшим та найбільшим значеннями вибірки:

$$R=x_{\max}-x_{\min}. \tag{3.4}$$

Найчастіше для характеристики варіації спостережуваних значень використовуються вибіркова дисперсія та середньоквадратичне відхилення.

Вибірковою дисперсією (англ. *sample variance*) називається середнє арифметичне квадратів відхилень значень вибірки від вибіркової середньої:

$$S^2=\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2=\frac{1}{n}\sum_{i=1}^{n}x_i^2-\overline{x}^2\ . \tag{3.5}$$

Поруч з цим використовується т.зв. *виправлена вибіркова дисперсія* (англ. *corrected* або *unbiased sample variance*):

$$s^2=\frac{n}{n-1}S^2=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x})^2\ . \tag{3.6}$$

Причини для цього будуть пояснені у наступному розділі. Поки що зауважимо, що при великому обсязі вибірки різниця між формулами (3.5) та

(3.6) є незначною.

Оскільки одиниці, в яких вимірюється дисперсія, є квадратами одиниць спостережуваної величини (наприклад, грн^2, м^2, л^2), то за показник розсіювання доцільно брати *вибіркове середньоквадратичне відхилення* (скорочено СКВ; англ. *sample standard deviation*), що дорівнює квадратному кореню з вибіркової дисперсії:

$$s = \sqrt{s^2}\,. \qquad (3.7)$$

Ці показники мають таку ж саму інтерпретацію, що й їх аналоги в теорії ймовірностей (див. п. 2.4).

Для даних з табл. 3.1 сума квадратів відхилень становить:

$$(3-2{,}5)^2 + (5-2{,}5)^2 + \ldots + (3{,}25-2{,}5)^2 + (1{,}5-2{,}5)^2 = 21.$$

Отже, вибіркова дисперсія дорівнює 21/10 = 2,1, виправлена дисперсія 21/9 ≈ 2,34; звичайне та виправлене СКВ становлять $\sqrt{2{,}1} \approx 1{,}45$ і $\sqrt{2{,}34} \approx 1{,}53$.

Також існують показники, які характеризують форму розподілу (зокрема, його асиметрію). З ними можна ознайомитись в посібниках з математичної статистики, наприклад, в [1, 3].

3.3 Оцінки параметрів розподілу

Повернемось ще раз до прикладу з п. 3.1. Якщо б інша людина фіксувала свої спостереження за часом очікування потягу, вона б отримала інші дані ніж ті, що наведені в табл. 3.1. В той же час зрозуміло, що об'єктивно існує певна періодичність руху потягів, яка обумовлює властивості випадкової величини «час очікування». Більш того, існують слушні підстави вважати, що ця величина (X) має рівномірний розподіл на відрізку $[0,b]$, де b – інтервал руху між потягами. Вибірку $x_1,...,x_{10}$ з табл. 3.1 можна тлумачити як 10 незалежних реалізацій випадкової величини $X \sim U[0,b]$. Параметр b може бути невідомим, і задача статистика полягає в тому, щоби знайти його оцінку по наявній виборці.

В загальному випадку в класичній статистиці вважається, що вибірка $x_1, x_2, ..., x_n$ є серією n незалежних спостережень за випадковою величиною із відомим законом розподілу $X \sim F(x, \boldsymbol{\theta})$ з невідомими параметрами, заданими вектором $\boldsymbol{\theta} = (\theta_1, ..., \theta_p)$.

Будь-яка функція вибірки $q(x_1, x_2, ..., x_n)$ називається *статистикою* (англ. *statistic*). Прикладами статистик можуть бути вибіркове середнє та вибіркова дисперсія, розглянуті в попередньому розділі. Оскільки значення $x_1, x_2, ..., x_n$ є випадковими величинами, то й сама статистика теж є випадковою величиною, розподіл і числові характеристики якої залежать від розподілу випадкової величини X.

Задача *точкової оцінки* параметрів (англ. *point estimate*) полягає у пошуку

таких статистик $\hat{\theta}_1(x_1,..,x_n),...,\hat{\theta}_p(x_1,..,x_n)$, які б надавали наближені значення невідомих параметрів $\theta_1,...,\theta_p$. *Інтервальна оцінка* (англ. *interval estimate*) надає діапазон значень, який містить справжнє значення параметру із заданою ймовірністю.

Оцінки параметрів мають бути «гарними» у певному сенсі. Важливими критеріями якості статистичних оцінок є незміщеність та ефективність.

Оцінка називається *незміщеною* (англ. *unbiased*), якщо її математичне сподівання дорівнює оцінюваному параметру: $M[\hat{\theta}]=\theta$.

Незміщеність означає відсутність у оцінки систематичної похибки. Наприклад, якщо в якості оцінки математичного сподівання досліджуваної випадкової величини брати максимальне з наявних спостережень, інтуїтивно очевидно, що «у середньому» оцінка вийде завищеною.

Точність оцінок характеризується показником ефективності.

Незміщена оцінка $\hat{\theta}_1$ називається більш *ефективною* (англ. *efficient*), ніж інша незміщена оцінка $\hat{\theta}_2$, якщо вона має меншу дисперсію: $D[\hat{\theta}_1]<D[\hat{\theta}_2]$.

Приклад 3.1. Нехай досліджувана випадкова величина X характеризується математичним сподіванням μ та дисперсією σ^2. Розглянемо дві альтернативні оцінки її математичного сподівання: $\hat{\theta}_1=x_1$ та $\hat{\theta}_2=\overline{x}$.

$$M[\hat{\theta}_1]=M[x_1]=\mu;$$

$$M[\hat{\theta}_2]=M[\overline{x}]=M\left[\frac{1}{n}\sum_{i=1}^{n}x_i\right]=\frac{1}{n}\sum_{i=1}^{n}\underbrace{M[x_i]}_{=\mu}=\frac{1}{n}n\mu=\mu.$$

Отже, обидві оцінки є незміщеними. Порівняємо тепер їх ефективність:

$$D[\hat{\theta}_1]=D[x_1]=\sigma^2;$$

$$D[\hat{\theta}_2]=D[\overline{x}]=D\left[\frac{1}{n}\sum_{i=1}^{n}x_i\right]=\frac{1}{n^2}\sum_{i=1}^{n}\underbrace{D[x_i]}_{=\sigma^2}=\frac{1}{n^2}n\sigma^2=\frac{\sigma^2}{n}.$$

З цього випливає, що оцінка $\hat{\theta}_2$ є ефективнішою, ніж $\hat{\theta}_1$. ■

Закономірно виникає питання: чи можна знайти таку оцінку, яка б мала найменшу дисперсію серед усіх незміщених оцінок? Позитивну відповідь на це запитання надає нерівність Крамера–Рао, яка встановлює нижню границю для дисперсії оцінок невідомих параметрів залежно від обсягу вибірки. Розгляд цієї нерівності виходить за рамки даного розділу. Він може бути знайдений, наприклад, в [1]. Проте, варто відмітити, що вибіркове середнє сягає цієї нижньої границі і, отже, може вважатися найкращою можливою оцінкою математичного сподівання будь-якої досліджуваної випадкової величини.

Розглянемо тепер дещо складніший приклад, який стосується оцінки інтервалу руху між потягами з п. 3.1.

Приклад 3.2. Нехай є вибірка $x_1, x_2, ..., x_n$ значень випадкової величини X із рівномірним розподілом на відрізку [0,b]. Знайти незміщену оцінку для невідомого параметру b.

Непоганою ідеєю виглядає використання в якості статистики для такої оцінки максимального значення у виборці $q = x_{(n)} = \max\{x_1, ..., x_n\}$. Закон розподілу цієї статистики можна знайти у такий спосіб. Нагадаємо, що для $X \sim U[0,b]$ $P\{X \le x\} = F_X(x) = x/b$ (див. п. 2.5). Тоді:

$$F_q(x) = P\{q \le x\} = P\{x_1 \le x, ..., x_n \le x\} = P\{x_1 \le x\} \times ... \times P\{x_n \le x\} = \left(\frac{x}{b}\right)^n.$$

Відповідно, щільність розподілу складатиме:

$$f_q(x) = \frac{dF_q(x)}{dx} = \frac{nx^{n-1}}{b^n}.$$

За формулою (2.22) математичне сподівання статистики q становить:

$$M[q] = \int_0^b x f_q(x) dx = \frac{n}{b^n} \int_0^b x^n dx = \frac{n}{b^n} \frac{x^{n+1}}{n+1}\Big|_0^b = \frac{n}{n+1} b.$$

Отже, q виявляється зміщеною (заниженою) оцінкою параметру b. Проте, це легко виправити, помноживши q на поправочний коефіцієнт:

$$\hat{b} = \frac{n+1}{n} q = \frac{n+1}{n} \max(x_1, ..., x_n).$$

За останньою формулою, оцінка інтервалу руху між потягами за даними табл. 3.1 складає $5 \times \frac{11}{10} = 5{,}5$ хвилин. ■

Ситуація, подібна до розглянутої в прикладі 3.2, виникає і при оцінці дисперсії. За формулою (2.26) дисперсія є математичним сподіванням, а найкращою оцінкою математичного сподівання є вибіркове середнє. Якщо б нам було відомо значення математичного сподівання генеральної сукупності μ, то найкращою оцінкою дисперсії дійсно було б

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^2.$$

Але, як правило, справжнє значення μ невідомо, і замість нього доводиться користатися його оцінкою $\bar{x}$, яка сама по собі є випадковою величиною. Завдяки цьому вибіркова дисперсія (3.5) виявляється зміщеною (заниженою) оцінкою дисперсії генеральної сукупності σ^2:

$$M[S^2] = M\left[\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2\right] = \frac{n}{n-1} \sigma^2 \qquad (3.8)$$

(виведення цієї формули є не складним, але дещо громіздким і залишається читачеві в якості вправи для самостійної роботи). Саме тому вводиться статистика виправленої вибіркової дисперсії (3.6), яка є незміщеною оцінкою

справжньої дисперсії σ^2:

$$M[s^2] = M\left[\frac{n}{n-1}S^2\right] = M\left[\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2\right] = \sigma^2. \qquad (3.9)$$

Знаменник $n-1$ в останній формулі називається кількістю *ступенів свободи* (англ. *degrees of freedom*, скорочено *df*). Сенс цього поняття полягає у наступному. Якщо нам відомо $\bar{x}$ та значення $x_1,...,x_{n-1}$, то можна знайти значення x_n як $n\bar{x} - \sum_{i=1}^{n-1} x_i$. Отже, з n змінних x_i в формулах (3.5) та (3.6) є тільки $n-1$ справжніх незалежних змінних. Поняття ступенів свободи є дуже поширеним у статистиці. Як правило, використання у формулах кожної додаткової статистики зменшує кількість ступенів свободи на одиницю.

Точкові оцінки є наближеними і можуть суттєво відрізнятися від справжніх значень оцінюваних параметрів. Щоб охарактеризувати точність і надійність знайдених оцінок, використовуються довірчі інтервали.

Довірчим інтервалом (англ. *confidence interval*) для невідомого параметру θ називається інтервал, який містить справжнє значення θ із ймовірністю γ:

$$P\{\underline{\theta}(x_1,...,x_n) \le \theta \le \bar{\theta}(x_1,...,x_n)\} = \gamma. \qquad (3.10)$$

Довжина довірчого інтервалу характеризує точність оцінки, а значення ймовірності γ – її надійність, або *рівень достовірності* (англ. *confidence level*). Останню величину часто подають у формі $\gamma = 1-\alpha$, де α називається *рівнем значущості* (англ. *significance level*). Найчастіше в статистиці використовується значення $\gamma = 0,95$ ($\alpha = 0,05$), хоча зустрічаються також значення $\gamma = 0,9$, $\gamma = 0,99$ та інші. Точність та надійність оцінки пов'язані зворотною залежністю: чим ширше довірчий інтервал, тим вище впевненість у результатах оцінювання та навпаки.

Основний прийом для побудови довірчих інтервалів полягає у пошуку *опорної статистики* (англ. *pivot, ancillary statistic*) $Q(x_1,...,x_n,\theta)$, розподіл якої 1) є відомим і 2) не залежить від вектора оцінюваних параметрів θ.

Приклад 3.3. Нехай є вибірка $x_1, x_2,..., x_n$ значень випадкової величини X із рівномірним розподілом на відрізку $[0,b]$. Знайти інтервальну оцінку для невідомого параметру b із рівнем значущості α.

Якщо $x_i \sim U[0,b]$, то $\frac{x_i}{b} \sim U[0,1]$, тобто підпорядковуються знаному розподілу, який не залежить від оцінюваного параметру b. Виходячи з цього і з результатів прикладу 3.2, спробуємо використати в якості опорної статистики $Q = \frac{x_{(n)}}{b} = \frac{\max(x_1,...,x_n)}{b}$. Встановимо її функцію розподілу:

$$F_Q(x) = P\{\frac{x_{(n)}}{b} \le x\} = P\{x_{(n)} \le bx\} = P\{x_1 \le bx\} \times ... \times P\{x_n \le bx\} = \left(\frac{bx}{b}\right)^n = x^n.$$

З рівняння $P\{Q \le c\} = \alpha$ знайдемо $c = \sqrt[n]{\alpha}$. Вочевидь, $P\{Q \le 1\} = 1$. Отже,

$$P\{c \le Q \le 1\} = P\{Q \le 1\} - P\{Q \le c\} = 1 - \alpha.$$

Але

$$P\{c \le Q \le 1\} = P\left\{c \le \frac{x_{(n)}}{b} \le 1\right\} = P\left\{1 \le \frac{b}{x_{(n)}} \le \frac{1}{c}\right\} = P\left\{x_{(n)} \le b \le \frac{x_{(n)}}{c}\right\}.$$

Зібравши все разом, отримуємо:

$$P\left\{x_{(n)} \le b \le \frac{x_{(n)}}{\sqrt[n]{\alpha}}\right\} = 1 - \alpha.$$

Наприклад, для даних з табл. 3.1 при $\alpha = 0{,}05$ маємо $5 \le b \le 6{,}75$. Зверніть увагу, що довірчий інтервал у цьому прикладі не є симетричним відносно незміщеної точкової оцінки $\hat{b} = 5{,}5$ з прикладу 3.2. ■

Найпоширенішим і найбільш дослідженим є випадок, коли вибірка породжена нормально розподіленою випадковою величиною. Відзначною рисою цього розподілу є те, що сума нормально розподілених випадкових величин також буде мати нормальний розподіл. Отже, якщо $x_i \sim N(\mu, \sigma^2)$, то $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \sim N(\mu, \frac{\sigma^2}{n})$. Це дозволяє для більшості практично важливих задач інтервального оцінювання побудувати допоміжну статистику, яка зводиться до стандартних розподілів математичної статистики.

Якщо дисперсія σ^2 є відомою, то величина $z = \frac{\overline{x} - \mu}{\sqrt{\sigma^2 / n}} = \frac{\sqrt{n}(\overline{x} - \mu)}{\sigma}$ буде мати стандартний нормальний розподіл $N(0,1)$. Нехай $P\{-c_\alpha \le z \le c_\alpha\} = 1 - \alpha$. Тоді:

$$P\{-c_\alpha \le \frac{\sqrt{n}(\overline{x} - \mu)}{\sigma} \le c_\alpha\} = P\left\{\overline{x} - c_\alpha \frac{\sigma}{\sqrt{n}} \le \mu \le \overline{x} - c_\alpha \frac{\sigma}{\sqrt{n}}\right\} = 1 - \alpha. \quad (3.11)$$

Значення c_α обирається виходячи із властивостей стандартного нормального розподілу (див. п.2.5). Для $\alpha = 0{,}05$ значення $c_{0{,}05} = 1{,}96$. Інші поширені значення c_α проілюстровані на рис. 3.8.

Якщо дисперсія σ^2 не є відомою, то замість неї доводиться використовувати її незміщену оцінку s (3.7). В цьому випадку величина $t = \frac{\overline{x} - \mu}{\sqrt{s^2 / n}} = \frac{\sqrt{n}(\overline{x} - \mu)}{s}$ буде мати t-розподіл Стьюдента із $df = n - 1$ ступенів свободи.

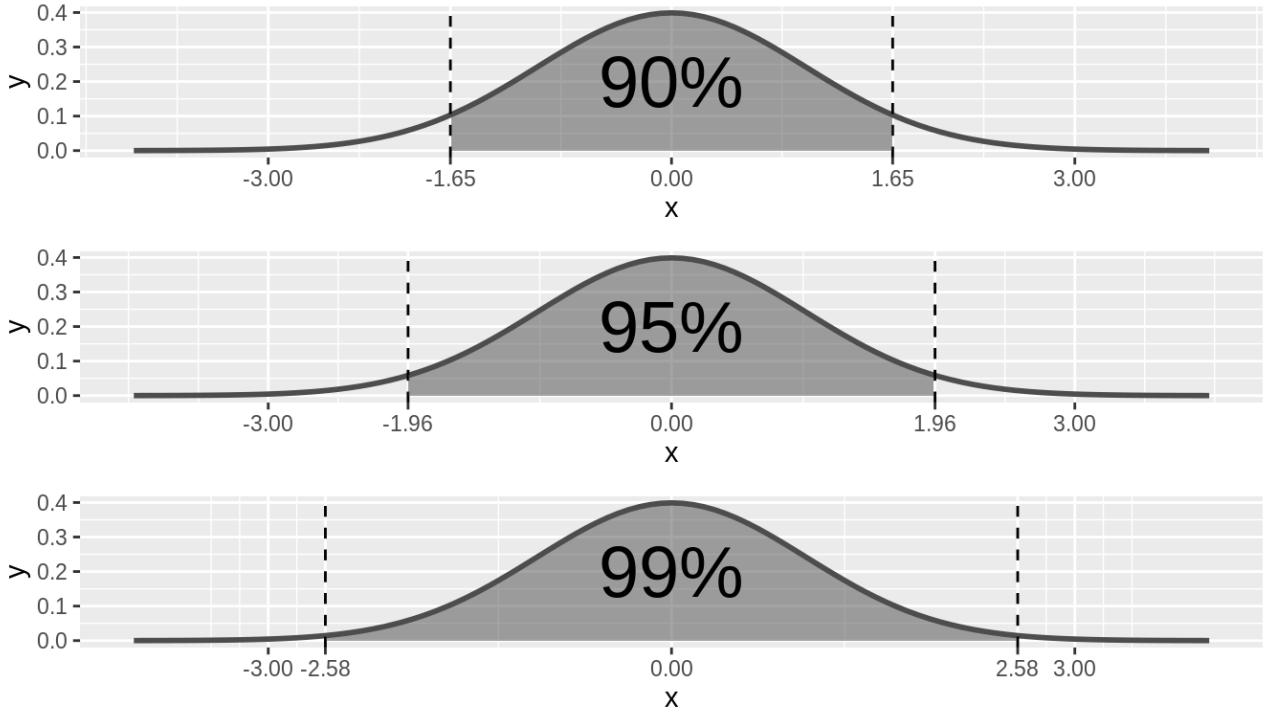


Рис. 3.8. Довірчі інтервали для стандартного нормального розподілу для типових рівнів значущості

Довірчий інтервал для вибіркової середньої має такий же вигляд, як і в формулі (3.11), але значення c_α слід обирати з таблиць розподілу Стьюдента для *df* ступенів свободи та бажаного рівня значущості. Воно буде дещо більшим, ніж значення c_α для нормального розподілу, але при великому обсязі вибірки ця різниця стає незначною.

Довірчий інтервал для дисперсії може бути побудований на базі опорної статистики $\chi^2 = \dfrac{(n-1)s^2}{\sigma^2}$, яка має розподіл χ^2 із $df = n-1$ ступенів свободи. Деталі та інші корисні довірчі інтервали можна знайти в [3, 4, 25, 57].

3.4 Перевірка гіпотез

Перевірка гіпотез – це сукупність методів, за допомогою яких робляться висновки про узгодженість наявних даних з висунутими припущеннями про властивості генеральної сукупності.

Прикладом може бути оцінка дієвості нового препарату для лікування певної хвороби. При клінічних випробуваннях виділяються дві групи пацієнтів. Першій групі піддослідних дають препарат, що тестується, а іншій – плацебо (нейтральну речовину без терапевтичного ефекту). Треба зробити висновок про дієвість препарату шляхом порівняння результатів лікування в двох групах.

У цілому *статистичною гіпотезою* (англ. *statistical hypothesis*) називають будь-яке припущення стосовно вигляду розподілу генеральної сукупності або числового значення параметрів цього розподілу.

Статистичні гіпотези поділяють на параметричні та непараметричні. Параметричні статистичні гіпотези стосуються значення параметрів закону розподілу генеральної сукупності, загальна форма якого вважається відомою. Усі інші статистичні гіпотези є непараметричними. Параметричні гіпотези розподіляються далі на прості та складні. Якщо статистична гіпотеза однозначно визначає розподіл генеральної сукупності, то вона називається *простою* (англ. *simple*). Якщо ж статистична гіпотеза робить твердження про належність розподілу генеральної сукупності до певної родини розподілів, то її називають складною (англ. *composite*). Складну гіпотезу можна розглядати як скінченну або нескінченну множину простих гіпотез.

Прикладами непараметричних гіпотез для вибірки з табл. 3.1 можуть бути твердження «у 80% випадків потяг доведеться чекати не більше 4 хвилин» або «середній час очікування потягу однаковий для всіх ліній метрополітену». Якщо можна вважати, що дані про час очікування з табл. 3.1 є реалізаціями випадкової величини з рівномірним розподілом $U[0,b]$, де b – невідомий параметр, то твердження «$b=5$» є прикладом простої гіпотези, а твердження «$b<6$» – складної. Іноді ці терміни суперечать інтуїції: так, у виборці з нормального закону $N(\mu,\sigma^2)$ гіпотеза «$\mu=3,\ \sigma=1$» є простою, бо однозначно фіксує криву розподілу, а простіша на вигляд гіпотеза «$\mu=3$» – складною, бо вона сумісна з будь-яким законом розподілу виду $N(3,\sigma^2)$.

Перевірка статистичних гіпотез завжди починається з визначення основної та альтернативної гіпотез. Основну гіпотезу називають *нульовою гіпотезою* (англ. *null hypothesis*) та позначають H_0. Паралельно розглядається гіпотеза, яка суперечить нульовій. Вона називається *альтернативною* (англ. *alternative hypothesis*) і позначається H_1. Наприклад, для гіпотези $H_0: \mu=3$ можливі такі варіанти альтернативних гіпотез: $H_1: \mu\neq 3$, $H_1: \mu>3$ та $H_1: \mu<3$.

Далі за допомогою визначених правил проводиться перевірка узгодженості даних з нульовою гіпотезою. Якщо наявні дані їй суперечать, нульова гіпотеза відкидається на користь альтернативної гіпотези. У іншому випадку нульова гіпотеза не відхиляється. Це не може вважатися доказом її правдивості; за висловом німецького статистика Лотара Закса, «гіпотеза може бути перевірена, але ніколи не може бути доведена» [46]. Насправді, оскільки вибірка вважається випадковою, то й будь-яке правило прийняття рішень на основі даних може призвести до різних висновків у різних вибірках. При цьому можливі помилки, які можна поділити на два типи:

– гіпотеза H_0 помилково відхиляється, коли вона насправді є правдивою – *помилка I типу* (англ. *type I error*);

– гіпотеза H_0 помилково приймається, коли насправді істинною є гіпотеза H_1 – *помилка II типу* (англ. *type II error*).

Наприклад, у кримінальному суді відповідач може бути винним або невинним у скоєнні злочину, в якому він підозрюється, а присяжні можуть засудити його або виправдати. Фактично, присяжні мають перевірити гіпотезу про винність підсудного на базі наявних доказів і прийняти відповідне рішення. За принципом презумпції невинності в якості нульової гіпотези логічно обрати твердження H_0: відповідач не винен. Альтернативною тоді буде гіпотеза H_1: відповідач винен. Помилкою першого типу буде засудження невинної людини, а помилкою другого типу – виправдання винного.

У зв'язку з поширеністю таких ситуацій прийняття рішень, у багатьох предметних областях існує власна термінологія для позначення помилок I та II типів. Наприклад, у медицині результат тесту вважається позитивним, якщо він свідчить про наявність у пацієнта хвороби. В цьому контексті помилки першого типу називають хибно позитивними результатами (англ. *false positives*), а другого – хибно негативними (англ. *false negatives*). В радіолокації помилки першого типу відомі як пропуск цілі, а другого – як хибна тривога. Під час контролю якості продукції ймовірність забракувати стандартні вироби називають ризиком виробника, а ймовірність визнати придатними браковані вироби – ризиком споживача і т.д.

Ймовірність помилки I типу $P(H_1 \mid H_0)$, яка виникає при застосуванні певної статистичної процедури перевірки гіпотез (тесту), називають *рівнем значущості* (англ. *significance level* або *size of the test*) і позначають через α. Ймовірність помилки II типу $P(H_0 \mid H_1)$ називають *потужністю* тесту (англ. *power of the test*) і позначають через β.

На жаль, цих помилок неможливо уникнути, а єдиним джерелом їх одночасного зменшення є збільшення обсягу вибірки. Зазвичай, прийнятний рівень значущості (найчастіше α = 0,05 чи α = 0,01) задається наперед.

Ситуації, які можуть виникнути при перевірці статистичних гіпотез, ілюструються таблицею 3.2.

Таблиця 3.2 – Термінологія статистичної перевірки гіпотез

		Вірна гіпотеза	
		$\boldsymbol{H_0}$	$\boldsymbol{H_1}$
Висновок	$\boldsymbol{H_0}$	H_0 вірно прийнята (ймовірність $1-\alpha$)	Помилка II типу: H_0 помилково прийнята (ймовірність β)
	$\boldsymbol{H_1}$	Помилка I типу: H_0 помилково відкинута (ймовірність α)	H_0 вірно відкинута (ймовірність $1-\beta$)

Процедура проведення статистичного тесту багато в чому подібна до методики інтервальної оцінки параметрів. Формується опорна статистика K (в

контексті перевірки гіпотез вона називається також тестовою), яка має знаний розподіл і не залежить від параметрів, що перевіряються (для параметричних гіпотез). Розподіл тестової статистики за нульовою гіпотезою поділяє діапазон її можливих значень на дві області: ту, яка наводить до відхилення H_0 (т.зв. *критична область*, англ. *critical region*) і ту, яка сумісна з H_0. Ймовірність того, що тестова статистика опиниться в критичному регіоні, повинна складати α. Нарешті, нульова гіпотеза відкидається, якщо спостережуване значення k^* статистики K буде знаходитись у критичній області.

Вид критичної області залежить від альтернативної гіпотези. Виділяють три види критичних областей (див. рис: 3.9):

– правостороння визначається інтервалом $(\overline{k},\infty)$, де ліва межа критичної області знаходиться із умови $P\{k \le \overline{k}\} = 1-\alpha$;

– лівостороння визначається інтервалом $(-\infty,\underline{k})$, де права межа критичної області знаходиться із умови $P\{k \le \underline{k}\} = \alpha$;

– двостороння визначається двома інтервалами $(-\infty,\underline{k}) \cup (\overline{k},\infty)$, де ліва межа знаходиться із умови $P\{k \le \underline{k}\} = \frac{\alpha}{2}$, а права – із умови $P\{k \le \overline{k}\} = 1-\frac{\alpha}{2}$.

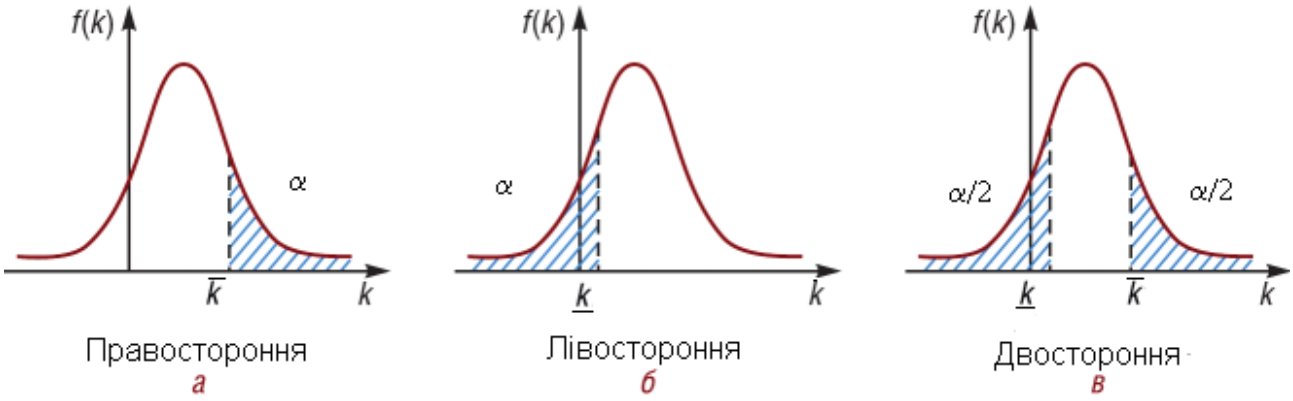


Рис. 3.9. Різновиди критичних областей

Приклад 3.4. Повернемось ще раз до даних з табл.3.1. Будемо вважати, що часи очікування $x_i \sim U[0,b]$ і перевіримо гіпотезу $H_0 : b = 6$ проти $H_1 : b < 6$ на рівні значущості α = 0,05. Нагадаємо, що обсяг вибірки складає $n = 10$.

В прикладі 3.3 було встановлено, що статистика $Q = \frac{x_{(n)}}{b} = \frac{\max(x_1,...,x_n)}{b}$ має функцію розподілу $F_Q(x) = x^n$ на $x \in [0,1]$ і, відповідно, щільність розподілу $f_Q(x) = \frac{dF_Q(x)}{dx} = nx^{n-1}$. Оскільки альтернативна гіпотеза сформульована у вигляді односторонньої нерівності $b < 6$, слід побудувати лівосторонню критичну область. Щоб дізнатись границю цієї області, слід вирішити рівняння $F_Q(x) = x^{10} = 0,05$, звідки $\underline{q} \approx 0,74$ (див. рис. 3.10, де критичну область виділено

блакитним кольором). В вибірці з табл. 3.1 $q^* = \frac{5}{6} \approx 0{,}83$ знаходиться за межами критичної області. Отже, нульова гіпотеза не відхиляється.

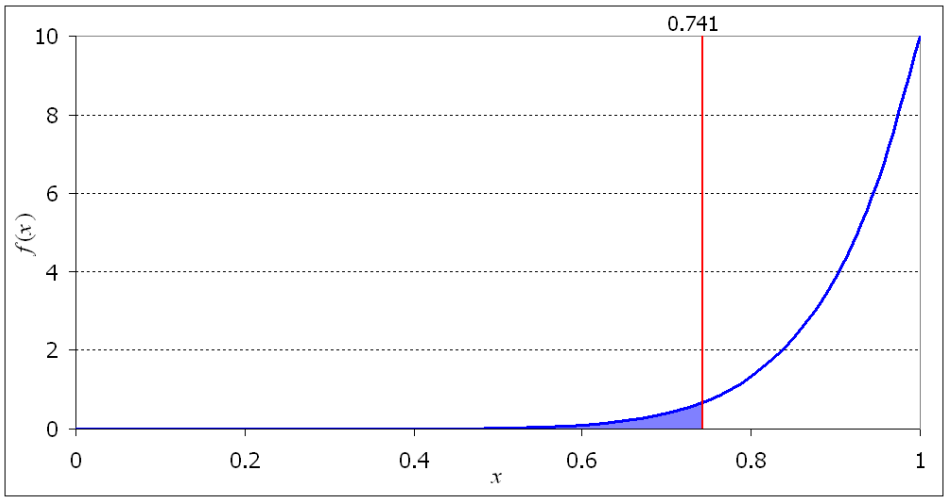


Рис. 3.10. Критична область для тестової статистики з прикладу 3.4.

Нульову гіпотезу не буде відхилено для будь яких значень тестової статистики Q в діапазоні $[0{,}74;1]$. Оскільки $Q = \frac{x_{(n)}}{b}$, нерівність $0{,}74 \le Q \le 1$ можна перетворити на інтервальну оцінку для b: $0{,}74 \le \frac{x_{(n)}}{b} \le 1 \Rightarrow x_{(n)} \le b \le \frac{x_{(n)}}{0{,}74}$. Якщо підставити у останню формулу значення $x_{(n)} = 5$ з табл. 3.1, отримаємо $5 \le b \le 6{,}75$, що збігається з результатом, отриманим в прикладі 3.3. ■

Останній приклад підкреслює тісний зв'язок між перевіркою гіпотез та інтервальними оцінками. Довірчий інтервал надає діапазон правдоподібних значень для оцінюваного параметру. Отже, якщо гіпотетичне значення параметру знаходиться за межами цього діапазону, то дані не узгоджуються з гіпотезою, і її слід відхилити.

Приклад 3.5 (перевірка гіпотез про ймовірність подій). За даними табл. 3.1 перевірити гіпотезу, що медіана розподілу M становить 3 хвилини проти альтернативної гіпотези $H_1 : M > 3$ при $\alpha = 0{,}05$.

Медіана розподілу M – це таке значення, що $P(X < M) = P(X \ge M) = 0{,}5$. Отже, слід перевірити гіпотезу $H_0: P\{X \ge 3\} = 0.5$ (тобто, у 50% випадків час очікування потягу складає 3 хвилини та вище) проти $H_1: P\{X \ge 3\} > 0.5$.

Оскільки емпіричним аналогом ймовірності є частота події, то в якості критерію природно обрати частоту спостережень події $\{X \ge 3\}$ у виборці. Позначимо абсолютну частоту цієї події через S. За даними табл. 3.1, $S = 4$. Збільшення ймовірності будь-якої події призводить до того, що вона трапляється частіше. Отже, занадто велике значення S свідчить на користь

альтернативної гіпотези, і треба обрати поріг $\overline{S}$, який би визначив, що означає «занадто».

Оскільки S не може перевищувати розміру вибірки n, то правостороння критична область матиме вигляд $[\overline{S}, n]$. Щоб знайти $\overline{S}$, слід вирішити рівняння

$$P\{\overline{S} \le S \le n \mid p = 0{,}5\} = \alpha, \qquad (3.12)$$

де p позначає «справжню» ймовірність події $\{X \ge 3\}$.

Випадкову змінну S можна трактувати як кількість успіхів в n випробуваннях Бернуллі з ймовірністю успіху p. Це означає, що S матиме біноміальний розподіл. Ймовірності окремих значень S згідно з формулою (2.31) становлять:

$$P\{S = i\} = C_n^i p^i (1-p)^{n-i}.$$

Ряд та інтегральну функцію розподілу для S при p = 0,5 надано в табл. 3.3.

Таблиця 3.3 – Ряд та функція розподілу для випадкової величини S

i	$P\{S = i\}$	$P\{S \le i\}$
0	0,0010	0,0010
1	0,0098	0,0107
2	0,0439	0,0547
3	0,1172	0,1719
4	**0,2051**	**0,3770**
5	0,2461	0,6230
6	0,2051	0,8281
7	0,1172	0,9453
8	0,0439	0,9893
9	0,0098	0,9990
10	0,0010	1,0000

Завдяки дискретності випадкової величини S виявляється, що точного рішення рівняння (3.12) не існує. Найближчим до бажаного рівня значущості буде значення $\overline{S} = 8$. Тоді критичною областю буде інтервал [8;10], виділений в табл. 3.3 червоним кольором, а ймовірність потрапляння в цей інтервал становитиме 0,0439 + 0,0098 + 0,0010 = 1 – 0,9453 = 0,0547. Значення S = 4 з табл. 3.1 потрапляє в «зелену» зону, отже нульова гіпотеза не відкидається.

При великому обсязі вибірки біноміальний розподіл буде наближуватись до нормального з $\mu = np$ та $\sigma^2 = np(1-p)$, що спрощує розрахунки. ■

Існує величезна кількість «готових» статистичних тестів для поширених на практиці ситуацій (пошук в Google надає 138 статей в англомовній Вікіпедії станом на 10.07.2023). Більшість цих тестів базується на припущенні про нормальний розподіл генеральної сукупності. В табл. 3.4 наводяться найбільш поширені тестові статистики.

Якщо існують вагомі підстави для припущення про інший розподіл даних,

то слід вивести розподіл тестової статистики, як це було зроблено в прикладі 3.3. Якщо ж розподіл даних слід вважати невідомим, то для знаходження критичних областей можна використати загальні властивості випадкових змінних, як, наприклад, нерівність Чебишева.

Таблиця 3.4 – Найбільш поширені тестові статистики

Найменування	**Формула**	**Розподіл**
Z–тест	$z = \frac{\sqrt{n}(\bar{x} - \mu)}{\sigma}$	$N(0,1)$
T–тест	$t = \frac{\sqrt{n}(\bar{x} - \mu)}{s}$	*t*–розподіл Стьюдента $df = n - 1$
Хі–квадрат	$\chi^2 = (n-1)\frac{s^2}{\sigma^2}$	Розподіл χ^2 $df = n - 1$
F–тест	$F = \frac{s_1^2}{s_2^2}$	*F*–розподіл Фішера для порівняння дисперсій двох вибірок ($df_1 = n_1 - 1,\ df_2 = n_2 - 1$)

Контрольні запитання

1. Дайте визначення поняттю «генеральна сукупність».
2. Дайте визначення поняттю «вибірка».
3. Що мається на увазі під статистичним рядом?
4. Як перетворити статистичний ряд на варіаційний?
5. Як побудувати гістограму на основі варіаційного ряду?
6. Що характеризує і як будується емпірична функція розподілу?
7. Дайте визначення основних показників центральної тенденції. В чому полягають їх переваги та недоліки?
8. Дайте визначення основних показників варіації даних.
9. Чим відрізняються інтервальні оцінки параметрів від точкових?
10. Розкрийте зміст поняття «статистика».
11. Назвіть та охарактеризуйте основні показники якості точкових оцінок.
12. Що саме виправляє виправлена вибіркова дисперсія?
13. Дайте визначення поняттю «довірчий інтервал».
14. В чому полягають основні етапи побудови інтервальних оцінок?
15. Як розподілено вибіркове середнє нормально розподілених змінних?
16. Як пов’язані між собою точність та надійність інтервальних оцінок?
17. Наведіть довірчі інтервали для параметру із стандартним нормальним розподілом при типових рівнях значущості.
18. Як побудувати довірчий інтервал для вибіркового середнього нормально розподілених змінних?
19. Як класифікуються статистичні гіпотези?
20. Чим відрізняються помилки I та II типу? Наведіть приклади.

21. Як визначити рівень значущості та потужність статистичного тесту?
22. Дайте визначення поняттю «критична область».
23. Які існують різновиди критичних областей?
24. Назвіть основні етапи перевірки статистичних гіпотез.
25. Як перевірити статистичну гіпотезу на базі інтервальної оцінки?
26. Назвіть основні типи тестів, що пов'язані із нормальним розподілом.

Завдання для самостійної роботи

3.1. В таблиці наведені результати виміру росту 20 чоловіків.

№	1	2	3	4	5	6	7	8	9	10
Зріст, см	174	168	177	184	183	186	163	175	183	169
№	11	12	13	14	15	16	17	18	19	20
Зріст, см	182	172	170	163	178	184	180	177	178	166

а) Побудуйте варіаційний ряд для вихідних даних.

б) Побудуйте гістограму та емпіричну функцію розподілу.

в) Розрахуйте середнє значення, дисперсію та СКВ.

г) Розрахуйте медіану розподілу, а також нижній та верхній квартилі.

3.2. У виборці з 500 спостережень із нормального розподілу $N(\mu,\sigma^2)$ 175 спостережень менше, ніж 2.1 і 275 спостережень менше, ніж 3.6. Знайдіть оцінки параметрів μ та σ^2.

3.3. Припустимо, що ми маємо дві незалежні незміщені оцінки $\hat{\theta}_1$, $\hat{\theta}_2$ деякого параметру θ. Дисперсії цих оцінок дорівнюють σ_1^2 та σ_2^2, відповідно. Знайдіть таку лінійну комбінацію цих двох оцінок $\hat{\theta} = w_1\hat{\theta}_1 + w_2\hat{\theta}_2$, яка буде незміщеною оцінкою параметру θ із мінімальною дисперсією.

3.4. Для прикладу 3.2, знайдіть дисперсію оцінки $\hat{b} = \frac{n+1}{n}\max(x_1,...,x_n)$.

3.5*. Доведіть формулу (3.8):

$$M[S^2] = M\left[\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2\right] = \frac{n}{n-1}\sigma^2.$$

3.6. Домашні апарати вимірювання артеріального тиску (тонометри) часто мають не дуже велику точність, внаслідок чого лікарі рекомендують зробити декілька вимірювань і взяти середнє значення. Припустимо, що результати вимірювання тиску тонометром мають нормальний розподіл із похибкою (СКВ) 5 мм рт. ст. Скільки вимірювань треба зробити, щоб бути на 95% впевненим, що справжнє значення тиску не відрізняється від середнього результату вимірювань більше, ніж на 4 мм рт. ст.?

3.7. Припустимо, що вибірка в завданні 3.1 взята з нормального розподілу $N(\mu,\sigma^2)$. Для 5% рівня значущості:

а) знайдіть інтервальну оцінку для параметрів μ та σ^2;

б) перевірте гіпотезу $\mu = 175$;

в) перевірте гіпотезу $\mu \geq 180$;

г) перевірте гіпотезу $\sigma^2 = 36$;

д) перевірте гіпотезу про рівність середніх для спостережень 1–10 та 11–20 (у припущенні, що дисперсії двох підвибірок співпадають).

4. ВИЯВЛЕННЯ ТА ВІЗУАЛІЗАЦІЯ ЗВ'ЯЗКІВ МІЖ ДАНИМИ

В переважної більшості випадків аналізовані дані не обмежуються однією змінною. Кожне спостереження в таких ситуаціях можна трактувати як *m*-вимірний вектор. Тому аналіз таких даних відносять до сфери *багатовимірної статистики* (англ. *multivariate statistics*).

Центральною темою багатовимірного статистичного аналізу є виявлення залежностей між даними. Найбільш наочно це можна зробити у випадку двовимірної статистики (англ. *bivariate statistics*), коли розглядаються всього дві змінні, одна з яких зазвичай вважається незалежною (X), а друга – залежною (Y). Типові методи двовимірного аналізу і будуть розглянуті в цьому розділі.

Наявність залежності між двома змінними X та Y в статистиці визначається шляхом подвійного заперечення, а саме як відсутність незалежності. При цьому, як розглядалося в п.2.7, існують три форми відсутності зв'язку між випадковими змінними – загальна незалежність, незалежність за математичним сподіванням та відсутність кореляції. Виявлення залежності між випадковими змінними в найбільш загальній формі є дуже складною задачею, але окремі види зв'язку встановлюються досить просто.

4.1 Залежність між категоріальними змінними: таблиці спряженості

Для категоріальних випадкових величин корисним інструментом для пошуку взаємозалежності є таблиці спряженості.

Таблиця спряженості (факторна таблиця, англ. *contingency table*) – це засіб представлення спільного розподілу двох змінних, призначений для дослідження зв'язку між ними. Рядки таблиці відповідають значенням однієї змінної, стовпці – значенням іншої. На перетині рядка та стовпця вказується частота спільної появи відповідних значень двох ознак. Сума частот по рядку називається маргінальною частотою рядка; сума частот за стовпцем – маргінальною частотою стовпця. Сума маргінальних частот надає граничний розподіл змінних, що утворюють рядки та стовпці таблиці. У таблицях спряженості можуть використовуватись як абсолютні, так і відносні частоти. Відносні частоти можуть розраховуватися стосовно: а) маргінальної частоти за рядком; б) маргінальної частоти за стовпцем; в) до обсягу вибірки.

Приклад 4.1. В таблиці 4.1 наведені деякі результати другого туру президентських виборів 2019 року в Україні по обраним областям і по країні в цілому за даними, наведеними в [Д2].

Перетворимо цю таблицю в таблицю сумісного розподілу змінних «область мешкання виборця» та «вибір кандидата». Для цього поділимо дані таблиці на загальну кількість врахованих бюлетенів (18 064 тис.). Результати зведено в табл. 4.2.

Таблиця 4.1 – Деякі результати другого туру президентських виборів в Україні в 2019 р., тис. осіб

Область / Кандидат	Львівська	Харківська	…	Усього
Зеленський	548	1126	…	13542
Порошенко	676	146	…	4522
Усього	1224	1272	…	18064

Таблиця 4.2 – Імовірнісна інтерпретація результатів другого туру президентських виборів в Україні в 2019 р.

Область / Кандидат	Львівська	Харківська	…	Усього
Зеленський	0.030	0.062	…	0.750
Порошенко	0.037	0.008	…	0.250
Усього	0.068	0.070	…	1.000

Дані в останньому стовпчику табл. 4.2 надають безумовний граничний розподіл голосів за кандидатами, а дані в нижньому рядку – безумовний граничний розподіл виборців за областю мешкання. Наприклад, в Харківській області мешкає біля 7% від загальної кількості громадян, які взяли участь в виборах.

Для того, щоб визначити вплив регіону на електоральні уподобання, знайдемо умовні ймовірності обрання того чи іншого кандидату з врахуванням області мешкання виборця. Для цього поділимо дані табл. 4.2 на граничний розподіл виборців за областями (нижній рядок таблиці). Результати зведено в табл. 4.3.

Таблиця 4.3 – Умовні ймовірності вибору кандидатів з врахуванням області мешкання виборця

Область / Кандидат	Львівська	Харківська
Зеленський	0.448	0.885
Порошенко	0.552	0.115
Усього	1.000	1.000

Як можна побачити з табл. 4.3, регіональні відмінності суттєво впливають на вибір кандидата. Так, шанси голосування за Порошенко в Львівській області майже в 5 разів перевищували аналогічні показники в Харківській області.

Спираючись на дані табл. 4.2, можна спробувати спрогнозувати місце мешкання виборця виходячи з його електоральних симпатій. Для цього в якості умовної змінної слід взяти обраного кандидата. Результати, отримані шляхом поділу даних табл. 4.2 на граничні ймовірності обрання кандидатів (останній стовпчик), зведені в табл. 4.4.

Таблиця 4.4 – Умовні ймовірності мешкання в певній області виходячи із обраного кандидата

Область / Кандидат	Львівська	Харківська	…	Усього
Зеленський	0.040	0.083	…	1.000
Порошенко	0.149	0.032	…	1.000

Таким чином, якщо про людину відомо, що в другому турі президентських виборів вона голосувала за Порошенко, то непогані шанси на те, що вона зі Львівської області.

Нарешті, визначимо, як виглядали би результати виборів, якби електоральні уподобання виборців були незалежними від місця їх мешкання. В цьому випадку кількість голосів за кожного кандидата визначалась би як: (гранична ймовірність голосування за певного кандидата) × (гранична ймовірність мешкання в певній області) × (загальна кількість виборців).

Результати розрахунків наведені в табл. 4.5.

Таблиця 4.5 – Очікувані результати виборів в припущенні незалежності електоральних уподобань від місця мешкання

Область / Кандидат	Львівська	Харківська
Зеленський	918	954
Порошенко	306	318
Усього	1224	1272

Легко бачити, що значення в табл. 4.5 суттєво відрізняються від реальних результатів, наведених в табл. 4.1. Наступне питання полягає в тому, наскільки суттєвою є ця різниця. Для цього використовуються різні статистичні тести, найбільш поширеним з яких є критерій згоди Пірсона хі-квадрат:

$$\chi^2 = \sum_{i=1}^{n}\sum_{j=1}^{m}\frac{(C_{ij} - E_{ij})^2}{E_{ij}},$$

де C_{ij} – абсолютна частота спостережень в i-му рядку j-го стовпця таблиці спряженості, а E_{ij} – очікувана частота спостережень цього елемента таблиці за умови незалежності. Ця статистика має розподіл хі-квадрат із $(n-1)\times(m-1)$ ступенями свободи. Гіпотеза про незалежність відхиляється, якщо обчислене значення χ^2 перевищує критичний рівень при заданому рівні значущості. ■

Таблиці спряженості можуть використовуватись також і для аналізу неперервних даних, але для цього кількісні шкали попередньо повинні бути згруповані в інтервали.

4.2 Залежності в кількісних даних: кореляційний та регресійний аналіз

Для візуалізації зв'язків між кількісними даними використовується *діаграма розсіювання* (англ. *scatter plot*), де на горизонтальній осі відкладаються значення змінної *X*, а на вертикальній – відповідні значення *Y*. Приклад такої діаграми для даних з табл. 3.1 наведено на рис. 4.1 (див. також рис. 1.1). Виглядає так, що час очікування у другій половині доби менше, ніж у першій.

Форма діаграми може вказувати на наявність статистичного зв'язку між досліджуваними змінними. Найпростішою формою такого зв'язку є лінійна.

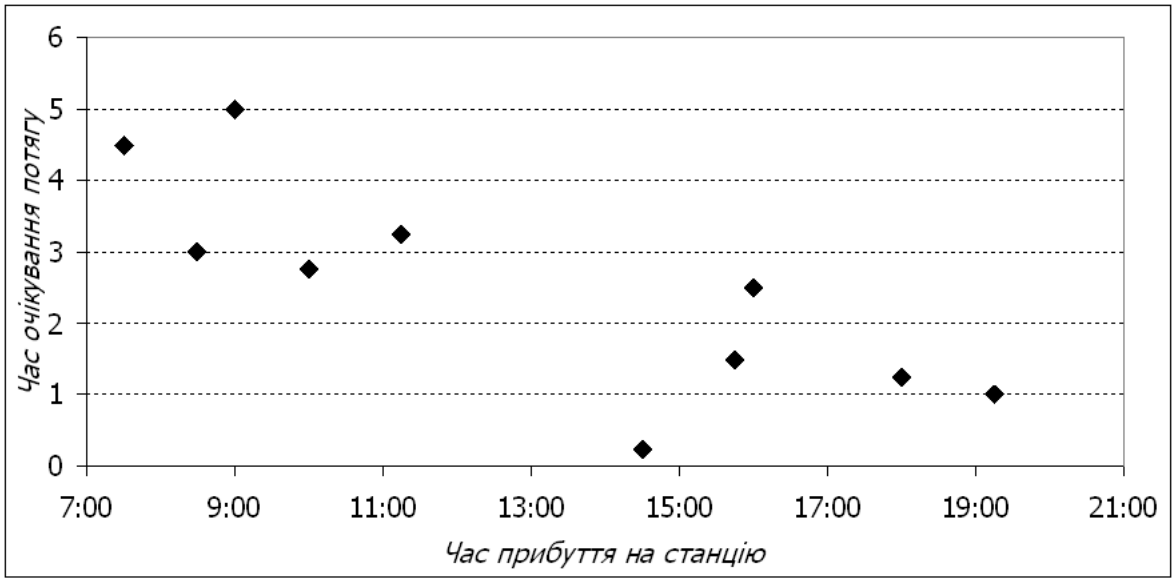


Рис. 4.1. Залежність часу очікування потягу від часу доби для даних з табл. 3.1

Сила лінійного зв'язку між двома змінними вимірюється за допомогою показників коваріації та кореляції (див. п. 2.7) . Їх емпіричними аналогами є *вибіркова коваріація* (англ. *sample covariance*)

$$s_{xy} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{n-1} \tag{4.1}$$

і *вибірковий коефіцієнт кореляції* (англ. *sample correlation coefficient*)

$$r_{xy} = \frac{s_{xy}}{s_x s_y}, \tag{4.2}$$

де $\bar{x}, \bar{y}$ – вибіркові середні змінних *X* та *Y*, s_x, s_y – їх виправлені СКВ, а n – розмір вибірки[2].

Для даних з табл. 3.1, зображених на рис. 4.1, вибірковий коефіцієнт кореляції складає –0,8083, що свідчить про доволі сильний негативний зв'язок між часом доби та часом очікування[3].

---

[2] Використання значення $n - 1$ замість n в знаменнику формули (4.1) пояснюється тими ж причинами, що і в формулі (3.6) для виправленої вибіркової дисперсії.

[3] При розрахунках час прибуття був перетворений в формат часу Excel, тобто в частку 24–годинного інтервалу. Наприклад 7:30=7,5/24=0,3125.

Важно розуміти, що наявність кореляції не означає причинно-наслідкового зв'язку між змінними. Наприклад, за великою кількістю пожеж до їх гасіння буде залучено багато пожежників, отже кореляція між кількістю пожеж та кількістю пожежників, вірогідно, вийде позитивною. Однак це не означає, що збільшення кількості пожежників призведе до зростання кількості пожеж.

При наявності сильного статистичного зв'язку між змінними логічно спробувати апроксимувати його за допомогою певної функції. Саме це і є метою регресійного аналізу даних.

Регресійний аналіз (англ. *regression analysis*) – це розділ математичної статистики, присвячений методам аналізу залежності однієї величини від інших за допомогою параметричної моделі цього зв'язку, вираженої у функції регресії. Залежну змінну називають також *відгуком* (англ. *response, outcome*), а незалежні змінні – *регресорами* або *предикторами* (англ. *covariates, predictors, explanatory variables*). Найчастіше припускається лінійна форма залежності між змінними, що призводить до моделі лінійної регресії. Властивості такої моделі досконально досліджені.

Більш формально, нехай y позначає залежну змінну, а X – вектор незалежних змінних. Нагадаємо, що математичне сподівання $M[y \mid X = x]$ називається регресією y на X (див. п.2.7). В регресійному аналізі вважається, що ця регресія може бути апроксимована функцією $f(X,\theta)$, визначеною з точністю до вектору невідомих параметрів θ, тобто:

$$M[y \mid X] = f(X,\theta). \tag{4.3}$$

Будь-яку випадкову змінну можна подати у вигляді

$$y = M[y \mid X] + \underbrace{y - M[y \mid X]}_{\varepsilon} = f(X,\theta) + \varepsilon, \tag{4.4}$$

де ε – випадкова похибка, така що $M[\varepsilon \mid X] = 0$.

Якщо використовується лише один регресор, а функція $f(X,\theta)$ є лінійною стосовно параметрів, то рівняння (4.4) спрощується до:

$$y = \alpha + \beta x + \varepsilon. \tag{4.5}$$

Модель (4.5) називають також простою або парною лінійною регресією.

Нехай в нас є вибірка значень $\{(x_i, y_i), i = 1,...,n\}$. Оцінка невідомих параметрів регресії α, β зводиться до підбору лінії, яка б найточніше описувала ці дані. Для цього найчастіше використовується *метод найменших квадратів* (скорочено МНК; англ. *ordinary least squares*, OLS). Він полягає у наступному.

Для будь-яких значень $\alpha = a, \beta = b$ можна обчислити оцінку $M[y_i \mid x_i]$ як

$$\hat{y}_i = a + bx_i. \tag{4.6}$$

Різниця між реальним значенням y_i та його оцінкою $\hat{y}_i$

$$e_i = y_i - \hat{y}_i = y_i - a - bx_i \tag{4.7}$$

називається *залишком* (англ. *residual*) (див. рис. 4.2).

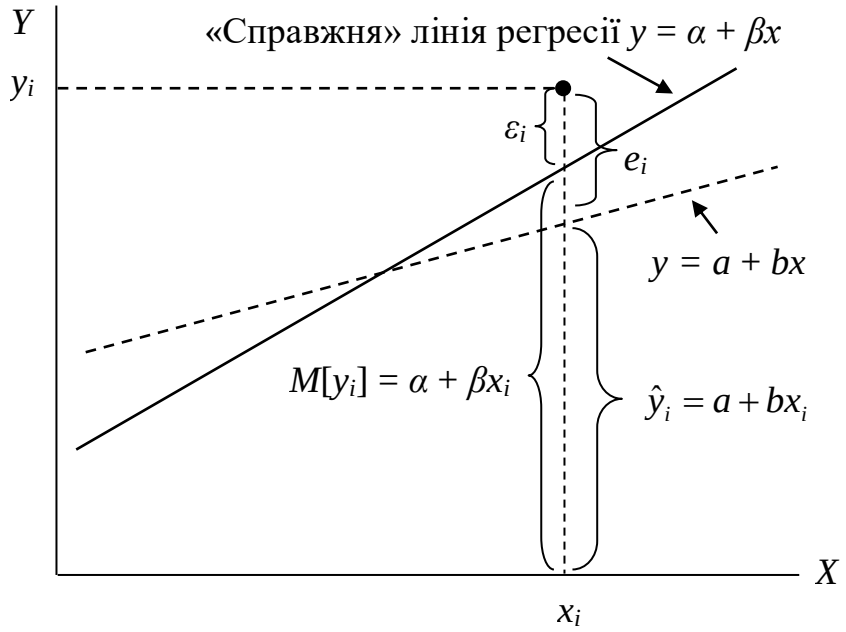


Рис. 4.2. Геометричний зміст термінології лінійної регресії

В якості критерію узгодженості даних з моделлю в МНК приймається сума квадратів залишків. Параметри обираються так, щоб мінімізувати цю суму:

$$S(a,b)=\sum_{i=1}^{n}e_i^2=\sum_{i=1}^{n}(y_i-a-bx_i)^2\underset{a,b}{\to}\min, \tag{4.8}$$

що, власне, і обумовлює назву методу. Необхідна умова мінімуму полягає в рівності нулю похідних функції $S(a,b)$:

$$\frac{\partial S(a,b)}{\partial a}=0:\sum_{i=1}^{n}\underbrace{(y_i-a-bx_i)}_{e_i}=0\Rightarrow\sum_{i=1}^{n}e_i=0; \tag{4.9}$$

$$\frac{\partial S(a,b)}{\partial b}=0:\sum_{i=1}^{n}\underbrace{(y_i-a-bx_i)}_{e_i}x_i=0\Rightarrow\sum_{i=1}^{n}x_ie_i=0. \tag{4.10}$$

Рівняння (4.9)–(4.10) називаються системою нормальних рівнянь. Вирішивши цю систему, отримуємо оцінки методу найменших квадратів:

$$\hat{\beta}=\frac{\sum_{i=1}^{n}(x_i-\overline{x})\cdot(y_i-\overline{y})}{\sum_{i=1}^{n}(x_i-\overline{x})^2}=\frac{s_{xy}}{s_x^2}=r_{xy}\frac{s_y}{s_x};\ \hat{\alpha}=\overline{y}-\hat{\beta}\overline{x}. \tag{4.11}$$

Наприклад, для даних з табл. 3.1, зображених на рис. 4.1, за формулою (4.11) отримуємо $\hat{\alpha}=6,2826$, $\hat{\beta}=-6,9968$. Графік отриманого рівняння регресії наведено на рис. 4.3, де також зображені залишки регресії.

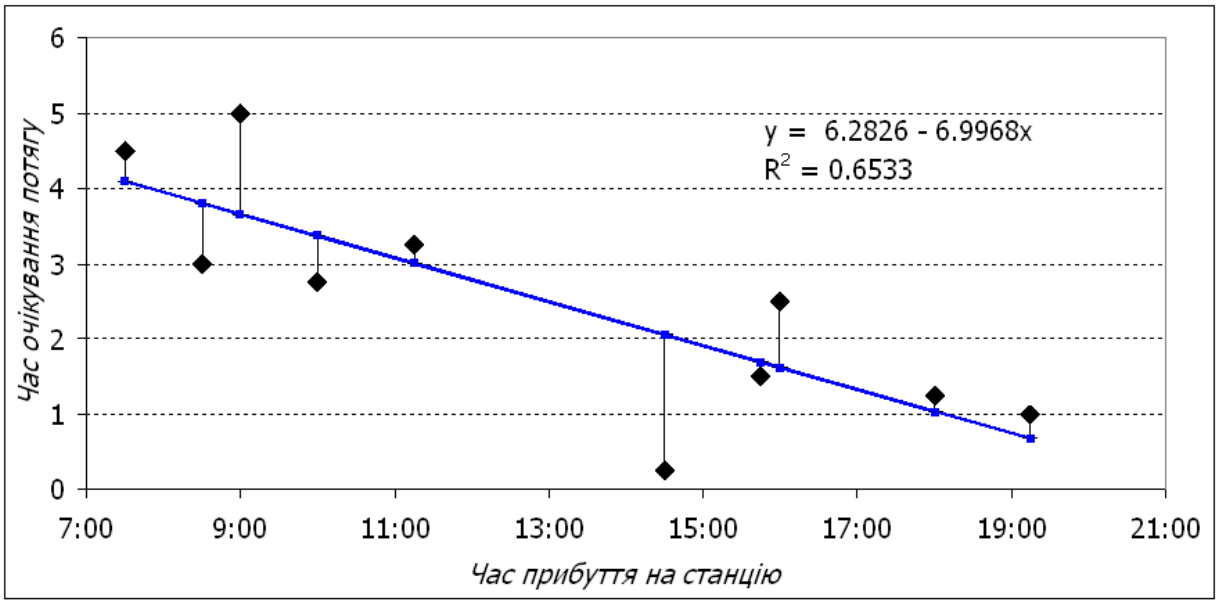


Рис. 4.3. Графік лінійної регресії для даних з табл. 3.1

Значення цільової функції МНК (4.8), на жаль, не може служити показником узгодженості моделі з даними, бо воно залежить від шкали вимірювання змінних. Можна поставити питання, у якій мірі варіація значень Y обумовлена варіацією значень X. Для цього розглянемо повну суму квадратів відхилень Y від середнього значення:

$$TSS = \sum_i (y_i - \bar{y})^2 = \sum_i (\hat{y}_i - \bar{y} + e_i)^2 = \sum_i (\hat{y}_i - \bar{y})^2 + 2\sum_i (\hat{y}_i - \bar{y})e_i + \sum_i e_i^2 .$$

Покажемо, що середній доданок в останній формулі дорівнює нулю:

$$\sum_i (\hat{y}_i - \bar{y})e_i = \sum_i (\underbrace{\bar{y} - \hat{\beta}\bar{x}}_{\hat{\alpha}} + \hat{\beta}x_i - \bar{y})e_i =$$

$$= \hat{\beta}\sum_i (x_i - \bar{x})e_i = \hat{\beta}\left(\sum_i x_i e_i - \bar{x}\sum_i e_i\right) = 0,$$

де останнє перетворення випливає із нормальних рівнянь (4.9)–(4.10). Отже,

$$TSS = \sum_i (\hat{y}_i - \bar{y})^2 + \sum_i e_i^2 = RSS + ESS, \qquad (4.12)$$

де $RSS = \sum_i (\hat{y}_i - \bar{y})^2 = \hat{\beta}^2 \sum_i (\hat{x}_i - \bar{x})^2$ – сума квадратів відхилень лінії регресії від середнього значення Y, а $ESS = \sum_i e_i^2$ – сума квадратів залишків[4]. Геометричний зміст такої декомпозиції дисперсії Y ілюструється рисунком 4.4.

На основі останнього рівняння можна сформувати інтегральний показник узгодженості рівняння регресії з даними, відомий як *коефіцієнт детермінації* (англ. *coefficient of determination*):

$$R^2 = \frac{RSS}{TSS} = 1 - \frac{ESS}{TSS}. \qquad (4.13)$$

[4] Назви змінних обумовлені скороченнями від англомовних термінів Total Sum of Squares, Regression Sum of Squares та Error Sum of Squares.

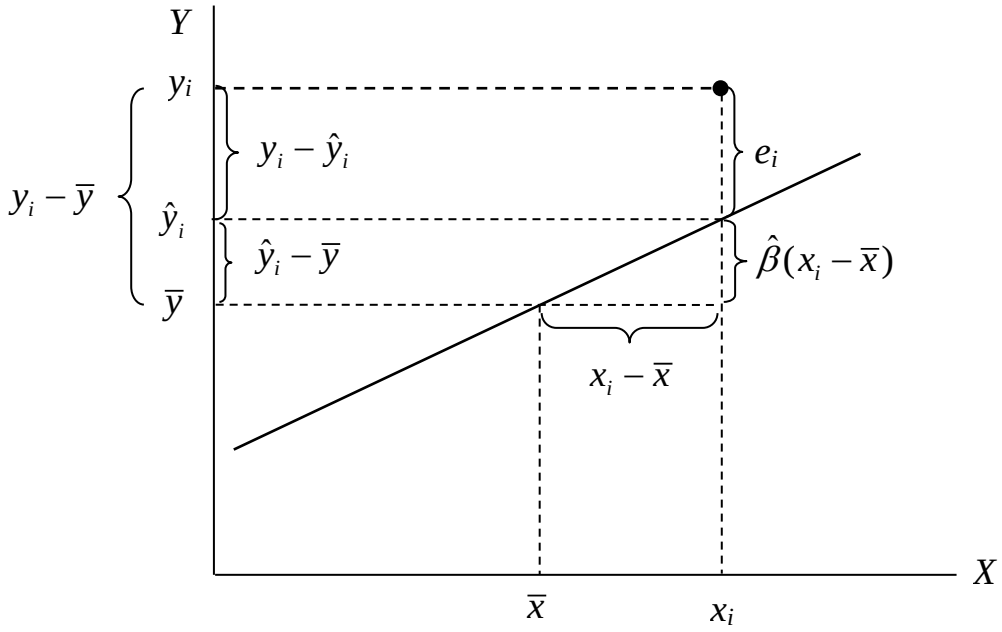


Рис. 4.4. Геометричний зміст декомпозиції дисперсії *Y*

Якщо лінія регресії точно «кладеться» на дані, то сума квадратів залишків дорівнюватиме нулю, і коефіцієнт детермінації дорівнюватиме одиниці. Якщо ж лінія регресії буде горизонтальною, тобто знання *X* ніяк не допомагає при визначенні *Y*, то R^2 дорівнюватиме нулю. Чим ближче R^2 до одиниці, тим краще рівняння регресії описує дані. Для даних з табл. 3.1 коефіцієнт детермінації становить 0,6533, тобто приблизно 65% варіації часу очікування може бути пояснено часом доби.

Нарешті, відзначимо, що підстановка значень коефіцієнтів МНК (4.11) в формулу (4.13) призводить до наступного корисного співвідношення:

$$R^2 = \frac{RSS}{TSS} = \hat{\beta}^2 \frac{\sum_i (\hat{x}_i - \bar{x})^2}{\sum_i (y_i - \bar{y})^2} = \left(r_{xy} \frac{s_y}{s_x} \right)^2 \frac{s_x^2}{s_y^2} = r_{xy}^2 . \qquad (4.14)$$

Перевіримо цю формулу для прикладу з табл. 3.1: $r_{xy}^2 = (-0{,}8083)^2 = 0{,}6533 = R^2$.

Деякі важливі властивості оцінок МНК випливають безпосередньо із системи нормальних рівнянь (4.9)–(4.10). Зокрема:

1) лінія регресії проходить через центральну точку даних $(\bar{x}, \bar{y})$:

$$\hat{y}(\bar{x}) = \hat{\alpha} + \hat{\beta}\bar{x} = \underbrace{\bar{y} - \hat{\beta}\bar{x}}_{\hat{\alpha}} + \hat{\beta}\bar{x} = \bar{y};$$

2) сума залишків і їх середнє значення дорівнює нулю (4.9);

3) залишки некорельовані з регресорами (4.10).

Інші властивості потребують припущень щодо структури моделі (4.5). Наприклад, якщо $\varepsilon \sim N(0, \sigma^2)$, $\mathrm{cov}(\varepsilon, x) = 0$, то $\hat{\beta} \sim N\left(\beta, \frac{\sigma^2}{(n-1)s_x^2} \right)$, на базі чого

можна будувати довірчі інтервали, перевіряти гіпотези тощо. Детальний розгляд статистичних властивостей оцінок МНК відкладемо до наступного розділу, де це буде зроблено у найбільш загальному вигляді. Зауважимо тільки, що з припущеннями треба бути обережними: так, у прикладі з часом очікування потягу немає жодних підстав вважати похибку нормально розподіленою.

Також слід бути обережним з прогнозуванням на базі побудованої регресійної моделі, особливо коли треба зробити прогноз за межами спостережуваних значень незалежної змінної (англ. *out-of-sample forecasts*). Наприклад, із моделі, наведеної на рис. 4.3, випливає, що у 22:00 середній час очікування потягу буде негативним, що, вочевидь, не може бути вірним.

4.3 Залежність кількісних змінних від якісних: дисперсійний аналіз

Інший погляд на ті ж самі дані з рис. 4.1 може полягати в тому, що між часом очікування та часом доби немає лінійної залежності, а просто інтервал руху між потягами є різним вранці і ввечері (що виглядає більш правдоподібним із суто логічних міркувань). На рис. 4.5 дані про час очікування потягу розбиті на дві групи: до та після 13:00. До цього рисунку також додано середній час очікування у першій групі спостережень (зелена лінія), у другій (синя), та загальне середнє (чорна).

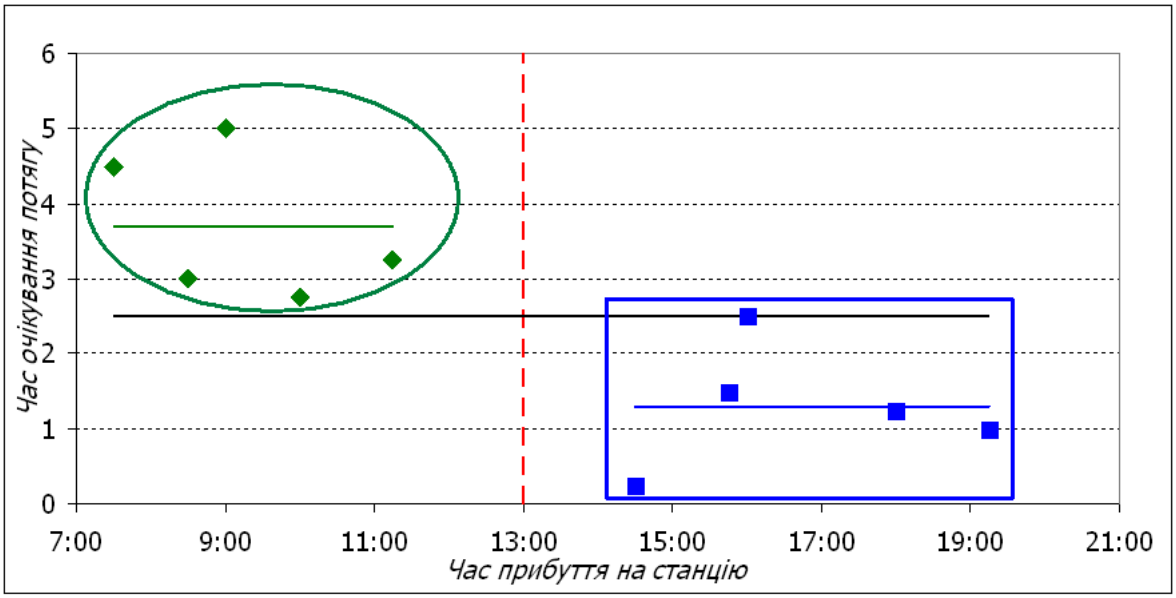


Рис. 4.5. Дані з табл. 3.1, розбиті на дві групи залежно від часу доби

Зауважимо, що усередині кожної з виділених груп немає помітної неозброєним оком залежності між часом доби та часом очікування потягу.

Різниця між середніми значеннями двох груп, виділених на рис. 4.5, виглядає суттєвою. На рис. 3.3 наводилось інше групування даних табл. 3.1 – за ознакою лінії метрополітену, де різниця в середніх виглядала незначною. Для того, щоб чітко визначити, наскільки значущими є різниці між групами, використовується дисперсійний аналіз.

Дисперсійний аналіз (англ. *analysis of variance* або скорочено ANOVA) – це розроблений Р. Фішером статистичний метод, спрямований на пошук залежностей в даних шляхом дослідження значущості різниць в середніх значеннях. Ідея методу полягає в тому, щоб визначити, яка частка загальної варіації даних обумовлена впливом систематичних факторів. В дисперсійному аналізі вважається, що залежна змінна є кількісною (вимірюється за шкалою інтервалів або відношень), а незалежні – номінативними, тобто визначають приналежність спостереження до певної групи. Дисперсійний аналіз базується на наступній декомпозиції загальної варіації даних.

Нехай загальна вибірка розбита на m груп $A_1,...,A_m$, що не перетинаються. Позначимо через $\bar{x}$ середнє значення досліджуваної ознаки у повній вибірці, а через $\bar{x}_j$ – відповідне середнє значення в групі A_j. Варіація ознаки у повній виборці, яка у дисперсійному аналізі вимірюється сумою квадратів відхилень від середнього може бути подана як:

$$SS_{total} = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{j=1}^{m}\sum_{i\in A_j}(x_i - \bar{x})^2 = \sum_{j=1}^{m}\sum_{i\in A_j}(x_i - \bar{x}_j + \bar{x}_j - \bar{x})^2 =$$

$$= \sum_{j=1}^{m}\sum_{i\in A_j}(x_i - \bar{x}_j)^2 + 2\sum_{j=1}^{m}\sum_{i\in A_j}(x_i - \bar{x}_j)(\bar{x}_j - \bar{x}) + \sum_{j=1}^{m}\sum_{i\in A_j}(\bar{x}_j - \bar{x})^2.$$

Середній доданок в останній формулі дорівнює нулю:

$$\sum_{j=1}^{m}\sum_{i\in A_j}(x_i - \bar{x}_j)(\bar{x}_j - \bar{x}) = \sum_{j=1}^{m}(\bar{x}_j - \bar{x})\sum_{i\in A_j}(x_i - \bar{x}_j) = \sum_{j=1}^{m}(\bar{x}_j - \bar{x})(n_j\bar{x}_j - n_j\bar{x}_j) = 0,$$

де n_j – кількість спостережень в групі A_j. Отже,

$$SS_{total} = \sum_{j=1}^{m}SS_j + \sum_{j=1}^{m}n_j(\bar{x}_j - \bar{x})^2, \qquad (4.15)$$

де SS_j – сума квадратів відхилень від середнього *j*–ї групи.

Першій доданок в формулі (4.15) називається *внутрішньогруповою варіацією* (англ. *within-group sum of squares*, SS_{within}). Вона відображає випадкову варіацію, тобто ту частину варіації, яка не залежить від фактору, на базі якого здійснювалось групування. Другий доданок в формулі (4.15) називається *міжгруповою варіацією* (англ. *between-group sum of squares*, $SS_{between}$). Вона характеризує систематичну варіацію, яка обумовлюється впливом групувальної ознаки. Таким чином,

$$SS_{total} = SS_{within} + SS_{between} \qquad (4.16)$$

(зверніть увагу на подібність цього співвідношення до декомпозиції дисперсії (4.12) у парній лінійній регресії).

Аналогічно розкладаються ступені свободи:

$$df_{total} = df_{within} + df_{between}, \qquad (4.17)$$

де $df_{total} = n-1,\ df_{within} = \sum_{j=1}^{m}(n-1) = n-m,\ df_{between} = m-1.$

Дисперсія кожного із джерел варіації, яка в термінології дисперсійного аналізу називається «середнім квадратом» (англ. mean square, скорочено MS) визначається тоді як відношення суми квадратів до відповідної кількості ступенів свободи:

$$MS_{total} = \frac{SS_{total}}{n-1};\ MS_{within} = \frac{SS_{within}}{n-m};\ MS_{between} = \frac{SS_{between}}{m-1}. \qquad (4.18)$$

Якщо додати припущення про нормальність розподілу ознак всередині груп, то відношення міжгрупової дисперсії до внутрішньогрупової

$$F = \frac{MS_{between}}{MS_{within}} \qquad (4.19)$$

матиме розподіл Фішера із $df_1 = m-1, df_2 = n-m$ (див. п.2.5). Якщо воно перевищує критичний рівень для бажаного рівня значущості, то відмінності між групами не можуть бути пояснені суто випадковими факторами.

Приклад 4.1. Для даних табл. 3.1, згрупованих згідно рис. 4.5:

A_1 = {4,5; 3; 5; 2,75; 3.25}; A_2 = {0,25; 1,5; 2,5; 1,25; 1}; $n_1 = n_2 = 5$;

$\overline{x} = 2,5;\ \overline{x}_1 = 3,7;\ \overline{x}_2 = 1,3;$

$SS_{total} = 21;\ SS_1 = 3,925;\ SS_2 = 2,675;\ SS_{within} = SS_1 + SS_2 = 6,6;\ SS_{between} = 14,4;$

$$MS_{total} = \frac{21}{10-1} \approx 2,34;\ MS_{within} = \frac{6,6}{10-2} = 0,825;\ MS_{between} = \frac{14,4}{2-1} = 14,4;$$

$$F = \frac{14,4}{0,825} \approx 17,45 > F_{crit}(\alpha = 0,05, df_1 = 1, df_2 = 8) \approx 5,32.$$

Таким чином, різниця між групами є статистично значущою (якщо вважати дані нормально розподіленими, що для даних табл. 3.1 є сумнівним). ■

Подібність формул (4.12) та (4.16) наводить до думки про спорідненість регресійного та дисперсійного аналізу. Насправді, у випадку двох груп *A* та *B*, як у прикладі 4.1, дисперсійний аналіз може бути зведений до моделі простої регресії (4.5), якщо в якості регресора використати двійкову змінну – індикатор приналежності спостереження до певної групи:

$$D_i = I\{i \in A\} = \begin{cases} 1, i \in A \\ 0, i \notin A \end{cases}.$$

В регресійному аналізі такі змінні називають *фіктивними* (англ. *dummy variables*). Рівняння

$$y_i = \alpha + \beta D_i + \varepsilon_i$$

зводиться тоді до системи

$$y_i = \begin{cases} \alpha + \varepsilon_i, i \notin A \\ \alpha + \beta + \varepsilon_i, i \in A \end{cases},$$

тобто залежна змінна буде описана кусково–лінійною функцією. Змістовно коефіцієнт β дорівнює різниці між груповими середніми, то ж про наявність систематичних відмінностей між двома групами можна судити на основі статистичної значущості оцінки $\hat{\beta}$. На рис. 4.6 наведено результати оцінки рівняння регресії для даних табл. 3.1, де в якості регресора обрано $I\{i \in A_2\}$.

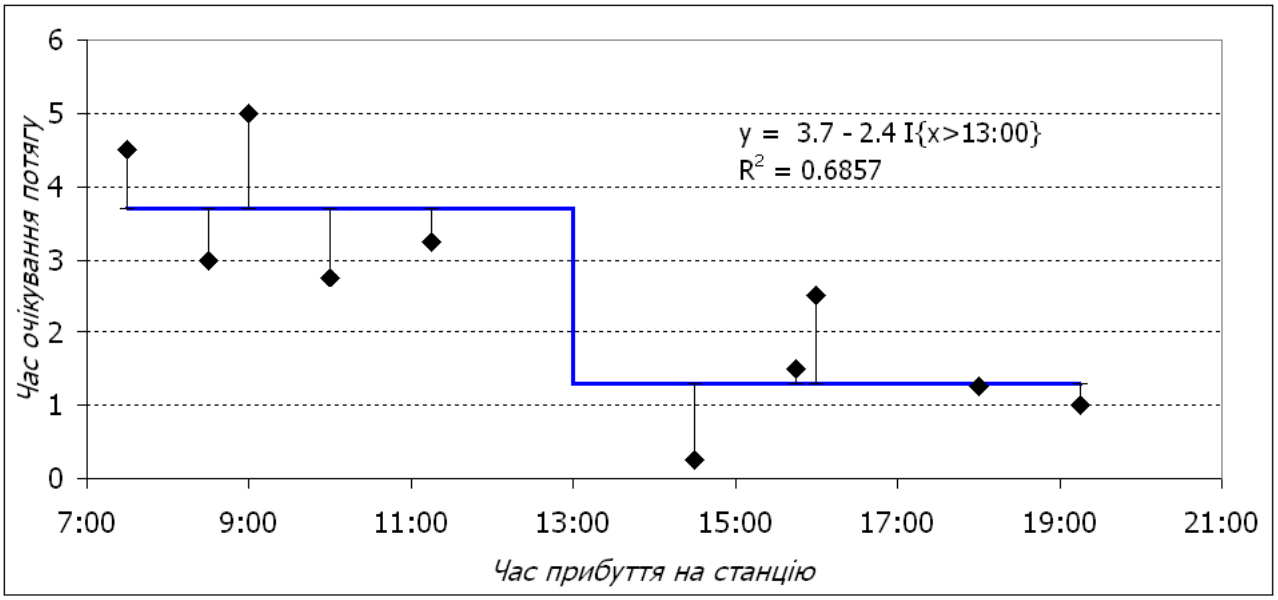


Рис. 4.6. Графік лінійної регресії для даних з табл. 3.1 із використанням фіктивної змінної

Легко бачити, що оцінки значень залежної змінної в останньому рівнянні відповідають груповим середнім: 3,7 для $i \in A_1$ і $3,7 - 2,4 = 1,3$ для $i \in A_2$. Коефіцієнт детермінації становить $R^2 \approx 0,69$, що трохи краще, ніж для регресії на рис. 4.3. Більш вагомим плюсом є те, що модель із фіктивною змінною не призводить до абсурдних прогнозів за межами діапазону варіації вибірки.

При трьох та більше групах спостережень дисперсійний аналіз не можна звести до простої регресійної моделі. Для врахування приналежності спостереження до певної групи знадобиться більш, ніж одна змінна, але тоді регресія перестане бути парною. Регресійна модель з декількома регресорами називається множинною. Ця модель буде докладно розглянута у наступному розділі.

***

Уважний читач може помітити, що залишається розглянути випадок, коли незалежна змінна є кількісною, а залежна – категоріальною. Така ситуація є окремим випадком задачі класифікації. Ці задачі грають надзвичайно важливу роль у інтелектуальному аналізі даних. Для їх вирішення розроблено величезну кількість методів і навіть їх стислий огляд вимагатиме окремого розділу. Перед цим варто ознайомитись із загальною теорією лінійної регресії.

Контрольні запитання

1. Для чого використовуються таблиці спряженості?
2. Що таке маргінальні частоти і як їх знайти за таблицею спряженості?
3. Як знаходяться умовні ймовірності у таблиці спряженості?
4. Як має виглядати таблиця спряженості при незалежності змінних?
5. Як перевірити гіпотезу про незалежність двох категоріальних змінних?
6. Для чого використовується і як розраховується вибірковий коефіцієнт кореляції?
7. Дайте визначення поняття «регресія».
8. В чому полягає сутність методу найменших квадратів?
9. Наведіть формули для оцінок коефіцієнтів у рівнянні парної лінійної регресії.
10. Як оцінки параметрів парної лінійної регресії пов'язані із коефіцієнтом кореляції між залежною та незалежною змінною?
11. Назвіть основні властивості оцінок МНК?
12. Як визначається і що характеризує коефіцієнт детермінації?
13. На які компоненти розкладається варіація залежної змінної у простій регресійній моделі?
14. Як коефіцієнт детермінації пов'язаний із коефіцієнтом кореляції між залежною та незалежною змінною?
15. Для чого використовується дисперсійний аналіз?
16. На які компоненти розкладається варіація досліджуваної змінної у моделі дисперсійного аналізу?
17. Дайте визначення внутрішньогрупової та міжгрупової дисперсії.
18. Як перевіряється наявність систематичних відмінностей між групами у моделі дисперсійного аналізу?
19. Що таке фіктивні змінні і для чого вони використовуються?
20. Як виконується дисперсійний аналіз за допомогою регресійної моделі?

Завдання для самостійної роботи

4.1. В таблиці наводяться результати опитування клієнтів компанії кабельного телебачення щодо їх улюблених телепрограм.

		Програма		
		Серіали	Спорт	Музика
Стать	Жінки	20	15	25
	Чоловіки	10	20	10

а) Створіть таблицю сумісного розподілу змінних *Стать* та *Програма*, а також ряди граничного розподілу для цих змінних.

б) Визначте умовний розподіл статі телеглядача в залежності від обраної програми і улюбленої програми в залежності від статі.

в) Як виглядала б ця таблиця, якщо б уподобання не залежали від статі?

г) Перевірте гіпотезу про незалежність уподобань телеглядачів від їх статі за допомогою критерію хі-квадрат.

4.2. В таблиці наводяться результати студентів групи на двох модульних контрольних роботах (за 100-бальною шкалою).

№	1	2	3	4	5	6	7	8	9	10	11	12
МК1	72	50	81	74	94	86	59	83	65	33	88	81
МК2	84	63	77	78	90	75	62	79	77	52	80	90

а) Побудуйте діаграму розсіяння для змінних *МК1* та *МК2*. Чи виглядає доречним припущення про наявність лінійного зв'язку між цими змінними?

б) Знайдіть вибірковий коефіцієнт кореляції між змінними.

в) Оцініть коефіцієнти рівняння $MK2 = a + bMK1 + \varepsilon$ за методом найменших квадратів і побудуйте його на тому ж графіку. Знайдіть коефіцієнт детермінації R^2. Про що свідчить отримане значення?

г) Спрогнозуйте оцінку на другій контрольній роботі для студента, який отримав 98 балів на першій.

д) В якому діапазоні будуть знаходитись прогнозні значення результатів другої контрольної роботи?

4.3. З рівняння $y = \alpha + \beta x$ випливає, що $x = -\frac{\alpha}{\beta} + \frac{1}{\beta} y$. Нехай b – оцінка кутового коефіцієнту в регресії y на x, а d – оцінка кутового коефіцієнту в регресії x на y (обидві оцінки отримані за методом найменших квадратів). Покажіть, що $d = \frac{1}{b}$ тоді і тільки тоді, коли $R^2 = 1$.

4.4. Розглянемо модель регресії на константу $y_i = \alpha + \varepsilon_i$, $i = 1,...,n$, де ε_i – незалежні випадкові величини із нормальним розподілом $N(0, \sigma^2)$.

а) Знайдіть оцінки МНК для α та σ^2.

б) Знайдіть дисперсію МНК–оцінки $\hat{\alpha}$.

в) Покажіть, що статистика $\frac{\hat{\alpha} - \alpha}{s_{\hat{\alpha}}} \sim t(n-1)$.

г) Чому буде дорівнювати коефіцієнт детермінації R^2?

4.5. В таблиці наведені результати ЗНО для випадково обраних випускників чотирьох шкіл.

Перевірте гіпотезу про рівність середніх для всіх чотирьох шкіл на 5% рівні значущості.

Школа 1	Школа 2	Школа 3	Школа 4
166	164	181	147
173	157	163	155
185	165	170	162
190	146	152	140
	167	159	153
			141

4.6. Вибірка містить дані про заробітну плату працівників двох професій, А та Б. Вибірка складається з 60 спостережень для працівників професії А та 40 спостережень для працівників професії Б. Середня заробітна плата для всієї вибірки становить 850 у. о., а для професії Б – 1000 у. о. Розглянемо регресію заробітної плати на константу та індикатор приналежності до професії Б.

а) Чому будуть дорівнювати МНК оцінки константи та коефіцієнта при індикаторній змінній?

б) СКВ заробітної плати для професії А складає 200 у.о., а для професії Б – 300 у.о. Знайдіть оцінки дисперсії коефіцієнтів регресії та коваріації між ними. Знайдіть значення коефіцієнта детермінації R^2.

5. МНОЖИННА ЛІНІЙНА РЕГРЕСІЯ

5.1 Загальна форма лінійної регресії

У більшості випадків є багато змінних, які потенційно можуть бути корисними при визначенні залежної змінної. Наприклад, попит на прохолодні напої може залежати від їх ціни, пори року, температури повітря, цін на продукцію конкурентів; артеріальний тиск залежить від віку людини, її ваги, статі, наявності певних хронічних захворювань тощо. Модель множинної регресії дозволяє вивчати зв'язок залежної змінної із декількома незалежними.

Загальна форма *множинної лінійної регресії* (англ. *multiple linear regression*) має вигляд:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} ... + \beta_m x_{mi} + \varepsilon_i . \qquad (5.1)$$

Вектор невідомих параметрів $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_m)$ має розмірність $k = m+1$. Для визначення цього вектору необхідно мати інформацію про значення змінних $Y, X_1, ..., X_m$. Можливі такі випадки.

1. Якщо обсяг вибірки n є таким, що $n < k$, то система рівнянь, що визначає регресійну модель (5.1) буде *невизначеною* (англ. *unidentified*) – в ній недостатньо даних для відновлення β.

2. Якщо обсяг вибірки $n = k$, то задача (5.1) зводиться до вирішення системи з n рівнянь із n невідомими (елементами вектору **β**), яка має унікальне рішення за умови, що змінні $X_1, ..., X_m$ є лінійно незалежними. В цьому випадку регресійна модель буде *визначеною* (англ. *just identified*), але неможливо встановити будь-які її статистичні властивості, оскільки відсутня база для порівняння оцінок моделі із реальними даними.

3. Найбільш типовим є випадок, де спостерігають $n > k$ точок даних. В такому випадку система рівнянь (5.1) є *перевизначеною* (англ. *overidentified*) і знайти її точне рішення неможливо. Треба знайти таке значення **β**, яке за певним критерієм підходило б найкраще в якості наближеного рішення. При цьому «зайві» спостереження використовуються для оцінки узгодженості регресійної моделі з даними. Кількість таких спостережень $n - k$ становить кількість ступенів свободи регресійного рівняння.

Найбільш поширеним способом рішення системи рівнянь (5.1) є метод найменших квадратів, де критерієм оптимальності є мінімальна евклідова відстань між реальними і прогнозованими згідно моделі значеннями залежної змінної.

Систему рівнянь (5.1) можна записати більш компактно у матричній формі. Для цього запишемо її по рядках:

$$\begin{bmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{bmatrix} = \beta_0 \begin{bmatrix} 1 \\ 1 \\ \dots \\ 1 \end{bmatrix} + \beta_1 \begin{bmatrix} x_{11} \\ x_{21} \\ \dots \\ x_{n1} \end{bmatrix} + \dots + \beta_m \begin{bmatrix} x_{1m} \\ x_{2m} \\ \dots \\ x_{nm} \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \dots \\ \varepsilon_n \end{bmatrix}$$

і введемо наступні позначення:

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{bmatrix}; \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \dots \\ \beta_m \end{bmatrix}; \mathbf{X} = \begin{bmatrix} 1 & x_{11} & \dots & x_{1m} \\ 1 & x_{21} & \dots & x_{2m} \\ \dots & \dots & \dots & \dots \\ 1 & x_{n1} & \dots & x_{nm} \end{bmatrix}; \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \dots \\ \varepsilon_n \end{bmatrix}.$$

Тоді

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}. \tag{5.2}$$

Формула (5.2) визначає матричну форму лінійної моделі. Слід підкреслити, що мається на увазі лінійність стосовно параметрів $\boldsymbol{\beta}$. Предиктори $X_1,...,X_m$ можуть бути будь-якими лінійними або нелінійними функціями відомих факторів. Наприклад, якщо i є номером спостереження, то $i, i^2, i^3, 1/i, \ln i$ є цілком припустимими регресорами.

В моделі (5.2) вектор $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_m)^T$ розмірності $k\times1$ є невідомим. Проте, для кожного його можливого значення b можна визначити прогнозоване значення $\mathbf{y}(\mathbf{b}) = \mathbf{X}\mathbf{b}$ і його відхилення від справжнього значення $\mathbf{e}(\mathbf{b}) = \mathbf{Y} - \mathbf{X}\mathbf{b}$ (вектор залишків). Метод найменших квадратів полягає у знаходженні такого вектору b, який мінімізував би суму квадратів цих відхилень, тобто евклідову відстань між справжніми і прогнозованими значеннями:

$$S(\mathbf{b}) = \sum_{i=1}^{n} e_i^2(\mathbf{b}) = \mathbf{e}^{\mathbf{T}}(\mathbf{b})\mathbf{e}(\mathbf{b}) = (\mathbf{y} - \mathbf{X}\mathbf{b})^{\mathbf{T}}(\mathbf{y} - \mathbf{X}\mathbf{b}) \underset{\mathbf{b}}{\rightarrow} \min. \tag{5.3}$$

Із матричної алгебри $(\mathbf{AB})^{\mathbf{T}} = \mathbf{B}^{\mathbf{T}}\mathbf{A}^{\mathbf{T}}$. Розкривши дужки в формулі (5.3) і застосувавши останню формулу, маємо:

$$S(\mathbf{b}) = (\mathbf{y} - \mathbf{X}\mathbf{b})^{\mathbf{T}}(\mathbf{y} - \mathbf{X}\mathbf{b}) = \mathbf{y}^{\mathbf{T}}\mathbf{y} - 2\mathbf{b}^{\mathbf{T}}\mathbf{X}^{\mathbf{T}}\mathbf{y} + \mathbf{b}^{\mathbf{T}}\mathbf{X}^{\mathbf{T}}\mathbf{X}\mathbf{b}. \tag{5.4}$$

Необхідною умовою мінімуму функції $S(\mathbf{b})$ є рівність нулю її похідної. Виконавши векторне диференціювання формули (5.4) по $\mathbf{b}$, отримаємо систему нормальних рівнянь:

$$-2\mathbf{X}^{\mathbf{T}}\mathbf{y} + 2\mathbf{X}^{\mathbf{T}}\mathbf{X}\mathbf{b} = \mathbf{0}, \tag{5.5}$$

рішенням якої і буде оцінка вектору параметрів $\boldsymbol{\beta}$ за МНК:

$$\hat{\boldsymbol{\beta}} = \mathbf{b} = (\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathbf{T}}\mathbf{y}. \tag{5.6}$$

Для того, щоб скористатися формулою (5.6), матриця $\mathbf{X}^{\mathbf{T}}\mathbf{X}$ має бути оберненою. З лінійної алгебри відомо, що для цього матриця $\mathbf{X}$ повинна мати повний ранг, тобто щоб жоден з її стовпців не міг бути виражений як лінійна комбінація інших. Порушення цієї умови відомо як *мультиколінеарність* (англ.

multicollinearity). Така проблема може виникнути, коли деякі з незалежних змінних пов'язані між собою відношенням тотожності. Наприклад, якщо x_1 – споживання електроенергії в грошових одиницях, а x_2 – її споживання у кіловат–годинах, то за постійної ціни електроенергії p $x_1 = px_2$. В цьому разі використання обох змінних x_1, x_2 як незалежних наведе до мультиколінеарності, і одну з цих змінних слід виключити з-поміж регресорів.

Часто зустрічається ситуація, коли деякі незалежні змінні виявляються сильно корельованими між собою. В останньому прикладі для змінних x_1, x_2 така ситуація матиме місце, якщо ціна електроенергії змінюватиметься, але незначно. В таких випадках (відомих як «майже мультиколинеарність») оцінки МНК за формулою (5.6) можуть бути обчислені, але їх точність вийде низькою. Це теж надає підстави для відкидання однієї з «проблемних» змінних.

Матриця **X** може вважатися як детермінованою, так і випадковою. У природничих науках спостереження залежної змінної часто отримуються у результаті експерименту, де значення регресорів є невипадковими і контрольованими (ба більше, існує окрема теорія, як обирати ці значення найкраще). В соціальних науках та у бізнес–середовищі рідко є можливість проводити контрольовані експерименти, отже регресори слід вважати випадковими. Формула МНК «працює» в обох випадках, однак доведення її властивостей простіше у разі невипадкових регресорів, то ж ми теж будемо так вважати.

Деякі властивості оцінок МНК випливають безпосередньо із системи нормальних рівнянь. Так, із формули (5.5)

$$-\mathbf{X^T y} + \mathbf{X^T X b} = -\mathbf{X^T}(\mathbf{y} - \mathbf{X b}) = -\mathbf{X^T e} = \mathbf{0}.$$

Це означає, що для кожного стовпця $\mathbf{x}_j$ матриці **X** виконується $\mathbf{x}_j^{\mathbf{T}}\mathbf{e} = 0$. Зокрема, якщо перший стовпець матриці **X** є стовпчиком одиниць (тобто в рівнянні регресії (5.1) є константа), то сума залишків регресії має дорівнювати нулю. Також у цьому випадку $\bar{y} = \bar{\mathbf{x}}^{\mathbf{T}}\mathbf{b}$, де $\bar{\mathbf{x}}$ – вектор середніх значень регресорів, тобто поверхня регресії проходить через середні значення всіх змінних[5]. Ці властивості аналогічні формулам (4.9), (4.10) для парної регресії.

5.2 Статистичні властивості оцінок методу найменших квадратів

Оскільки в рівнянні (5.2) вектор **ε** є випадковим, то вектор значень залежної змінної **y**, а отже і вектор оцінок параметрів **b**, теж слід вважати випадковими. Для аналізу їх властивостей слід зробити певні припущення щодо властивостей послідовності випадкових похибок **ε**. Найчастіше вважається, що вона створює так званий *білий шум* (англ. *white noise*). Так називається вектор випадкових величин **ε**, який відповідає наступним критеріям:

[5] Це випливає з рівняння $[1 \quad \ldots \quad 1](\mathbf{y} - \mathbf{X b}) = 0$, якщо поділити його на n та розкрити дужки.

П1. Математичне сподівання похибки дорівнює нулю:

$$M[\varepsilon_i] = 0,\ i = 1,...,n. \tag{5.7}$$

П2. Дисперсія похибки є постійною для всіх елементів вибірки:

$$D[\varepsilon_i] = M[\varepsilon_i^2] = \sigma^2 = const,\ i = 1,...,n. \tag{5.8}$$

Ця властивість відома як *однорідність дисперсії* або *гомоскедастичність* (англ. *homogeneity of variance* або *homoscedasticity*). Порушення умови (5.8) називається *гетероскедастичністю* (англ. *heteroscedasticity*).

П3. Відсутність автокореляції (англ. *autocorrelation*) :

$$\mathrm{cov}(\varepsilon_i, \varepsilon_j) = M[\varepsilon_i \varepsilon_j] = 0,\ i \neq j;\ i, j = 1,...,n. \tag{5.9}$$

Останні дві умови можна поєднати і записати більш компактно у матричній формі із використанням *дисперсійно–коваріаційної матриці* (англ. *variance-covariance matrix*). Для випадкового вектору $\mathbf{x}$ так називається матриця

$$D[\mathbf{x}] = M[(\mathbf{x} - \overline{\mathbf{x}})(\mathbf{x} - \overline{\mathbf{x}})^{\mathbf{T}}] = \begin{bmatrix} D[x_1] & \mathrm{cov}(x_1, x_2) & ... & \mathrm{cov}(x_1, x_n) \\ \mathrm{cov}(x_2, x_1) & D[X_2] & ... & \mathrm{cov}(x_2, x_n) \\ ... & ... & ... & ... \\ \mathrm{cov}(x_n, x_1) & \mathrm{cov}(x_n, x_2) & ... & D[x_n] \end{bmatrix}.$$

З використанням цієї позначки умови (5.8) і (5.9) можна об’єднати як

$$D[\boldsymbol{\varepsilon}] = M[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^{\mathbf{T}}] = \sigma^2 \mathbf{I}, \tag{5.10}$$

де $\mathbf{I}$ – одинична матриця розмірності $n \times n$.

Ці три умови часто замінюються наступним припущенням:

П4. Випадкові похибки мають нормальний розподіл:

$$\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}), \tag{5.11}$$

яке не є обов’язковим для можливості застосування МНК, але дозволяє провести значно глибший аналіз оцінок параметрів та їх значущості. Також це припущення є досить розумним, коли нема вагомих підстав вважати інакше.

За цих припущень можна довести наступні твердження.

1. За умови П1 оцінка МНК є незміщеною. Насправді,

$$M[\hat{\boldsymbol{\beta}}] = M[(\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathbf{T}}\mathbf{y}] = M[(\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathbf{T}}(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon})] = \boldsymbol{\beta} + (\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathbf{T}} M[\boldsymbol{\varepsilon}] = \boldsymbol{\beta}.$$

2. За умов П1–П3 дисперсійно–коваріаційна матриця оцінки $\hat{\boldsymbol{\beta}}$ становить:

$$D[\hat{\boldsymbol{\beta}}] = \sigma^2 (\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-1}. \tag{5.12}$$

Дійсно,

$$D[\mathbf{b}] = M[(\mathbf{b} - \boldsymbol{\beta})(\mathbf{b} - \boldsymbol{\beta})^{\mathbf{T}}] = M[(\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathbf{T}}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^{\mathbf{T}}\mathbf{X}(\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-1}] =$$

$$= (\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathbf{T}} M[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^{\mathbf{T}}]\mathbf{X}(\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-1} = (\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathbf{T}} \sigma^2 \mathbf{I} \mathbf{X}(\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-1} = \sigma^2 (\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-1}.$$

3. За умов П1–П3 оцінка методу найменших квадратів (5.6) є найкращою незміщеною лінійною оцінкою вектора параметрів $\boldsymbol{\beta}$. Це твердження відомо як *теорема Гауса–Маркова*, доведення якої можна знайти, наприклад, в [1]. Зокрема, теорема стверджує, що дисперсія оцінки МНК (5.12) є мінімальною

можливою серед усіх лінійних оцінок. Якщо ж додати умову П4, то оцінка МНК вийде найкращою серед усіх можливих незміщених оцінок вектора **β**.

Англійською мовою теорема Гауса–Маркова більш відома у меметичній формі “OLS is BLUE”, що є скороченням від Best Linear Unbiased Estimator.

Слід зауважити, що якщо умова П4 не виконується, то можуть існувати більш ефективні (нелінійні) оцінки параметрів **β**, ніж оцінки МНК.

4. Справжнє значення дисперсії похибки σ^2 в формулі (5.12), як правило, невідомо. Незміщеною оцінкою цього параметру є величина

$$s^2 = \frac{\mathbf{e}^{\mathbf{T}}\mathbf{e}}{df}, \tag{5.13}$$

де $df = n-(m+1)$ – кількість ступенів свободи регресії.

Із використанням формули (5.13) отримаємо наступну формулу для стандартних похибок вектору коефіцієнтів регресії $\hat{\boldsymbol{\beta}}$:

$$se(\hat{\boldsymbol{\beta}}) = \sqrt{\mathrm{diag}\left(s^2(\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-1}\right)}, \tag{5.14}$$

де $\mathrm{diag}(A)$ позначає головну діагональ матриці A.

5. За умови П4 відношення

$$t_k = \frac{\hat{\beta}_k - \beta_k}{se(\hat{\beta}_k)} \tag{5.15}$$

має t–розподіл Стьюдента із $df = n-(m+1)$ ступенями свободи. Це надає можливість застосовувати до оцінок коефіцієнтів регресії стандартний інструментарій математичної статистики.

Зокрема, для перевірки гіпотези про незначущість коефіцієнту регресії β_k, тобто гіпотези $H_0 : \beta_k = 0$ проти $H_1 : \beta_k \neq 0$ відношення

$$t_k = \frac{\hat{\beta}_k}{se(\hat{\beta}_k)} \tag{5.16}$$

слід порівняти із критичним значенням $\overline{t}$ при *df* ступенях свободи і бажаному рівні значущості α. Значення $\overline{t}$ визначається за формулою

$$P\{-\overline{t} \le t_k \le \overline{t}\} = 1-\alpha \tag{5.17}$$

і знаходиться із статистичних таблиць (або за допомогою стандартних функцій, які реалізовані в усіх програмних середовищах для вирішення задач прикладної статистики). Аналогічно можуть бути виконані односторонні t–тести.

Із використанням критичного значення $\overline{t}$ можна знайти також довірчий інтервал для коефіцієнту регресії β_k:

$$\hat{\beta}_k - \overline{t} \times se(\hat{\beta}_k) \le \beta_k \le \hat{\beta}_k + \overline{t} \times se(\hat{\beta}_k) \tag{5.18}$$

Слід зазначити, що із зростанням кількості ступенів свободи t–розподіл Стьюдента асимптотично наближується до нормального розподілу, а критичне

значення $\bar{t}$ у двосторонніх тестах – до 1,96 (при $\alpha = 0{,}05$). Тому при значній кількості спостережень можна грубо оцінити довірчий інтервал для коефіцієнту β_k як $\hat{\beta}_k \pm 2 \times se(\hat{\beta}_k)$. Якщо цей інтервал містить нуль, то відповідний коефіцієнт слід вважати статистично незначущим.

6. Як і у випадку парної регресії (4.5), для перевірки загальної узгодженості моделі з даними (адекватності моделі) використовується коефіцієнт детермінації

$$R^2 = \frac{RSS}{TSS} = 1 - \frac{ESS}{TSS} = 1 - \frac{\mathbf{e}^{\mathbf{T}}\mathbf{e}}{(\mathbf{y} - \bar{\mathbf{y}})^{\mathbf{T}}(\mathbf{y} - \bar{\mathbf{y}})}. \qquad (5.19)$$

Цей коефіцієнт приймає можливі значення в інтервалі від 0 до 1, якщо рівняння регресії (5.1) містить константу (тобто перший стовпець матриці регресорів є стовпцем одиниць).

Часто перевіряють гіпотезу про значущість рівняння регресії у цілому, тобто $H_0 : \beta_1 = \beta_2 = ... = \beta_m = 0$. За умови П4 статистика

$$F = \frac{R^2 / m}{(1 - R^2) / (n - (m + 1))} \qquad (5.20)$$

має *F*–розподіл Фішера із ступенями свободи m та $n - (m + 1)$. Гіпотезу H_0 відкидають, якщо значення *F* перевищує критичне.

Чим більше значення R^2, тим краще регресійна модель узгоджується з даними. Хоча об'єктивних підходів для визначення критичного значення R^2 не існує, часто модель вважають «доброю», якщо $R^2 \geq 0{,}75$.

Проте, при використанні R^2 для аналізу узгодженості моделі з даними виникає наступна проблема. Можна довести, що R^2 ніколи не знижується при включенні до рівняння регресії (5.1) додаткової незалежної змінної [1]. Тому виникає спокуса додавати незалежні змінні, доки R^2 не досягне «прийнятного» рівня. При цьому збільшення коефіцієнту детермінації відбувається за рахунок точності вимірювання окремих коефіцієнтів рівняння регресії. Щоб уникнути цієї проблеми, часто використовується скоригований коефіцієнт детермінації $\bar{R}^2$, який можна обчислити за формулою

$$\bar{R}^2 = 1 - \frac{n-1}{n-df}(1 - R^2). \qquad (5.21)$$

На відміну від коефіцієнта детермінації R^2, скоригований коефіцієнт $\bar{R}^2$ може зменшуватись при додаванні в рівняння регресії нових змінних. Можна довести, що $\bar{R}^2$ зменшиться (підвищиться) при виключенні із регресії змінної *x*, якщо *t*–відношення (5.16) для цієї змінної більше (менше) одиниці.

МНК може також застосовуватися для оцінки параметрів функції регресії загального виду $\mathbf{y} = f(\mathbf{X}, \boldsymbol{\beta})$. В цьому разі аналітичну формулу для параметрів рівняння регресії отримати, як правило, неможливо. Визначення параметрів

здійснюється за допомогою чисельної мінімізації функції (5.3). Також значно ускладнюється оцінка стандартних похибок параметрів. Особливості застосування нелінійного МНК детально розглядаються в [25, 37, 39].

5.3 Вибір функціональної форми моделі

Як вже зазначалось, під терміном «лінійна модель» в регресійному аналізі мається на увазі лінійність відносно коефіцієнтів, а не відносно регресорів. Це дає змогу зводити до формулі (5.1) досить складні види функціональної залежності між змінними. Наприклад, поліноміальна модель зв'язку між залежною змінною та предиктором цілком укладається в рамки лінійної моделі

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + ... + \varepsilon_i,$$

яка повністю еквівалентна моделі (5.1), де $x_{1i} = x_i, x_{2i} = x_i^2, x_{3i} = x_i^3, ...$

Однім з найпоширеніших перетворень змінних в регресійному аналізі є їх логарифмування. Це доречно в тих випадках, коли залежність між змінними носить мультиплікативний характер. Наприклад, інфляція проявляється в тому, що ціни мають тенденцію зростати з плином часу. Це може бути виражено наступною залежністю:

$$P_{t+1} = P_t(1+i),$$

де P_t – рівень цін у періоді *t*, а i – темпи інфляції (швидкість зростання цін). Послідовними підстановками отримаємо $P_1 = P_0(1+i);\ P_2 = P_1(1+i) = P_0(1+i)^2; ...$ і для довільного періоду *t*:

$$P_t = P_0(1+i)^t = \alpha_0 \times \alpha_1^t. \qquad (5.22)$$

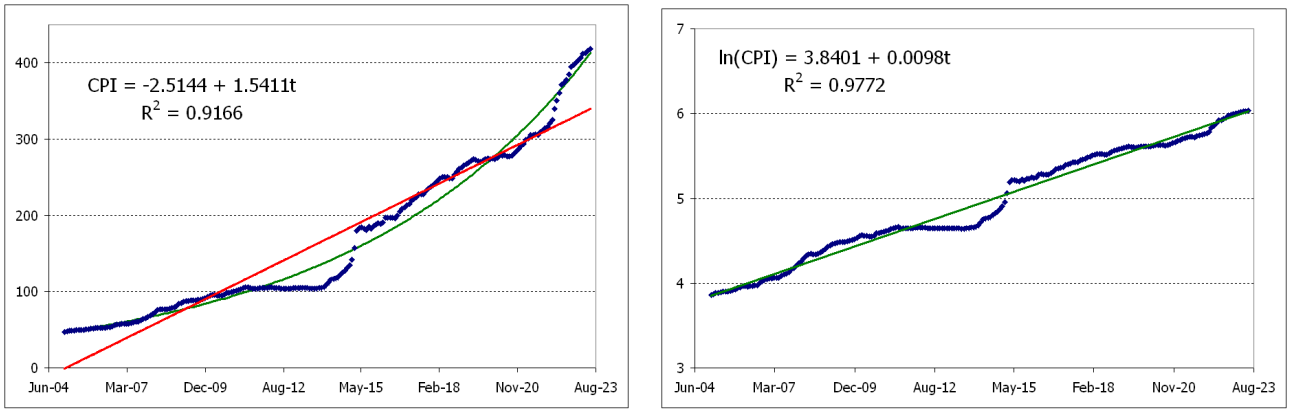
Модель (5.22) не є лінійною відносно параметрів α_0, α_1, але легко зводиться до неї шляхом логарифмування: $\ln P_t = \ln P_0 + t\ln(1+i)$ або

$$\ln P_t = \beta_0 + \beta_1 t, \qquad (5.23)$$

де $\beta_0 = \ln P_0$, $\beta_1 = \ln(1+i) \approx i$ (при невеликих значеннях i останнє наближення випливає з апроксимації Тейлора першого порядку). Це рівняння може бути оцінено за допомогою МНК, а прогнозовані значення отримані як $\hat{P}_t = \exp(\hat{p}_t)$.

Приклад 5.1. Державна служба статистики України [Д1] щомісяця публікує оцінки *індексу споживчих цін* (англ. *consumer price index*, CPI). На рис. 5.1*а* наведені значення цього індексу за період з 2005 до червня 2023 року у відношенні до грудня 2010 р. Червоною лінією на рис. 5.1а показано лінійний *тренд* (англ. trend). Так називається окремий випадок залежності між двома змінними, коли в ролі предиктора виступає час. На рис. 5.1б показані ті ж самі дані, але залежною змінною обрано натуральний логарифм CPI. Легко бачити, що на другому графіку залежність між змінними є набагато ближчою до лінійної, що підтверджується також значеннями коефіцієнта детермінації у відповідних регресіях. Зелена лінія на рис. 5.1 відповідає прогнозованим

значенням CPI, які випливають із регресійної моделі на рис. 5.1*б*. Значення коефіцієнта β_1 = 0,0098 відповідає середнім темпам інфляції біля 1% в місяць. ■



а *б*

Рис. 5.1. Інфляція в Україні за період з січня 2005 р. по червень 2023 р.:
а – індекс споживчих цін; *б* – натуральний логарифм цього показника

Приклад 5.1 ілюструє іншу проблему, яка виникає при описі залежності між змінними лінійною функцією. Регресійна модель, що наведена на рис. 5.1а, прогнозує від'ємне значення CPI для першого періоду: –2,5144 + 1,5411 = –0,9733, але індекс цін може приймати тільки позитивні значення. Використання логарифмів в якості залежної змінної дозволяє уникнути цієї неприємної можливості для змінних , які за визначенням є суто позитивними.

В більш загальному випадку, модель виду

$$Y_i = aX_{1i}^{\beta_1} X_{2i}^{\beta_1} \ldots X_{mi}^{\beta_1} (1+\varepsilon_i) \qquad (5.24)$$

логарифмуванням зводиться до лінійної моделі

$$\ln Y_i = \beta_0 + \beta_1 \ln X_{1i} + \beta_2 \ln X_{2i} \ldots + \beta_m \ln X_{mi} + \varepsilon_i, \qquad (5.25)$$

яка після цього може бути оцінена за допомогою МНК. Таку модель називають *лог–лінійною* (англ. *loglinear*).

Варто сказати кілька слів щодо інтерпретації коефіцієнтів в моделях (5.1) та (5.25). В моделі (5.1) коефіцієнт $\beta_j = \dfrac{dy}{dx_j}$ визначає швидкість зміни залежної змінної у відповідь на зміну *j*-го предиктора. Завдяки лінійній формі залежності можна сказати, що зростання *j*–ї незалежної змінної на одну одиницю призводить до зміни залежної змінної на β_j одиниць. Коефіцієнти β_j мають розмірність, яка дорівнює відношенню одиниць вимірювання змінної *y* до змінної x_j.

В лог–лінійній моделі (5.25)

$$\beta_j = \frac{d \ln Y}{d \ln X_j} = \frac{(dY/Y)\times 100\%}{(dX_j/X_j)\times 100\%} = \frac{X_j}{Y}\frac{dY}{dX_j}. \qquad (5.26)$$

Відношення $(dY/Y)\times 100\%$ характеризує приріст змінної *Y* у відсотках від її поточного рівня. Те саме вірно для змінної у знаменнику формули (5.26).

Отже, можна сказати, що у лог–лінійній моделі збільшення j–ї незалежної змінної на один відсоток наводить до зміни залежної змінної на β_j відсотків.

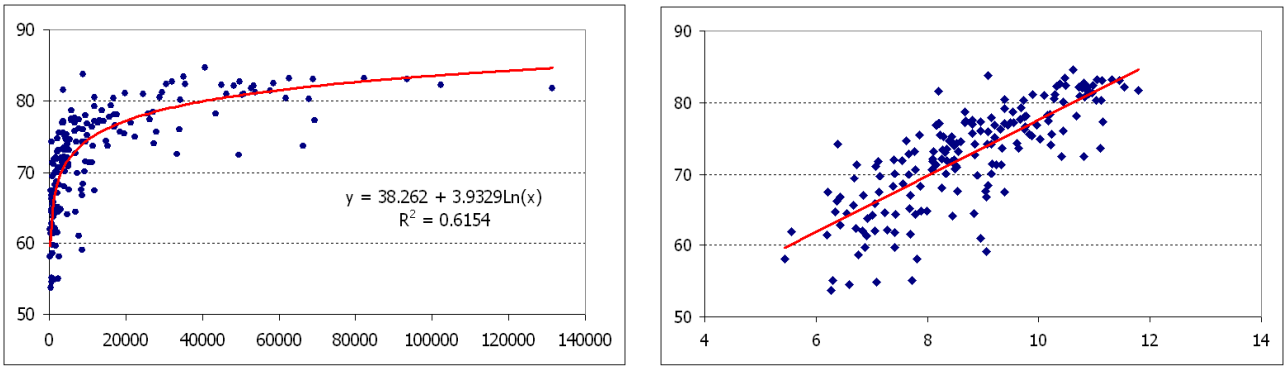
Величина, що визначається формулою (5.26), називається *еластичністю* (англ. *elasticity*) змінної Y по відношенню до змінної X_j. Отже, коефіцієнти лог–лінійної регресії (5.25) можна інтерпретувати як еластичності. Еластичності є безрозмірними величинами, адже і чисельник, і знаменник в формулі (5.26) вимірюються у частках або відсотках.

Можлива також ситуація, коли, як в рівнянні (5.23), лише одна із пов'язаних змінних (залежна або незалежна) перетворена до логарифму. В цьому випадку відповідний коефіцієнт β називають *напів–еластичністю* (англ. *semi-elasticity*). Він надає співвідношення між зміною однієї із змінних (логарифмованої) в процентах та зміною іншої у абсолютних одиницях.

Приклад 5.2. На рис. 5.2а наведено діаграму розсіяння очікуваної тривалості життя при народженні проти ВВП на душу населення для 180 країн світу [Д10]. З графіку видно, що ця залежність не є лінійною. Для порівняно бідних країн зростання доходів суттєво збільшує тривалість життя, але, на жаль, після досягнення певного рівня добробуту ця залежність майже зникає.

На рис. 5.2б незалежна змінна перетворена на логарифм, що наближує залежність до лінійної. На рис. 5.2а червоною лінією показано прогнозовану тривалість життя за напів–логарифмічною моделлю $y \approx 38{,}3 + 3{,}9\ln(x)$.

Це рівняння можна інтерпретувати так: зростання ВВП на душу населення вдвічі (тобто на 100%) збільшує очікувану тривалість життя на 3,9 роки. ■



а *б*

Рис. 5.2. Очікувана тривалість життя як функція від:
а – ВВП на душу населення; *б* – натурального логарифму цього показника

При моделюванні процесів та явищ, для яких характерно насичення, можуть стати в нагоді інші функціональні форми, наприклад, $y = a + bx^{-1}$, $y = a + bx^{-1/2}$ тощо. Узагальненням таких прийомів є *перетворення Бокса–Кокса* (англ. *Box-Cox transformation*):

$$x'_\lambda = \begin{cases} \dfrac{x^\lambda - 1}{\lambda}, \ \lambda \neq 0 \\ \ln(x), \ \lambda = 0 \end{cases}. \qquad (5.27)$$

Значення λ може бути обрано шляхом максимізації коефіцієнта детермінації в парній регресії залежної змінної на x'_λ (який, згідно з формулою (4.14) є квадратом кореляції між y та x'_λ). В прикладі 5.2 оптимальним значенням λ є –0,095, що практично співпадає з логарифмічним перетворенням.

Трансформація Бокса–Кокса застосовується також для наближення розподілу перетворюваної змінної до нормального [47].

5.4 Використання фіктивних змінних

Часто на значення залежної змінної можуть впливати категоріальні фактори, які не вимірюються у кількісних шкалах. Для оцінки їх впливу існують дві стратегії. Перша полягає в тому, щоб розділити дані на декілька підвибірок залежно від значень категоріальних змінних. Це не дуже зручно, особливо коли є декілька категоріальних змінних, і зазвичай наводить до зменшення ступенів свободи у кожній окремій підвиборці. Альтернативна стратегія полягає у врахуванні впливу якісних факторів за допомогою фіктивних змінних.

Фіктивні змінні вже згадувалися при описі дисперсійного аналізу в п. 4.3. Вони приймають одне з двох значень, 0 чи 1, в залежності від приналежності спостереження до певної категорії. Використання таких змінних дозволяє описувати поведінку залежної змінної за допомогою кусково-лінійних функцій.

Розглянемо рівняння регресії

$$y_i = \beta_0 + \beta_1 d_i + \beta_2 x_i + \varepsilon_i, \qquad (5.28)$$

де d_i – фіктивна змінна. Рівняння (5.28) по суті описує дві лінії регресії:

$$y_i = \begin{cases} \beta_0 + \beta_2 x_i + \varepsilon_i, \ d_i = 0 \\ \beta_0 + \beta_1 + \beta_2 x_i + \varepsilon_i, \ d_i = 1, \end{cases} \qquad (5.29)$$

які різняться точкою перетину з віссю y (див. рис. 5.3*а*). При цьому статистична значущість коефіцієнта β_1 свідчить про наявність систематичних відмінностей між спостереженнями двох категорій.

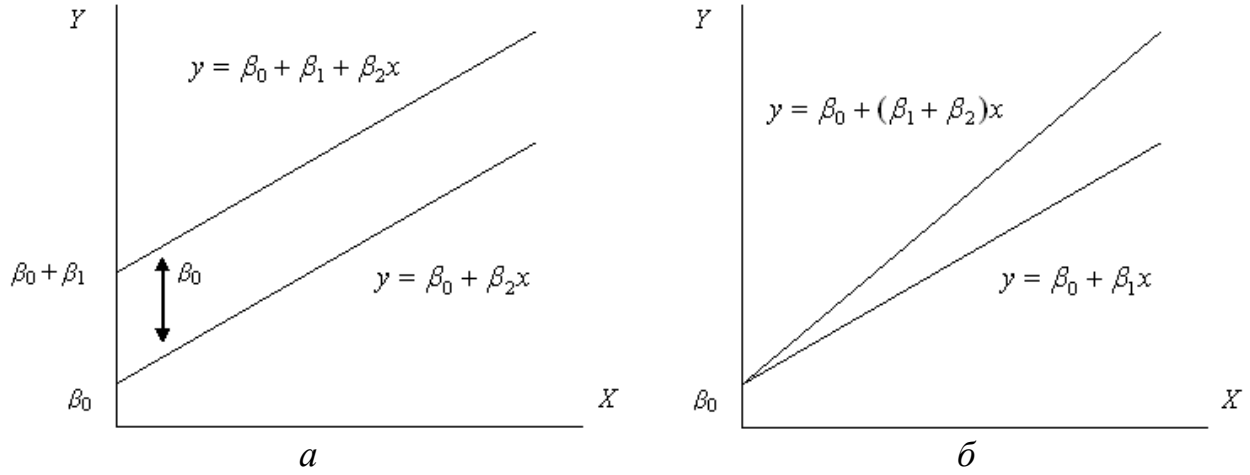
Дещо складніше врахувати потенційний вплив категоріальної приналежності на коефіцієнт нахилу лінії регресії. Для цього треба створити нову змінну – добуток між фіктивною та кількісною змінними, як у наступному рівнянні:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i d_i + \varepsilon_i. \qquad (5.30)$$

Це рівняння еквівалентно системі

$$y_i = \begin{cases} \beta_0 + \beta_1 x_i + \varepsilon_i, \ d_i = 0 \\ \beta_0 + (\beta_1 + \beta_2) x_i + \varepsilon_i, \ d_i = 1 \end{cases} \qquad (5.31)$$

(див. рис. 5.3*б*). Звісно, можна комбінувати обидва варіанти, що дозволяє створювати як завгодно складні кусково–лінійні функції.



а *б*

Рис. 5.3. Зміна рівняння регресії за допомогою фіктивних змінних:
а – зміна точки *y*–перетину; *б* – зміна коефіцієнта нахилу

Проте, при введенні фіктивних змінних слід бути обережним, щоб уникнути мультиколінеарності. Наприклад, при аналізі часових рядів часто вводять змінні, які відповідають місяцям, порам року або календарним кварталам. Якщо в останньому випадку $Q1_t, Q2_t, Q3_t, Q4_t$ – фіктивні змінні, введені для кожного кварталу, то $Q1_t + Q2_t + Q3_t + Q4_t = 1$ для всіх спостережень. Отже, з цих чотирьох змінних незалежними є тільки три. В термінах лінійної алгебри, сума чотирьох стовпців, які відповідають квартальним змінним, буде дорівнювати першому (стовпчик одиниць), і матриця регресорів буде виродженою. Рішення цієї проблеми полягає в тому, щоби виключити одну із змінних $Q1$–$Q4$ з рівняння регресії.

Можливі й інші джерела мультиколінеарності. Наприклад, ціна на каву може залежати від її біологічного виду (арабіка/робуста) та країни–виробника, поруч з іншими факторами. Але більшість країн[6] спеціалізується на вирощуванні тільки одного виду кави, то ж якщо ввести фіктивні змінні для виду кави і окремих країн її походження, то, ймовірно, виникне мультиколінеарність.

Фіктивні змінні можуть використовуватись для дисперсійного аналізу (як правило, у панельних даних). Наприклад, якщо є набір спостережень за N об'єктами протягом T періодів, то модель із *постійними індивідуальними ефектами* (англ. *fixed effects model*) задається як:

$$y_{it} = \alpha_i + \beta_1 x_{1,it} + \beta_2 x_{2,it} + ... + \beta_m x_{m,it} + \varepsilon_{it}. \qquad (5.32)$$

В цій моделі змінні α_i називаються індивідуальними ефектами і відбивають системні відмінності між досліджуваними об'єктами. Модель (5.32) може бути оцінена за допомогою МНК, хоча для неї існують більш ефективні з

[6] За виключенням Бразилії.

обчислювальної точки зору методи, які враховують спеціальну структуру матриці регресорів в моделі (5.32). Більш докладно із статистичними методами дослідження панельних даних можна ознайомитись в [25].

5.5 Загальна схема регресійного аналізу

Класична схема регресійного аналізу складається з наступних етапів.

1. Формулювання задачі. На цьому етапі формуються попередні гіпотези про залежність досліджуваних явищ.

2. Визначення залежної змінної і незалежних змінних, які можуть бути корисними для пояснення її поведінки.

3. Збір статистичних даних. Дані повинні бути зібрані для кожної із змінних, що включені в регресійну модель.

4. Формулювання гіпотези про форму залежності між змінними (проста або множинна, лінійна або нелінійна регресія).

5. Оцінка параметрів функції регресії полягає в розрахунку їх чисельних значень за допомогою формули (5.6) у випадку лінійної регресії або чисельної мінімізації функції (5.3) у випадку нелінійної регресії.

6. Оцінка точності регресійного аналізу здійснюється шляхом визначення значущості окремих коефіцієнтів регресії (t–статистики) і узгодженості моделі з даними у цілому (коефіцієнт детермінації, F–статистика).

7. Інтерпретація отриманих результатів. Результати регресійного аналізу порівнюються з висунутими гіпотезами. Оцінюється коректність і правдоподібність отриманих результатів. При цьому доцільно знайти відповіді на наступні запитання (чотири “S” регресійного аналізу):

– (sign) чи мають коефіцієнти очікуваний знак?

– (significance) чи є вони статистично значущими?

– (size) чи є правдоподібним оцінений масштаб впливу окремих регресорів на залежну змінну?

– (sense) чи має сенс побудована модель у цілому?

Негативна відповідь на деякі з цих запитань може привести до перегляду моделі і повторення етапів 2–6.

8. Прогнозування невідомих значень залежної змінної. Ця задача, у свою чергу, зводиться до вирішення задачі одного з двох наступних типів:

– оцінка значень залежної змінної всередині розглянутого інтервалу вихідних даних, тобто пропущених значень; при цьому вирішується задача *інтерполяції*;

– оцінка значень залежної змінної поза заданого інтервалу вихідних даних; при цьому вирішується задача *екстраполяції*.

Обидві задачі вирішуються шляхом підстановки в рівняння регресії знайдених оцінок параметрів і відповідних значень незалежних змінних. Проте,

оцінки, отримані шляхом екстраполяції, є значно менш надійними, адже вони носять суто гіпотетичний характер.

В інтелектуальному аналізі даних ця схема має істотні відмінності.

По–перше, аналізується наявна база даних організації, тому відсутній етап пошуку даних. Це не виключає використання зовнішніх джерел інформації, але вони мають допоміжний характер.

По–друге, оскільки мова йде про виявлення неочікуваних закономірностей, відсутні попередні гіпотези про зв'язки між змінними. В ідеалі гіпотези формулюються і оцінюються автоматично, і відбираються ті з них, які мають високу статистичну стійкість. При цьому причини виявленого зв'язку не є першорядною метою аналізу. Відповідно, сьомий пункт процесу регресійного аналізу в ІАД має рудиментарний характер.

По–третє, в класичній статистиці і прикладних науках робиться наголос на встановленні статистичних залежностей між досліджуваними змінними, навіть якщо пояснювальна сила (англ. explanatory power) такого зв'язку є незначною. Наприклад, якщо певний медикамент знижує ймовірність серцевого нападу на 5%, це може бути достатньою причиною для призначення цього препарату. В ІАД акцент робиться на розробці моделей, які мають високу узгодженість з даними у цілому. Дещо спрощуючи, можна сказати, що в класичній статистиці дивляться перед усім на значущість змінних, тобто на *t*–статистику, а в ІАД – на коефіцієнт детермінації та *F*–статистику.

Але при всіх цих відмінностях добре розуміння теорії регресійного аналізу необхідне для правильної інтерпретації одержаних результатів. Так, у згаданому вище випадку майже мультиколінеарності модель може мати високий коефіцієнт детермінації, але погану прогностичну силу внаслідок низької точності оцінок впливу предикторів на залежну змінну.

Приклад 5.3. Розглянемо процес регресійного аналізу на прикладі прогнозування ціни товару в залежності від його споживчих характеристик[7]. Для цього використаємо дані щодо цін та технічних характеристик комп'ютерних принтерів з роботи [36], наведені в файлі `printers.csv`.

За технологією друку переважна більшість сучасних принтерів належить до однієї з двох категорій – лазерні та струменеві (англ. inkjets). Лазерні принтери мають високу швидкість та якість друку, але є переважно монохромними. Струменеві принтери натомість підтримують кольоровий друк. Завдяки спільним конструктивним елементам, принтери часто поєднують із сканером та ксероксом; такі пристрої називають багатофункціональними. Найкрупнішим та найвідомішим виробником принтерів є американська компанія Hewlett–Packard. Іншими крупними виробниками є Canon, Brother та Epson [Д1].

Файл `printers.csv` містить інформацію про ціни (*P*), обсяги продажів (*S*)

[7] Такий аналіз називають гедонічною регресією (англ. hedonic regression) [48].

та технічні характеристики для 19 моделей лазерних та 64 моделей струменевих принтерів. Атрибути принтерів наведені в табл. 5.1.

Таблиця 5.1 – Характеристики комп'ютерних принтерів

Позначення	Опис	Шкала	Примітка
PpmBW	швидкість друку у чорно–білому режимі, стор./хв.	відношень	
PpmC	швидкість друку у кольоровому режимі, стор./хв.	відношень	струменеві
Res	роздільна здатність друку, точок на дюйм	відношень	
RAM	обсяг вбудованої пам'яті, КБ	відношень	
PHC	ємність лотка для паперу, сторінок	відношень	
Laser	1, якщо принтер лазерний	бінарна	
MFC	1, якщо багатофункціональний пристрій	бінарна	
PS	1, якщо підтримує Adobe PostScript	бінарна	лазерні
Vendor	фірма–виробник	номінальна	

На базі цих атрибутів можуть бути створені інші. В подальшому використовується фіктивна змінна HP, яка дорівнює одиниці, якщо виробником є Hewlett–Packard. Безперечно, важливою характеристикою є можливість кольорового друку. Але всі наявні у виборці лазерні принтери є монохромними, а струменеві – кольоровими, тож введення такої фіктивної змінної призвело б до мультиколінеарності.

Щоб обрати змінні, які варто включити до регресійної моделі, розрахуємо кореляційну матрицю для наявних змінних (табл. 5.2).

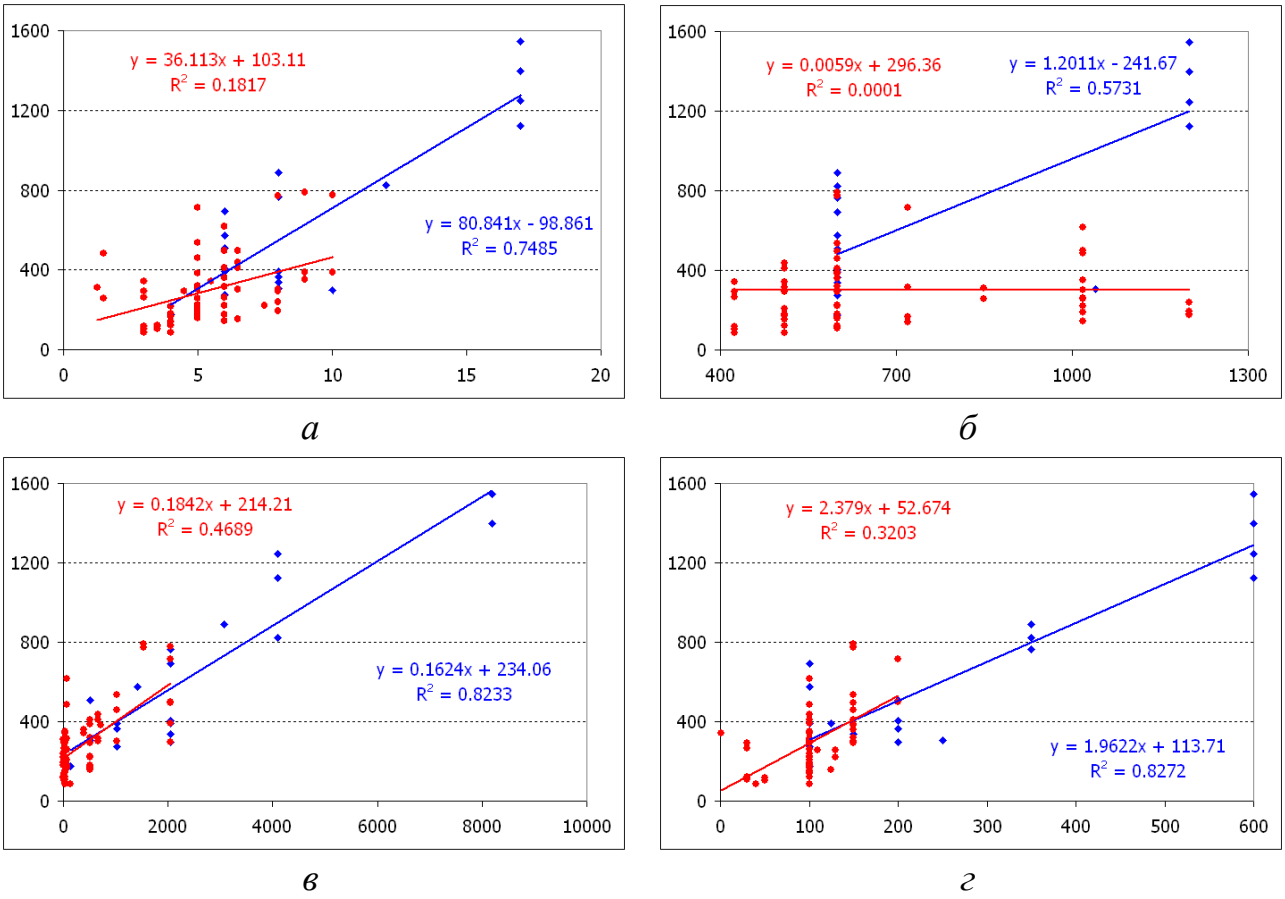
Таблиця 5.2 – Кореляційна матриця змінних

	P	PpmBW	PpmC	Res	RAM	PHC	Laser	MFC	HP
P	1.00	0.78	–0.15	0.32	0.88	0.86	0.53	0.15	0.44
PpmBW	0.78	1.00	0.04	0.43	0.81	0.83	0.52	–0.08	0.33
PpmC	–0.15	0.04	1.00	0.20	–0.25	–0.29	–0.61	–0.03	–0.07
Res	0.32	0.43	0.20	1.00	0.28	0.44	0.10	–0.22	–0.21
RAM	0.88	0.81	–0.25	0.28	1.00	0.87	0.60	0.04	0.43
PHC	0.86	0.83	–0.29	0.44	0.87	1.00	0.61	0.00	0.34
Laser	0.53	0.52	–0.61	0.10	0.60	0.61	1.00	–0.05	0.13
MFC	0.15	–0.08	–0.03	–0.22	0.04	0.00	–0.05	1.00	0.03
HP	0.44	0.33	–0.07	–0.21	0.43	0.34	0.13	0.03	1.00

Інтуїтивно здається, що ціна принтера повинна залежати в першу чергу від його швидкості та від якості друку (яка відбивається роздільною здатністю). Проте, виявляється, що це не так і найкращим предиктором ціни принтера є обсяг вбудованої пам'яті. Ймовірно, це пояснюється тим, що великий обсяг пам'яті потрібен як для швидкодії принтеру, так і для якості друку. Нагадаємо, що виявлення саме таких неочікуваних закономірностей є однією з головних задач інтелектуального аналізу даних.

Далі доцільно побудувати діаграми розсіювання для візуальної оцінки форми зв'язку між залежною змінною та кількісними факторами (або відсутності такого зв'язку). Вплив категоріальних змінних можна відобразити за допомогою кольору. На рис. 5.4 наводяться діаграми розсіяння для ціни принтера і кількісних факторів, які суттєво з нею корелюють. Сині точки відповідають лазерним принтерам, а червоні – струменевим. На графіки додано також рівняння парної регресії між відповідними змінними і значення коефіцієнта детермінації окремо для двох категорій принтерів.

Рис. 5.4 підтверджує закономірності, помітні в табл. 5.2. Ціни принтерів найкраще корелюють з обсягом пам'яті та з ємністю лотка для паперу. Важливо також, що форма цієї залежності є приблизно однаковою як для лазерних, так і для струменевих принтерів. Несподівано, роздільна здатність друку майже ніяк не впливає на ціну струменевих принтерів. З графіків видно також, що ціни лазерних принтерів видаються більш прогнозованими, ніж ціни струменевих. Це пояснюється тим, що моделі лазерних принтерів демонструють значно більшу розбіжність у своїх технічних характеристиках. Нарешті, виходячи з виду графіків, нема підстав для перетворення змінних.



а *б* *в* *г*

Рис. 5.4. Кореляційні діаграми зв'язку між ціною принтера та його атрибутами:
а – швидкість друку PpmBW; *б* – роздільна здатність друку Res;
в – обсяг пам'яті RAM; *г* – ємність лотка для паперу PHC.
Синій колір відповідає лазерним принтерам, червоний – струменевим.

Попередній аналіз даних дає можливість обрати змінні, які варто включити до регресійної моделі. Результати оцінки рівняння регресії зведені в табл. 3.3, яка є стандартною формою презентації результатів регресійного аналізу.

Таблиця 5.3 – Регресія ціни принтера на його атрибути

Змінна	Оцінка	Стандартна похибка	*t*-статистика	Довірчий інтервал Нижні 95%	Верхні 95%
Const	110.660	37.294	2.967	36.350	184.970
Laser	79.755	124.113	0.643	–167.545	327.054
MFC	57.904	38.812	1.492	–19.431	135.238
PS	127.281	123.222	1.033	–118.244	372.806
HP	43.897	31.456	1.396	–18.779	106.574
RAM	0.101	0.021	4.815	0.059	0.143
PHC	1.105	0.353	3.132	0.402	1.807
Laser×PpmBW	–29.093	20.375	–1.428	–69.692	11.505
Laser×Res	0.129	0.236	0.546	–0.342	0.600
Спостережень	83		R^2	0.841	
df	74		$\bar{R}^2$	0.816	

Оскільки на рис. 5.4*а,б* залежність ціни від швидкості друку та роздільної здатності помітна тільки для лазерних принтерів, в регресію з табл. 5.3 були введені перехресні добутки цих змінних з фіктивною змінною Laser, як у формулі (5.30). Коефіцієнт детермінації регресії непоганий, але в регресії виявилось багато незначущих на 95% рівні змінних, які виділені в табл. 5.3 світло–сірим кольором. Цього слід було очікувати, виходячи із високої кореляції між незалежними змінними, очевидну з табл. 5.2. В таких випадках виключення з регресії незначущих змінних може підвищити точність оцінки параметрів без суттєвого впливу на узгодженість моделі з даними.

Результати оцінки модифікованої регресійної моделі після виключення окремих змінних наведені в табл. 5.4. Змінні HP та Laser були об’єднані у комбіновану змінну HP×Laser, оскільки премія за бренд HP виявилась помітною лише для лазерних принтерів.

Таблиця 5.4 – Модифікована регресія ціни принтера на його атрибути

Змінна	Оцінка	Стандартна похибка	*t*-статистика	Довірчий інтервал Нижні 95%	Верхні 95%
Const	141.633	24.346	5.817	93.163	190.102
MFC	113.831	34.578	3.292	44.991	182.671
HP×Laser	132.124	62.731	2.106	7.237	257.011
RAM	0.088	0.018	4.772	0.051	0.124
PHC	0.826	0.226	3.662	0.377	1.275
Спостережень	83		R^2	0.838	
df	78		$\bar{R}^2$	0.830	

Після модифікації моделі решта змінних стали значущими на 95% рівні, а скоригований коефіцієнт детермінації підвищився. Всі коефіцієнти мають очікуваний знак. Залишається порівняти прогнозні значення, які випливають із останньої моделі, з реальними даними (рис. 5.5).

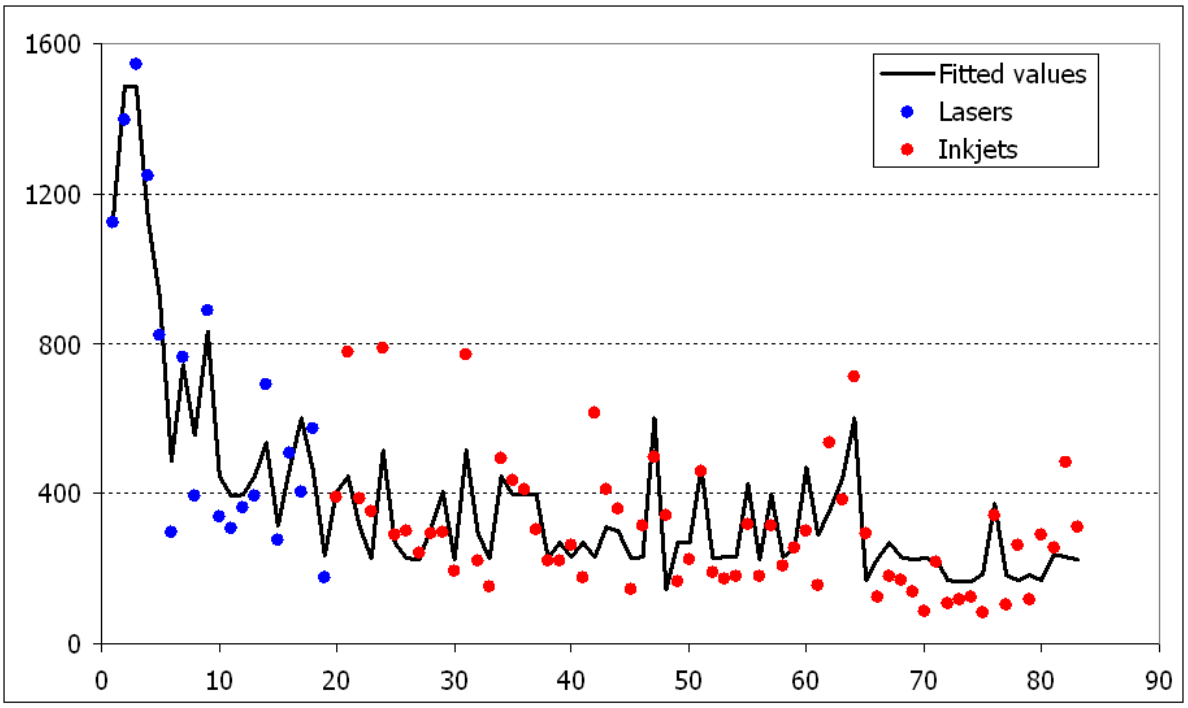


Рис. 5.5. Порівняння цін принтерів із прогнозованими значеннями

На базі результатів, наведених в табл. 5.4 та на рис. 5.5 можна зробити висновок про прийнятну узгодженість побудованої моделі з даними. Видно також, що модель працює краще для лазерних принтерів. ■

Практичне застосування регресійних моделей може бути досить різноманітним. Так, модель, подібна до розробленої в прикладі 5.3, може використовуватись, зокрема, для таких цілей:

– встановлення прийнятного цінового діапазону для нових товарів;
– оцінка конкурентоспроможності окремих товарів та фірм–виробників;
– оцінка вартості брендів;
– виявлення вигідних та невигідних пропозицій.

Останній вид аналізу є популярним у біржовій торгівлі, де на базі зіставлення реальної ціни активу з його модельною вартістю приймаються рішення чи надаються рекомендації стосовно його купівлі/продажу. До речі, цей приклад ілюструє також можливість застосування регресійного аналізу для вирішення задач класифікації. Наприклад, якщо ціна активу значно нижче прогнозної, він заноситься в категорію «купувати»; якщо значно вище – в категорію «продавати»; в інших випадках – в категорію «нічого не робити».

Застосування регресійного аналізу до вирішення задач класифікації буде більш детально розглянуто нижче.

Контрольні запитання

1. Як визначається кількість ступенів свободи множинної регресії?
2. В чому полягає критерій оптимальності методу найменших квадратів?
3. Чому в рівняння множинної регресії завжди варто додавати константу (вільний член)?
4. В чому полягає проблема мультиколінеарності?
5. Які характеристики має білий шум?
6. За яких умов оцінки МНК будуть незміщеними?
7. В чому полягає теорема Гауса–Маркова?
8. Дайте визначення дисперсійно–коваріаційної матриці.
9. Наведіть порядок розрахунку дисперсійно–коваріаційної матриці для вектору коефіцієнтів рівняння лінійної регресії.
10. Як знайти довірчий інтервал для коефіцієнтів лінійної регресії за умови нормальності похибки?
11. Як визначається адекватність регресійної моделі?
12. Для чого використовується скоригований коефіцієнт детермінації?
13. В чому полягає сенс перетворення змінних регресійної моделі на логарифми?
14. Як інтерпретуються коефіцієнти регресії в лог–лінійній моделі?
15. Наведіть приклади, коли доцільно використання напів–логаріфмичних моделей.
16. Для чого в регресійніх моделях використовуються фіктивні змінні?
17. Як надати можливість зміни коефіцієнту нахилу регресійної поверхні в залежності від значення категоріальної змінної?
18. В чому полягає і для чого використовується модель із постійними індивідуальними ефектами?
19. В чому полягають основні етапи регресійного аналізу в класичній статистиці?
20. В чому полягають відмінності у методиці регресійного аналізу між класичною статистикою та інтелектуальним аналізом даних?

Завдання для самостійної роботи

5.1. Наведіть два приклади практичних задач, для вирішення яких потрібна оцінка рівняння множинної регресії.

5.2. Як зміняться оцінки коефіцієнтів рівняння регресії $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$, якщо:

а) замінити змінну x_1 на змінну $X_1 = x_1 / 100$;

б) замінити змінну x_2 на змінну $X_2 = ax_2$, де a – деяка константа;

в) замінити змінну y на змінну $Y = cy$, де с – деяка константа?

5.3. Розглянемо два рівняння регресії:

$$\ln w_i = \beta_0 + \beta_1 \ln h_i + \beta_2 s_i + \varepsilon_i;$$

$$\ln(w_i / h_i) = \gamma_0 + \gamma_1 \ln h_i + \gamma_2 s_i + \varepsilon_i,$$

де w_i – тижнева заробітна плата і-го працівника, h_i – кількість відпрацьованих годин протягом тижня, s_i – кількість років, витрачених працівником на освіту.

а) Покажіть, що МНК–оцінки параметрів цих двох рівнянь будуть задовольняти наступним співвідношенням: $\hat{\gamma}_0 = \hat{\beta}_0; \hat{\gamma}_1 = \hat{\beta}_1 - 1; \hat{\gamma}_2 = \hat{\beta}_2$.

б) Покажіть, що залишки цих двох регресій співпадають.

в) За яких умов коефіцієнт детермінації R^2 першої регресії буде більшим, ніж в другій?

5.4. Регресія залежної змінної y на три незалежні змінні на базі 30 спостережень навела до наступних результатів:

$y =$	20.0 + 1.2x_1 + 1.0x_2 – 0.5x_3			
Стандартні помилки	(2.1)	(1.5)	(1.3)	(0.1)
t–значення	9.5			
95% довірчий інтервал	±4.3			

Заповніть відсутні елементи.

5.5. В таблиці наводяться дані щодо заробітної плати та інших характеристик спеціалістів певної професії.

№	Освіта, років	Стаж, років	Стать	Зарплата, грн.
i	s_i	e_i	g_i	w_i
1	17	12	Ч	16265
2	16	4	Ч	14086
3	11	6	Ч	13223
4	15	5	Ч	13805
5	18	19	Ж	18009
6	16	21	Ч	18445
7	15	10	Ч	14824
8	16	1	Ч	12249
9	18	25	Ж	19635
10	15	4	Ж	12375
11	20	16	Ж	16881
12	20	1	Ж	12469
13	13	16	Ж	15631
14	19	5	Ж	13442
15	14	17	Ж	15516
16	8	3	Ж	10337
17	14	5	Ч	14238
18	17	29	Ч	21563
19	21	25	Ч	19675
20	14	27	Ж	19154

а) Оцініть рівняння регресії $w_i = \beta_0 + \beta_1 s_i + \beta_2 e_i + \varepsilon_i$.

б) Оцініть заробітну плату працівника із освітою 12 років та стажем 4 роки.

в) Як зміниться заробітна плата працівника через 3 роки безперервної праці на тій же посаді?

г) Побудуйте графік прогнозних значень заробітної плати поруч з реальними значеннями.

г) Знайдіть суму квадратів залишків, коефіцієнт детермінації та оцінить дисперсію регресії.

5.6. Зробимо припущення про нормальність розподілу помилок в регресії з задачі 5.5.

а) Які з коефіцієнтів регресії є значущими на 95% рівні?

б) Побудуйте довірчі інтервали для коефіцієнтів регресії.

в) Перевірте гіпотези $\beta_1 = 200$; $\beta_1 = \beta_2$; $\beta_0 = \beta_1 = \beta_2 = 0$.

5.7. Додайте до регресії з двох попередніх задач фіктивну змінну «стать працівника» g_i і оцініть рівняння регресії $w_i = \beta_0 + \beta_1 s_i + \beta_2 e_i + \beta_3 g_i + \varepsilon_i$.

а) Як додання нової змінної вплинуло на інші коефіцієнти моделі?

б) Чи є коефіцієнт β_3 значущим? Надайте змістовну інтерпретацію значенню цього коефіцієнта.

в) Існує точка зору, що стартова заробітна плата для чоловіків та жінок не відрізняється, але чоловіки швидше просуваються кар'єрними сходами. Як перевірити цю гіпотезу за допомогою регресійної моделі? Оцініть відповідне рівняння регресії і зробіть висновок, чи підтверджується ця гіпотеза вихідними даними задачі 5.5.

5.8. Оцінка двох регресійних моделей на базі 40 спостережень навела до наступних результатів (в дужках надані стандартні помилки коефіцієнтів):

$$\underset{}{w} = \underset{(5.0)}{20} + \underset{(0.1)}{0.8}\, s + \underset{(1.3)}{3.7}\, e,\ R^2 = 0.4$$

$$\ln w = \underset{(3.0)}{3.2} + \underset{(0.01)}{0.1} \ln s + \underset{(0.03)}{0.2} \ln e,\ R^2 = 0.7,$$

де w – заробітна плата працівника, s – кількість років освіти, e – стаж роботи.

а) Зробіть прогноз заробітної плати працівника з 15 роками освіти та 10 роками стажу за двома моделями.

б) Порівняйте дві регресії з точки зору їх якості та прогностичної сили.

в) Дайте інтерпретацію коефіцієнтам при змінних s та e в першому та другому рівняннях.

6. ЗАДАЧІ КЛАСИФІКАЦІЇ: ЗАГАЛЬНІ ПОЛОЖЕННЯ

6.1 Постановка і типи задач класифікації

Класифікація (англ. *classification*) – це процес упорядкування за певним критерієм об'єктів, які мають ознаки для визначення подібності або відмінності між ними.

Типова задача класифікації полягає у наступному. Задана вибірка об'єктів та їх атрибутів, для яких відома приналежність до певного класу або категорії. До якого класу належать інші об'єкти, невідомо. Необхідно побудувати таке правило (функцію, алгоритм), яке б дозволило визначити клас довільного об'єкту з відомими атрибутами. Атрибути об'єкта можуть бути номінативними, порядковими, кількісними або комбінацією всіх цих типів. Різноманітність варіантів обумовлює велику кількість існуючих методів класифікації.

Класифікація вимагає дотримання наступних правил:

– повнота: кожний об'єкт має належати до якогось класу;

– чистота: множини об'єктів, віднесених до різних класів, не можуть перехрещуватися [10, 40].

Оскільки залежна змінна в задачах класифікації є позначкою класу, про успішність того чи іншого методу можна судити за відсотком правильно розставлених позначок. Це значно простіше інтерпретувати, ніж дисперсії, довірчі інтервали та інші міри узгодженості моделі з даними, які використовуються в математичній статистиці. Це пояснює популярність моделей класифікації, зокрема, в галузі машинного навчання.

Машинне навчання (англ. *machine learning*) – це підгалузь штучного інтелекту, яка застосовує статистичні прийоми для надання комп'ютерам здатності «навчатися» (тобто, поступово покращувати продуктивність у вирішенні певних задач) замість того, щоб бути програмованими явно. Згідно з Томом Мітчеллом, комп'ютерна програма вчиться по відношенню до якогось класу задач та визначеної міри продуктивності, якщо її продуктивність у вирішенні таких задач покращується з досвідом [37].

В якості прикладу можна навести системи визначення спаму, реалізовані в усіх розповсюджених системах електронної пошти. Це типова задача бінарної класифікації, в якій кожний з листів, які надходять до поштової скриньки, має бути віднесений до однієї з двох категорій: «спам» або «не спам». Традиційний підхід до вирішення цієї задачі полягає у веденні «чорних списків» для певних адрес відправників та ключових слів у тексті і вимагає постійної роботи програмістів. Системі з машинним навчанням замість цього надається велика вибірка листів, класифікованих як спам чи ні, на базі чого система має самостійно визначати ознаки спаму. Ця база даних постійно поповнюється (для цього користувачам надається можливість помітити лист як спам), що має

вдосконалити процес фільтрації небажаних листів із набуттям «досвіду».

Задачі машинного навчання поділяють на дві великі групи, залежно від того, чи надається системі, що навчається, зворотний зв'язок:

– *навчання з учителем* (кероване навчання, англ. *supervised learning*) полягає в тому, що системі надаються приклади правильного перетворення входів на виходи, на базі яких має бути розроблено загальне правило такого перетворення;

– *навчання без учителя* (спонтанне навчання, англ. *unsupervised learning*): системі доводиться самостійно знаходити структуру у своїх входах.

Задачі класифікації відносяться до стратегії навчання з учителем. Множина класів вважається заздалегідь визначеною. Якщо класи, до яких належать об'єкти, заздалегідь невідомі, то говорять про задачу кластеризації.

Найпростішим видом класифікації є *бінарна* класифікація (англ. *binary* classification), коли визначено лише два класи об'єктів. Крім вищезгаданої фільтрації спаму, прикладами задач бінарної класифікації може бути прогнозування переможця в баскетбольному матчі, визначення здатності споживача сплатити наданий кредит, діагностика наявності у пацієнта певної хвороби тощо.

Коли визначено три або більше класів, говорять про *багатокласову* або *мультиноміальну* класифікацію (англ. *multinomial classification*). Прикладом таких задач може бути прогнозування результату футбольного матчу (на відміну від баскетболу, у футболі можлива нічия) або розпізнавання рукописних букв. Вочевидь, бінарна класифікація є окремим випадком мультиноміальної. У свою чергу, багатокласова класифікація може бути зведена до послідовності задач бінарної класифікації. На першому етапі вирішується, які об'єкти належать до першого класу. Решта об'єктів перевіряється на приналежність до другого класу і так далі, доки всі об'єкти не будуть класифіковані.

Нарешті, в деяких випадках об'єкту може бути присвоєно одночасно декілька позначок, які не є взаємовиключними. Прикладом такого підходу є система Google Photos, яка може віднести одну й ту саму фотографію користувача до декількох категорій, скажімо, «море», «світанок» та «хмари». Такий різновид задач називається класифікацією з *кількома мітками* (англ. *multi-label classification*). Вони можуть бути зведені до серії задач бінарної класифікації (чи відноситься об'єкт до кожної з можливих категорій) або до задачі мультиноміальної класифікації (якщо створити окремий клас для кожної з можливих комбінацій міток).

Класифікація може бути одновимірною (за однією ознакою) і багатовимірною (за двома і більше ознаками). Розрізняють також *ординарну* (англ. *deterministic*) класифікацію, коли кожному об'єкту просто призначається мітка класу, і *ймовірнісну* (англ. *probabilistic*), коли оцінюється імовірність

належності об’єкту до кожного із визначених класів. Тоді прогнозований клас об’єкту визначається як найімовірніший з можливих.

Наведемо простий приклад задачі класифікації. Припустимо, є база даних про клієнтів туристичного агентства з інформацією про вік, стать, сімейний стан, місячний дохід тощо. Агентство пропонує автобусні екскурсії вихідного дня поза містом, екскурсійні тури Україною і пакетні авіатури з пляжним відпочинком за кордоном. Споживачі відповідних турів визначені як класи 1–3. Слід визначити, до якого класу належить новий клієнт і який з трьох видів рекламних матеріалів йому варто відсилати.

Випадкові фактори, безумовно, впливають на результат класифікації (наприклад, на платоспроможність клієнта банку). Проте, не завжди виправдано вважати їх нормально розподіленими, а іноді взагалі неможливо побудувати імовірнісну модель такого впливу. В таких ситуаціях для перевірки надійності розроблених правил класифікації часто використовуються евристичні методи[8]. Одним із таких методів, особливо поширених у машинному навчанні, є розбиття вихідних даних на дві підмножини: тренувальну і тестову.

Тренувальна множина (англ. *training set*) містить дані, що використовуються для конструювання (навчання) моделі. *Тестова множина* (англ. *test set*) використовується для перевірки працездатності моделі, тобто її здатності коректно класифікувати нові дані. Для цього відомі позначки класів з тестової множини порівнюються з прогнозами розробленої моделі. Рівень точності моделі визначається, наприклад, відсотком правильно класифікованих об’єктів в тестовій множині.

Для більшої впевненості у надійності класифікаційного правила розбиття вибірки на навчальну та тренувальну може виконуватись декілька разів різними способами, впорядковано або випадково, а результати перевірки на тестовій множині усереднюватись. Наприклад, спостереження з номером $i = 1,\ldots,n$ може використовуватись як тестова множина, а решта спостережень $1,\ldots,i-1,i+1,\ldots,n$ – як тренувальна. Усереднення результатів прогнозування за n такими розбиттями і визначатиме точність класифікації. Група методів, які оцінюють якість класифікації за такою або подібною схемою, називається *перехресним затвердженням* (англ. *cross-validation*).

Якщо точність моделі достатня, стає можливим її практичне використання для оцінки платоспроможності, медичної діагностики тощо.

Оцінювання методу класифікації слід проводити, виходячи з таких характеристик:

– швидкість – характеризує час, який потрібний на створення моделі і її використання;

[8] Тобто практичні методи вирішення задач, які дозволяють отримати «розумне» рішення, але не мають строго доведених властивостей і не обов’язково є оптимальними чи досконалими.

– робастність – характеризує стійкість до будь-яких порушень вихідних припущень і означає можливість роботи з помилками і пропущеними значеннями в вихідних даних;

– інтерпретованість – характеризує простоту розуміння моделі;

– надійність – характеризує коректність класифікації, зокрема, при наявності в даних шумів і викидів;

Для вирішення задач класифікації розроблено багато методів, зокрема:

– дерева рішень;
– випадковий ліс;
– байєсівська класифікація;
– метод опорних векторів;
– метод найближчого сусіда;
– логістична регресія;
– лінійний дискримінантний аналіз;
– штучні нейронні мережі тощо.

Найбільш поширені з них будуть розглянуті більш докладно нижче.

6.2 Оцінка якості класифікації

Найпростішим показником якості роботи алгоритму класифікації є його точність, яка визначається як відношення правильно спрогнозованих класів до обсягу вибірки. Проте, часто цього недостатньо для оцінки дієвості алгоритму при вирішенні практичних задач. Розглянемо проблеми з вибором критерію оцінки якості на прикладі задач бінарної класифікації.

Будемо вважати, що є два класи, які будемо називати позитивним (+) та негативним (–) (термінологія запозичена із медичної діагностики, де «позитивний» результат тесту означає наявність захворювання, а «негативний» – відсутність). При визначенні класу об’єкту можливі помилки I та II типу (див. п.3.4). Ціна цих похибок може суттєво різнитись. Наприклад, у протиповітряній обороні ціна помилки I типу (пропуск цілі) значно вища, ніж ціна помилки II типу (хибна тривога). Отже, надійність виявлення загрози може бути більш пріоритетною, ніж загальна точність класифікаційного правила.

Інша проблема пов’язана з поширеністю класів у генеральній сукупності. Якщо певне захворювання зустрічається лише у 1% населення, то правило, згідно з яким для кожного пацієнта прогнозується відсутність такої хвороби, буде мати точність 99%. Проте, зрозуміло, що використання такого правила з метою діагностики є безглуздим.

Наочну інформацію про якість бінарної класифікації можна отримати, якщо звести результати в таблицю, яка називається *матрицею помилок* або *матрицею невідповідностей* (англ. *confusion matrix*). Структура цієї матриці і позначки її елементів наведені у табл. 6.1.

Таблиця 6.1 – Результати класифікації, зведені у матрицю помилок

		Справжній клас (СК)	
		–	+
Прогнозований клас (ПК)	–	*TN* істинно негативний (true negative)	*FN* хибно негативний (false negative) помилка II типу
	+	*FP* хибно позитивний (false positive) помилка I типу	*TP* істинно позитивний (true positive)
Кількість у виборці	$n = N + P$	$N = TN + FP$	$P = FN + TP$

За виключенням позначок та термінології, ця таблиця є ідентичною з таблицею 3.2 в розділі про перевірку гіпотез (п. 3.4). Бінарна класифікація дійсно має багато спільних рис з перевіркою статистичних гіпотез, але є й суттєві відмінності [33] .

По–перше, метою перевірки гіпотез є формування висновків із наявних даних (англ. *inference*), в той час як метою класифікації є прогнозування приналежності об’єкта певному класу. Перевірка гіпотез використовується, наприклад, для оцінки дієвості нового медикаменту у лікуванні певної хвороби, а класифікація – для рішення про призначення медикаменту конкретному пацієнту.

По–друге, тестування гіпотез є більш загальною задачею, ніж класифікація. Наприклад, за допомогою перевірки гіпотез можна спростувати або підкріпити даними твердження, що жінки у середньому мають більшу тривалість життя, ніж чоловіки, що не є задачею класифікації. Специфічність задач бінарної класифікації дозволяє використання для їх вирішення спеціальних прийомів, таких як тренування та тестування, які не мають великого сенсу у загальних задачах перевірки гіпотез.

По–третє, перевірка гіпотез «тяжіє» до консервативної, нульової гіпотези. Згідно стандартної процедури, відхилення нульової гіпотези потребує 95% впевненості у її неузгодженості з даними. В задачах класифікації альтернативні класи або розглядаються симетрично, або зважуються відповідно до вимог конкретної задачі.

Нарешті, класичний апарат перевірки статистичних гіпотез відштовхується від певних припущень щодо закону розподілу даних у генеральній сукупності. Хоча, як ілюструє приклад 3.4, перевірка гіпотез не вимагає припущення про нормальність розподілу, воно використовується у більшості стандартних тестів математичної статистики (табл. 3.4). Для задач класифікації такі припущення часто або непотрібні, або не узгоджуються з даними.

На базі відношень елементів табл. 6.1 до сум за рядками та стовпцями можна сформувати наступні показники [32, 53].

1. *Поширеність* (англ. *prevalence*) $p = \frac{P}{n}$ характеризує частку генеральної сукупності, яка належить до позитивного класу.

2. *Істинно–позитивний рівень* , або *чутливість* (англ. *true positive rate, sensitivity*) $TPR = \frac{TP}{P} = \frac{TP}{TP + FN}$. Це частка правильно ідентифікованих представників позитивного класу. Його доповненням є *хибно–негативний рівень* (англ. *false negative rate*) $FNR = \frac{FN}{P} = \frac{FN}{TP + FN}$.

3. *Істинно–негативний рівень*, або *специфічність* (англ. *true negative rate, specificity*) $TNR = \frac{TN}{N} = \frac{TN}{TN + FP}$. Це частка правильно ідентифікованих представників негативного класу. Його доповненням є *хибно–позитивний рівень* (англ. *false positive rate*) $FPR = \frac{FP}{N} = \frac{FP}{TN + FP}$.

4. *Прогностична значущість позитивного результату*, або *влучність* (англ. *positive predictive value, precision*) $PPV = \frac{TP}{TP + FP}$. Це частка позитивних «діагнозів», які є правильними. Його доповненням є *рівень хибного виявлення* (англ. *false discovery rate*) $FDR = \frac{FP}{TP + FP}$.

5. *Прогностична значущість негативного результату* (англ. *negative predictive value*) $NPV = \frac{TN}{TN + FN}$. Це частка правильних «діагнозів» серед негативних. Його доповненням є *рівень хибного пропускання* (англ. *false omission rate*) $FOR = \frac{FN}{TN + FN}$.

6. *Точність* (англ. *accuracy*) $ACC = \frac{TP + TN}{n} = \frac{TP + TN}{TP + TN + FP + FN}$ – це частка правильних прогнозів серед усіх зроблених.

Точність правила залежить від поширеності класу, в той час як чутливість і специфічність – ні. Тому ці два показники вважаються більш інформативними, ніж загальна точність кваліфікаційного правила. Для того, щоб охарактеризувати середню надійність класифікації, використовують *збалансовану точність* (англ. *balanced accuracy*) $BA = \frac{TPR + TNR}{2}$.

Інший підхід до аналізу матриці помилок полягає в тому, щоб розглядати її як таблицю сумісного розподілу двох бінарних випадкових змінних ПК та СК. Чим надійніше класифікація, тим більше елементів вибірки будуть розташовані на головній діагоналі матриці помилок. Отже, про якість класифікації можна

судити, виходячи з кореляції між змінними ПК та СК. Із визначення (2.63) після спрощень отримаємо формулу:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}, \qquad (6.1)$$

яка називається *коефіцієнтом кореляції Меттьюза* (англ. *Matthews correlation coefficient*). Значення $MCC = 1$ свідчить про досконалу класифікацію, значення $MCC = 0$ – про те, що класифікаційне правило не краще, ніж вгадування класу шляхом підкидання монети. Деякі фахівці вважають *MCC* найкращим показником якості класифікації у контексті матриці помилок [20].

Часто бінарна класифікація здійснюється на базі порівняння деякої кількісної величини з граничним рівнем[9]. Наприклад, для здорової людини рівень білка у крові може бути нормально розподіленим із середнім значенням 1 г/дл, а для хворої – із середнім 2 г/дл. Люди можуть бути кваліфіковані як хворі, якщо їх рівень білка перевищує критичний рівень *K*. Зміна цього рівня приведе до змін у кількості двох типів помилок (див. рис. 6.1).

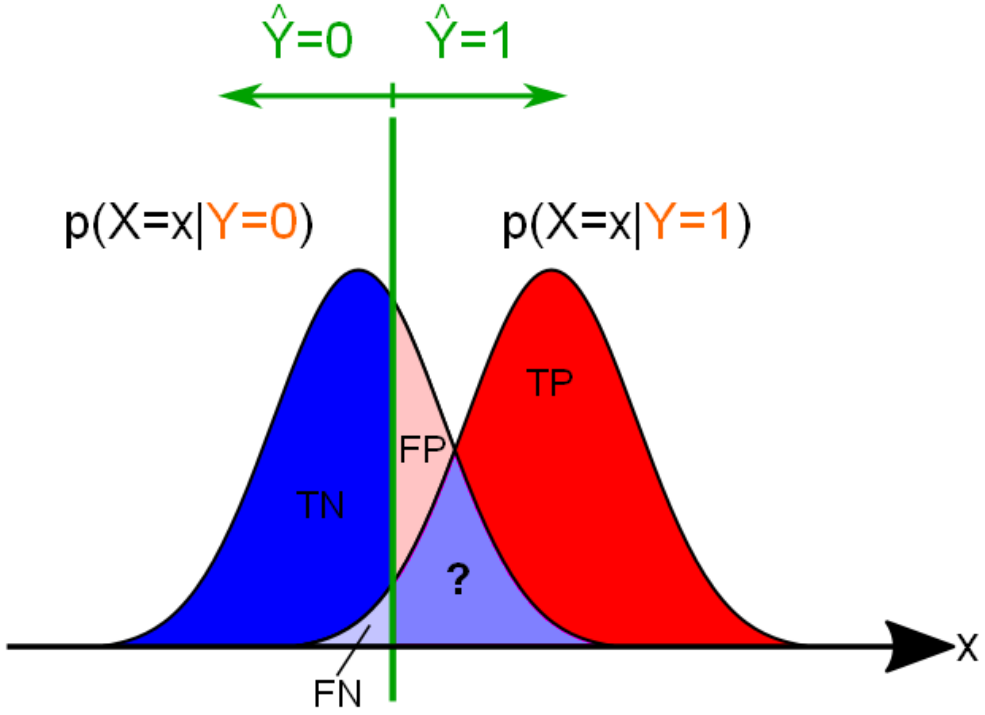


Рис. 6.1. Бінарна класифікація на базі кількісної змінної

Встановлення *K* на рівні 0 наведе до кваліфікації всіх людей як хворих ($FPR = 1$, $TPR = 1$); встановлення $K = \infty$ – до кваліфікації всіх як здорових ($FPR = 0$, $TPR = 0$). Збільшення *K* зменшує хибно–позитивний рівень за рахунок збільшення хибно–негативного. Цей компроміс можна відобразити графічно. Для цього на горизонтальній осі відкладається хибно–позитивний рівень правила класифікації (*FPR*), а на вертикальній – істинно–позитивний (*TPR*) при

[9] Нагадаємо, що у математичній статистиці нульова гіпотеза відкидається на користь альтернативної, якщо значення тестової статистики перевищує критичний рівень.

параметричній зміні критичного рівня *K*. Результатом буде крива помилок, яку частіше називають *ROC–кривою* (англ. *ROC curve*, скорочення від Receiver Operating Characteristic – робоча характеристика приймача[10]). Приклади таких кривих наведені на рис. 6.2.

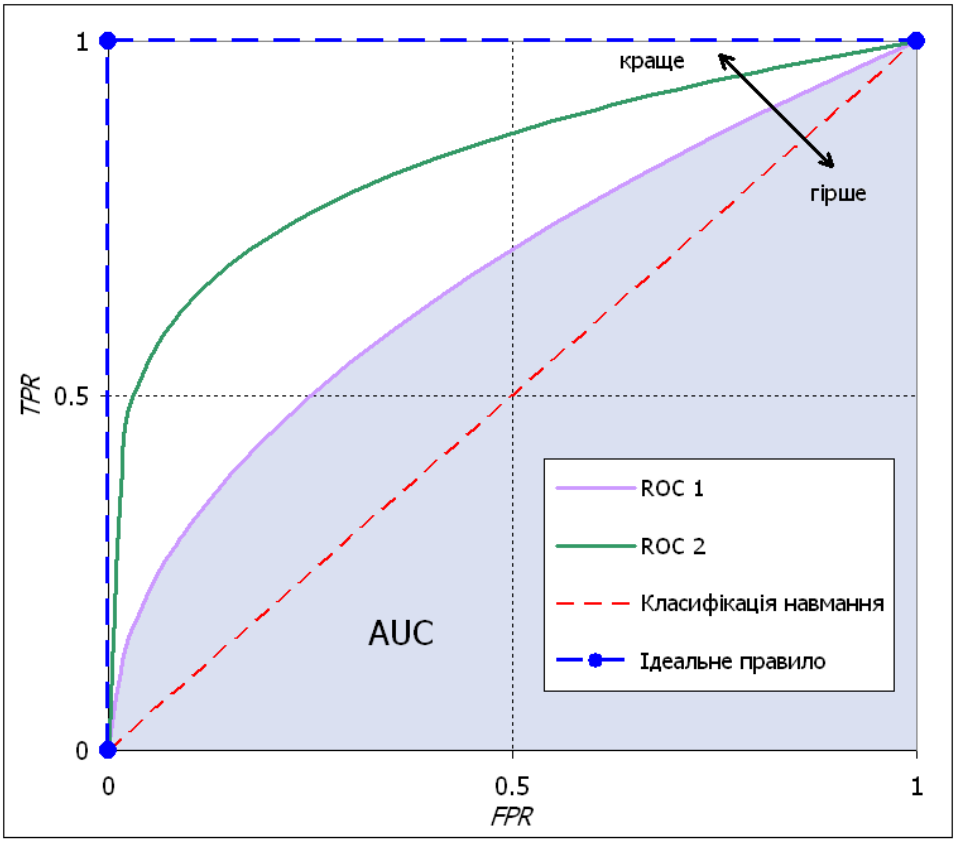


Рис. 6.2. Приклади ROC–кривих

Ідеальне класифікаційне правило має $FPR = 0$, $TPR = 1$ (синя точка у лівому верхньому куті графіка на рис. 6.2). Визначення класу навмання дає рівні шанси на правильне та неправильне його визначення, тобто $TPR = FPR$. Це штрих–пунктирна лінія під кутом 45º на рис. 6.2. Чим краще класифікаційне правило, тим більше його ROC–крива буде вигнута в бік точки (0,1). Отже, про якість класифікації можна судити за площею під ROC–кривою (англ. *area under curve* або скорочено *AUC*):

$$AUC = \int_0^1 TPR(FPR)dFPR\,.$$

Для ідеального правила $AUC = 1$, для випадкового $AUC = 0{,}5$.

Матриці помилок можна використовувати також і в задачах багатокласової класифікації. Наприклад, матриця помилок при розпізнаванні рукописного тексту може виглядати подібно до наведеної в табл. 6.2.

[10] Назва походить з радіолокації, де такі криві почали використовуватись ще з часів другої світової війни для підвищення якості виявлення повітряних цілей.

Таблиця 6.2 – Приклад матриці помилок для багатокласової класифікації

Розпізнано \ Рукопис	а	б	…	я
а	**70**	1	…	0
б	2	**56**	…	1
…	…	…	…	…
я	0	0	…	**34**

Багато показників якості бінарної кваліфікації природно узагальнюються на випадок багатокласової. Наприклад, якщо позначити елементи матриці помилок через x_{ij}, влучність правила відносно *i*–го класу визначається як

$$prec_i = \frac{x_{ii}}{\sum_j x_{ij}}, \tag{6.2}$$

тобто як частка правильних прогнозів серед всіх об'єктів, віднесених до класу *i* (аналог прогностичної значущості). *Покриттям* (англ. *coverage*, *recall*) *i*–го класу називається частка правильно визначених представників цього класу:

$$cvr_i = \frac{x_{ii}}{\sum_j x_{ji}}. \tag{6.3}$$

(аналог чутливості та специфічності). Хорошим показником якості класифікації у цілому, крім точності *ACC*, є коефіцієнт кореляції Меттьюза, який у випадку багатокласової класифікації розраховується як

$$MCC = \frac{ACC \times n^2 - \sum_i t_i p_i}{\sqrt{\left(n^2 - \sum_i t_i^2\right)\left(n^2 - \sum_i p_i^2\right)}}, \tag{6.4}$$

де t_i – справжня кількість елементів *i*–го класу у виборці, а p_i – прогнозована кількість елементів *i*–го класу.

6.3 Одновимірна класифікація: алгоритм One Rule

Це найпростіший з методів класифікації, який класифікує об'єкти за одним єдиним атрибутом (що обумовлює назву алгоритму One Rule або 1R, тобто єдине правило).

Алгоритм генерує одне правило для кожного атрибута даних (предиктора), а потім вибирає правило з найменшою загальною помилкою як своє «єдине правило». Щоб створити правило для атрибута, треба створити частотну таблицю, яка кожному його можливому значенню ставить у відповідність кількість об'єктів певного класу. Клас, який зустрічається найчастіше, і буде очікуваним значенням класу для цього атрибуту. Таким чином, класифікаційне правило формується у вигляді:

`if` значення атрибуту X == x `then` клас = найчастіший клас для x. (6.5)

Точність правила оцінюється за відсотком правильно обраних класів. Цей процес повторюється для всіх предикторів, і обирається атрибут із найменшою відносною похибкою. Отже, схема алгоритму 1R виглядає наступним чином.

Для кожного предиктора:

1) визначити, як часто з'являється кожне значення класу для кожного значення предиктора;

2) знайти найчастіший клас;

3) сформувати правило, яке призначає найчастіший клас кожному значенню предиктора, тобто у формі (6.5);

4) обчислити сумарну похибку правил для кожного предиктора;

5) обрати предиктор із найменшою сумарною похибкою.

Приклад 6.1. Таблиця 6.3 наводить фрагмент бази даних про клієнтів компанії кабельного телебачення. В останньому стовпці відзначені ті з них, які підписані на платний канал «Євроспорт». Треба за допомогою алгоритму 1R побудувати класифікаційне правило для прогнозування підписки на цей канал на базі інформації з решти стовпців табл. 6.3.

Таблиця 6.3 – Вихідні дані про клієнтів компанії кабельного телебачення

№	Стать	Дохід	Студент?	Євроспорт?
1	Ч	високий	так	так
2	Ж	середній	ні	ні
3	Ч	низький	так	ні
4	Ж	низький	ні	ні
5	Ч	середній	ні	так
6	Ж	високий	ні	так
7	Ж	середній	так	так
8	Ч	середній	так	так
9	Ж	низький	так	ні
10	Ж	середній	ні	ні

Створимо частотні таблиці для кожного з трьох предикторів «Стать», «Дохід» та «Студент». Результати зведені в табл. 6.4.

Таблиця 6.4 – Частотні таблиці для різних предикторів

Предиктор	Значення	Підписує «Євроспорт»?		Точність
		так	ні	
Стать	Ч	**3**	1	70%
	Ж	2	**4**	
Дохід	високий	**2**	0	80 %
	середній	**3**	2	
	низький	0	**3**	
Студент	так	**3**	2	60%
	ні	2	**3**	

Дохід обирається в якості найкращого предиктора, бо має найменшу похибку (2/10). Остаточне правило має вигляд:

`if` Дохід == низький `then` Євроспорт = ні `else` Євроспорт = так.

Прогнози, побудовані за цим правилом, наведені в табл. 6.5. Матриця помилок і основні показники якості класифікації наведені в табл. 6.6.

Коефіцієнт кореляції Меттьюза складає приблизно 0,65. Отримати це значення можна або за формулою (6.1), або розрахувавши коефіцієнт кореляції між стовпцями «Євроспорт» та «Прогноз» таблиці 6.5. В якості вправи для самостійної роботи переконайтесь, що результати будуть співпадати. ■

Таблиця 6.5 – Точність прогнозування для побудованого правила

№	Стать	Дохід	Студент?	Євроспорт?	Прогноз
1	Ч	високий	так	так	так
2	Ж	середній	ні	ні	так
3	Ч	низький	так	ні	ні
4	Ж	низький	ні	ні	ні
5	Ч	середній	ні	так	так
6	Ж	високий	ні	так	так
7	Ж	середній	так	так	так
8	Ч	середній	так	так	так
9	Ж	низький	так	ні	ні
10	Ж	середній	ні	ні	так

Таблиця 6.6 – Матриця помилок для побудованого правила

		Підписує «Євроспорт»?		
		ні	так	прогностична значущість:
Прогноз	ні	3	0	негативного результату $NPV = 3/3 = 1$
	так	2	5	позитивного результату $PPV = 5/7 \approx 0{,}71$
		специфічність $TNR = 3/5 = 0{,}6$	чутливість $TPR = 5/5 = 1$	точність $ACC = 8/10 = 0{,}8$

Перевагами алгоритму 1R є його простота та легкість інтерпретації. Недоліки теж очевидні:

1) не використовується вся інформація, що міститься в вихідних даних;
2) алгоритм розрахований на випадок категоріальних атрибутів.

З другою проблемою можна впоратись, перетворивши кількісні атрибути на категоріальні (власне, в прикладі 6.1 так і зроблено для змінної «Дохід»), але при цьому теж втрачається частина інформації.

Більш досконалі алгоритми класифікації будуть розглянуті у наступних розділах.

6.4 Межі точності класифікації

Якщо в даних існують об'єкти з однаковими атрибутами, які належать до різних класів, то клас об'єкту не може бути однозначно встановлено. Тому існує верхня границя точності, яка не може бути перевищена будь-яким алгоритмом класифікації.

У відповідності з типовими припущеннями математичної статистики, будемо вважати об'єкти досліджуваної вибірки реалізаціями багатовимірної випадкової величини *X*. Позначимо через $P(C_k \mid X = \mathbf{x})$ умовну ймовірність того, що об'єкт з атрибутами **x** належить до класу $C_k, k = 1,...,m.$ Якщо нам була б відомою ця система умовних ймовірностей, то найкращим предиктором класу для цього об'єкту буде $C_{B(\mathbf{x})}$, де

$$B(\mathbf{x}) = \arg\max_{k=1,...,m} P(C_k \mid X = \mathbf{x}). \qquad (6.6)$$

Ця формула визначає так званий *байєсівський класифікатор* (англ. *Bayes classifier*). При його використанні можлива помилка у визначенні справжнього класу об'єкта, ймовірність якої складає:

$$BE(\mathbf{x}) = \sum_{k=1, k\neq B(\mathbf{x})}^{m} P(C_k \mid \mathbf{x}) = 1 - P(C_{B(\mathbf{x})} \mid \mathbf{x}), \qquad (6.7)$$

де останнє перетворення випливає із повноти та взаємної виключності класів: $\sum_{k=1}^{m} P(C_k \mid \mathbf{x}) = 1.$ Ймовірність неправильного визначення класу, яка задається формулою (6.7), називається *байєсівською частотою помилок* (англ. *Bayes error rate*). Байєсівський класифікатор максимізує $P(C_{B(\mathbf{x})} \mid \mathbf{x})$ і, відповідно, має мінімальну частоту помилок.

Байєсівська частота помилок буде відмінною від нуля, якщо існує ненульова ймовірність, що об'єкти різних класів матимуть однакові атрибути. У цьому випадку при класифікації неможливо уникнути помилок. Оскільки справжній закон розподілу випадкової величини *X* зазвичай невідомий, будь-який алгоритм класифікації буде помилятися з частотою вище байєсівської.

Контрольні запитання

1. В чому полягає задача класифікації?
2. Назвіть основні види класифікації.
3. Як задачі класифікації пов'язані із машинним навчанням?.
4. Як перевіряється надійність алгоритму класифікації?
5. В чому полягає стратегія перехресної перевірки?
6. Чим відрізняються задачі бінарної класифікації від перевірки статистичних гіпотез?
7. Як будується матриця помилок для задач бінарної класифікації?
8. Як визначаються специфічність та чутливість бінарної класифікації?

9. Що мається на увазі під прогностичною значущістю бінарної класифікації?

10. Як розраховується і інтерпретується коефіцієнт кореляції Меттьюза?

11. Що характеризує ROC–крива і як її побудувати?

12. Як оцінити якість бінарної класифікації за допомогою ROC–кривої?

13. Як будується матриця помилок для задач багатокласової класифікації?

14. За якими формулами розраховуються покриття та влучність в задачах багатокласової класифікації?

15. Що таке класифікаційне правило?

16. Поясніть спосіб побудови частотних таблиць в алгоритмі OneRule.

17. Як визначається клас об'єкту в алгоритмі OneRule?

18. Дайте визначення байєсівського класифікатору

19. Як визначається байєсівська частота помилок?

20. Чому байєсівська частота помилок задає верхню планку для точності алгоритму класифікації?

Завдання для самостійної роботи

6.1. Наведіть два приклади практичних задач, які потребують вирішення задачі класифікації.

6.2. Покажіть, що точність класифікатора є зваженим середнім чутливості та специфічності, а саме:

$$ACC = \frac{P}{P+N} TPR + \frac{N}{P+N} TNR.$$

6.3. Доведіть формулу (6.1).

6.4. Розглянемо наступний тренувальний набір даних:

№	X	Y	Z	Клас
1	15	1	A	1
2	20	3	B	0
3	25	2	A	1
4	30	4	A	1
5	35	2	B	0
6	25	4	A	1
7	15	2	B	0
8	20	3	B	0
9	10	2	A	0
10	20	1	B	1
11	30	3	A	0
12	40	2	B	0
13	15	1	B	1

а) Побудуйте частотні таблиці для кожного з предикторів *X*, *Y*, *Z*. Для атрибуту *X* розгляньте всі можливі точки поділу значень на дві частини за схемою $X_i < c_j$ та $X_i > c_j$, де $c_j = \frac{x_{(j)} + x_{(j+1)}}{2}$, а $x_{(j)}$ позначає j-й елемент варіаційного ряду для *X*.

б) Оберіть найкраще класифікаційне правило за критерієм найменшої сумарної помилки.

в) Побудуйте матрицю помилок для обраного класифікаційного правила за схемою табл. 6.6.

г) Розрахуйте коефіцієнт кореляції Меттьюза для обраного правила.

6.5*. Рівень глюкози в крові натщесерце для здорових людей має нормальний розподіл із середнім значенням 4,8 ммоль/л і середньоквадратичним відхиленням 0,5 ммоль/л. На початковій стадії цукрового діабету параметри розподілу складають 6,5 та 0,6 ммоль/л, відповідно (цифри приблизні). Будемо вважати розподіл рівня глюкози нормальним як для здорових, так і для хворих.

а) Побудуйте на одному графіку криві розподілу рівня глюкози в крові для здорових і для хворих. Яке значення цього показника Ви б обрали в якості гранично допустимого, щоб вважати людину здоровою?

б) Розглянемо тест, який діагностує у людини діабет, якщо рівень глюкози в крові перевищує g ммоль/л. Розрахуйте чисельно ймовірності істинно-позитивного, істинно-негативного, хибно-позитивного та хибно-негативного результатів такого тесту при $g = 4; 4{,}1; \ldots; 6{,}9; 7$.

в) За результатами попереднього пункту розрахуйте чутливість і специфічність тесту і побудуйте криву помилок (ROC-криву).

г) Чисельно розрахуйте значення площі під кривою помилок (AUC).

д) Проаналізуйте, як залежить форма кривої помилок та показник AUC від різниці між середніми значеннями двох груп та величини СКВ.

7. ОРДИНАРНІ МЕТОДИ КЛАСИФІКАЦІЇ

7.1 Дерева рішень

Класифікаційне правило, яке було збудовано в прикладі 6.1, коректно визначає, що клієнти з високим доходом завжди підписуються на платний преміум канал, а клієнти з низьким доходом – ні. Проблематичним є визначення перспективних передплатників для клієнтів із середнім доходом. Логічно припустити, що в цьому випадку доцільно буде переглянути інші атрибути клієнта і побудувати за ними додаткові правила. Ця ідея реалізована в групі методів класифікації, спільно відомих як дерева рішень.

Дерева рішень (англ. *decision trees*) є ієрархічними деревовидними структурами, що складаються з правил рішень виду «якщо – тоді». Правила в деревах рішень обираються шляхом узагальнення множини окремих спостережень (навчальних прикладів). Тому їх називають індуктивними правилами по аналогії з відповідним методом логічного висновку, а сам процес навчання – *індукцією* дерев рішень (англ. *decision tree induction*).

Дерева рішень були запропоновані як засіб комп'ютерного моделювання поведінки людей в кінці 50–х років минулого сторіччя в роботах Ховленда [30], Ханта, Маріна та Стоуна [31]. В середині 1980–х Джон Росс Куінлен розробив алгоритм ID3 і його вдосконалення C4.5 [42, 43], а Лео Брейман, Джером Фрідман, Чарльз Стоун та Річард Олшен. – алгоритм CART [21]. Ці алгоритми стали найпоширенішими методами побудови дерев класифікації.

Дерево рішень – це спосіб представлення правил в ієрархічній, послідовній структурі. Для класифікації об'єкта потрібно відповісти на ряд питань, які знаходяться в *вузлах* (англ. *nodes*) цього дерева, починаючи з його кореня. В найпростішому випадку є всього два варіанти відповіді на кожне запитання, «так» чи «ні». При позитивній відповіді на запитання здійснюється перехід до лівої гілки дерева, при негативному – до правої. Далі йдуть наступні питання, доки не буде досягнутий кінцевий вузол дерева, що є вузлом рішення, або *листом* (англ. *leaf*). Лист призначає клас кожному об'єкту, що потрапив в нього. Оскільки шлях в дереві до кожного листа єдиний, то кожний об'єкт може потрапити тільки в один лист, що гарантує єдиність рішення.

В якості прикладу на рис. 7.1 наведено дерево рішень для розподілу студентів школи Хогвартс за коледжами Сортувальним Капелюхом в серії книг та фільмів про Гаррі Поттера (в інтерпретації автора). Листя дерева відмічено зеленим кольором.

На етапі навчання моделі будується дерево класифікації, тобто створюється набір правил для прийняття рішення. Правилом є умовний оператор «якщо, то». Кожне правило перевіряє один із атрибутів, які називають прогнозуючими або *атрибутами розгалуження* (англ. *splitting attribute*).

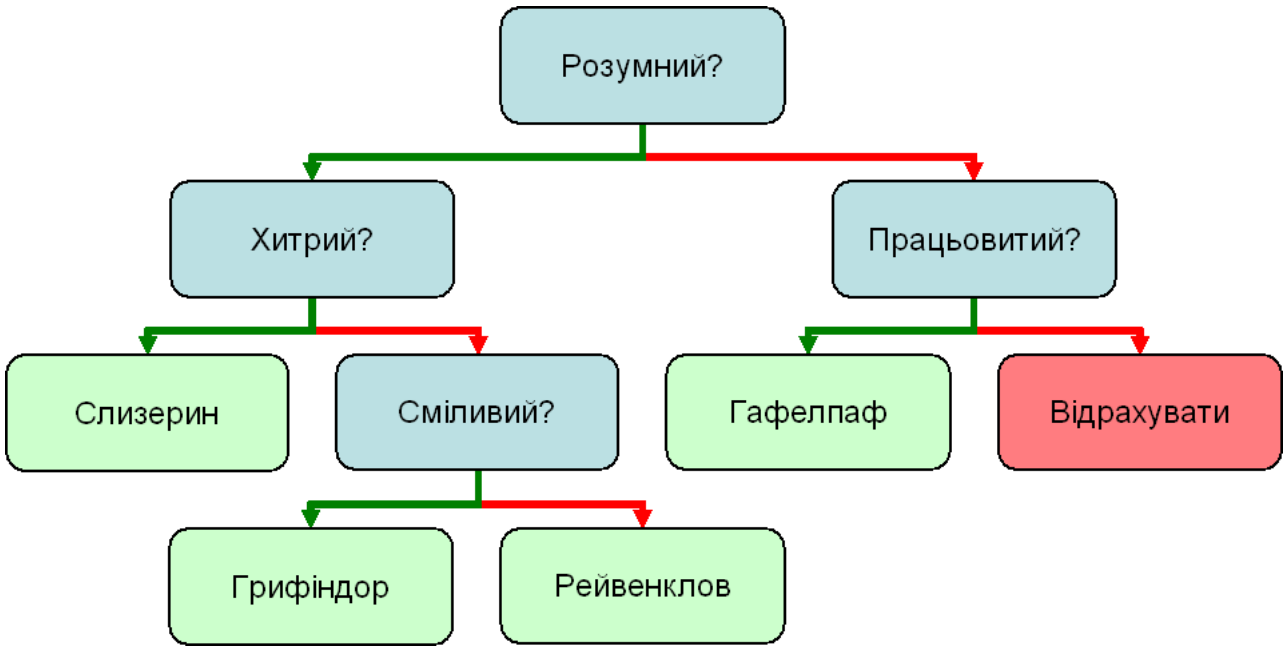


Рис. 7.1. Дерево рішень для розподілу студентів за коледжами в Хогвартс

Кожна гілка дерева, що йде від внутрішнього вузла, відзначена предикатом розгалуження. Найчастіше використовуються бінарні предикати «так» чи «ні», але можуть бути більш, ніж два предикати. Об'єднана інформація про атрибути і предикати розгалуження називається *критерієм розгалуження* (англ. *splitting criterion*). Якість побудованого дерева рішення в значній мірі залежить від правильного вибору критерію розгалуження.

На етапі використання моделі побудоване дерево обходиться від його кореня до одного з листів, щоб вирішити задачу класифікації (у прикладі на рис. 7.1, відповісти на запитання «до якого коледжу розподілити студента?»).

Процес побудови дерев рішень полягає в послідовному, рекурсивному розбитті навчальної вибірки на підмножини із застосуванням вирішальних правил у вузлах. Процес розбиття триває доти, доки усі кінцеві вузли не вважатимуться листям. Визначення вузла листом відбувається у наступних випадках:

– коли відповідна вузлу підмножина містить об’єкти тільки одного класу;

– коли відповідна вузлу підмножина не містить жодного об’єкту;

– коли вичерпані всі доступні для розгалуження атрибути;

– коли досягнуто задану розробником умову зупинки (наприклад, мінімально допустиму кількість об’єктів у вузлі або максимальну глибину дерева).

Найбільш популярні алгоритми побудови дерев рішень ID3, C4.5 та CART відносяться до категорії так званих *жадібних* (англ. *greedy*) алгоритмів. Так називаються алгоритми, які на кожному кроці обирають локально–оптимальне рішення. Надалі ці рішення не переглядаються. Жадібні алгоритми не гарантують знаходження глобально–оптимального рішення, але є швидкими і простими до розуміння. У контексті дерев рішень «жадібність» полягає у тому,

що на кожному кроці для розгалуження обирається атрибут, який наводить до найкращого у певному сенсі впорядковування вихідних даних.

Алгоритми ID3, C4.5 та CART мають однакову структуру і розрізняються лише критерієм для обрання атрибута розгалуження. Позначимо через $\boldsymbol{D}$ тренувальний набір даних $(\mathbf{X}, \mathbf{y})$, де y_i, $i = 1,...n$ – позначка класу *i*–об'єкту, а $x_{ij}, i = 1,...,n,\ j = 1,...,m$ – *j*-й атрибут *i*–го об'єкту. Позначимо через $A = \{1,...,m\}$ перелік доступних атрибутів. Для спрощення будемо поки вважати всі дані категоріальними. Загальна схема алгоритму побудови класифікаційного дерева `GenerateTree(`$\boldsymbol{D}$`,` A`)` полягає у наступному.

1. Створити вузол *Node*.
2. Якщо всі об'єкти в $\boldsymbol{D} \in C$, повернути *Node* як лист з поміткою класу *C*.
3. Якщо $m = 0$, повернути *Node* як лист з поміткою найпоширенішого класу в $\boldsymbol{D}$.
4. Обрати найкращий атрибут для розгалуження $s \in \{1,...m\}$.
5. Помітити *Node* критерієм розгалуження для атрибута *s*.
6. Для кожного можливого значення *k* атрибуту *s*:
   6.1. Створити підмножину об'єктів $D_k = (\mathbf{X}_k, \mathbf{y}_k) \subseteq D$, для яких $x_{is} = k$.
   6.2. Якщо $D_k = \varnothing$,
   приєднати до *Node* лист з позначкою найпоширенішого класу в $\boldsymbol{D}$;
   інакше приєднати до *Node* вузол `GenerateTree(`$\boldsymbol{D_k}$`,` $A \setminus \{s\}$`)`.
7. Повернути *Node*.

Обрання атрибуту для розгалуження вимагає окремого алгоритму. Ідея полягає в тому, щоб досягнути максимальної впорядкованості підмножин, що утворюються після розгалуження, але оцінка здійснюється по різному.

7.1.1 Алгоритм ID3 (скорочення від англ. Iterative Dichotomiser).

Критерієм вибору атрибуту є мінімальна зважена ентропія розбиття. Поняття ентропії було введено засновником теорії інформації Клодом Шенноном як міра невизначеності або хаотичності досліджуваної системи. Формально, *ентропія* (англ. *entropy*) дискретної випадкової величини *X* з можливими значеннями $\{x_1,...,x_m\}$ визначається як:

$$H(X) = -\sum_{i=1}^{m} p_i \log_2 p_i , \qquad (7.1)$$

де $p_i = P(X = x_i)$. Легко довести наступні твердження [27] :

1. $H(X) \geq 0$.
2. Ентропія приймає мінімальне значення 0, якщо існує значення *j*, для якого $p_j = 1$ (тобто *X* є детермінованою величиною)[11].

[11] При цьому вважається, що $p \log_2 p = 0$ при $p = 0$. Це випливає з того, що $p \log_2 p \to 0$ при $p \to 0$.

3. Для заданого m ентропія максимальна, якщо всі значення рівноймовірні. Вона дорівнює в цьому випадку:

$$H(X) = -\sum_{i=1}^{m} \frac{1}{m} \log_2 \frac{1}{m} = \log_2 m . \qquad (7.2)$$

Чим ближче ентропія до нуля, тим менше невизначеності, і навпаки.

Якщо вважати клас об'єкту випадковим і застосувати частотну інтерпретацію ймовірності, то формулу (7.1) можна застосувати для оцінки впорядкованості множини класів. Наприклад, для множини $A = \{1,0,1,1\}$ ентропія складатиме $-\left(0,75\log_2 0,75 + 0,25\log_2 0,25\right) \approx 0,81$, а для множини $B = \{1,0,0,1\}$ $-\left(0,5\log_2 0,5 + 0,5\log_2 0,5\right) = 1$. Отже, множина A є більш впорядкованою.

Тепер припустимо, що ми розіб'ємо множину $\boldsymbol{D}$ за деяким атрибутом s, який має l можливих значень. В результаті отримаємо l підмножин $\boldsymbol{D}_1, ..., \boldsymbol{D}_l$. В ідеалі таке розбиття вирішувало б задачу класифікації, тобто кожна підмножина містила б об'єкти тільки одного класу і її ентропія дорівнювала б нулю. Проте, це малоймовірно. Для того, щоб оцінити, наскільки розбиття близько до ідеального, можна розрахувати його середньозважену ентропію

$$H(\boldsymbol{D}, s) = \sum_{i=1}^{l} \frac{n(\boldsymbol{D}_i)}{n(\boldsymbol{D})} H_i(\boldsymbol{D}_i), \qquad (7.3)$$

де $n(A)$ позначає кількість елементів множини A.

В алгоритмі ID3 для розгалуження обирається такий атрибут s, який наводить до найменшого значення середньозваженої ентропії (7.3) або, еквівалентно, до найбільшого *інформаційного виграшу* (англ. *information gain*)

$$IG(\boldsymbol{D}, s) = H(\boldsymbol{D}) - H((\boldsymbol{D}, s) . \qquad (7.4)$$

Приклад 7.1. Застосуємо алгоритм ID3 для класифікації клієнтів компанії кабельного телебачення з прикладу 6.1. Для зручності дані зведені в табл. 7.1.

Таблиця 7.1 – Вихідні дані про клієнтів компанії кабельного телебачення

№	Стать	Дохід	Студент?	Євроспорт?
1	Ч	високий	так	так
2	Ж	середній	ні	ні
3	Ч	низький	так	ні
4	Ж	низький	ні	ні
5	Ч	середній	ні	так
6	Ж	високий	ні	так
7	Ж	середній	так	так
8	Ч	середній	так	так
9	Ж	низький	так	ні
10	Ж	середній	ні	ні

На першому етапі слід обрати серед трьох атрибутів – кандидатів для розгалуження: стать, дохід та студент. В табл. 6.4 ми вже створювали частотні таблиці для цих атрибутів, що стане в нагоді при обчисленні ентропії.

Результати обчислень інших параметрів зведені в табл. 7.2.

Таблиця 7.2 – Вибір атрибуту розгалуження на першому етапі ID3

Атрибут	Значення	«Євроспорт»?		Позначка	Ентропія	Зважена ентропія
		так	ні			
Стать	Ч	**3**	1	$\boldsymbol{D}_1$	0,811	0,875
	Ж	2	**4**	$\boldsymbol{D}_2$	0,918	
Дохід	високий	**2**	0	$\boldsymbol{D}_1$	0,000	0,485
	середній	**3**	2	$\boldsymbol{D}_2$	0,971	
	низький	0	**3**	$\boldsymbol{D}_3$	0,000	
Студент	так	**3**	2	$\boldsymbol{D}_1$	0,971	0,971
	ні	2	**3**	$\boldsymbol{D}_2$	0,971	

Для розгалуження обирається атрибут «дохід», оскільки він має найнижчу ентропію розбиття. Результуючі підмножини $\boldsymbol{D}_1$ та $\boldsymbol{D}_3$ є чистими (складаються з об'єктів лише одного класу) і стають листям класифікаційного дерева. Для підмножини $\boldsymbol{D}_2$ повторюємо розгалуження за атрибутами, що залишилися. Результати розрахунків ентропії зведені в табл. 7.3.

Таблиця 7.3 – Вибір атрибуту розгалуження на другому етапі ID3

Атрибут	Значення	«Євроспорт»?		Позначка	Ентропія	Зважена ентропія
		так	ні			
Стать	Ч	**2**	0	$\boldsymbol{D}_1$	0,000	0,551
	Ж	1	**2**	$\boldsymbol{D}_2$	0,918	
Студент	так	**2**	0	$\boldsymbol{D}_1$	0,000	0,551
	ні	1	**2**	$\boldsymbol{D}_2$	0,918	

В даному випадку ентропії розбиття по обох атрибутах співпадають, так що можна обрати будь-який з них, наприклад, «стать». При цьому підмножина $\boldsymbol{D}_1$ стає новим листям дерева, а підмножину $\boldsymbol{D}_2$ можна розбити далі за останнім атрибутом, що залишився («студент»). Результати наведені в табл. 7.4.

Таблиця 7.4 – Результати розгалуження на третьому етапі ID3

Атрибут	Значення	«Євроспорт»?		Позначка	Ентропія	Зважена ентропія
		так	ні			
Студент	так	**1**	0	$\boldsymbol{D}_1$	0,000	0,000
	ні	0	**2**	$\boldsymbol{D}_2$	0,000	

Побудоване дерево рішень наведено на рис. 7.2, де листя виділено фоновим кольором.

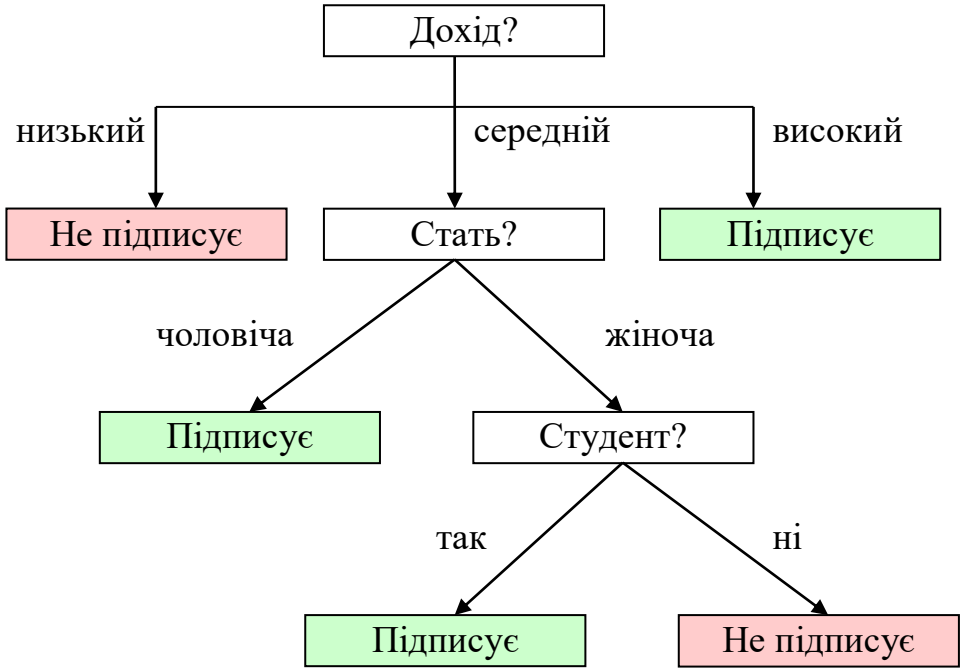


Рис. 7.2. Класифікаційне дерево для даних з табл. 7.1 за алгоритмом ID3

Класифікаційне дерево можна також подати у вигляді таблиці, яка послідовно відсортована за атрибутами розгалуження, як в табл. 7.5. Фоновим кольором виділені атрибути, які задіяні в процесі класифікації.

Таблиця 7.5 – Подання дерева класифікації у табличному вигляді

№	Дохід	Стать	Студент?	Євроспорт?
1	високий	Ч	так	так
6	високий	Ж	ні	так
3	низький	Ч	так	ні
9	низький	Ж	так	ні
4	низький	Ж	ні	ні
8	середній	Ч	так	так
5	середній	Ч	ні	так
7	середній	Ж	так	так
2	середній	Ж	ні	ні
10	середній	Ж	ні	ні

Отже, в даному прикладі нам вдалося побудувати класифікаційне дерево, яке здійснює класифікацію зі 100% точністю (у тренувальному наборі даних).

Недоліком алгоритму ID3 є те, що критерій мінімуму ентропії розбиття надає перевагу атрибутам з великою кількістю можливих значень. Наприклад, якби ми використали в якості атрибуту розгалуження номер клієнта, то кожний клієнт потрапив би в окремий лист. Ентропія такого розбиття мінімальна, але зрозуміло, що прогностична цінність такої моделі невелика.

7.1.2 Алгоритм C4.5

Щоб подолати наведену вище схильність алгоритму ID3 до обрання для розгалуження атрибутів з багатьма значеннями, у його вдосконаленій версії C4.5 робиться спроба нормалізувати інформаційний виграш відносно до розміру утворюваних підмножин.

При розбитті множини $\boldsymbol{D}$ за деяким атрибутом s на підмножини $\boldsymbol{D}_i$ ймовірність того, що об'єкт потрапить до $\boldsymbol{D}_i$, становить $\frac{n(\boldsymbol{D}_i)}{n(\boldsymbol{D})}$. Ентропія такого розбиття, яку Куінлен (з точністю до навпаки) назвав *інформацією розбиття* (англ. *split information*) [43] складає:

$$SI(\boldsymbol{D}, s) = -\sum_i \frac{n(\boldsymbol{D}_i)}{n(\boldsymbol{D})} \log_2 \frac{n(\boldsymbol{D}_i)}{n(\boldsymbol{D})}. \tag{7.5}$$

Це значення надає потенційну інформацію, яку можна отримати шляхом розподілу даних за значеннями атрибуту s. На відміну від цього, інформаційний виграш (7.4) вказує, наскільки те ж саме розбиття допомагає у вирішенні задачі класифікації. В алгоритмі C4.5 пропонується обрати атрибут для розгалуження за критерієм максимального *відношення інформаційного виграшу* до інформації розбиття (англ. *information gain ratio*):

$$IGR(\boldsymbol{D}, s) = \frac{IG(\boldsymbol{D}, s)}{SI(\boldsymbol{D}, s)}. \tag{7.6}$$

Приклад 7.2. Для даних з табл. 7.1, при розбитті тренувального набору даних за статтю (gender):

$$H(\boldsymbol{D}) = -\left(\frac{5}{10}\log_2\frac{5}{10} + \frac{5}{10}\log_2\frac{5}{10}\right) = 1;$$

$$H(\boldsymbol{D}, gender) \approx 0{,}875;$$

$$IG(\boldsymbol{D}, gender) = H(\boldsymbol{D}) - H(\boldsymbol{D}, gender) \approx 0{,}125;$$

$$SI(\boldsymbol{D}, gender) = -\left(\frac{4}{10}\log_2\frac{4}{10} + \frac{6}{10}\log_2\frac{6}{10}\right) \approx 0{,}971;$$

$$IGR(\boldsymbol{D}, gender) = \frac{0{,}125}{0{,}971} \approx 0{,}129.$$

При розбитті за доходом (income):

$$H(\boldsymbol{D}) = -\left(\frac{5}{10}\log_2\frac{5}{10} + \frac{5}{10}\log_2\frac{5}{10}\right) = 1;$$

$$H(\boldsymbol{D}, income) \approx 0{,}485;$$

$$IG(\boldsymbol{D}, income) = H(\boldsymbol{D}) - H(\boldsymbol{D}, income) \approx 0{,}515;$$

$$SI(\boldsymbol{D}, gender) = -\left(\frac{2}{10}\log_2\frac{2}{10} + \frac{5}{10}\log_2\frac{5}{10} + \frac{3}{10}\log_2\frac{3}{10}\right) \approx 1{,}485;$$

$$IGR(\boldsymbol{D}, income) = \frac{0,515}{1,485} \approx 0,347.$$

В даному випадку зміна критерію не вплинула на вибір атрибуту для розгалуження, але різниця між альтернативами трохи зменшилась.

Решта розрахунків виконується аналогічно. ■

На жаль, показник *IGR* теж не позбавлений недоліків. На відміну від інформаційного виграшу, він надаватиме перевагу атрибутам, які мають меншу кількість окремих значень.

В алгоритмі C4.5 додано також можливість роботи з атрибутами, значення яких є безперервними величинами [44]. Представимо їх у вигляді варіаційного ряду $x_{(1)} < x_{(2)} < ... < x_{(n)}$ (див. п. 3.2). Для розщеплення за умовою «менше-більше» слід обрати деяке граничне значення у цьому інтервалі. В якості кандидатів оберемо середини інтервалів між послідовними значеннями: $\frac{x_{(1)} + x_{(2)}}{2}$, $\frac{x_{(2)} + x_{(3)}}{2}$ тощо. В результаті отримуємо $n - 1$ кандидатів для розбиття, серед яких простим перебором обирається найкращий за критерієм відношення інформаційного виграшу.

Структура алгоритму при наявності атрибутів із неперервними значеннями є дещо складнішою, оскільки:

1) треба зберігати знайдені граничні значення;

2) можливе подальше розгалуження за тим же самим атрибутом.

7.1.3 Алгоритм CART (від англ. Classification and Regression Trees)

В цьому алгоритмі для вибору атрибуту розгалуження замість ентропії та інших концепцій з теорії інформації використовується *індекс Джині* (англ. *Gini index*), запозичений з соціології та економіки [19]. Стосовно задач класифікації, він розраховується як:

$$G(\boldsymbol{D}) = 1 - \sum_{i=1}^{m} p_i^2, \qquad (7.7)$$

де p_i – ймовірність того, що елемент множини $\boldsymbol{D}$ належить до i–го класу. Згідно частотної інтерпретації ймовірності, p_i розраховується як відносна частота класу у виборці $\boldsymbol{D}$. Легко бачити, що індекс Джині приймає своє мінімальне значення 0, коли всі елементи належатимуть до одного класу. Також нескладно довести, що значення індексу буде максимальним і дорівнюватиме $1 - \frac{1}{m}$ в тому випадку, коли всі класи у виборці рівноймовірні. Отже, індекс Джині теж можна розглядати як показник впорядкованості множини.

В алгоритмі CART розглядаються тільки бінарні розбиття. Після цього утворюються дві підмножини $\boldsymbol{D}_1$ та $\boldsymbol{D}_2$, а зважений індекс Джині складатиме:

$$G(\boldsymbol{D}_1 \cup \boldsymbol{D}_2) = \frac{n(\boldsymbol{D}_1)}{n(\boldsymbol{D})} G(\boldsymbol{D}_1) + \frac{n(\boldsymbol{D}_2)}{n(\boldsymbol{D})} G(\boldsymbol{D}_2). \quad (7.8)$$

Обирається таке розбиття, яке матиме найменше значення індексу Джині. З бінарними атрибутами все просто. Континуальні атрибути обробляються так само, як і в алгоритмі C4.5. Що ж стосується категоріальних атрибутів з $l > 2$ можливих значень, то розглядаються всі можливі варіанти бінарного розбиття. Так, у прикладі 7.1 атрибут «дохід» має три можливих значення: високий, середній та низький. З них можна утворити три бінарні розбиття:

1. {високий} ∪ {середній, низький}
2. {середній} ∪ {високий, низький}
3. {низький} ∪ {високий, середній}.

З цих трьох варіантів обирається найкращий за індексом Джині.

Якщо атрибут має l можливих значень, то з них може бути утворено 2^l підмножин. Два з них, множина всіх можливих значень та пуста множина, не наводять до розбиття на дві непустих підмножини. Серед решти половина підмножин буде дублюватися. Отже, всього існує $\frac{2^l - 2}{2} = 2^{l-1} - 1$ бінарних розбиттів. Це означає, що кількість комбінацій, які підлягають оцінці, зростає експоненційно із збільшенням l, що може призвести до неприйнятної кількості обчислень.

Приклад 7.4. Розглянемо ще раз перший етап розгалуження для даних з табл. 7.1, тепер за методикою алгоритму CART.

Індекс Джині для вихідного набору даних складає:

$$G = 1 - \left(\frac{5}{10}\right)^2 - \left(\frac{5}{10}\right)^2 = \frac{1}{2}.$$

Після розбиття за доходом на категорії {високий} та {середній, низький} в першій групі опиняться два клієнти компанії, кожен з яких підписує «Євроспорт», а в другій – 8 клієнтів, серед яких 3 передплатника. Індекс Джині після такого розщеплення буде дорівнювати:

$$G' = \frac{2}{10} \times 0 + \frac{8}{10} \times \left(1 - \left(\frac{3}{8}\right)^2 - \left(\frac{5}{8}\right)^2\right) = \frac{3}{8} = 0{,}375.$$

Решта обчислень за тією ж схемою наведена в табл. 7.6. Найкращим виявляється розбиття за атрибутом «дохід» на значення «низький» та інші. Підмножина, що відповідає низькому доходу, є чистою і стає листом дерева. Для другої підмножини індекс Джині становить:

$$G = 1 - \left(\frac{5}{7}\right)^2 - \left(\frac{2}{7}\right)^2 = \frac{20}{49} \approx 0{,}408.$$

Наступні етапи роботи алгоритму зведені в табл. 7.7–7.9.

Таблиця 7.6 – Вибір атрибуту розгалуження на першому етапі CART

Атрибут	Значення	«Євроспорт»?		Індекс Джині	Зважений індекс Джині
		так	ні		
Стать	Ч	**3**	1	0,375	0.417
	Ж	2	**4**	0,444	
Дохід	високий	**2**	0	0,000	0.375
	середній, низький	3	**5**	0,469	
	середній	**3**	2	0,480	0.480
	високий, низький	2	**3**	0,480	
	низький	0	**3**	0,000	0.286
	високий, середній	**5**	2	0,408	
Студент	так	**3**	2	0,480	0.480
	ні	2	**3**	0,480	

Таблиця 7.7 – Вибір атрибуту розгалуження на другому етапі CART

Атрибут	Значення	«Євроспорт»?		Індекс Джині	Зважений індекс Джині
		так	ні		
Стать	Ч	**3**	0	0,000	0.286
	Ж	2	2	0,500	
Дохід	високий	**2**	0	0,000	0.343
	середній	**3**	2	0,480	
Студент	так	**3**	0	0,000	0.286
	ні	2	2	0,500	

Таблиця 7.8 – Вибір атрибуту розгалуження на третьому етапі CART

Атрибут	Значення	«Євроспорт»?		Індекс Джині	Зважений індекс Джині
		так	ні		
Дохід	високий	**1**	0	0.000	0.333
	середній	1	**2**	0.444	
Студент	так	**1**	0	0.000	0.333
	ні	1	**2**	0.444	

Таблиця 7.9 – Вибір атрибуту розгалуження на четвертому етапі CART

Атрибут	Значення	«Євроспорт»?		Індекс Джині	Зважений індекс Джині
		так	ні		
Студент	так	**1**	0	0.000	0.000
	ні	0	**2**	0.000	

На другому і третьому етапах виявилось, що два варіанти розбиття мають однаковий виграш. В обох випадках для розгалуження було обрано атрибут, що вище розташований у таблиці. Остаточне дерево рішень наведено на рис. 7.3. ■

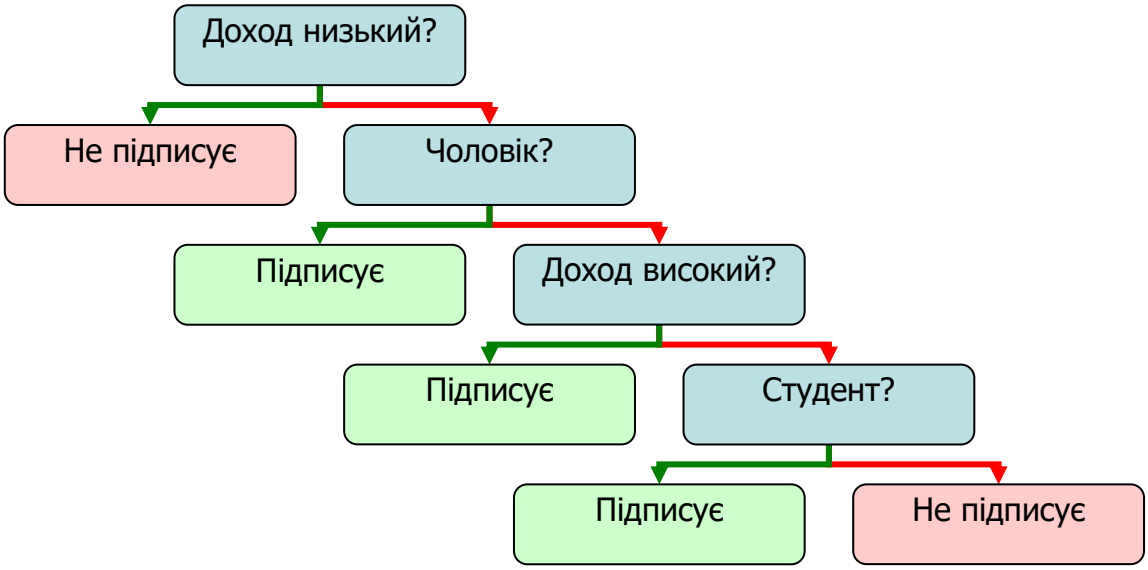


Рис. 7.3. Кваліфікаційне дерево для даних з табл. 7.1 за алгоритмом CART

Використання в CART тільки бінарного розбиття є одночасно його сильною та слабкою стороною. З одного боку, це позбавляє алгоритм проблем із схильністю до обрання атрибутів за кількістю їх значень, властивих ID3 та C4.5. Також це робить алгоритм однотипним і універсальним. З іншого боку, довжина дерев збільшується, а обчислювальна складність алгоритму може стати надмірною для дискретних атрибутів з великою кількістю значень.

В цілому, безперечною перевагою дерев рішень є прозорість алгоритму класифікації і простота його інтерпретації. Правила формулюються практично на природній мові, їх просто зрозуміти і пояснити іншим. Легко виділити умови, які критично впливають на результати класифікації (наприклад, пояснити клієнту банка, що йому потрібно зробити для отримання кредиту). Корисні правила можна отримати навіть при малому обсязі даних.

Недоліком дерев рішень є їх нестабільність: невеликі зміни у даних можуть радикально змінити структуру дерева. Точність класифікації не завжди є високою (особливо при обробці кількісних даних), а обчислення можуть бути дуже складними.

Незважаючи на ці недоліки, дерева рішень широко використовуються для вирішення задач класифікації. Згідно [56], алгоритм C4.5 є найпоширенішим алгоритмом інтелектуального аналізу даних взагалі.

7.2 Метод *k*–найближчих сусідів

Метод k–найближчих сусідів (англ. *k–nearest neighbor*, скорочено *kNN*) – це непараметричний метод навчання з учителем, розроблений Фіксом та Ходжесом у 1951 році і розвинутий пізніше [14]. На відміну від дерев рішень та НБК, він орієнтований на аналіз кількісних даних, хоча може бути адаптований і до категоріальних. Метод використовується як для класифікації, так і для регресії. В обох випадках вхідні дані складаються з k найближчих об’єктів у

виборці. Результат залежить від того, використовується kNN для класифікації чи для регресії.

В задачах класифікації результатом є призначення об'єкту мітки класу. Об'єкт класифікується «більшістю голосів» його сусідів, тобто відноситься до найпоширенішого класу серед своїх k найближчих сусідів. Якщо $k = 1$, то об'єкту присвоюється клас свого єдиного найближчого сусіда.

В задачах регресії результатом є числове значення певної ознаки об'єкта, яке отримується усередненням значень таких ознак для k найближчих сусідів. Якщо $k = 1$, то об'єкт успадковує значення ознаки свого найближчого сусіда.

Алгоритм kNN є прикладом *ледачого навчання* (англ. *lazy learning*), коли узагальнення тренувальних даних відбувається тільки тоді, коли необхідно класифікувати конкретний об'єкт. Це контрастує зі стратегією *охочого навчання* (англ. *eager learning*), яке вимагає узагальнення тренувальних даних перш ніж класифікувати нові об'єкти (як у деревах рішень). Алгоритм kNN не створює загальних класифікаційних правил або функцій, а намагається лише дізнатися значення такої функції у конкретній точці, тобто знаходить локальну апроксимацію.

Ледаче навчання взагалі і алгоритм kNN зокрема корисні в тих випадках, коли навчальні дані швидко застарівають. Наприклад, рекомендаційні онлайн системи часто працюють за схемою: «люди, які переглянули/купили/слухали певний фільм/книгу/мелодію, також були зацікавлені в ...». Оскільки постійно з'являються нові блокбастери/бестселери/хіти, навчання на старих даних швидко втрачає цінність.

Схема алгоритму дуже проста. Нехай задано тренувальний набір даних $\boldsymbol{D} = (\mathbf{X}, \mathbf{y})$ з n елементів і тестовий об'єкт z з вектором характеристик $\mathbf{z} \in \boldsymbol{R}^{\mathrm{m}}$. Обчислимо відстань між тестовим об'єктом та усіма елементами тренувальної множини $d(\mathbf{z}, \mathbf{x}_i), i = 1, ..., n.$ Позначимо через $D_z(k)$ множину k найближчих до z елементів тренувального набору. Тоді клас об'єкту z визначається як найчастіший клас у цій множині:

$$y^* = \arg\max_j \sum_{(\mathbf{x}_i, y_i) \in D_z(k)} I\{y_i = j\}, \tag{7.9}$$

де $I\{\cdot\}$ – індикаторна функція, яка дорівнює 1, якщо її аргумент істинний і 0 в іншому випадку.

Для роботи алгоритму необхідні дві складові: значення k і функція відстані між об'єктами $d(\mathbf{z}, \mathbf{x})$.

Вплив значення k на роботу алгоритму ілюструється на рис. 7.4. Тестовий набір даних складається з об'єктів двох класів: сині квадрати та червоні трикутники. При $k = 3$ тестовий об'єкт (зелене коло) буде віднесено до трикутників, а при $k = 5$ – до квадратів.

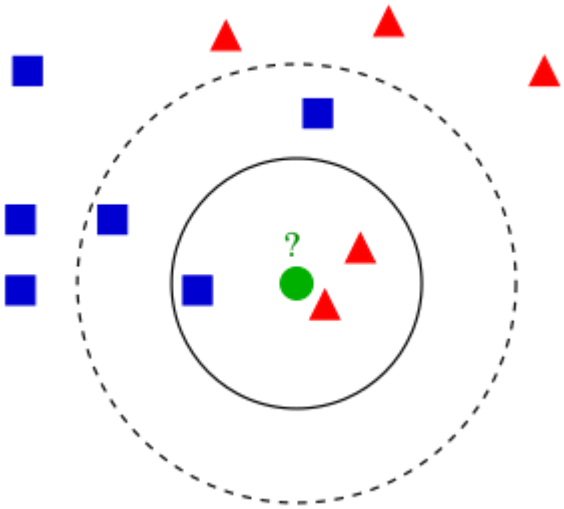


Рис. 7.4. Вплив значення *k* на результати класифікації.

Найкращий вибір *k* залежить від даних; як правило, більші значення *k* зменшують вплив шуму на класифікацію, але роблять межі між класами менш чіткими.

В якості функції відстані між об'єктами найчастіше використовується евклідова відстань:

$$d(\mathbf{z},\mathbf{x}) = \|\mathbf{z}-\mathbf{x}\| = \sqrt{\sum_{j=1}^{m}(z_j - x_j)^2} \qquad (7.10)$$

(або її квадрат). Інші поширені варіанти включають:

– відстань Чебишева, яка визначається як найбільша різниця серед значень всіх атрибутів:

$$d_C(\mathbf{z},\mathbf{x}) = \max_{j=1,\ldots,m} |z_j - x_j|; \qquad (7.11)$$

– Манхеттенську відстань, названу так за аналогією з довжиною шляху таксі між двома точками у місті з паралельно–перпендикулярними вулицями:

$$d_M(\mathbf{z},\mathbf{x}) = \sum_{j=1}^{m} |z_j - x_j|; \qquad (7.12)$$

– відстань Геммінга, яка використовується при аналізі текстів і визначається як кількість позицій, де різняться символи у порівнюваних рядках;

– при аналізі часових рядів сусідами визначаються просто найближчі за часом спостереження. В цьому випадку регресія за методом kNN еквівалентна прогнозуванню за методом *ковзного середнього* (англ. *moving average*).

Оскільки різні атрибути можуть вимірюватись у різних фізичних одиницях та/або мати дуже різні масштаби, то нормалізація навчальних даних може значно підвищити точність роботи алгоритму kNN. Для кількісних даних використовуються два основних прийоми нормалізації.

1. Мінімаксна нормалізація зводить змінну до інтервалу [0,1]:

$$x' = \frac{x - \min\limits_i x_i}{\max\limits_i x_i - \min\limits_i x_i}. \qquad (7.13)$$

Її недоліком є те, що при наявності в даних викидів вони розтягуватимуть діапазон, в результаті чого більшість перетворених даних буде сконцентровано у вузькому вікні нормалізованого діапазону.

2. Нормалізація середнім, або *z*–нормалізація:

$$x' = \frac{x - \overline{x}}{s_x}, \qquad (7.14)$$

де $\overline{x}$ та s_x – середнє значення та СКВ змінної *x*. Перетворена змінна буде мати нульове середнє та одиничне СКВ і приблизно стандартний нормальний розподіл, якщо розподіл перетворюваної змінної є близьким до нормального.

Класифікація може бути неточною, якщо найближчі сусіди сильно відрізняються за своєю відстанню до об'єкта. Тому корисним може бути використання вагових коефіцієнтів, щоб зробити внесок у середнє найближчих сусідів більшим, ніж у віддалених. Наприклад, кожному сусіду можна надати вагу, обернено пропорційну відстані до тестового об'єкту: $w_i = 1/d(\mathbf{z}, \mathbf{x}_i)$. Тоді правило (7.9) модифікується як:

$$y^* = \arg\max_j \sum_{(\mathbf{x}_i, y_i) \in D_z(k)} w_i I\{y_i = j\}. \qquad (7.15)$$

Класифікація за зваженим kNN (7.15) зазвичай є набагато менш чутливою до вибору кількості сусідів *k* [56].

7.3 Метод опорних векторів

Метод опорних векторів (англ. *Support Vector Machine*, скорочено *SVM*) був розроблений в 1990х рр. Володимиром Вапником та його колегами [17, 22]. Він відноситься до групи граничних методів, які визначають класи за допомогою кордонів областей. В базовій формі метод опорних векторів використовується для бінарної класифікації об'єктів.

В основі методу лежить поняття площин рішень. Площина рішення розділяє об'єкти з різною класової приналежністю.

На рис. 7.5 наведено приклад, в якому беруть участь об'єкти двох типів. Кожен з об'єктів характеризується двома атрибутами (x_1, x_2), що дозволяє зобразити їх у вигляді точок на площині. Мета методу опорних векторів полягає в тому, щоби знайти лінію (або, в загальному випадку, гіперплощину[12]), яка

[12] Гіперплощина є узагальненням поняття прямої лінії на випадок розмірності $m > 2$. Рівняння прямої лінії можна записати у вигляді $w_1 x_1 + w_2 x_2 = b$. Рівняння гіперплощини має вигляд $w_1 x_1 + ... + w_m x_m = b$, або, у векторній нотації, $w^T x - b = 0$.

б надійно розділяла об’єкти на два класи так, щоб об’єкти з різних боків лінії належали б до різних класів.

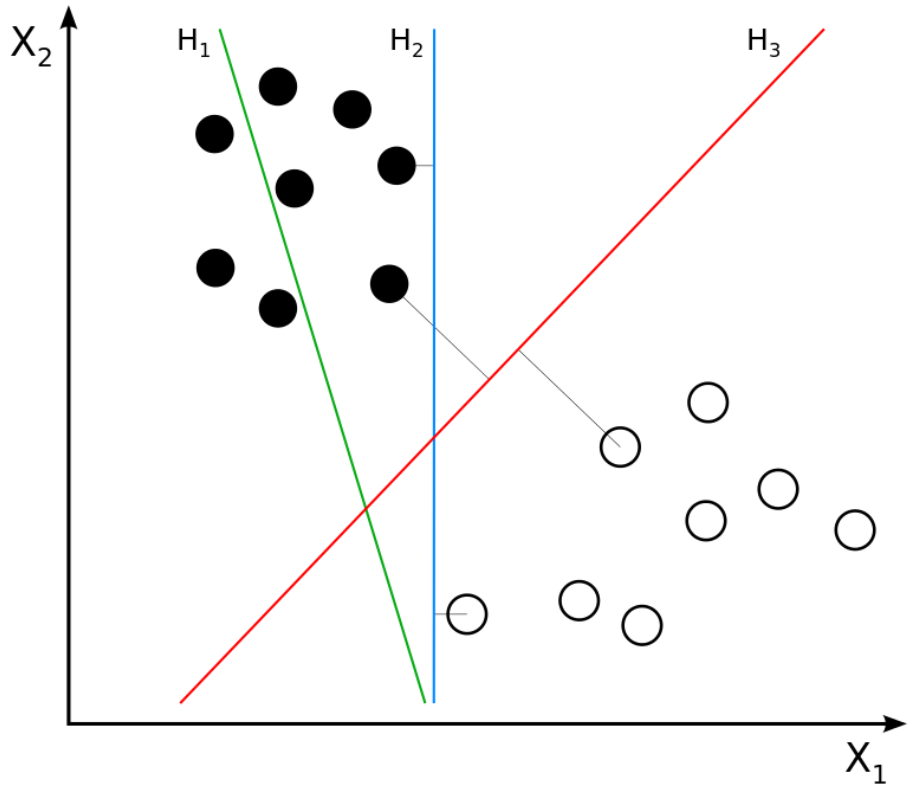


Рис. 7.5. Розподіл класів прямою лінією (джерело: commons.wikimedia.org/wiki/File:Svm_separating_hyperplanes.png)

Існує багато гіперплощин, які могли би розділяти одні й ті ж дані. В якості найкращої розумно обрати таку, яка забезпечувала б найбільшу відстань між двома класами. Тож гранична лінія (гіперплощина) обирається таким чином, щоб відстань від неї до найближчих точок даних з кожного боку була максимальною. Така гіперплощина, якщо вона існує, називається *максимальною розділовою гіперплощиною* (англ. *maximum-margin hyperplane*).

Ідею методу пояснено на рис. 7.5. Лінія H_1 взагалі не розділяє два виділених класи. Лінія H_2 розділяє об’єкти, але з невеликим віддаленням. Лінія H_3 розділяє їх із максимальним віддаленням.

З рис. 7.5 зрозуміло, що мінімальна відстань між об’єктами різних класів залежить лише від об’єктів, розташованих найближче до лінії розподілу. Такі об’єкти і називаються *опорними векторами* (англ. *support vectors*).

В найпростішій лінійній SVM для побудови розділової гіперплощини використовується тренувальний набір даних $(\mathbf{x}_1, y_1),...,(\mathbf{x}_n, y_n)$, де $\mathbf{x}_i \in \boldsymbol{R}^m$ – вектор характеристик i–го об’єкту, а $y_i \in \{1,-1\}$ вказує клас об’єкту. Треба знайти максимальну розділову гіперплощину таким чином, щоби відстань до неї найближчих точок з двох класів була максимальною.

Рівняння гіперплощини має вигляд

$$w^T x - b = 0. \tag{7.16}$$

Дві множини X_1 та X_2 називаються *лінійно роздільними* (англ. *linearly separable*), якщо існують такі числа $w_1,...,w_m,b$, що кожна точка $x \in X_1$ задовольняє $\sum_j w_j x_j > b$, а кожна точка $x \in X_2$ задовольняє $\sum_j w_j x_j \le b$. Якщо дані є лінійно роздільними, то можна обрати дві паралельні гіперплощини, які розділяли б два класи даних таким чином, щоб відстань між ними була б якомога більшою. Область, обмежена цими двома гіперплощинами, називається *розділенням* (англ. *margin*), а максимально розділова гіперплощина є гіперплощиною, яка лежить посередині між ними (рис. 7.6).

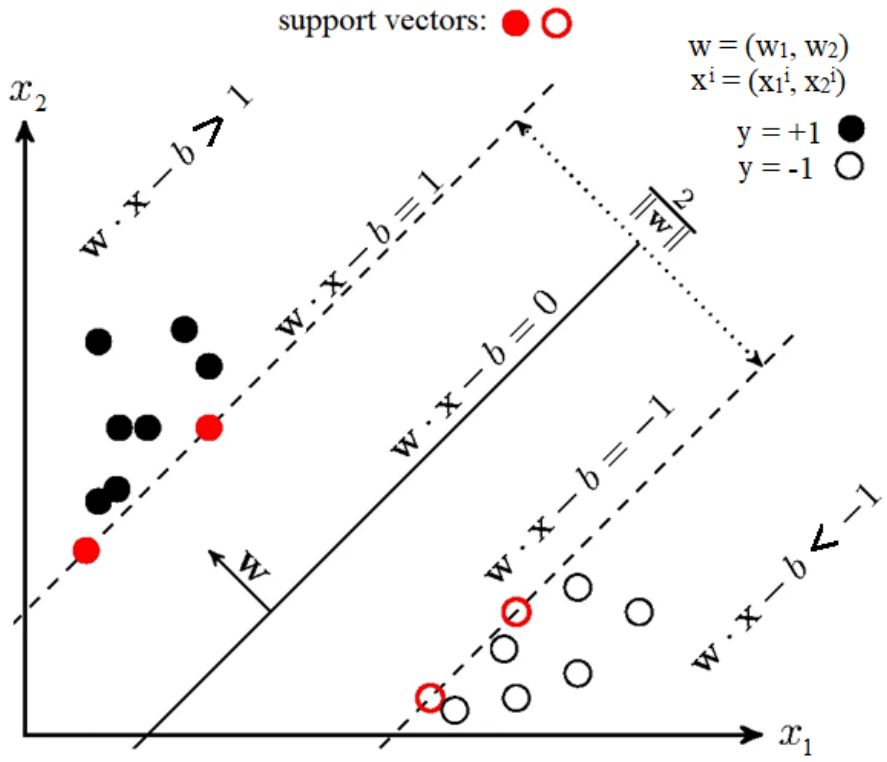


Рис. 7.6. Побудова максимальної розділової гіперплощини

Ці гіперплощини задаються рівняннями:

$$w^T x - b = 1; \tag{7.17}$$

$$w^T x - b = -1. \tag{7.18}$$

З лінійної алгебри відомо, що відстань від точки з координатами $(x_1,...,x_m)$ до гіперплощини $w^T x + c = 0$ дорівнює:

$$d = \frac{|w^T x + c|}{\sqrt{(w_1^2 + ... + w_m^2)}} = \frac{|w^T x + c|}{\|w\|} \tag{7.19}$$

(див., напр., [6, с.93–94, с.119–120]). За цією формулою відстань від будь-якої точки на гіперплощині (7.16) до гіперплощин (7.17) та (7.18) становить $1/\|w\|$,

де $\|w\|$ – норма вектору w. Отже, розділення між гіперплощинами (7.17) та (7.18) становить $2/\|w\|$.

Щоб зробити відстань між класами якнайбільшою, треба вирішити наступну оптимізаційну задачу:

$$\|w\| \underset{w,\,b}{\rightarrow} \min \tag{7.20}$$

при обмеженнях

$$\begin{aligned} y_i = 1\colon\ & w^T x_i - b \geq 1; \\ y_i = -1\colon\ & w^T x_i - b \leq -1. \end{aligned} \tag{7.21}$$

Обмеження (7.21) можна скомбінувати, щоб записати однією нерівністю:

$$y_i (w^T x_i - b) \geq 1. \tag{7.22}$$

Рівняння (7.20)–(7.22) задають задачу квадратичного програмування, яку можна вирішити чисельними методами [38, 39]. Після її вирішення лінійним класифікатором буде $\text{sgn}(w^T x - b)$.

Приклад 7.3. В цьому прикладі використовується набір даних `iris.csv`, відомий як «іриси Фішера». Він став широко відомим після того, як англійський статистик та біолог Рональд Фішер в 1936 році використав його для демонстрації роботи розробленого ним методу лінійного дискримінантного аналізу. Набір даних містить результати вимірювань чотирьох кількісних характеристик для трьох видів ірисів – *Iris setosa*, *Iris virginica* і *Iris versicolor*:

- довжина чашолистка (sepal length);
- ширина чашолистка (sepal width);
- довжина пелюстка (petal length);
- ширина пелюстка (petal width).

Для кожного виду наводиться 50 спостережень. Вихідні дані можна знайти, наприклад, в [Д7]. На рис. 7.7 наведені фотографії представників цих видів з ілюстрацією вимірюваних параметрів.

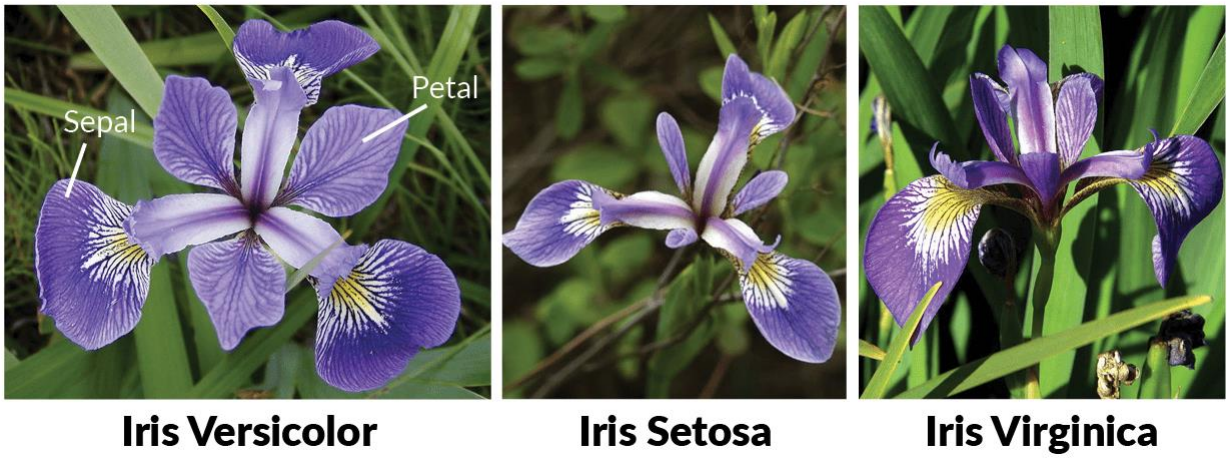


Рис. 7.7. Іриси Фішера [Д7]

На рис. 7.8 наведена діаграма розсіяння довжини та ширини чашолистка для видів *I. setosa* (синій колір) та *I. versicolor* (червоний).

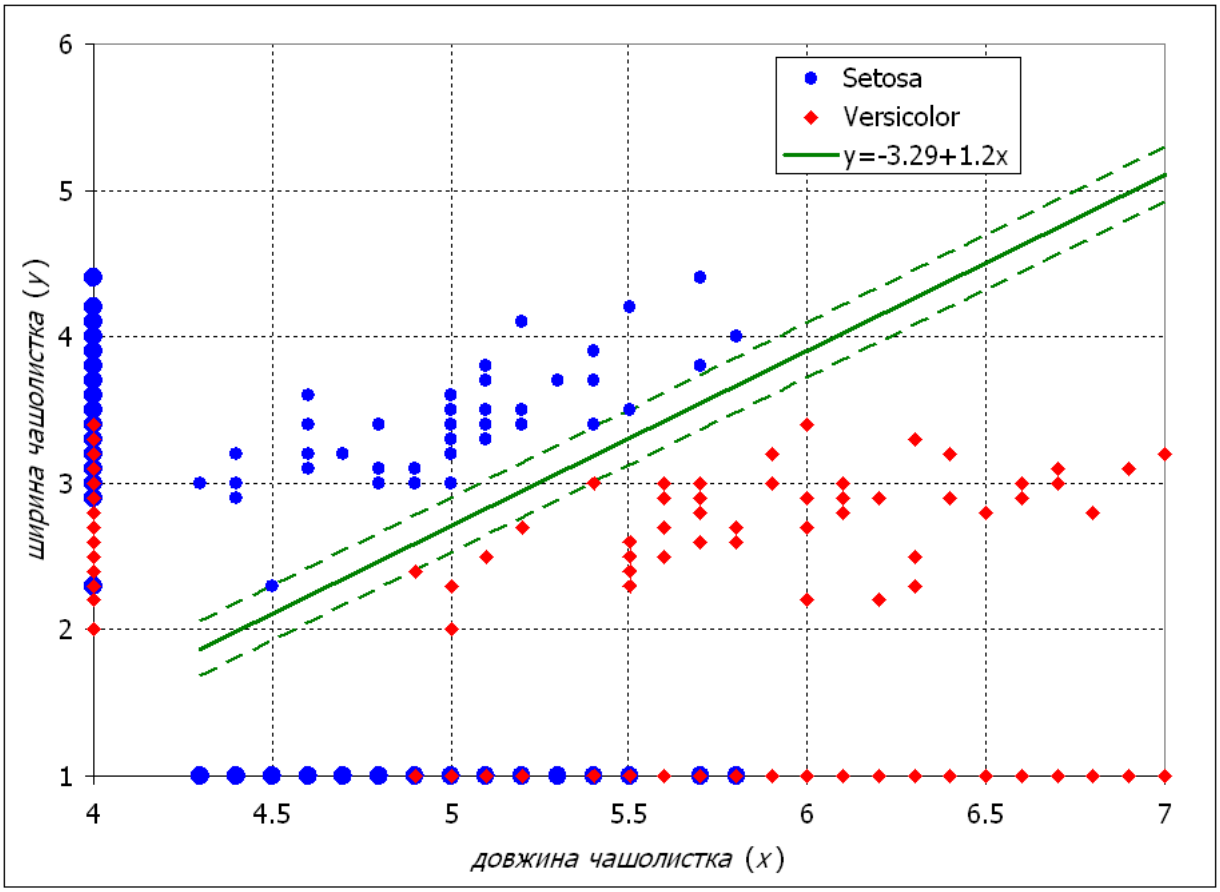


Рис. 7.8. Максимальна розділова лінія для двох видів ірисів Фішера

Можна побачити, що *I. setosa* у середньому має більшу ширину чашолистка, але меншу довжину. Проте, жодна з цих характеристик окремо недостатня для того, щоб надійно розділити два класи, як видно з проекцій набору даних на координатні осі. Але комбінація двох атрибутів виявляється достатньою для лінійного розділення двох класів. Максимальна розділова лінія і границі розділення показані на рис. 7.8 зеленими лініями. Рівняння розділової лінії має вигляд $-1.2x_1 + x_2 + 3.29 = 0$.

Щоб знайти коефіцієнти цієї лінії, треба вирішити задачу квадратичної оптимізації (7.20)–(7.21). Для цього можна скористатись інструментом «пошук рішення» в Excel, або вбудованими функціями інших програмних середовищ, таких як Optimization Toolbox в Matlab, scipy.optimize в Python (метод SLSQP), тощо. Слід відзначити, що навіть для такого порівняно невеликого набору даних, як `iris.csv`, нерівності (7.21) створюють в оптимізаційній задачі 100 обмежень. Це доволі багато і тому пряма реалізація SVM на базі (7.20)–(7.21), вірогідно, буде дуже повільною. Насправді, більшість спостережень в цьому методі виявляється зайвою, бо в класифікації безпосередньо приймають участь тільки чотири опорних вектори (виділені на рис. 6.9 збільшеним розміром), які лежать на демаркаційних лініях. Існує багато алгоритмів для попереднього

відсіву зайвих внутрішніх точок, які підвищують швидкість класифікації методом SVM. ■

Метод опорних векторів можна поширити на випадки, коли тренувальні дані не є лінійно роздільними.

Введемо *функцію нев'язки* (англ. *hinge loss function*) як:

$$e_i = \max\left\{0, 1 - y_i(w^T x_i - b)\right\}. \tag{7.23}$$

Ця функція дорівнюватиме нулю, якщо виконуються обмеження (7.21), тобто x_i лежить із правильного боку розділення. В іншому випадку значення цієї функції буде пропорційним відстані від розділення.

Далі мінімізуємо функцію втрат

$$L(w,\lambda) = \frac{1}{n}\sum_{i=1}^{n} e_i + \lambda\|w\|^2, \tag{7.24}$$

де параметр λ визначає компроміс між збільшенням розміру розділення та правильною класифікацією. Класифікація за критерієм (7.24) називається «м'якою» на відміну від розглянутої вище «жорсткої» класифікації. Для достатньо малих значень λ ці класифікатори будуть практично співпадати, якщо вхідні дані є лінійно роздільними. Якщо ж ні, то м'який класифікатор поступово навчиться життєздатному правилу класифікації.

Також існують узагальнення методу опорних векторів, які дозволяють використовувати в якості границі між класами нелінійні функції.

Метод опорних векторів має свої сильні і слабкі сторони, які слід враховувати при виборі даного методу. Недоліки методу полягають в тому, що він:

– призначений тільки для бінарної класифікації;
– орієнтований на роботу з кількісними атрибутами;
– є досить повільним;
– результати складно інтерпретувати.

При цьому метод має і суттєві переваги:

– висока точність класифікації;
– робастність до викидів та аномалій, оскільки вони ніяк не впливають на розташування опорних векторів;
– не потребує значних обсягів даних;
– висока ефективність при роботі з даними великої розмірності;
– не схильний до перенавчання, внаслідок чого має високу прогностичну здатність.

Контрольні запитання

1. Що називають деревом рішень?
2. Як виглядає загальна схема побудови дерева рішень?

3. Яким чином визначається листя дерева рішень?

4. Як обирається атрибут розгалуження в алгоритмах ID3, C4.5 та CART?.

5. Що характеризує і як розраховується ентропія Шеннона для дискретної випадкової величини?

6. В яких випадках ентропія Шеннона набуває мінімального та максимального значення?

7. Як визначається середньозважена ентропія розбиття множини на підмножини?

8. Як розраховується інформаційний виграш при обранні критерію розгалуження?

9. Чим відрізняються інформаційний виграш та відношення інформаційного виграшу?

10. В чому полягає перевага використання відношення інформаційного виграшу порівняно з абсолютним значенням інформаційного виграшу?

11. Як здійснюється розгалуження в алгоритмі C4.5 для кількісних атрибутів?

12. Як розраховується індекс Джині?

13. В чому полягають переваги та недоліки алгоритму CART порівняно з алгоритмом C4.5?

14. Як подати дерево рішень у табличному вигляді?

15. В чому полягають стратегії ледачого та охочого навчання?

16. В яких випадках доцільно використання стратегій ледачого навчання?

17. Наведіть загальну схему методу k–найближчих сусідів.

18. Які види відстані між об’єктами використовуються в алгоритмі kNN?

19. Для чого необхідна нормалізація даних в алгоритмі kNN?

20. Охарактеризуйте найбільш поширені способи нормалізації даних.

21. В чому полягають переваги використання методу зваженого kNN?

22. Що таке розділова гіперплощина?

23. Як визначається максимальна розділова гіперплощина?

24. Що називається опорним вектором?

25. В чому полягають умови лінійної роздільності двох множин?

26. Що мається на увазі під розділенням двох лінійно роздільних множин? Як воно розраховується?

27. Як звести задачу максимального розділення двох множин до задачі квадратичного програмування?

28. Чим відрізняється «жорстка» класифікація від «м’якої»?

29. Як модифікується метод опорних векторів у випадку, коли два класи не є лінійно роздільними?

30. В чому полягають переваги та недоліки методу опорних векторів?

Завдання для самостійної роботи

7.1. В таблиці наводяться частотності букв англійського алфавіту. Якщо вважати букви алфавіту незалежними випадковими величинами, то чому буде дорівнювати ентропія однієї букви? Порівняйте із значенням ентропії, якщо б всі букви були рівноймовірними.

Буква	E	T	A	O	I	N	S	H	R	D	L	C	U
Частота,%	12,7	9,06	8,17	7,51	6,97	6,75	6,33	6,09	5,99	4,25	4,03	2,78	2,76
Буква	M	W	F	G	Y	P	B	V	K	X	J	Q	Z
Частота,%	2,41	2,36	2,23	2,02	1,97	1,93	1,49	0,98	0,77	0,15	0,15	0,1	0,05

7.2. В таблиці нижче наводяться результати ігор місцевої футбольної команди в чемпіонаті. Атрибут «суперники» вказує їх місце у турнірній таблиці відносно місцевої команди.

Побудуйте дерево рішень для прогнозування результату за алгоритмом ID3. Спрогнозуйте результат гри для атрибутів (нижче, у гостях, ні).

№	Суперники	Гра	Дощ?	Перемога?
1	Вище	Вдома	Так	Ні
2	Вище	Вдома	Ні	Так
3	Вище	Вдома	Ні	Ні
4	Нижче	Вдома	Ні	Так
5	Нижче	У гостях	Ні	Ні
6	Нижче	Вдома	Так	Так
7	Вище	У гостях	Так	Ні

7.3. Нарисуйте графік ентропії для випадкової величини із біноміальним розподілом $X \sim B(n, p)$ як функції від ймовірності успіху p для $n = 1, 2, 4.$ При якому значенні p ентропія буде максимальною? Поясніть, чому.

7.4. Покажіть, що для незалежних дискретних випадкових величин X, Y $H(X \cdot Y) = H(X) + H(Y)$ (це так звана властивість адитивності ентропії).

7.5. Розглянемо наступний набір даних:

№	a_1	a_2	a_3	Клас
1	*T*	*T*	5.0	1
2	*T*	*T*	7.0	1
3	*T*	*F*	8.0	0
4	*F*	*F*	3.0	1
5	*F*	*T*	7.0	0
6	*F*	*T*	4.0	0
7	*F*	*F*	5.0	0
8	*T*	*F*	6.0	1
9	*F*	*T*	1.0	0

Сформуйте корінь дерева рішень за критеріями:

а) інформаційного виграшу (ID3);

б) відношення інформаційного виграшу (C4.5);

в) найменшого зваженого індексу Джині (CART).

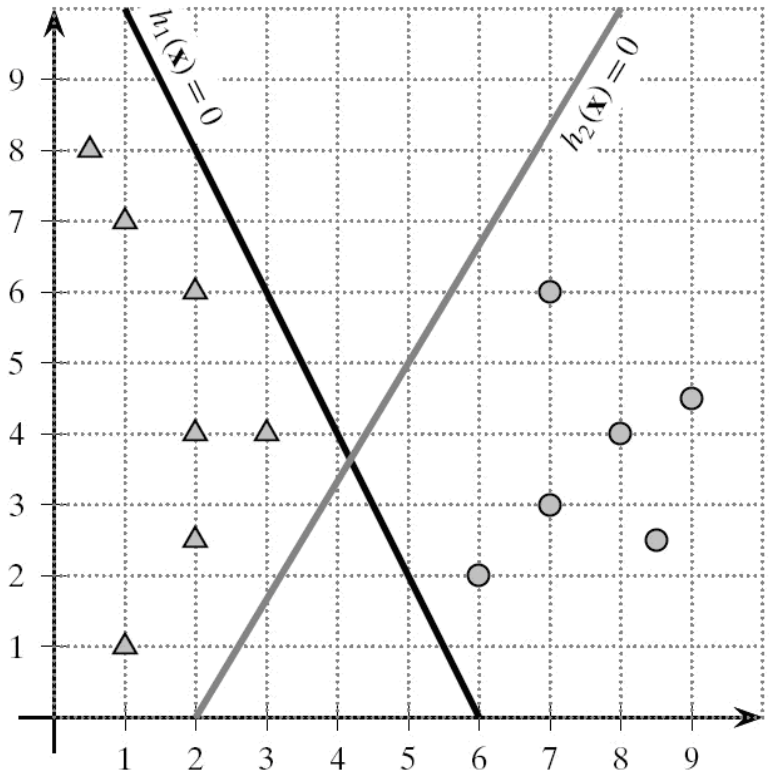
7.6. В таблиці наводяться антропометричні показники чоловіків та жінок.

Зріст, м	Вага, кг	Розмір ноги, см	Стать
1.83	82	30	Ч
1.80	86	28	Ч
1.70	77	30	Ч
1.80	75	25	Ч
1.52	45	15	Ж
1.68	68	20	Ж
1.65	59	18	Ж
1.75	68	23	Ж

а) Нормалізуйте показники із використанням z–нормалізації.

б) За допомогою методу kNN з евклідовими відстанями визначте стать людини з показниками (1.83, 80, 20) для значень $k = 1$ та $k = 3$.

7.7. Розгляньте представлений на рисунку набір даних із двома класами (трикутники та кола).



а) Знайдіть рівняння для гіперплощин $h_1(x)$ та $h_2(x)$.

б) Знайдіть усі опорні вектори для $h_1(x)$ и $h_2(x)$.

в) Яка з двох гіперплощин краще підходить для поділу класів за критерієм найбільшої відстані?

г) Знайдіть рівняння для найкращої роздільної гіперплощини для даного набору даних. Визначте відповідні опорні вектори.

7.8. Відстань Мінковського між точками $\mathbf{x} = (x_1, ..., x_n)$ та $\mathbf{y} = (y_1, ..., y_n)$ визначається як

$$d(\mathbf{x}, \mathbf{y}) = \left(\sum_{i=1}^{n} | x_i - y_i |^{\rho} \right)^{\frac{1}{\rho}}$$

Покажіть, що:

а) при $\rho = 1$ ця відстань еквівалентна манхеттенській;

б) при $\rho = 2$ – евклідовій;

в*) при $\rho \to \infty$ – відстані Чебишова.

8. ІМОВІРНІСНІ МЕТОДИ КЛАСИФІКАЦІЇ

Імовірнісні методи класифікації відрізняються від ординарних тим, що надають оцінку ймовірності приналежності кожного об'єкта до кожного із класів. Це, безумовно, є дуже корисним, оскільки така оцінка є показником впевненості у результатах класифікації. При прийнятті рішень ймовірність помилки є прямим індикатором пов'язаного з ними ризику. Кількісна оцінка таких ризиків є основою для існування таких галузей, як страхування, букмекерство тощо. Наприклад, у страховій справі ризик є визначальним елементом при встановленні страхової премії.

8.1 Байєсівські методи класифікації

Так називається група статистичних методів, що пов'язані з використанням для класифікації даних теореми Байєса. Вони набули популярності як засіб формалізації знань в експертних системах, а зараз застосовуються також при вирішенні задач ІАД.

Байєсівські методи відносяться до категорії імовірнісних, тобто вони оцінюють імовірність приналежності об'єкта до кожного класу. Найпростішим, і водночас найпоширенішим з цієї групи методів є так званий *наївний байєсівський класифікатор* (скорочено НБК; англ. *naïve Bayes classifier*). "Наївність" методу полягає у припущенні про взаємну незалежність атрибутів.

Метод заснований на теоремі Байєса (2.6):

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)},$$

де A і B – дві події, а $P(A|B)$ – умовна імовірність настання події A за умови B.

Стосовно до задачі класифікації, нехай об'єкт X характеризується вектором m атрибутів (незалежних змінних): $\mathbf{x} = (x_1,...,x_m)$. Припустимо для простоти, що всі атрибути категоріальні, тобто приймають тільки дискретні значення.

Нехай множина об'єктів розділена на s класів $y_1,...,y_s$. Для довільного заданого об'єкта X з невідомою міткою класу потрібно визначити ймовірності його належності до кожного з класів $P(y_k|x_1,...,x_n) = P(y_k|\mathbf{x})$. Об'єкт буде віднесений до того класу, якому відповідає найбільша ймовірність.

За теоремою Байєса апостеріорна ймовірність класу дорівнює:

$$P(y_k|\mathbf{x}) = \frac{P(\mathbf{x}|y_k)P(y_k)}{P(\mathbf{x})}, \qquad (8.1)$$

де $P(y_k)$ – апріорна ймовірність того, що об'єкт X належить до класу y_k;

$P(\mathbf{x}|y_k)$ – ймовірність спостереження вектору атрибутів $\mathbf{x}$, якщо об'єкт належить до класу y_k;

$P(\mathbf{x})$ – апріорна ймовірність спостереження вектору атрибутів $\mathbf{x}$.

На практиці, знаменник у формули (8.1) не має великого значення, бо він не залежить від класу об'єкту і відіграє роль нормуючої константи.

Обчислення $P(\mathbf{x} \mid y_k)$ у загальному випадку є дуже складним завданням. Але якщо вважати всі атрибути незалежними, то задача значно спрощується. В цьому випадку

$$P(\mathbf{x} \mid y_k) = \prod_{i=1}^{m} P(x_i \mid y_k). \tag{8.2}$$

Значення $P(x_i \mid y_k)$, $P(y_k)$, $P(\mathbf{x})$ можна обчислити за тренувальною вибіркою наступним чином:

$$P(x_i \mid y_k) = \frac{s_{ki}(x_i)}{s_k}; \tag{8.3}$$

$$P(y_k) = \frac{s_k}{n}; \tag{8.4}$$

$$P(\mathbf{x}) = \sum_{k=1}^{s} P(\mathbf{x} \mid y_k) P(y_k) \tag{8.5}$$

де $s_{ki}(x_i)$ – кількість об'єктів класу y_k у тренувальному наборі, для яких i–й атрибут дорівнює x_i;

s_k – загальна кількість об'єктів класу y_k;

n – обсяг тренувальної вибірки.

Класифікація об'єкта відбувається шляхом присвоєння йому найбільш вірогідного класу y^*, де

$$y^* = \arg\max_{k=1,\ldots,s} P(y_k) \prod_{i=1}^{m} P(x_i \mid y_k). \tag{8.6}$$

В якості ілюстрації роботи методу повернемось до даних з прикладу 6.1, які для зручності перенесені до табл. 8.1.

Таблиця 8.1 – Вихідні дані про клієнтів компанії кабельного телебачення

№	Стать	Дохід	Студент?	Євроспорт?
1	Ч	високий	так	1
2	Ж	середній	ні	0
3	Ч	низький	так	0
4	Ж	низький	ні	0
5	Ч	середній	ні	1
6	Ж	високий	ні	1
7	Ж	середній	так	1
8	Ч	середній	так	1
9	Ж	низький	так	0
10	Ж	середній	ні	0

Приклад 8.1. За допомогою НБК слід визначити, чи підпишеться на канал «Євроспорт» людина із характеристиками **x** = (Ч, середній, так).

В нашому приклади визначено два класи за ознакою підписки на платний канал (1 – так, 0 – ні). Визначимо умовні ймовірності спостереження окремих атрибутів для кожного класу:

P(Ч | 1) = 3/5; P(Ч | 0) = 1/5;

P(середній | 1) = 3/5; P(середній | 0) = 2/5;

P(так | 1) = 3/5; P(так | 0) = 2/5;

Імовірності кожного з класів дорівнюють:

P(1) = 5/10 = 0,5; P(0) = 5/10 = 0,5.

Умовні ймовірності спостереження об'єкта з вектором атрибутів **x** для визначених класів становлять:

$P(\mathbf{x} \mid 1) = P(\text{Ч} \mid 1)\, P(\text{середній} \mid 1)\, P(\text{так} \mid 1) = 3/5 \times 3/5 \times 3/5 = 0{,}216;$

$P(\mathbf{x} \mid 0) = P(\text{Ч} \mid 0)\, P(\text{середній} \mid 0)\, P(\text{так} \mid 0) = 1/5 \times 2/5 \times 2/5 = 0{,}032.$

Безумовна імовірність спостереження об'єкта з характеристиками **x** = (Ч, середній, так) складає:

$P(\mathbf{x}) = P(\mathbf{x} \mid 1) \times P(1) + P(\mathbf{x} \mid 0) \times P(0) = 0{,}216 \times 0{,}5 + 0{,}032 \times 0{,}5 = 0{,}124.$

Тоді імовірності придбання комп'ютера складатимуть:

$P(1 \mid \mathbf{x}) = P(\mathbf{x} \mid 1) \times P(1) / P(\mathbf{x}) = 0{,}216 \times 0{,}5 / 0{,}124 \approx 0{,}871;$

$P(0 \mid \mathbf{x}) = P(\mathbf{x} \mid 0) \times P(0) / P(\mathbf{x}) = 0{,}032 \times 0{,}5 / 0{,}124 \approx 0{,}129.$

Отже, ми визначаємо, що людина з характеристиками **x** = (Ч, середній, так) підпишеться на канал «Євроспорт».

Результати обчислень за цією схемою для всього набору даних зведені в табл. 8.2.

Таблиця 8.2 – Прогнозування підписки на «Євроспорт» за методом НБК

№	Стать	Дохід	Студент?	Євроспорт?	Ймовірність підписки	Прогноз
1	Ч	високий	так	так	1,000	так
2	Ж	середній	ні	ні	0,333	ні
3	Ч	низький	так	ні	0,000	ні
4	Ж	низький	ні	ні	0,000	ні
5	Ч	середній	ні	так	0,750	так
6	Ж	високий	ні	так	1,000	так
7	Ж	середній	так	так	0,529	так
8	Ч	середній	так	так	0,871	так
9	Ж	низький	так	ні	0,000	ні
10	Ж	середній	ні	ні	0,333	ні

На відміну від, наприклад, дерев рішень, наочно видно достовірність прогнозів. Так, для студентки із середнім рівнем доходів шанси на підписку лише трохи краще, ніж 50:50, що безперечно є корисною інформацією. ■

Методологічна проблема із застосуванням НБК полягає в тому, що припущення про незалежність ознак рідко відповідає дійсності. Наприклад, дохід споживача у реальності залежить від його статі, віку, роду занять та інших атрибутів (а рід занять «студент», у свою чергу, залежить від віку). За це алгоритм іноді називають навіть «ідіотським». Проте, досвід застосування НБК свідчить про те, що навіть у випадках сильної кореляції між незалежними змінними він часто демонструє непогані результати [29].

При застосуванні формули (8.6) часто виникають такі проблеми.

1. Проблема переповнення. Обчислення пов'язані з перемноженням великої кількості ймовірностей, багато з яких будуть дуже маленькими. Це може призвести до арифметичного переповнення знизу. Щоб уникнути цієї проблеми, можна взяти логарифм від виразу (8.6). Оскільки логарифм – монотонна функція, значення максимуму при цьому не зміниться. Отримаємо:

$$y^* = \arg\max_{k=1,\dots,s}\left[\ln P(y_k) + \sum_{i=1}^{m} \ln P(x_i \mid y_k)\right]. \qquad (8.7)$$

2. Проблема відсутніх атрибутів. Деякі значення атрибутів можуть бути відсутні в навчальній вибірці. Наприклад, у тексті, що класифікується, може зустрітися слово, яке було відсутнє в тренувальному наборі. В цьому випадку $s_{ki}(x_i) = 0$, а, отже, і ймовірність $P(x_i \mid y_k)$ виявиться рівною нулю для всіх класів. Це призводить до неможливості класифікації, так як значення виразу під знаком максимуму в (8.6) виявиться рівним нулю для всіх класів. Типовим вирішенням цієї проблеми є так зване *адитивне згладжування* (англ. *additive smoothing*). Ідея полягає в тому, що ми «прикидаємося», ніби бачили кожне значення атрибуту на один раз більше, ніж насправді, тобто додаємо одиницю до частоти кожного значення:

$$P(x_i \mid y_k) = \frac{s_{ki}(x_i) + 1}{s_k + l}, \qquad (8.8)$$

де l – кількість можливих значень i–го атрибута.

3. Кількісні атрибути із безперервними значеннями в НБК створюють менше проблем, ніж у деревах рішень. Якщо нема вагомих причин для інших припущень, можна вважати, що значення такого атрибуту мають нормальний розподіл. Його параметри $\mu(y_k), \sigma(y_k)$ оцінюються як арифметичні середні та СКВ, розраховані окремо для кожного класу (див. п. 3.3). Тоді «ймовірність»[13] отримання значення x розраховується як $f\left(x, \mu(y_k), \sigma(y_k)\right)$, де $f(x, \mu, \sigma)$ – щільність нормального розподілу $N(\mu, \sigma^2)$. Після такого перетворення алгоритм працює так само, як і для категоріальних атрибутів.

НБК як метод класифікації має багато переваг.

---

[13] З математичної точки зору так казати некоректно, бо для безперервних випадкових величин $P\{X = x\} = 0$. Проте, $P\{x < X \le x + dx\} \approx f(x, \mu, \sigma)dx$. Детальніше див. п.2.4.

1. Алгоритм класифікації є досить швидким – складність оцінювання лінійно зростає із збільшенням числа класів та атрибутів, тоді як в багатьох інших методах складність зростає експоненційно.

2. В моделі визначаються залежності між усіма змінними, що дозволяє легко обробляти ситуації, в яких значення деяких змінних невідомі.

3. З точки зору прозорості НБК поступається деревам рішень, але все ж залишається порівняно легким для інтерпретації. Завдяки припущенню про незалежність ознак, легко виділити значення атрибутів, які впливають на віднесення об'єкту до певного класу (фактори успіху чи фактори ризику в залежності від сфери застосування). Чим більше таких факторів, тим більше ймовірність класифікації у цей клас.

4. Байєсівський підхід дозволяє природним чином поєднувати закономірності, знайдені в даних, і експертні оцінки (суб'єктивні ймовірності), отримані в явному вигляді.

5. Якщо певні атрибути не є релевантними для вирішення задачі класифікації, то вони практично не вплинуть на результати класифікації. На відміну від цього, дерева рішень та деякі інші алгоритми можуть побудувати складні правила на базі таких атрибутів. Це наводить до надмірного ускладнення моделі без суттєвого поліпшення її прогностичної здатності – проблема, відома як *перенавчання* (англ. *overfitting*). НБК уникає цієї проблеми.

Найбільш серйозною проблемою НБК є ігнорування зв'язків між атрибутами. Крім цього, можна назвати також такі недоліки.

1. На результат класифікації в НБК впливають тільки індивідуальні значення вхідних змінних. Комбінований вплив пар або трійок значень різних атрибутів ніяк не враховується.

2. Повинно бути достатньо даних, щоб НБК міг коректно оцінити всі умовні ймовірності. Це може створювати проблеми, особливо коли класи є незбалансованими.

3. Обробка кількісних змінних вимагає припущень про параметричну форму їх розподілу (наприклад, нормальність), які можуть не відповідати дійсності.

Байєсівські класифікатори знайшли широке застосування на практиці. Найчастіше вони використовуються для категоризації текстових документів (реклама чи ділова кореспонденція, спорт чи політика і т.д.). В якості атрибутів при цьому виступають слова в аналізованому тексті. Також байєсівські класифікатори використовуються в медичній діагностиці, задачах розпізнавання образів, тощо.

НБК є найбільш популярним методом фільтрації спаму. Ефективній роботі саме байєсівських методів сприяють деякі властивості предметної області.

По–перше, у класифікованих об'єктів є дуже велика кількість ознак. Як правило, в якості атрибутів виступають всі слова листів користувача, за

винятком зовсім коротких і таких, що дуже рідко зустрічаються. Деякі слова та їх комбінації («виграш», «спадщина», «встигніть», «унікальний» тощо) набагато частіше зустрічаються у спамі, ніж у звичайній кореспонденції, на чому й ґрунтується принцип класифікації.

По–друге, досить легко здійснювати перенавчання класифікатора шляхом поповнення набору «спам–не спам». Це добре працює навіть в умовах локальних поштових клієнтів, так як потік регулярної кореспонденції у кінцевого клієнта є сталим або таким, що повільно змінюється. Загальні правила визначити важче, бо те, що є спамом для одного користувача, може не бути спамом для іншого.

8.2 Регресійні моделі бінарної класифікації: логіт і пробіт

В наївному байєсівському класифікаторі вплив кожного атрибуту на клас об'єкту розглядається окремо. В багатьох випадках більш імовірно, що результат класифікації залежить від суми балів, набраних об'єктом за окремими ознаками. Наприклад, при прийомі на роботу важливими факторами можуть бути попередній досвід претендента, академічна успішність, володіння іноземною мовою, комп'ютерна грамотність, наявність рекомендацій тощо. При цьому низькі результати за одними показниками можна компенсувати високими балами за іншими. Тож логічно припустити, що вирішальним фактором є деяка зважена сума оцінок претендента за різними кількісними та якісними показниками[14].

Щоб дізнатись вагові коефіцієнти з даних, логічно звернутись до апарату множинної регресії (розділ 5). Однак той факт, що залежна змінна може приймати тільки два значення, не дозволяє успішно застосувати метод найменших квадратів для оцінки параметрів моделі й прогнозування.

Дійсно, розглянемо звичайну модель лінійної регресії

$$y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \varepsilon_i, i = 1,...,n, \tag{8.9}$$

де $\mathbf{x}_i$ – вектор характеристик *i*–го об'єкту. Оскільки $y_i \in \{0,1\}$, а $M[\varepsilon_i] = 0$, то:

$$M[y_i] = 1 \times P(y_i = 1) + 0 \times P(y_i = 0) = P(y_i = 1) = \mathbf{x}_i^T \boldsymbol{\beta}.$$

Отже, модель (8.9) може бути еквівалентно записана як

$$P(y_i = 1) = \mathbf{x}_i^T \boldsymbol{\beta}, \tag{8.10}$$

завдяки чому її називають *лінійною моделлю ймовірності* (англ. *linear probability model*). Фатальний недолік цієї моделі полягає в тому, що прогнозні значення ймовірності $\mathbf{x}_i^T \boldsymbol{\beta}$ можуть лежати поза межами інтервалу [0, 1], що, звичайно, є нісенітницею.

Щоб уникнути цієї проблеми, можна припустити, що

[14] Стосовно задачі оцінки платоспроможності позичальника, така схема називається *кредитним скорингом* (англ. *credit scoring*).

$$P(y_i = 1) = p_i = F(\mathbf{x}_i^T \boldsymbol{\beta}), \tag{8.11}$$

де $F(\cdot)$ приймає значення в інтервалі [0, 1]. Зокрема, в якості $F(\cdot)$ можна обрати функцію розподілу деякої випадкової величини. Двома найбільш поширеними виборами для $F(\cdot)$ є:

1) функція стандартного нормального розподілу (2.37)

$$F(u) = \Phi(u) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{u} \exp\left(-\frac{x^2}{2}\right) dx, \tag{8.12}$$

в якому випадку модель (8.11) називають *пробіт–моделлю* (англ. *probit model*);

2) функція логістичного розподілу (2.43)

$$F(u) = \Lambda(u) = \frac{e^u}{1+e^u}, \tag{8.13}$$

в якому випадку модель (8.11) називають *логіт–моделлю* або *логістичною регресією* (англ. *logit model*, *logistic regression*).

Функції $\Phi(u)$ та $\Lambda(u)$ мають подібну форму, але логістичний розподіл має значно товстіші «хвости» (див. рис. 8.1). Якщо незалежні змінні мають помірну варіацію, то результати класифікації за допомогою обох моделей не сильно відрізнятимуться. В той же час виявляється, що параметри $\boldsymbol{\beta}$ простіше оцінити і інтерпретувати в моделі логіт, ніж в моделі пробіт. Ці два фактори зумовили значно більшу популярність логістичної регресії порівняно з моделлю пробіт, принаймні у прикладних дослідженнях.

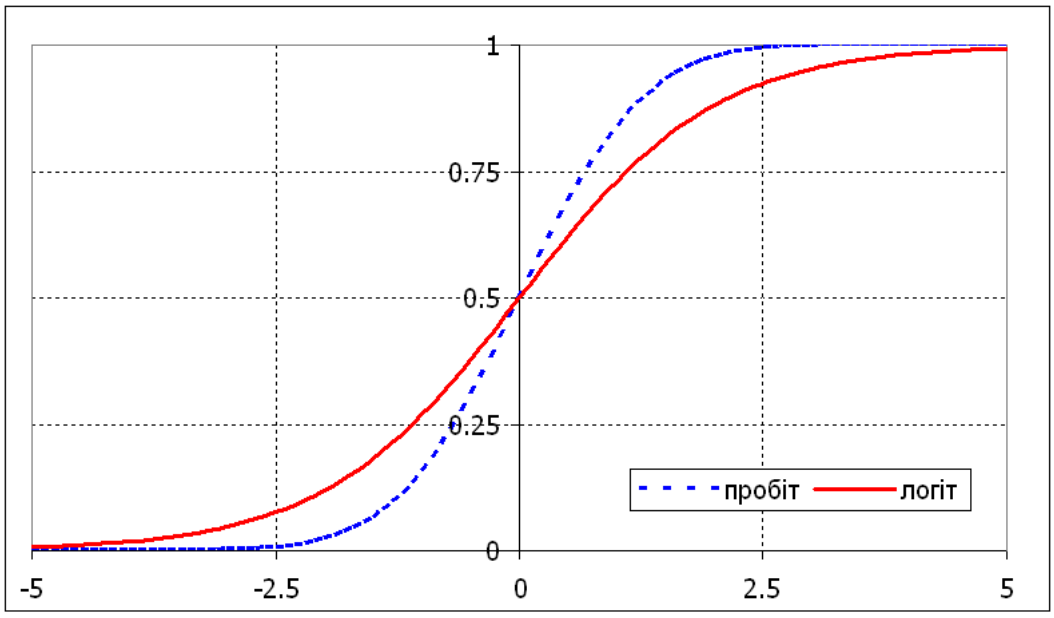


Рис. 8.1. Вид функції $F(u)$ для моделей пробіт та логіт

Нелінійний характер моделі (8.11) не дозволяє скористатись формулою методу найменших квадратів (5.6). Для оцінки параметрів моделі зазвичай використовують метод максимальної правдоподібності [1, 46]. Ідея цього методу полягає у створенні функції правдоподібності, яка описує ймовірність

отримання наявного набору даних для заданого вектора параметрів $\boldsymbol{\beta}$ з її подальшою максимізацією.

В моделі (8.11) бінарна змінна y_i має розподіл Бернуллі із ймовірністю «успіху» p_i (див. п. 2.5.1):

$$P(y_i) = \begin{cases} p_i, \; y_i = 1 \\ 1 - p_i, \; y_i = 0 \end{cases}. \tag{8.14}$$

Вираз (8.14) можна більш компактно записати як $p_i^{y_i}(1-p_i)^{1-y_i}$. Тоді функція правдоподібності для всього набору даних матиме вигляд:

$$L = \prod\nolimits_i p_i^{y_i}(1-p_i)^{1-y_i} \; . \tag{8.15}$$

Для максимізації (8.15) зручно взяти логарифм від останнього виразу (це не вплине на точку максимуму, бо логарифм є монотонним перетворенням):

$$l = \ln L = \sum\nolimits_i \left[y_i \ln p_i + (1-y_i)\ln(1-p_i) \right]. \tag{8.16}$$

Необхідною умовою максимуму функції є рівність нулю її похідної. Отже, для (8.16) отримуємо:

$$\frac{\partial l}{\partial \beta_k} = \sum\nolimits_i \left(\frac{y_i}{p_i} - \frac{1-y_i}{1-p_i} \right) \frac{\partial p_i}{\partial \beta_k} = 0, \; k = 0,1,...,m. \tag{8.17}$$

Із формули (8.11) $\dfrac{\partial p_i}{\partial \beta_k} = \dfrac{\partial F(\mathbf{x}_i^T \boldsymbol{\beta})}{\partial \beta_k} = f(\mathbf{x}_i^T \boldsymbol{\beta}) x_k$, де $f(u) = F'(u)$. В розділі 2.5.2 було показано, що для логістичного розподілу $f(u) = \Lambda(u)(1-\Lambda(u))$. Із використанням цієї властивості (8.17) спрощується до:

$$\sum\nolimits_i \left(y_i - p_i \right) x_k = 0, \; k = 0,1,...,m \, , \tag{8.18}$$

де $p_i = \Lambda(\mathbf{x}_i^T \boldsymbol{\beta})$. Якщо рівняння регресії містить константу, то з системи рівнянь (8.18) випливає, що $\dfrac{\sum_i p_i}{n} = \dfrac{\sum_i y_i}{n}$, тобто середні прогнозовані частоти класів мають дорівнювати спостережуваним частотам.

Оцінки коефіцієнтів рівняння логістичної регресії можна отримати або максимізацією функції (8.16), або рішенням системи нелінійних рівнянь (8.18). В обох випадках доведеться звернутися до використання чисельних методів. В багатьох програмних середовищах є вбудовані функції для рішення цієї задачі; наприклад, `LogisticRegression` в пакеті `scikit-learn` в Python.

Приклад 8.2. Застосуємо логістичну регресію для задачі класифікації ірисів Фішера з прикладу 7.3. Розставимо позначки класу $y_i = 1$ для *Iris setosa* і $y_i = 0$ для *Iris versicolor*. Максимізація логарифму функції правдоподібності (8.16) методом Ньютона призводить до рівняння:

$$u = 344.17 - 125.78 x_1 + 105.10 x_2 \, ,$$

де x_1 – довжина чашолистка, а x_2 – ширина чашолистка.

Квітка буде класифікована як *Iris setosa*, якщо

$$P(y_i = 1) = \frac{e^u}{1+e^u} > \frac{1}{2} \Leftrightarrow u > 0.$$

З останньої нерівності можна знайти рівняння лінії, яка розподіляє класи:

$$x_2 = -\frac{344.17}{105.10} + \frac{125.78}{105.10} x_1 = -3.275 + 1.197 x_1 .$$

(див. рис. 8.2). Можна помітити, що рівняння роздільної лінії з точністю до тисячних співпадає з тим, що було отримано методом опорних векторів. ■

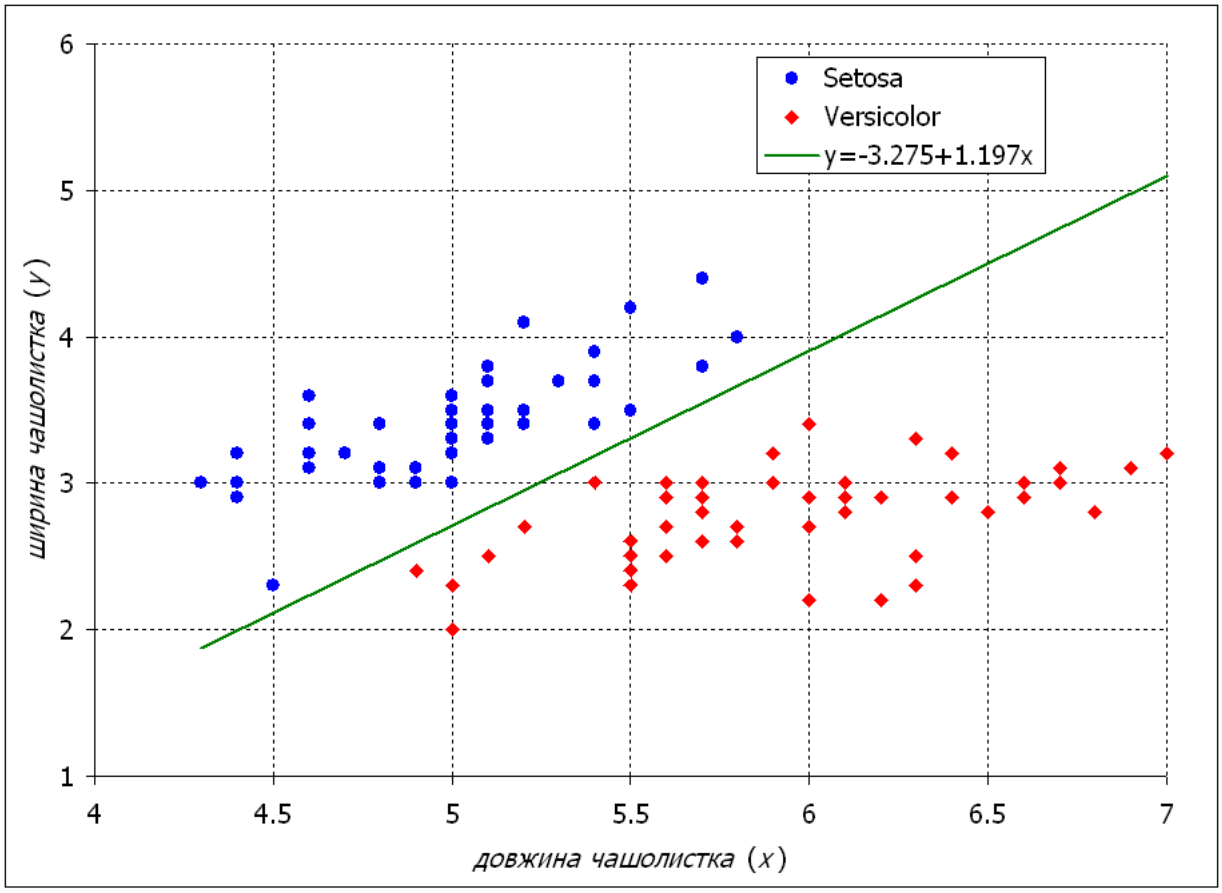


Рис. 8.2. Класифікація ірисів Фішера за допомогою логістичної регресії

Те, що в прикладах 7.3 і 8.2 результати класифікації виявились практично ідентичними, не зовсім випадково. Метод опорних векторів і логістичну регресію поєднує те, що в обох методах шукається спосіб розділення двох класів прямою лінією (або гіперплощиною, якщо кількість предикторів > 2). Логістична регресія максимізує ймовірність отримання наявного набору даних; чим далі знаходиться точка даних від роздільної гіперплощини, тим вищим буде значення цільової функції. Метод опорних векторів натомість є суто геометричним; його ідея полягає в створенні якомога ширшої «нейтральної полоси» між класами. Точки, що знаходяться далеко від цієї межи (не є опорними векторами) взагалі не впливають на рівняння роздільної поверхні. Тому результати двох методів можуть розрізнятись при наявності викидів.

Також логістична регресія не дуже надійно працює, коли класи є лінійно роздільними, бо невеликі зміни параметрів майже не впливатимуть на значення функції правдоподібності. Метод опорних векторів, навпаки, гірше працює, коли класи перемішані між собою. Але в багатьох випадках обидва методи будуть давати схожі результати. Набагато докладніше порівняння цих методів можна знайти в [52].

З формул (8.11) та (8.13) випливає, що

$$P(y_i = 0) = 1 - P(y_i = 1) = 1 - \frac{e^u}{1+e^u} = \frac{1}{1+e^u}. \tag{8.19}$$

Тоді *відношення шансів* (англ. *odds ratio*) становить

$$\frac{P(y_i = 1)}{P(y_i = 0)} = e^u, \tag{8.20}$$

а логарифм цього відношення –

$$\ln \frac{P(y_i = 1)}{P(y_i = 0)} = u = \mathbf{x}_i^T \boldsymbol{\beta}. \tag{8.21}$$

Отже, логістичну регресію можна розглядати як напівлогарифмічну лінійну модель стосовно відношення шансів. Згідно з інтерпретацією таких моделей (див. п. 5.3), коефіцієнт β_k в рівнянні (8.21) вказує, на скільки відсотків зміниться відношення шансів при зміні предиктора x_k на одиницю.

8.3 Мультиноміальна логістична модель

Мультиноміальна логістична модель (надалі МНЛ, англ. *multinomial logit*) є узагальненням логістичної регресії на випадок декількох класів. Ця модель була розроблена нобелівським лауреатом Деніелом Макфадденом в рамках теорії імовірнісного вибору, яка намагається пояснити, як люди обирають одну з наявних альтернатив в умовах невизначеності [35]. Прикладами таких ситуацій є вибір між політичними партіями, конкуруючими брендами однотипних товарів тощо.

МНЛ ґрунтується на припущенні про *незалежність непов'язаних альтернатив* (англ. *independence of irrelevant alternatives*). Це припущення є типовим для теорії раціонального вибору і полягає в тому, що шанси віддати перевагу однієї з альтернатив над іншою не залежать від наявності інших, «непов'язаних» альтернатив. Наприклад, відносна ймовірність обрати для поїздки автомобіль або автобус не зміниться, якщо додати варіанти скористатися велосипедом або взагалі відмовитись від поїздки. Це дозволяє змоделювати вибір серед k альтернатив як серію $k - 1$ незалежних порівнянь цих альтернатив з однією, обраною в якості «опорної».

Нехай існує s альтернатив. Будемо вважати, що виграш[15] індивіда i від обрання альтернативи j, U_{ij}, є сумою двох компонент: детермінованої u_{ij} та стохастичної ε_{ij}, тобто $U_{ij} = u_{ij} + \varepsilon_{ij}$. Детермінована компонента може залежати від характеристик індивіда та атрибутів альтернативи: $u_{ij} = \mathbf{x}_{ij}^T \boldsymbol{\beta}$. Стохастична компонента відбиває вплив настрою індивіда, його попереднього досвіду, естетичної привабливості альтернативи та подібних факторів, які відсутні в базі даних чи взагалі не можуть спостерігатися.

Приклад 8.3. Нехай мова йде про вибір споживачем автомобіля. Модель корисності споживача i від покупки автомобіля марки j може виглядати як:

$$U_{ij} = \beta_0 + \beta_1 Income_i + \beta_2 HP_j + \beta_3 Age_i Sport_j + \beta_4 Rural_i Trunk_j + ... + \varepsilon_{ij}, \quad (8.22)$$

де $Income_i$ – дохід споживача i,

Age_i – вік споживача i,

$Rural_i$ = 1, якщо споживач i мешкає у сільській місцевості і 0 – в іншому разі;

HP_j – міцність двигуна автомобіля марки j у кінських силах;

$Sport_j$ = 1 для спортивних автомобілів і 0 – для моделей інших типів;

$Trunk_j$ – об'єм багажника автомобіля марки j в літрах;

ε_{ij} – неспостережений компонент корисності споживача i від вибору автомобіля марки j.

Зверніть увагу, що предиктори можна поділити на три категорії:

– ті, що залежать тільки від характеристик споживача;

– ті, що залежать тільки від характеристик альтернативи;

– ті, що залежать від взаємодії між ними. Так, модель (8.22) передбачає, що на вибір спортивних автомобілів може впливати вік споживача, а об'єм багажника може бути більш важливим для сільських мешканців.

Виходячи з цього, детерміновану компоненту корисності можна записати більш детально як:

$$u_{ij} = \mathbf{x}_{ij}^T \boldsymbol{\beta} = \mathbf{z}_j^T \boldsymbol{\alpha} + \mathbf{u}_i^T \boldsymbol{\gamma} + \mathbf{v}_{ij}^T \boldsymbol{\delta}. \quad (8.23)$$

Звісно, не всі категорії предикторів можуть бути доступними. Насправді, найчастіше в наявності є тільки характеристики окремих альтернатив. ■

Індивід i обере альтернативу j, якщо $u_{ij} + \varepsilon_{ij} > u_{ik} + \varepsilon_{ik}$ для всіх $k \neq j$. Отже,

$$P(y_i = j) = P(\varepsilon_{ij} > \max_{k=1,...,s;\, k \neq j} (u_{ik} - u_{ij} + \varepsilon_{ik})). \quad (8.24)$$

Для довільного розподілу ε_{ij} (зокрема, для нормального) знаходження ймовірностей вибору альтернатив за формулою (8.24) призводить до обчислення складних багатовимірних інтегралів, що робить модель непрактичною. Проте, Макфадден знайшов такий розподіл ε_{ij}, для якого

[15] Корисність (англ. utility) в економічній термінології.

формула (8.24) має аналітичне рішення. Він довів, що якщо помилки ε_{ij} є незалежними випадковими величинами із розподілом екстремальних значень (див. п. 2.5.2), то:

$$P(y_i = j) = \frac{e^{u_{ij}}}{\sum_k e^{u_{ik}}} = \frac{\exp(\mathbf{x}_{ij}^T \boldsymbol{\beta})}{\sum_k \exp(\mathbf{x}_{ik}^T \boldsymbol{\beta})},\ j = 1,...,s.. \qquad (8.25)$$

Якщо підставити декомпозицію (8.23) в формулу (8.25), отримаємо:

$$P(y_i = j) = \frac{\exp(\mathbf{z}_j^T \boldsymbol{\alpha} + \mathbf{u}_i^T \boldsymbol{\gamma} + \mathbf{v}_{ij}^T \boldsymbol{\delta})}{\sum_k \exp(\mathbf{z}_k^T \boldsymbol{\alpha} + \mathbf{u}_i^T \boldsymbol{\gamma} + \mathbf{v}_{ik}^T \boldsymbol{\delta})} = \frac{\exp(\mathbf{u}_i^T \boldsymbol{\gamma}) \exp(\mathbf{z}_j^T \boldsymbol{\alpha} + \mathbf{v}_{ij}^T \boldsymbol{\delta})}{\exp(\mathbf{u}_i^T \boldsymbol{\gamma}) \sum_k \exp(\mathbf{z}_k^T \boldsymbol{\alpha} + \mathbf{v}_{ik}^T \boldsymbol{\delta})}.$$

Спільний множник $\exp(\mathbf{u}_i^T \boldsymbol{\gamma})$ скорочується. Отже, МНЛ не дозволяє оцінити безпосередній вплив індивідуальних характеристик на вибір тієї чи іншої альтернативи. Це пояснюється тим, що для кожного індивіда при прийнятті рішень власні характеристики є константою. З точки зору задач класифікації, це не дуже приємна властивість МНЛ, адже ми намагаємось пояснити, які характеристики об'єднують індивідуальні спостереження в один клас чи, навпаки, роз'єднують в різні класи.

Проте, модель (8.25) дозволяє оцінити вплив факторів, які відбивають взаємодію між характеристиками індивіда і атрибутами альтернативи. Якщо альтернатив небагато, то в якості останніх можна обрати просто $I\{y_i = j\}$. Іншими словами, для кожної альтернативи введемо власний набір коефіцієнтів $\boldsymbol{\beta}$ при індивідуальних характеристиках: $u_{ij} = \mathbf{x}_i^T \boldsymbol{\beta}_j$.

Далі, якщо в формулі (8.25) поділити чисельник і знаменник на ту саму величину $e^{u_{i1}}$, вийде:

$$P(y_i = j) = \frac{\mathrm{e}^{u_{ij} - u_{i1}}}{\sum_k \mathrm{e}^{u_{ik} - u_{i1}}} = \frac{\mathrm{e}^{u_{ij} - u_{i1}}}{1 + \sum_{k=2}^{s} \mathrm{e}^{u_{ik} - u_{i1}}} = \frac{\exp(\mathbf{x}_i^T (\boldsymbol{\beta}_j - \boldsymbol{\beta}_1))}{1 + \sum_{k=2}^{s} \exp(\mathbf{x}_i^T (\boldsymbol{\beta}_k - \boldsymbol{\beta}_1))}. \qquad (8.26)$$

Ця ймовірність залежить лише від $s - 1$ різниць в рівнях корисності між першою та j–ю альтернативами. Це безпосередньо випливає з того, що ймовірності вибору мають підсумовуватися в одиницю і в формулі (8.25) лише $s - 1$ рівнянь є лінійно незалежними. Отже, можна оцінити тільки різниці коефіцієнтів $\boldsymbol{\beta}_j$ порівняно з відповідними коефіцієнтами деякого опорного варіанта. Зручно перенумерувати класи, починаючи з нуля і прийняти $\boldsymbol{\beta}_0 = 0$. Наприклад, в маркетингових та соціологічних дослідженнях зазвичай «нульовою» альтернативою вважається відмова від всіх пропонованих альтернатив (наприклад, відмова від участі в виборах), яка дає індивіду нульову корисність.

При такому нормуванні отримуємо наступну систему рівнянь:

$$P(y_i = 0) = \frac{1}{1 + \sum_{k=1}^{s} \exp(\mathbf{x}_i^T \boldsymbol{\beta}_k)};$$

$$P(y_i = j) = \frac{\exp(\mathbf{x}_i^T \boldsymbol{\beta}_j)}{1 + \sum_{k=1}^{S} \exp(\mathbf{x}_i^T \boldsymbol{\beta}_k)},\ j = 1, \ldots, S \qquad (8.27)$$

(де $S = s - 1$). При $s = 2$ модель (8.27) еквівалентна бінарній логістичній регресії, розглянутій у попередньому розділі.

На базі формули (8.27) можна розрахувати S відношень шансів, які мають лог–лінійну форму:

$$\ln \frac{P(y_i = j)}{P(y_i = 0)} = \mathbf{x}_i^T \boldsymbol{\beta}_j, \qquad (8.28)$$

або, якщо обрати як базу порівняння іншу альтернативу:

$$\ln \frac{P(y_i = j)}{P(y_i = k)} = \mathbf{x}_i^T (\boldsymbol{\beta}_j - \boldsymbol{\beta}_k). \qquad (8.29)$$

Останнє відношення ніяк не залежить від наявності інших альтернатив, крім j та k, що є наслідком властивості незалежності непов'язаних альтернатив.

Для оцінки МНЛ, як і для бінарної логістичної регресії, використовується метод максимальної правдоподібності. Клас `LogisticRegression` з пакету `scikit-learn` в Python може виконувати МНЛ за допомогою налаштування `multi_class='multinomial'`.

Приклад 8.4. Застосуємо МНЛ до класифікації ірисів Фішера з прикладів 7.3 та 8.2. В цих прикладах використовувалась інформація тільки про два біологічних види, *I.Setosa* та *I.Versicolor*, які навіть візуально легко розрізнити один від одного внаслідок їх лінійної відокремленості. З третім наявним в базі даних видом, *I.Virginica*, не все так добре, бо характеристики цих квіток значно перетинаються з характеристиками *I.Versicolor*.

Щоб результати класифікації були порівняними з попередніми та наступними прикладами, скористуємось для класифікації всього двома атрибутами даних – довжиною та шириною чашолистка (*Sepal_Length, SL* та *Sepal_Width, SW*). «Нульовим» класом оберемо *I.Versicolor*.

Результати оцінювання МНЛ моделі наведені в табл. 8.3. Для оцінки параметрів використовувався онлайн–калькулятор [13].

Таблиця 8.3 – Параметри МНЛ–моделі для класифікації ірисів Фішера

Позначка	Клас	Змінна	Значення	СКВ
$\beta_{00}, \beta_{01}, \beta_{02}$	*I.Versicolor*	(опорний)	0	0
β_{10}	*I.Setosa*	*Const*	5.95	1.74
β_{11}		*Sepal_Length*	–4.67	0.20
β_{12}		*Sepal_Width*	6.36	0.38
β_{20}	*I.Virginica*	*Const*	–12.51	1.74
β_{21}		*Sepal_Length*	1.70	0.20
β_{22}		*Sepal_Width*	0.65	0.38

Для інтерпретації коефіцієнтів скористаємось відношенням шансів. З (8.28)

$$\ln\frac{P(y_i = Setosa)}{P(y_i = Versicolor)} = 5.95 - 4.67 Sepal\_Length + 6.36 Sepal\_Width.$$

Якщо це відношення більше одиниці (і, відповідно, його логарифм більше нуля), то об'єкт скоріше відноситься до виду *I.Setosa*, ніж до виду *I.Versicolor*. Отже, роздільна лінія між цими двома класами має вигляд

$$5.95 - 4.67 Sepal\_Length + 6.36 Sepal\_Width = 0,$$

або

$$Sepal\_Width = -\frac{5.95}{6.36} + \frac{4.67}{6.36} Sepal\_Length = -0.94 + 0.74 Sepal\_Length.$$

Аналогічно для порівняння *I.Virginica* з *I.Versicolor* отримаємо роздільну лінію у вигляді

$$Sepal\_Width = 19.19 - 2.61 Sepal\_Length.$$

Залишається порівняти *I.Setosa* з *I.Virginica*. Скористуємось тим, що

$$\ln\frac{P(y_i = Setosa)}{P(y_i = Virginica)} = \ln\frac{P(y_i = Setosa)}{P(y_i = Versicolor)} - \ln\frac{P(y_i = Virginica)}{P(y_i = Versicolor)} =$$

$$= 5.95 - 4.67 SL + 6.36 SW - (12.51 + 1.7 SL + 0.65 SW) = 18.46 - 6.38 SL + 5.71 SW.$$

З останнього рівняння знаходимо:

$$Sepal\_Width = -3.23 + 1.12 Sepal\_Length.$$

Ці три лінії наведені на рис. 8.3. Їх опукла оболонка (виділена напівжирними лініями) і визначає правило класифікації.

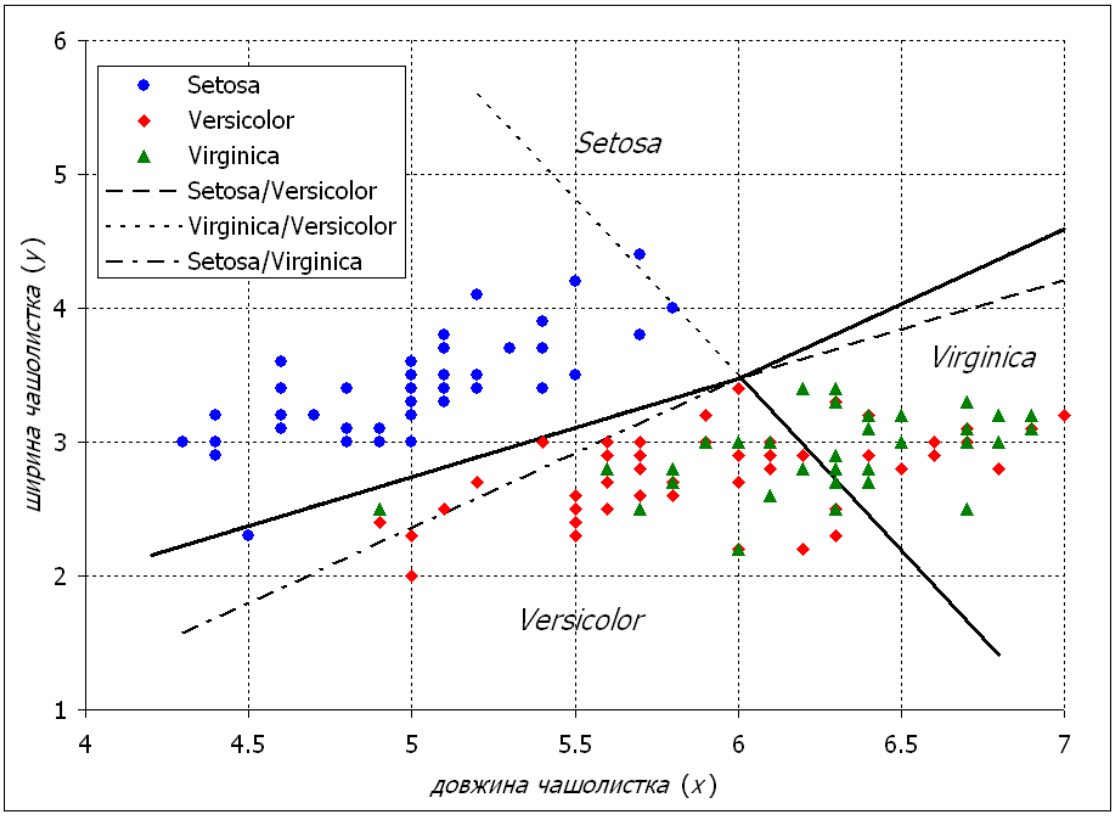


Рис. 8.3. Класифікація ірисів Фішера за допомогою МНЛ–моделі

В таблиці 8.4 наводиться матриця помилок для результатів класифікації за допомогою МНЛ моделі.

Таблиця 8.4 – Матриця помилок для МНЛ–класифікації ірисів Фішера

Справжній вид / Кластер	*versicolor*	*setosa*	*virginica*	Влучність, %
versicolor	**38**	1	14	≈72
setosa	0	**49**	0	100
virginica	12	0	**36**	75
Покриття, %	76	98	72	Точність: 82

Як можна бачити, МНЛ досить надійно розпізнає *I.Setosa*. Для двох інших видів ірисів процент помилок досить високий, що в принципі очікувано, виходячи із розподілу їх характеристик. Але в випадках невірної класифікації МНЛ чесно попереджає про невисоку впевненість в результатах класифікації. Для ілюстрації на рис. 8.4 наводиться графік ймовірності найбільш вірогідного виду ірисів в залежності від характеристик. Видно, що вздовж «каньйону» між *I.Virginica* та *I.Versicolor* достовірність класифікації є низькою. ■

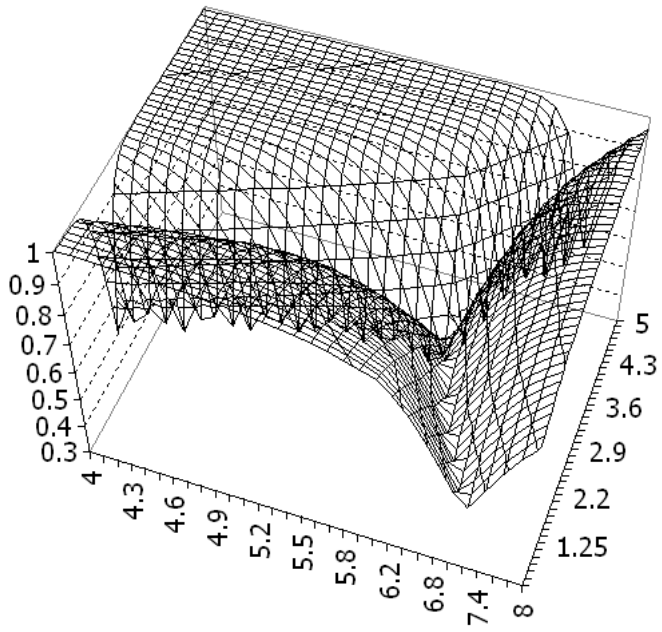


Рис. 8.4. Ймовірність найбільш вірогідного виду ірисів Фішера в залежності від довжини та ширини чашолистка згідно з МНЛ–моделлю

На закінчення зауважимо, що функція (8.25) у вигляді

$$s(j,u_1,...,u_s)=\frac{\exp(u_j)}{\sum_{k=1}^{s}\exp(u_k)} \tag{8.30}$$

широко використовується у машинному навчанні, де вона відома як функція *м'якого максимуму* (англ. *softmax*). Назва пояснюється тим, що експоненціальне

перетворення аргументів збільшує різницю між ними, тож функція (8.30) поверне значення, близьке до одиниці, коли $j = \arg\max(u_1,...,u_s)$. Це дозволяє використовувати м'який максимум як апроксимацію індикаторної функції $I\{\arg\max(u_1,...,u_s) = j\}$, яка є гладкою й диференційованою, що є важливим для багатьох алгоритмів.

Контрольні запитання

1. Як визначаються апріорні ймовірності класів в байєсівських методах класифікації?
2. Як пов'язані апріорні ймовірності класів з апостеріорними?
3. В чому полягає «наївність» наївного байєсівського класифікатора?
4. Охарактеризуйте принцип роботи наївного байєсівського класифікатора.
5. Для чого в алгоритмах байєсівської класифікації застосовується адитивне згладжування?
6. Як в байєсівських методах класифікації опрацьовуються кількісні атрибути?
7. В чому полягають переваги і недоліки наївного байєсівського класифікатора?
8. Як застосувати регресійну модель до вирішення задач класифікації?
9. Як виглядає лінійна модель ймовірності? В чому полягають її недоліки?
10. В чому полягають і чим відрізняються моделі логіт та пробіт?
11. Як виглядає функція правдоподібності для задач бінарної класифікації в логістичній моделі?
12. Що таке відношення шансів і як воно пов'язано з логістичною регресією?
13. Як інтерпретуються коефіцієнти в рівнянні логістичної регресії?
14. В чому полягає властивість незалежності непов'язаних альтернатив?
15. Вплив яких факторів можна і не можна визначити в мультиноміальній логістичній моделі?
16. Як знаходяться ймовірності приналежності до класів у мультиноміальній логістичній моделі?
17. Чому необхідна нормалізація коефіцієнтів рівняння регресії в мультиноміальній логістичній моделі?
18. Як виглядає відношення шансів у мультиноміальній логістичній моделі?
19. Як знайти границі розмежування класів у мультиноміальній логістичній моделі?
20. Що таке «м'який максимум» і для чого він застосовується?

Завдання для самостійної роботи

8.1. В таблиці нижче наводяться найбільш поширені в Україні жіночі та чоловічі імена станом на 2022 р. [Д3]

№	Жіночі	Чоловічі	№	Жіночі	Чоловічі
1	Анна/Ганна	Володимир	11	Віра	Анатолій
2	Марія	Микола	12	Надія	Михайло
3	Софія	Олександр	13	Христина	Петро
4	Олена	Іван	14	Марина	Артем
5	Ірина	Василь	15	Ольга	Владислав
6	Світлана	Сергій	16	Катерина	Марко
7	Тетяна	Андрій	17	Лідія	Юрій
8	Наталія	Дмитро	18	Оксана	Олексій
9	Вікторія	Максим	19	Олександра	Григорій
10	Анастасія	Степан	20	Вероніка	Віталій

а) Розробіть наївний байєсівський класифікатор, який би визначав стать людини за морфологічними характеристиками її імені. В якості атрибутів використовуйте останню та передостанню букви імені (та, якщо знадобиться, інші атрибути – першу букву, довжину імені тощо).

б) Побудуйте матрицю помилок для розробленого у попередньому пункті класифікатора. Розрахуйте показники точності класифікації, які були розглянуті в п. 6.2 – чутливість, специфічність, прогностичну значущість, загальну та збалансовану точність, коефіцієнт кореляції Меттьюза.

в) Визначте за допомогою розробленого НБК стать людини для імен *Людмила, Микита, Любов*. Наведіть імовірності приналежності цих імен «жіночому» та «чоловічому» класу.

8.2. Для наведених в таблиці даних класифікуйте точку (*T*, *F*, 1.0) за допомогою НБК. Вважайте, що атрибут a_3 є нормально розподіленим. Наведіть розрахунки ймовірностей.

№	a_3	a_3	a_3	Клас
1	*T*	*T*	5.0	1
2	*T*	*T*	7.0	1
3	*T*	*F*	8.0	0
4	*F*	*F*	3.0	1
5	*F*	*T*	7.0	0
6	*F*	*T*	4.0	0
7	*F*	*F*	5.0	0
8	*T*	*F*	6.0	1
9	*F*	*T*	1.0	0

8.3. В таблиці наводяться антропометричні показники декількох чоловіків та жінок.

Зріст, м	Вага, кг	Розмір ноги, см	Стать
1,83	82	30	Ч
1,80	86	28	Ч
1,70	77	30	Ч
1,80	75	25	Ч
1,52	45	15	Ж
1,68	68	20	Ж
1,65	59	18	Ж
1,75	68	23	Ж

а) Розрахуйте середні значення та середньоквадратичні відхилення антропометричних показників, а також їх кореляційну матрицю. Наскільки узгодженим з даними є припущення НБК про незалежність ознак?

б) Нарисуйте криві нормального розподілу окремих показників для чоловіків та жінок. Який з показників є найкращим предиктором статі людини?

в) За допомогою НБК визначте стать людини з антропометричними показниками (1,83; 59; 20).

8.4. Оцініть рівняння логістичної регресії за даними із задачі 8.3. Класифікуйте на базі отриманого рівняння людину з антропометричними показниками (1,83; 59; 20).

8.5*. Покажіть, що для задач бінарної класифікації з бінарними атрибутами НБК є частковим випадком логістичної регресії.

8.6. Який метод класифікації Ви обрали б для вирішення задачі 8.1 – НБК чи логістичну регресію? Поясніть, чому.

8.7. Вирішіть задачу класифікації ірисів Фішера за допомогою мультиноміальної логістичної моделі із використанням всіх чотирьох наявних атрибутів в файлі `iris.csv` (довжина чашолистка; ширина чашолистка; довжина пелюстка; ширина пелюстка). Оцініть значущість коефіцієнтів моделі і ранжуйте атрибути за прогностичною здатністю. Сформуйте матрицю помилок моделі за зразком табл. 8.4 і зробіть висновки про надійність класифікації.

8.8. Припустимо, що мешканці деякого міста мають два альтернативних способи дістатися до його центру – власним автомобілем або автобусом. Корисність обох альтернатив співпадає, отже, ймовірності обрання обох видів транспорту дорівнює ½. Припустимо далі, що автоперевізник, який обслуговував маршрути міста, розділився на дві компанії. Автобуси першої компанії тепер мають червоний колір, а автобуси другої – блакитний. В

іншому, з точки зору мешканців міста, нічого не змінилось. Як зміняться варіанти обрання способів транспортування в мультиноміальній логістичній моделі при нових варіантах вибору {автомобіль, червоний автобус, блакитний автобус}? Чи цей результат збігається з Вашими очікуваннями щодо нових ймовірностей вибору? Яка властивість МНЛ є відповідальною за таку реакцію моделі на додання нових альтернатив?

9. АСОЦІАТИВНІ ПРАВИЛА

Пошук асоціативних правил (англ. *association rule learning*) – це сфера машинного навчання, присвячена знаходженню зв'язків між подіями у формі правил «якщо А, то В». Правила, які спостерігаються досить часто, вважаються цікавими. Такі правила дають можливість прогнозувати події, ґрунтуючись на виявлених шаблонах. Вони також допомагають у прийнятті рішень.

Асоціативні правила були вперше запропоновані Ракешем Агравалом та ін. в 1993 р. стосовно задачі *аналізу ринкового кошика* (англ. *market basket analysis*) [11, 12]. Сучасні касові термінали дають можливість зберігати інформацію про кожну транзакцію, здійснену в торгівельній точці. Кожний елемент такої бази даних містить, зокрема, інформацію про товари, одночасно куплені споживачем (його ринковий кошик). Це дає можливість виявити поширені комбінації товарів у формі правил виду «якщо споживач купує в'ялену рибу, то він також купує пиво». Таке правило є досить тривіальним, але інші можуть бути цікавими та неочевидними (як асоціація між пивом та пелюшками, згадувана в першому розділі). Ця інформація може бути корисною менеджерам для встановлення знижок, розміщення товарів та інших маркетингових заходів.

У подальшому асоціативні правила знайшли застосування в таких областях, як медична діагностика, аналіз відвідуваності веб–сторінок (web mining), біоінформатика, кібербезпека тощо.

Термін «асоціативні правила» застосовується зазвичай для аналізу подій, порядок яких неважливий. При аналізі послідовності подій на кшталт «якщо споживач купує смартфон, то протягом року він також купує бездротові навушники» говорять про *пошук послідовних шаблонів* (англ. *sequence mining*).

9.1 Основні визначення

Об'єктом аналізу при пошуку асоціативних правил є база даних транзакцій (або операційна база даних). Це двовимірна таблиця, яка складається з номеру транзакції (tid) і переліку товарів, які були придбані під час цієї транзакції.

Нехай $I = \{i_1, ..., i_m\}$ – множина всіх наявних товарів або інших об'єктів (англ. *items*). Будь-яка його підмножина $X \subseteq I$ називається *набором об'єктів* (англ. *itemset*). Транзакція (англ. transaction) – це пара $\langle i, X \rangle$, де i – це унікальний ідентифікатор транзакції, а X – набір об'єктів. Множина таких транзакцій $D = \{t_1, ..., t_n\}$ називається *операційною базою даних* (англ. *transaction database*).

Операційна база даних може бути подана також іншими способами. У бінарному представленні кожна транзакція – це двійковий вектор, k–й елемент якого дорівнює 1, якщо в i–й транзакції присутній k–й об'єкт і 0 – в іншому випадку. Саме такий спосіб використовувався у піонерській статті Агравала, Імелинські та Свамі [11]. Іноді буває зручно представити базу даних у

вертикальній формі, де для кожного об’єкта наводиться перелік транзакцій, в яких він був присутній.

Рис. 9.1 ілюструє різні варіанти подання операційної бази даних[16].

t	X
1	$ABDE$
2	BCE
3	$ABDE$
4	$ABCE$
5	$ABCDE$
6	BCD

а

t \ x	A	B	C	D	E
1	1	1	0	1	1
2	0	1	1	0	1
3	1	1	0	1	1
4	1	1	1	0	1
5	1	1	1	1	1
6	0	1	1	1	0

б

A	B	C	D	E
1	1	2	1	1
3	2	4	3	2
4	3	5	5	3
5	4	6	6	4
	5			5
	6			

в

Рис. 9.1. Приклад подання операційної бази даних у різних формах:
а – стандартна, *б* – бінарна, *в* – вертикальна.

Асоціативне правило (англ. *association rule*) – це імплікація у формі $X \Rightarrow Y$, де $X, Y \subseteq I$, $X \bigcap Y = \varnothing$. Кожне правило складається з двох частин: лівої, яку називають *приводом* (англ. *antecedent*) та правої, яку називають *наслідком* (англ. *consequent*). В роботі [11] розглядався окремий випадок, коли наслідок складався лише з одного об’єкта.

Для того, щоб правило було цікавим, воно має зустрічатись в базі даних досить часто і мати високу надійність. Ці вимоги формалізуються за допомогою наступних конструкцій.

Підтримка (англ. *support*) набору об’єктів X визначається як частка транзакцій у базі даних D, які містять X:

$$\mathrm{supp}(X) = \frac{n\{t \in D, X \subseteq t\}}{n\{D\}}, \tag{9.1}$$

де $n\{A\}$ позначає кількість елементів множини A.

Набір об’єктів називається *поширеним* (англ. *large*, *frequent*), якщо його підтримка перевищує деякий наперед заданий рівень $s_{\min}$.

В прикладі з рис. 9.1, якщо встановити $s_{\min} = 0{,}5$, тобто розглянути набори, які зустрічаються щонайменше у 3 транзакціях, можна виділити наступні поширені k–елементні набори об’єктів:

$F_1 = \{A, B, C, D, E\}$;

$F_2 = \{AB, AD, AE, BC, BD, BE, CE, DE\}$;

$F_3 = \{ABD, ABE, ADE, BCE, BDE\}$;

$F_4 = \{ABDE\}$.

Якщо відсортувати ці набори за поширеністю, отримаємо табл. 9.1.

[16] Цей приклад і деякі ілюстрації до нього взяті з [57].

Таблиця 9.1 – Поширені набори даних для прикладу з рис. 9.1

Підтримка	Набори об’єктів
6/6	*B*
5/6	*E*, *BE*
4/6	*A*, *C*, *D*, *AB*, *AE*, *BC*, *BD*, *ABE*
3/6	*AD*, *CE*, *DE*, *ABD*, *ADE*, *BCE*, *BDE*, *ABDE*

Достовірність (англ. *confidence*) правила $X \Rightarrow Y$ вказує, наскільки часто воно виконується і визначається як частка транзакцій, які містять *X*, які також містять *Y*:

$$\text{conf}(X \Rightarrow Y) = \frac{\text{supp}(X \bigcup Y)}{\text{supp}(X)}. \qquad (9.2)$$

Чисельник в формулі (9.2) називають також підтримкою правила $X \Rightarrow Y$.

Для прикладу з рис. 9.1 достовірність правила $BC \Rightarrow E$ складає:

$$\text{conf}(BC \Rightarrow E) = \frac{\text{supp}(BCE)}{\text{supp}(BC)} = \frac{3}{4}.$$

Правило називається *сильним* (англ. *strong*), якщо його достовірність перевищує деякий наперед заданий рівень $c_{\min}$. Правила, які є одночасно поширеними та сильними, вважаються *цікавими* (англ. *interesting*).

Нескладно надати ймовірнісну інтерпретацію формулам (9.1) та (9.2). Підтримка набору *X* – це його відносна частота в операційній базі даних, а достовірність правила $X \Rightarrow Y$ – це частотна оцінка умовної ймовірності $P(Y \subset t \mid X \subset t)$. Варто відзначити, що $\text{supp}(X \bigcup Y)$ це не те саме, що $P(X \bigcup Y)$, бо в першому випадку мова йдеться про відносну частоту об’єднання $X \bigcup Y$ в базі даних, а у другому – як часто зустрічаються в ній набори *X* або *Y*.

Використовують також інші показники цікавості асоціативних правил.

Підйом (англ. *lift*) правила $X \Rightarrow Y$ визначається як:

$$\text{lift}(X \Rightarrow Y) = \frac{\text{supp}(X \bigcup Y)}{\text{supp}(X) \times \text{supp}(Y)}. \qquad (9.3)$$

Знаменник у формулі (9.3) надає частотну оцінку ймовірності одночасно зустріти у транзакції *X* та *Y*, якщо б події $X \subset t$ та $Y \subset t$ були незалежними (див. п. 2.7, 4.1). Чисельник надає спостережувану відносну частоту комбінації $X \bigcup Y$. Отже, одиничне значення підйому говорить на користь незалежності наборів *X* та *Y*. Значення більше одиниці вказує на те, що набори *X* та *Y* сумісно трапляються частіше, ніж можна було б пояснити випадковим збігом. Значення менше одиниці свідчить про те, що наявність одного набору негативно впливає на наявність іншого. В контексті аналізу ринкового кошику останнє означає, що товари є замінниками.

Переконливість (англ. *conviction*) правила – альтернативний показник його міцності, запропонований Сергієм Бріном, більш відомим як засновник Google [18]. Переконливість визначається як:

$$\mathrm{conv}(X \Rightarrow Y) = \frac{1-\mathrm{supp}(Y)}{1-\mathrm{conf}(X \Rightarrow Y)}. \qquad (9.4)$$

Підставивши у формулу (9.4) визначення достовірності (9.2), її можна переписати так:

$$\mathrm{conv}(X \Rightarrow Y) = \frac{1-\mathrm{supp}(Y)}{1-\frac{\mathrm{supp}(X \bigcup Y)}{\mathrm{supp}(X)}} = \frac{\mathrm{supp}(X)-\mathrm{supp}(X)\mathrm{supp}(Y)}{\mathrm{supp}(X)-\mathrm{supp}(X \bigcup Y)}. \qquad (9.5)$$

Чисельник в цій формулі показує, як часто зустрічалося б X без Y, якщо б події $X \subset t$ та $Y \subset t$ були незалежними. Знаменник показує, як часто зустрічається X без Y насправді, тобто спостережувану частоту помилок правила $X \Rightarrow Y$. Наприклад, переконливість 1,5 означає, що правило $X \Rightarrow Y$ помилялось би на 50% частіше, якщо б зв'язок між X та Y був би чисто випадковим. На відміну від підйому, переконливість залежить від напрямку правила: $\mathrm{conv}(X \Rightarrow Y) \neq \mathrm{conv}(Y \Rightarrow X)$.

Нарешті, як згадувалось в п. 4.1, статистично обґрунтованим способом перевірки незалежності між приводом та наслідком є тест хі–квадрат. У введених вище позначеннях він матиме вигляд [16]:

$$\chi^2(X \Rightarrow Y) = n\{D\} \times \sum_{a \in \{X, \bar{X}\}} \sum_{b \in \{Y, \bar{Y}\}} \frac{((\mathrm{supp}(a \bigcup b) - \mathrm{supp}(a) \times \mathrm{supp}(b))^2}{\mathrm{supp}(a) \times \mathrm{supp}(b)}. \qquad (9.6)$$

Проте, слід мати на увазі, що тест хі–квадрат допомагає у визначенні наявності зв'язку між змінними, але нічого не говорить про силу такого зв'язку.

Пошук цікавих асоціативних правил зводиться до вирішення двох задач:

1) пошук всіх поширених наборів об'єктів;

2) виявлення в них сильних правил.

Складність першого етапу обумовлена розмірністю множини I. Кожна підмножина I може бути поширеним набором об'єктів, а кількість таких підмножин складає $2^{n\{I\}}$ (див. рис. 9.2). Якщо $n\{I\} = 100$, це становить $2^{100} \approx 10^{30}$ варіантів, а для типової мережі супермаркетів кількість номенклатурних позицій вимірюється десятками тисяч. Отже, перегляд наборів методом «грубої сили» можливий хіба що в ілюстративних прикладах.

Другий етап є відносно простим. Для кожного поширеного набору об'єктів X слід знайти всі його непусті підмножини A, згенерувати правила у формі $A \Rightarrow X \setminus A$ і обрати ти з них, для яких достовірність перевищує $c_{\min}$. Якщо обмежитись випадком, коли наслідок складається лише з одного елементу, то кількість таких правил складає $n\{A\}$, тобто складність алгоритму лінійна. У

будь–якому разі $n\{X\} \square\ n\{I\}$, тобто складність загальної задачі визначається першим етапом.

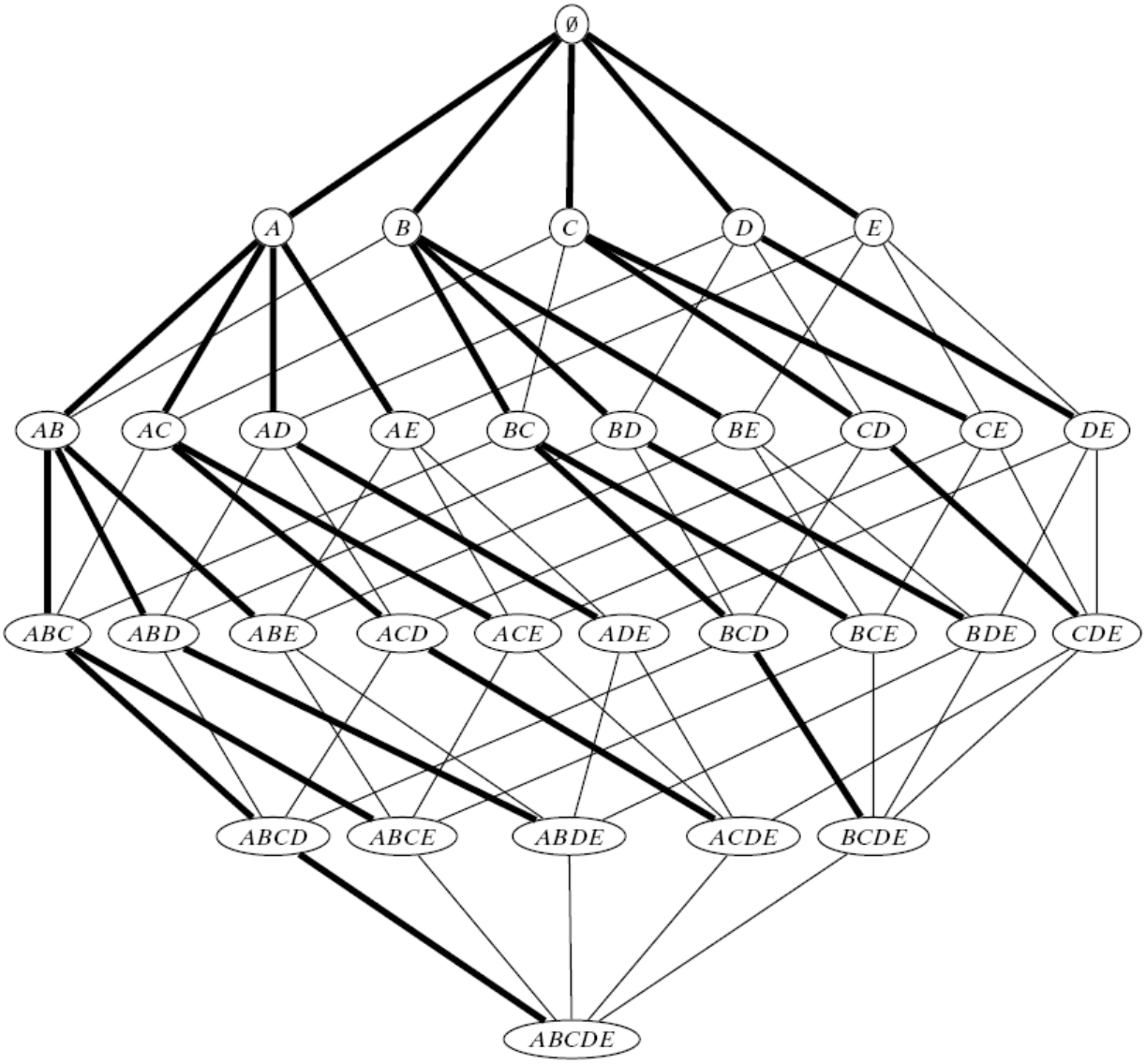


Рис. 9.2. Набори об’єктів для прикладу з рис. 9.1. Жирними лініями виділено дерево пошуку методом «грубої сили».

9.2 Алгоритм Apriori

Для вирішення проблеми з надвеликою розмірністю дерева пошуку наборів об’єктів Р. Агравалом і Р. Срікантом у 1994 р. був запропонований основоположний алгоритм Apriori [12]. Назва алгоритму пояснюється тим, що він використовує попередні (апріорні) знання про властивості поширених наборів об’єктів.

Нехай $X, Y \subseteq I$ – два довільних набори об’єктів. Якщо $X \subseteq Y$, то $\text{supp}(X) \geq \text{supp}(Y)$. З цього випливають два наступних спостереження:

1) якщо Y є поширеним, то будь-яка його підмножина $X \subseteq Y$ також є поширеною;

2) якщо X не є поширеним, то будь-яка його надмножина $Y \supseteq X$ не може бути поширеною (тобто ніякий поширений набір об’єктів не може містити X).

Рис. 9.3 показує, як у наведеному вище прикладі остання властивість (яку називають також *анти–монотонністю*, англ. *anti-monotonicity*) допомагає відсікти багато гілок дерева пошуку на ранній стадії. Заштриховані вузли відповідають непоширеним наборам об’єктів, тоді як штрих–пунктирні вузли та лінії вказують на всі відсічені вузли та гілки. Суцільні лінії позначають поширені набори об’єктів.

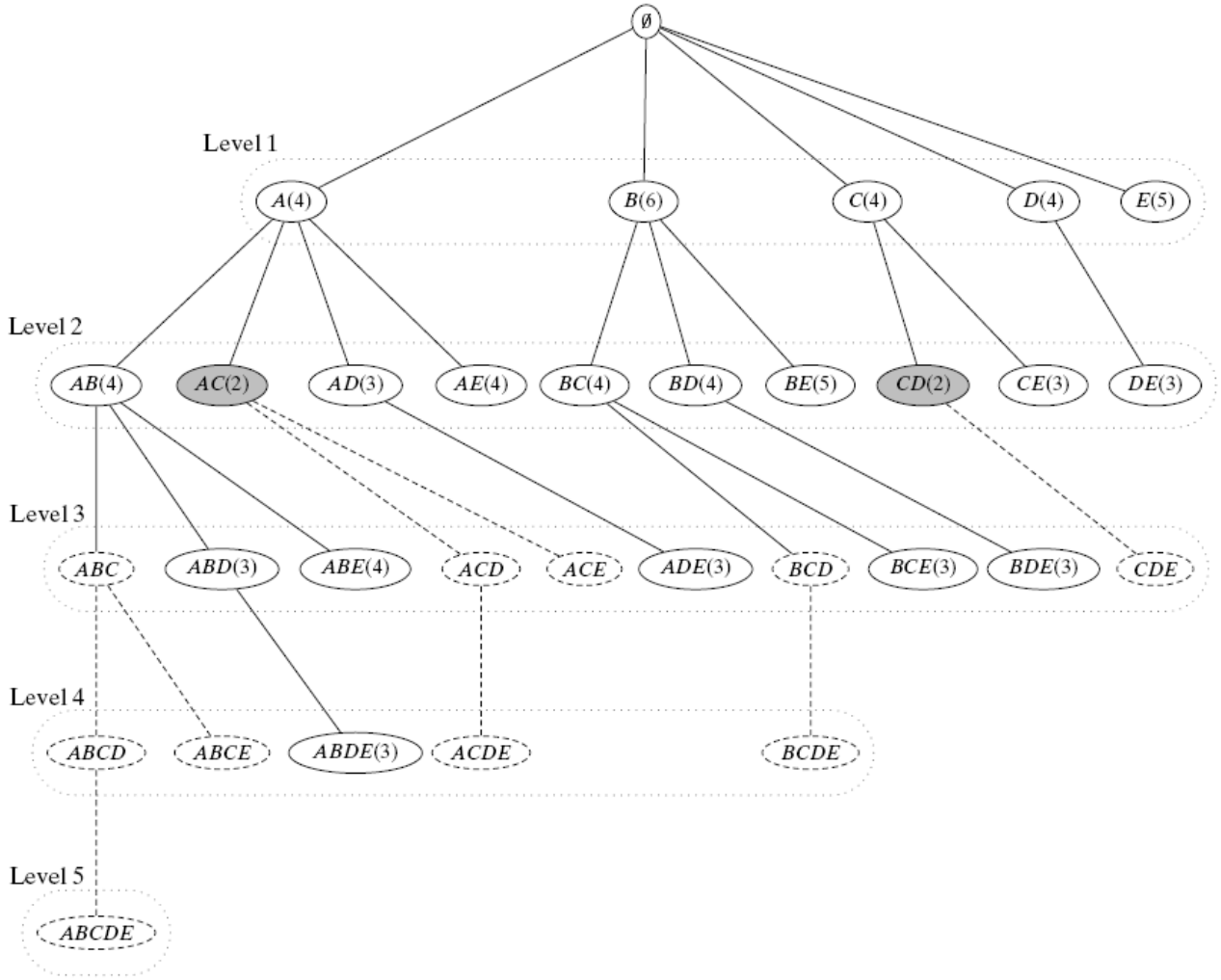


Рис. 9.2. Робота алгоритму Apriori для прикладу з рис. 9.1

Алгоритм Apriori використовує ітераційний пошук за рівнями, в якому k–елементні набори об’єктів використовуються для побудови (k+1)–елементних наборів. Позначимо множину поширених k–елементних наборів через F_k, а множину кандидатів для відбору – через C_k. Алгоритм Apriori спочатку сканує базу даних і створює множину F_1 з тих об’єктів, які задовольняють вимогу щодо мінімальної підтримки. Далі послідовно повторюються наступні три кроки, доки не будуть знайдені всі поширені набори об’єктів:

1. Згенерувати C_{k+1}, множину кандидатів до включення в F_{k+1}, із поширених k–елементних наборів F_k.

2. Обчислити підтримку для кожного кандидата з множини C_{k+1}.

3. Додати набори елементів з підтримкою не нижче мінімальної до F_{k+1}.

Вважається, що всі набори об'єктів відсортовані у лексикографічному порядку. Генерація кандидатів складається з двох етапів:

1. [*Об'єднання*, англ. *joining*]: новий кандидат створюється шляхом об'єднання двох поширених k–елементних наборів, які мають $k-1$ спільних перших елементів. Наприклад, якщо *ABC* та *ABD* – поширені 3–елементні набори, то поширеним 4–елементним набором може бути *ABCD*.

2. [*Обрізка*, англ. *pruning*]: відкинути кандидата, якщо хоча б одна з його k–елементних підмножин не має мінімальної підтримки. Так, можна відкинути набір *ABCD*, якщо його складові *ACD* або *BCD* не є поширеними.

Загальна схема алгоритму Apriori на псевдокоді наведена на рис. 9.3.

```
function Apriori(D, s_min):
    F_1 = {поширені об'єкти};
    k = 2;
    while F_{k-1} ≠ ∅
        C_k = GetCandidates(F_{k-1}, k);
        for each транзакція t ∈ D
            L_t = {c ∈ C_k | c ⊆ t};
            for each кандидат c ∈ L_t count[c]++;
        F_k = {c ∈ C_k | count(c) ≥ n{D} × c_min};
        k++;
    return F = ⋃_k F_k.
function GetCandidates(F, k):
    list = ∅;
    for each p,q ∈ F where p_1 = q_1, p_2 = q_2, ..., p_{k-2} = q_{k-2}, p_{k-1} < q_{k-1}
        кандидат c = p ⋃ {q_{k-1}};
        if кожна (k–1)–елементна підмножина c ∈ F
            list = list ⋃ c;
 return list.
```

Рис. 9.3. Загальна структура алгоритму Apriori на псевдокоді

Звичайно, в цій схемі опущено багато деталей. З точки зору швидкодії, найважливішою частиною програмної реалізації алгоритму є структури даних, що використовуються для зберігання кандидатів і підрахунку їх частот. Також є

багато досить очевидних шляхів для вдосконалення алгоритму. Наприклад, якщо певна транзакція не містить поширених *k*–елементних наборів об'єктів, то вона не може містити поширених (*k*+1)–елементних наборів. Отже, нема необхідності в її подальшому скануванні.

На жаль, у найгіршому випадку складність алгоритму експоненційна.

Після виділення всіх поширених наборів об'єктів, наступним кроком є генерація асоціативних правил. Для цього слід переглянути всі поширені набори і розрахувати достовірність правил, які можуть бути з них створені. Якщо є поширений набір $Z \in F$, то для будь-якої його непустої підмножини $X \subset F$ можна сформувати асоціативне правило у формі $X \Rightarrow Y$, де $Y = Z \setminus X$. Кожне таке правило має бути поширеним, оскільки $\text{supp}(X \Rightarrow Y) = \text{supp}(Z)$. Залишається перевірити достовірність такого правила. За формулою (9.2),

$$c = \text{conf}(X \Rightarrow Y) = \frac{\text{supp}(X \bigcup (Z \setminus X))}{\text{supp}(X)} = \frac{\text{supp}(Z)}{\text{supp}(X)}.$$

Якщо $c \geq c_{\min}$, це правило є сильним. З іншого боку, якщо $c < c_{\min}$, то $\text{conf}(W \Rightarrow Z \setminus W) < c_{\min}$ для всіх підмножин $W \subseteq X$, оскільки $\text{supp}(W) \geq \text{supp}(X)$. Отже, будь-які підмножини *X* не можуть бути приводом сильного правила і їх можна викреслити з подальшого перебору варіантів.

Розглянемо найбільший поширений набір для даних з рис. 9.1, *ABDE*. Встановимо $c_{\min} = 0{,}9$. Множина приводів для асоціативних правил матиме вигляд (в дужках вказані абсолютні частоти наборів):

X = {*ABD*(3), *ABE*(4), *ADE*(3), *BDE*(3),
AB(4), *AD*(3), *AE*(4), *BD*(4), *BE*(5), *DE*(3), *A*(4), *B*(6), *D*(4), *E*(5)}.

Перший елемент *X* це *ABD*, а достовірність правила $ABD \Rightarrow E$ дорівнює 3/3 = 1, тож маємо перше сильне правило. Наступним набором є *ABE*, але відповідне правило $ABE \Rightarrow D$ слабке, оскільки $\text{conf}(ABE \Rightarrow D) = 3/4 = 0{,}75$. Отже, можна видалити з *X* усі підмножини *ABE*. Оновлена множина приводів містить такі елементи:

X = {*ADE*(3), *BDE*(3), *AD*(3), *BD*(4), *DE*(3), *D*(4)}.

Далі обираємо *ADE*, що дає сильне правило $ADE \Rightarrow B$, як і набори *BDE* та *AD*. Для набору *BD* $\text{conf}(BD \Rightarrow AE) = 3/4 = 0{,}75$, так що можна видалити з *X* всі підмножини *BD*. Після цього залишається тільки правило $DE \Rightarrow AB$, яке теж виявляється сильним. У підсумку отримуємо такі сильні правила:

$ABD \Rightarrow E$, $ADE \Rightarrow B$, $BDE \Rightarrow A$, $AD \Rightarrow BE$, $DE \Rightarrow AB$.

Для всіх цих правил достовірність складає 1.

9.3 Алгоритм ECLAT

Кожний раз при збільшенні кількості об'єктів в наборі на одиницю, алгоритм Apriori заново сканує операційну базу даних для обчислення

підтримки. Для того, щоб підрахувати підтримку, треба створити підмножини кожної транзакції та перевірити, чи існують вони в у пошуковому дереві. Це наводить до великої кількості зайвих обчислень, бо генерується багато підмножин, які не існують у дереві префіксів.

Етап підрахунку підтримки можна значно покращити, якщо реорганізувати базу даних так, щоб прискорити обчислення частот наборів. Одним із таких способів є алгоритм ECLAT (скорочення від англ. Equivalence Class Clustering and bottom-up Lattice Traversal), запропонований М. Закі та ін. [58]. Алгоритм використовує вертикальну організацію операційної бази даних (див. рис. 9.1*в*), що дозволяє значно спростити підрахунок частот.

Позначимо через $t(X)$ множину транзакцій, які містять X (в [58] для неї використовується термін tid–множина, англ. *tidset*). Абсолютна частота набору X дорівнює кількості елементів цієї множини, $n\{t(X)\}$. При об'єднанні двох наборів X та Y

$$t(X \bigcup Y) = t(X) \bigcap t(Y), \tag{9.7}$$

а частота об'єднаного набору дорівнює просто кількості елементів цієї множини. Наприклад, для даних з рис. 9.1 $t(A) = \{1,3,4,5\}$, $t(C) = \{2,4,5,6\}$, і $t(AC) = \{4,5\}$. Отже, частота набору AC складає 2. Це дозволяє позбавитись сканування бази даних для визначення підтримки нових наборів об'єктів.

Алгоритм ECLAT об'єднує tid–множини поширених наборів лише якщо вони мають спільний префікс. Набори об'єктів з однаковим префіксом називають *префіксним класом* (англ. *prefix equivalence class*). Прикладом префіксного класу може бути $P_A = \{AB, AC, AD, AE\}$, оскільки всі елементи мають спільний префікс A. Алгоритм ECLAT проходить дерево пошуку в глибину, рекурсивно опрацьовуючи префіксні класи. Схема алгоритму наведена на рис. 9.4.

```
procedure ECLAT(P, s_min, F):
    for each ⟨X_a, t(X_a)⟩ ∈ P
        F = F ⋃ {(X_a, supp(X_a)};
        P_a = ∅;
        for each ⟨X_b, t(X_b)⟩ ∈ P | X_b > X_a
            X_ab = X_a ⋃ X_b;
            t(X_ab) = t(X_a) ⋂ t(X_b);
            if sup p(X_ab) ≥ s_min then
                P_a = P_a ⋃ {⟨X_ab, t(X_ab)⟩}
    if P_a ≠ ∅ then ECLAT(P, s_min, F).
```

Рис. 9.4. Загальна структура алгоритму ECLAT на псевдокоді

Вважається, що перший аргумент P містить тільки поширені набори даних. При першому зверненні до процедури $F = \varnothing$, а P містить tid–множини для всіх поширених базових об'єктів: $P = \{\langle i, t(i)\rangle \mid i \in I, n\{t(i)\} \geq s_{\min}\}$. Для заданого класу еквівалентності префікса P, алгоритм намагається поєднати набір елементів $X_a \in P$ з усіма іншими наборами $X_b \in P$, створюючи набір-кандидат $X_{ab} = X_a \bigcup X_b$. Щоб визначити, чи є цей набір поширеним, користуємось формулою (9.7). Якщо так, то кандидат додається до нового класу еквівалентності P_a, який буде містити усі поширені набори елементів з префіксом X_a. Рекурсивний виклик ECLAT знаходить далі усі продовження гілки X_a в дереві пошуку. Цей процес триває, доки не вичерпаються можливі продовження в усіх гілках дерева пошуку. Рис. 9.5 пояснює принцип роботи алгоритму. Затінені прямокутники позначають непоширені набори, які далі не розгалужуються.

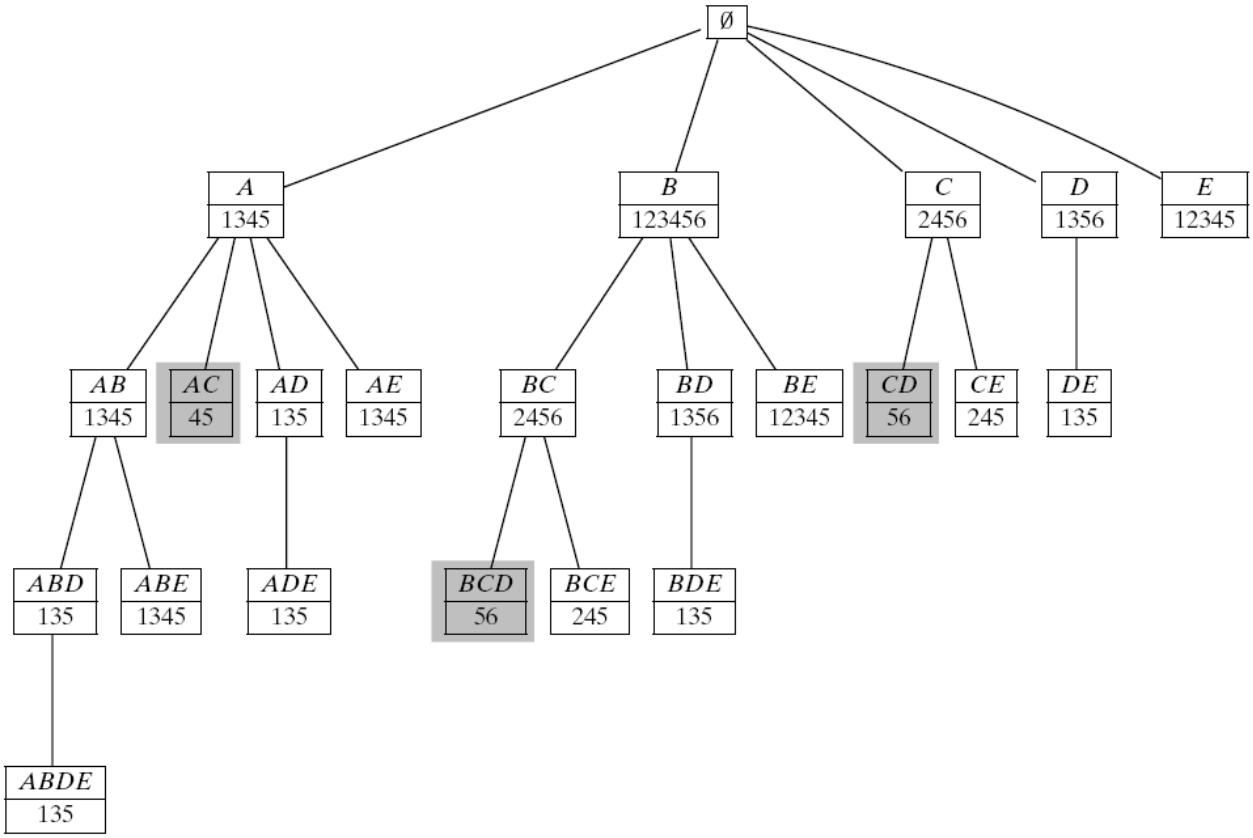


Рис. 9.5. Робота алгоритму ECLAT для прикладу з рис. 9.1.

9.4 Вдосконалення та розширення базової моделі

Властивість анти–монотонності, яка вдало використана в алгоритмі Apriori, дозволяє суттєво зменшити витрати часу на пошук асоціативних правил. Проте, при великій кількості об'єктів і низькому рівні мінімальної підтримки, обчислювальна складність задачі залишається дуже високою. З моменту появи алгоритму було зроблено багато спроб підвищити його ефективність і розширити функціональність.

Крім наведеного вище алгоритму ECLAT, слід відмітити алгоритм

FP–growth [28], авторам якого вдалося усунути етап генерації кандидатів. Алгоритм складається з двох етапів:

1) стиснення бази даних, під час якого поширені набори збираються у дерево поширених шаблонів (FP-дерево), яке зберігає всю важливу інформацію;

2) декомпозиція стисненої бази даних на множину умовних баз даних, пов'язаних з одним поширеним об'єктом, і пошук поширених наборів в кожній такій базі окремо.

Цей метод працює на порядок швидше, ніж оригінальний алгоритм Apriori, хоча у гіршому випадку його складність залишається експоненційною.

В залежності від встановлених граничних рівнів мінімальної підтримки та достовірності, всі наведені вище алгоритми можуть створювати велику кількість асоціативних правил, багато з яких будуть подібними. Це ускладнює їх інтерпретацію аналітиком. Отже, представляють інтерес методи для подання знайдених правил у найбільш компактній формі.

Одним із таких способів є знаходження *максимальних поширених наборів об'єктів* (англ. *maximal frequent itemsets*), тобто таких, які не є частиною інших поширених наборів. Множина максимальних наборів

$$M = \left\{ X \in F \mid \nexists \left(Y \supset X \,\&\, Y \in F \right) \right\} \tag{9.8}$$

є стиснутим поданням всієї множини поширених наборів F, оскільки будь-яка підмножина максимального набору теж має бути поширеною. Наприклад, для даних з рис. 9.1 є всього два максимальних поширених набори: *ABDE* та *BCE*, і всі інші набори можуть бути виведені з них. Існують також інші способи стиснутого подання множини F і група алгоритмів, які знаходять такі подання (GenMax, MaxMiner, Charm тощо). Деталі їх реалізації можна знайти в [57].

Розглянуті вище асоціативні правила встановлюють лише наявність чи відсутність асоціації, тому їх називають також *логічними* або *булевими* (англ. *Boolean association rules*). Операційні бази даних, як правило, містять також кількісну інформацію (ціни, кількість куплених однотипних товарів, дата та час транзакції тощо), яка теж може бути корисною при виявлені шаблонів поведінки споживачів. Для врахування такої інформації використовуються *кількісні асоціативні правила* (англ. *quantitative association rules*). Вони мають більш складний формат:

$$\left\langle A_1 = x_1, ..., A_{j-1} = x_{j-1}, A_{j+1} = x_{j+1}, ..., A_n = x_n \right\rangle \Rightarrow A_j = x_j,$$

де $A_i, i = 1..., n$ – атрибути транзакції. Для використання кількісних правил безперервні кількісні атрибути перетворюються в категоріальні шляхом розбиття на інтервали. Для цього використовуються ті ж самі методи, що й при вирішенні аналогічної проблеми в деревах рішень [50].

Пошук асоціативних правил на рівні окремих товарних одиниць[17] може не виявити цікавих правил внаслідок надмірної деталізації. Наприклад, правило "вівсяні пластівці “Геркулес”, виробник ТОВ “Добродія”, об’єм 500г» ⇒ «молоко “Яготинське” тривалого зберігання, тетрапак, 900 мл» може не мати високої підтримки, на відміну від узагальненого правила «вівсяні пластівці» ⇒ «молоко». Для врахування таксономії об’єктів використовуються *багаторівневі асоціативні правила* (англ. *multilevel association rules*). При побудові таких правил об’єкти групуються згідно з ієрархією і пошук ведеться на найвищому її рівні, що також знижує обчислювальну складність задачі [49].

Контрольні запитання

1. Що розуміється під асоціативним правилом?
2. Що таке операційна база даних і які бувають форми її подання?
3. Що таке набір об’єктів і як визначається його підтримка?
4. Як визначається достовірність асоціативного правила?
5. В чому полягає сенс підйому асоціативного правила?
6. Як інтерпретується переконливість асоціативного правила?
7. Який статистичний тест використовується для перевірки залежності між приводом та наслідком асоціативного правила?
8. В чому полягає властивість анти–монотонності і як вона допомагає спростити процес пошуку асоціативних правил?
9. В чому полягають основні етапи алгоритму Apriori?
10. Скільки разів сканується операційна база даних при виконанні алгоритму Apriori?
11. Як відбувається генерація кандидатів в алгоритмі Apriori?
12. В чому полягають переваги вертикальної організації операційної бази даних порівняно з горизонтальною?
13. Що мається на увазі під префіксною організацією дерева пошуку?
14. В чому полягають основні етапи алгоритму ECLAT?
15. Як генеруються кандидати в алгоритмі ECLAT і як розраховується їх підтримка?
16. Як із множини поширених об’єктів генеруються асоціативні правила?
17. Що розуміється під максимальним поширеним набором об’єктів?
18. Чому множини максимальних поширених наборів об’єктів достатньо для генерації всіх цікавих асоціативних правил?
19. Як подаються та формуються кількісні асоціативні правила?
20. В чому полягають переваги використання багаторівневих асоціативних правил?

[17] Для них зазвичай використовується термін SKU – скорочення від англ. stock keeping unit, одиниця складського обліку.

Завдання для самостійної роботи

9.1. Наведіть два приклади практичних задач, для вирішення яких був би корисним пошук асоціативних правил.

9.2. Доведіть, що будь-яка непорожня підмножина поширеного набору об’єктів також має бути поширеною.

9.3. Доведіть, що підтримка будь-якої непорожньої підмножини S набору елементів X має бути не меншою, ніж підтримка X.

9.4. Нехай $S' \subset S \subset X$, де X – поширений набор елементів. Покажіть, що достовірність правила $S' \Rightarrow X \setminus S'$ не може перевищувати достовірність правила $S \Rightarrow X \setminus S$.

9.5. Для наведеної в таблиці операційної бази даних

t	X
1	*КАРЛ*
2	*У КЛАРИ*
3	*УКРАВ*
4	*КОРАЛИ*
5	*КЛАРА*
6	*У КАРЛА*
7	*ВКРАЛА*
8	*КЛАРНЕТ*

кожна буква і символ пробілу відповідають окремим об’єктам. Якщо в транзакції зустрічається декілька однакових об’єктів, то вони розглядаються як один об’єкт.

а) Знайдіть всі набори об’єктів із підтримкою вище 4/8.

б) Знайдіть всі правила з достовірністю не нижче 0,9.

в) Розрахуйте підйом та переконливість знайдених сильних правил.

9.6. Для операційної бази даних із завдання 9.5:

а) Перетворіть базу даних до вертикального формату.

б) Знайдіть всі набори об’єктів із підтримкою вище 4/8 за допомогою алгоритму ECLAT.

9.7. Один із варіантів алгоритму Apriori спершу поділяє операційну базу даних D на n частин, що не перекриваються. Доведіть, що будь-який поширений набір елементів D має бути поширеним принаймні в одній із її частин.

9.8*. Більшість алгоритмів пошуку поширених шаблонів не враховують кількість однакових елементів у наборі. Проте, ця інформація може бути важливою для виявлення асоціативних правил. Наприклад, правило «пиво $\Rightarrow$ піца» може мати меншу достовірність, ніж правило «4 банки пива $\Rightarrow$ піца».

Запропонуйте спосіб модифікації алгоритму Apriori (або іншого), який дозволив би враховувати подібні ситуації.

9.9. У наведеній нижче таблиці спряженості узагальнено дані про кількість транзакцій супермаркету, які містять (або не містять) певні комбінації товарів.

		Чай		Всього
		Так	Ні	
Печиво	Так	2000	500	2500
	Ні	1000	1500	2500
Всього		3000	2000	5000

а) Чи буде асоціативне правило «чай $\Rightarrow$ печиво» сильним, якщо мінімальна підтримка становить 25%, а мінімальна достовірність – 50%? Обґрунтуйте свою відповідь.

б) Чи є покупка чаю незалежною від покупки печива? Якщо ні, то як вони корельовані між собою?

в) Розрахуйте підйом та переконливість цього асоціативного правила, а також коефіцієнт кореляції Меттьюза для покупок печива та чаю.

10. ЗАДАЧІ КЛАСТЕРИЗАЦІЇ

Кластеризація (англ. *clustering*) – це процес розбиття вибірки на підмножини, які називають кластерами (англ. *clustering*). Об'єкти всередині кожного кластера мають бути подібними, а об'єкти різних кластерів – суттєво відрізнятися. Якщо дані вибірки подати як точки у просторі атрибутів, то задача кластеризації зводиться до визначення «згущень точок». Об'єкти одного кластеру можна розглядати разом як одну групу, і тому кластеризація є формою стиснення даних та підведення підсумків.

Задача кластеризації подібна до задачі класифікації, але, на відміну від неї, класи досліджуваного набору даних заздалегідь не визначені. Тому кластеризація відноситься до задач навчання без вчителя. Відсутність прикладів правильно класифікованих об'єктів означає, що не існує об'єктивного критерію для визначення якості кластеризації. Подібність об'єктів всередині кластера теж можна визначити різними способами (див. рис. 10.1). Популярні поняття кластерів включають групи з малими відстанями між членами кластерів, щільні області у просторі атрибутів, або певні багатовимірні статистичні розподіли. Кількість кластерів теж заздалегідь невідома. Тому кластеризацію можна вважати багатокритеріальною оптимізаційною задачею.

Все це обумовлює величезну кількість алгоритмів кластеризації (більше 100 алгоритмів). Далі будуть розглянуті тільки найбільш поширені методи.

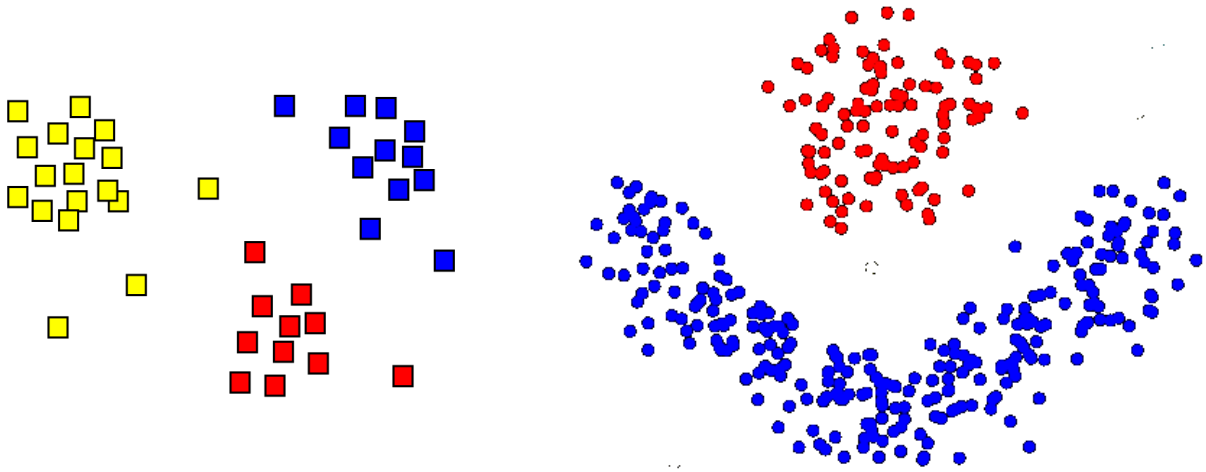


Рис. 10.1. Приклади кластерів із різним визначенням подібності

10.1 Основні визначення і класифікація методів

Формально задача кластеризації полягає у наступному. Задано набір даних з n об'єктів з m атрибутами $D=\{\mathbf{x}_i\}_{i=1}^{n}$. Треба розбити цей набір на $k<n$ взаємовиключних класів (кластерів) $\{C_1,...,C_k\}$ так, щоб кожен кластер містив хоча б один об'єкт, а кожний об'єкт належав в точності до одного кластеру.

Метод «грубої сили» стосовно задачі кластеризації полягає в тому, щоб спробувати всі варіанти розбиття n об'єктів на k кластерів, обчислити деякий

критерій якості такого розбиття і обрати найкращий варіант. В комбінаториці кількість невпорядкованих розбиттів n–елементної множини на k підмножин називається числом Стірлинга другого роду і задається формулою [55]:

$$S(n,k) = \frac{1}{k!}\sum_{j=0}^{k}(-1)^{k+j}C_k^j(k-j)^n. \qquad (10.1)$$

Для $n = 100$, $k = 2$ це становить приблизно 10^{30} комбінацій. Зрозуміло, що такий підхід є нездійсненним для будь-яких практичних задач. Тому всі алгоритми кластеризації є евристичними.

Основні методи кластеризації можна розділити на такі категорії.

1. *Методи розбиття* (англ. *partitioning methods*) є ітеративними. Виходячи із наперед заданої кількості кластерів k, вони створюють деяке початкове розбиття об’єктів на кластери. Потім послідовно здійснюються спроби покращити поточне розбиття за певним критерієм якості шляхом переміщення об’єктів з однієї групи в іншу. Найбільш відомим представником цієї групи методів є алгоритм k–середніх, в якому кожному кластеру ставиться у відповідність репрезентативна точка – середнє значення включених до кластеру об’єктів:

$$\bar{\mathbf{x}}_i = \frac{1}{n\{C_i\}}\sum\nolimits_{j\in C_i}\mathbf{x}_j. \qquad (10.2)$$

Ці методи добре працюють для пошуку кластерів сферичної форми в базах даних малого та середнього розміру. Щоб знайти кластери зі складними формами та для кластеризації великих наборів даних потрібні інші методи.

2. *Ієрархічні методи* (англ. *hierarchical methods*) створюють ієрархічну, деревоподібну декомпозицію набору об’єктів бази даних. Таку декомпозицію називають також *таксономією* (англ. *taxonomy*). Класичним прикладом таксономії на основі схожості є класифікація живих істот, запропонована Карлом Ліннеєм у середині XVIII ст. У сучасному уявленні біологічна ієрархія включає сім основних рівнів: царство, тип, клас, ряд, родина, рід, вид.

Ієрархічні методи, у свою чергу, поділяються на об'єднувальні та розділювальні в залежності від того, як формується ієрархічна декомпозиція. *Об'єднувальний,* або *агломератний* підхід (англ. *agglomerative approach*) створює таксономію методом «знизу–вгору». Спочатку кожний об’єкт розміщується у власний кластер, а далі схожі кластери об’єднуються, доки всі групи не об’єднаються в одну (найвищий рівень ієрархії). *Розділювальний підхід* (англ. *divisive approach*) працює «згори–вниз». Спочатку всі об’єкти знаходяться в одному кластері. На кожній наступній ітерації кластер розбивається на менші кластери, поки врешті-решт кожен об’єкт не потрапить у власний кластер (див. рис. 10.2).

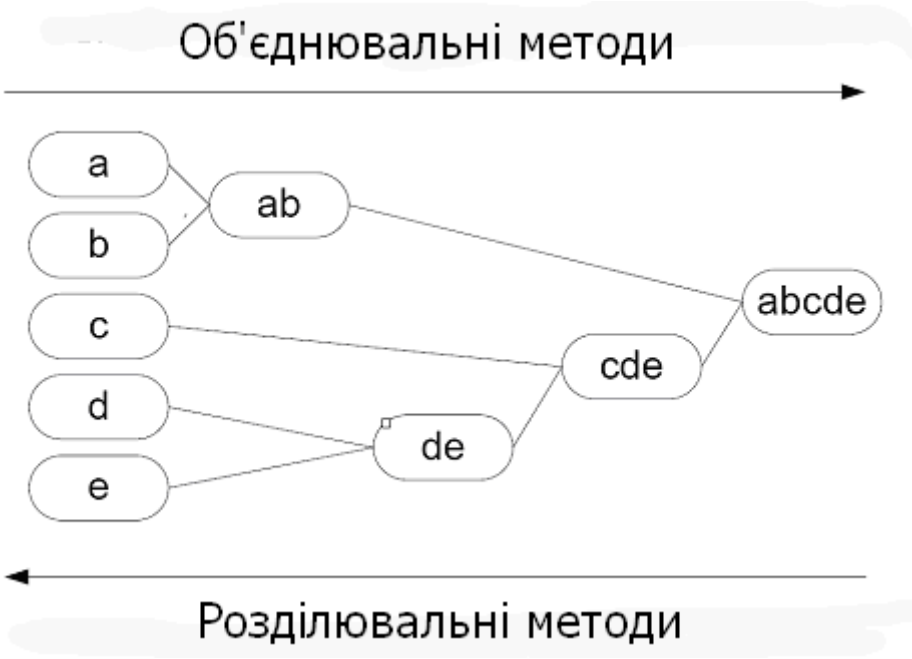


Рис. 10.2. Порівняння об’єднувальних та розділювальних методів кластеризації

Особливість ієрархічних методів полягає в тому, що після виконання наступного кроку (злиття або розбиття) його результати не можна скасувати. З одного боку, це приводить до суттєвого зменшення обчислювальних витрат. З іншого боку, немає можливості виправити попередні рішення у разі помилки.

3. *Методи на основі щільності* (англ. *density-based methods*). Більшість методів розбиття групують об’єкти на основі відстані між ними. Такі методи можуть знайти лише сферичні скупчення і непридатні для виявлення кластерів довільної форми. Для вирішення останньої задачі застосовуються методи, які базуються на виділенні зв’язаних областей високої щільності. Їх загальна ідея – продовжувати зростання обраного кластера доки кількість точок даних в його «околиці» (щільність) перевищує деякий поріг. Приклад кластерів складної форми наведено на рис. 10.1. Також ця група методів успішно використовується для пошуку викидів в даних.

4. *Решіткові методи* (англ. *grid-based methods*) розбивають простір атрибутів на дискретні комірки, які утворюють структуру решітки. Операції кластеризації виконуються на цьому дискретизованому просторі. Основна перевага такого підходу полягає у високій швидкодії: час обробки даних залежить не від обсягу вибірки, а від кількості комірок дискретизованого простору.

5. *Модельні методи* (англ. *model-based methods*) будують модель поведінки даних і знаходять таке розбиття об’єктів на кластери, яке найкраще відповідає цій моделі. Наприклад, алгоритм може знаходити кластери шляхом побудови функції густини ймовірності, яка відображає розподіл об’єктів у просторі атрибутів. Це дозволяє автоматично визначати кількість кластерів, спираючись на стандартний апарат математичної статистики, відфільтровувати «шум» та

викиди, що робить модельні методи робастними. Найбільш відомим представником цього класу є алгоритм максимізації очікувань (EM–алгоритм) для розділення суміші нормально–розподілених випадкових величин.

Окремо слід відзначити задачі кластеризації багатовимірних даних. Деякі прикладні задачі вимагають аналізу об'єктів, які характеризуються дуже великою кількістю атрибутів. Наприклад, атрибутами текстових документів є тисячі та десятки тисяч слів, які вони містять. Кластеризація таких даних є складною через «прокляття розмірності». Зі збільшенням кількості вимірів простору атрибутів дані стають все більш розрідженими, тому вимірювання відстані між об'єктами стають безглуздими, а середня щільність точок у будь-якому місці даних ймовірно буде низькою. Багато атрибутів можуть бути нерелевантними. Тому багатовимірні дані вимагають окрему методологію кластеризації. CLIQUE і PROCLUS є двома впливовими методами, які шукають кластери в підмножинах вимірів даних, а не в усьому просторі. Іншим підходом є кластеризація на основі поширених шаблонів, які використовуються для групування об'єктів і генерування значущих кластерів. Прикладом такої методології є алгоритм pCluster [57].

Більшість алгоритмів кластеризації групують об'єкти на базі відстані між ними. Формально, для будь-яких об'єктів $i, j \in D$ має бути визначена функція відстані $d(i, j)$, яка задовольняє наступним вимогам:

1. $d(i, j) \geq 0$: відстань є невід'ємним числом.
2. $d(i,i) = 0$: відстань від об'єкта до самого себе дорівнює нулю.
3. $d(i, j) = d(j,i)$: відстань є комутативною функцією.
4. $d(i, j) \leq d(i,k) + d(k, j)$: відстань від i до j не перевищує довжини «об'їзного маршруту» через проміжний об'єкт k (нерівність трикутника).

Якщо для всіх об'єктів $i, j \in D$ задана така функція, то D називають *метричним простором* (англ. *metric space*).

Поняття відстані між об'єктами використовується також в алгоритмах kNN та SVM, тому всі зауваження стосовно проблем із її вимірюванням, описані у відповідних розділах (п. 7.2, 7.3), застосовні і тут. Якщо всі атрибути об'єктів вимірюються за кількісними шкалами (інтервалів або відношень), то в якості функції відстані зазвичай використовується евклідова відстань, іноді – манхеттенська або Чебишева. Для категоріальних змінних можуть використовуватись інші показники близькості [27, 389–398].

Інформацію про подібність різних об'єктів іноді зручно надати у формі матриці попарних відстаней між ними з елементами $d_{ij} = d(i, j),\ i, j = 1,...,n$.

10.2 Алгоритм *k*–середніх

Для будь-якого розбиття об'єктів на кластери $C = \{C_1, ..., C_k\}$ треба визначити функцію оцінки якості такого розподілу. Найчастіше використовується оцінка за сумою квадратів помилок:

$$SSE(C) = \sum_{i=1}^{k} \sum_{\mathbf{x}_j \in C_i} \left\| \mathbf{x}_j - \bar{\mathbf{x}}_i \right\|^2, \tag{10.3}$$

де $\bar{\mathbf{x}}_i$ – середнє *i*–го кластеру за формулою (10.2), $\|\mathbf{a} - \mathbf{b}\|$ – евклідова відстань між векторами **a** і **b**. Ціль полягає у пошуку такого розбиття, яке б мінімізувало цю функцію:

$$C^* = \arg\min_C SSE(C). \tag{10.4}$$

Алгоритм *k–середніх* (англ. *k–means*) використовує жадібну стратегію пошуку оптимального розбиття. Алгоритм починається з вибору *k* середніх точок (які називають також *центроїдами*, англ. *centroids*) у просторі атрибутів. Це можна зробити такими способами:

1) згенерувати *k* псевдовипадкових точок, рівномірно розподілених у діапазоні варіації даних за кожним виміром;

2) ініціалізувати кластери навмання обраними об'єктами.

Кожна ітерація алгоритму складається з двох кроків:

1) розподіл об'єктів за кластерами;

2) оновлення центроїдів.

На першому кроці кожному об'єкту $\mathbf{x}_i \in D$ призначається найближчий кластер C_{j^*}, де

$$j^* = \arg\min_{j=1,...,k} \left\| \mathbf{x}_i - \bar{\mathbf{x}}_j \right\|^2. \tag{10.5}$$

На другому кроці оновлюються значення середніх точок для кожного кластеру. Ці два кроки повторюються, доки центроїди не перестануть змінюватись між ітераціями (і, відповідно, стабілізується склад кожного кластеру). Наприклад, критерій збіжності алгоритму можна визначити як:

$$\sum_{j=1}^{k} \left\| \bar{\mathbf{x}}_j^t - \bar{\mathbf{x}}_j^{t-1} \right\|^2 \le \varepsilon, \tag{10.6}$$

де ε – поріг збіжності, а $\bar{\mathbf{x}}_j^t$ – значення середнього *j*–го кластеру на ітерації *t*.

Приклад 10.1. В файлі `grades.csv` наведено суму балів за 100–бальною шкалою, набрану студентами при вивченні одного з курсів, що викладав автор. Треба розбити їх на 4 кластери, які б умовно відповідали оцінкам «відмінно», «добре», «задовільно» і «незадовільно».

Результати знаходяться в діапазоні від 48 до 96 балів; їх розподіл показаний чорними точками на рис. 10.3. На першому кроці генеруємо 4

псевдовипадкових числа, рівномірно розподілених в діапазоні [48, 96]. Ними виявились 60, 72, 90 та 85. Результати розподілу студентів за цими чотирма кластерами показані кольором на рівні першої ітерації на рис. 10.3; ромбами виділено центроїди кластерів. Після початкового розподілу перераховуються середини для кожного із кластерів, які тепер становлять 51, 64, 79 и 91. Новий перерозподіл показаний на рівні другої ітерації. Третя ітерація нічого не змінює, збіжність досягнута. Фінальні результати: 48...54 – незадовільно, 60..70 – задовільно, 72..84 – добре, 86..96 – відмінно. ■

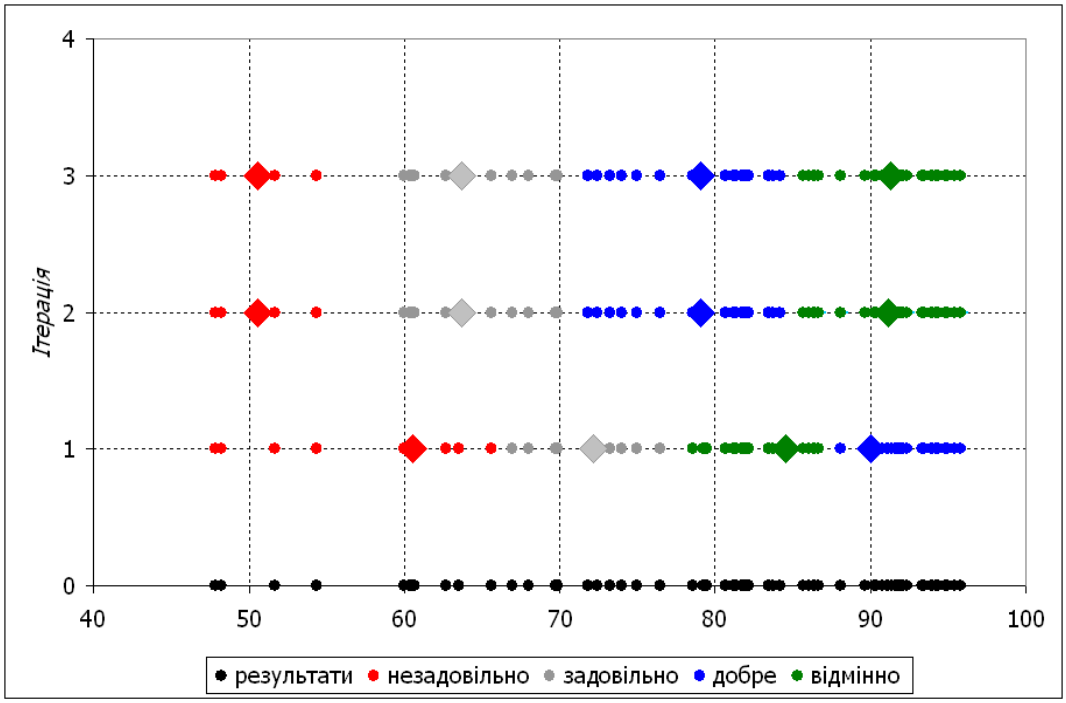


Рис. 10.3. Результати роботи алгоритму *k*–середніх при формуванні оцінок

Приклад 10.2. Розглянемо тепер застосування алгоритму до аналізу двовимірних даних на прикладі набору даних `iris.csv`, який вже неодноразово використовувався вище (пп. 7.3, 8.3). Нагадаємо, що цей набір даних містить по 50 спостережень для чотирьох кількісних характеристик трьох видів ірисів, *Iris setosa*, *Iris virginica* і *Iris versicolor*. Для кластеризації використаємо перші дві з них, довжину і ширину чашолистка. Діаграма розсіювання в цих двох координатах наведена на рис. 10.4*а*.

В ідеалі алгоритм кластеризації створив би три кластери, які точно відповідали би біологічним видам, але одного погляду на рис. 10.4а достатньо, щоб зрозуміти, що це нереально. В той час як *Iris setosa* візуально відокремлений від двох інших класів, два інших види досить сильно перемішані один з одним, принаймні за обраними характеристиками.

Для ініціалізації стартових кластерів скористуємось генератором псевдовипадкових чисел. Отримаємо такі центроїди початкових кластерів: $\bar{\mathbf{x}}_1 = (6{,}88; 3{,}20)$, $\bar{\mathbf{x}}_2 = (5{,}69; 3{,}87)$, $\bar{\mathbf{x}}_3 = (4{,}39; 2{,}54)$. Розподіл даних на першій ітерації проілюстрований на рис. 10.4*б*.

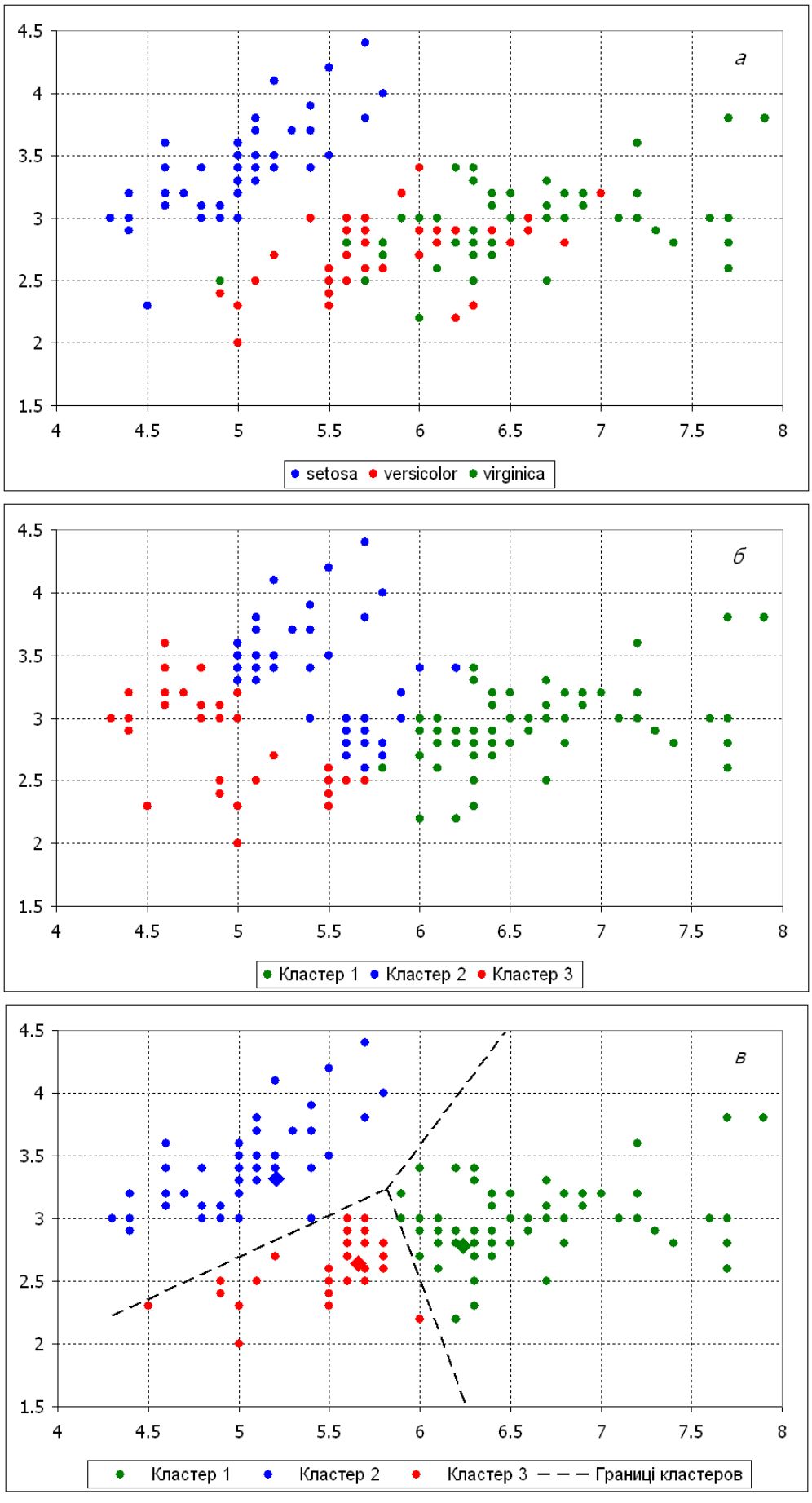


Рис. 10.4. Робота алгоритму *k*–середніх для кластеризації ірисів Фішера: *а* – початковий розподіл; *б* – після першої ітерації; *в* – фінальний розподіл

Алгоритм збігається на 9–й ітерації; фінальний розподіл показаний на рис. 10.4*в*. Ромбами показані центроїди сформованих кластерів: $\bar{\mathbf{x}}_1 = (6,24; 2,78)$, $\bar{\mathbf{x}}_2 = (5,21; 3,32)$, $\bar{\mathbf{x}}_3 = (5,66; 2,64)$.

Щоб надати геометричну інтерпретацію методу *k*–середніх, на рис. 10.4*в* штрих–пунктирними лініями показано границі між кластерами. Для визначення демаркаційної лінії між кластерами із центроїдами $\bar{\mathbf{x}}_i$ та $\bar{\mathbf{x}}_j$ треба знайти геометричне місце рівновіддалених від них точок, тобто вирішити рівняння $d(\bar{\mathbf{x}}_i, \mathbf{x}) = d(\mathbf{x}, \bar{\mathbf{x}}_j)$. Для двовимірних даних $\mathbf{x} = (x, y)$ це призводить до рівняння:

$$(x_i - x)^2 + (y_i - y)^2 = (x_j - x)^2 + (y_j - y)^2.$$

Звідси знаходимо, що границя є лінійною функцією $y = a_{ij} - b_{ij}x$, де:

$$a_{ij} = \frac{x_i^2 + y_i^2 - x_j^2 - y_j^2}{2(y_i - y_j)}; \; b_{ij} = \frac{x_i - x_j}{y_i - y_j}.$$

Для трьох кластерів отримаємо рівняння трьох прямих ліній, які ділять простір атрибутів на 3 області, як показано на рис. 10.4в. При більшій розмірності ми отримали б роздільні площини та гіперплощини. Такий спосіб візуалізації відомий як діаграма Вороного для центрів кластерів [15].

В таблиці 10.1 наводиться матриця помилок для результатів кластеризації, де для зручності кластери названі відповідно до домінуючої квітки.

Таблиця 10.1 – Матриця помилок для кластеризації ірисів Фішера

Справжній вид / Кластер	*virginica*	*setosa*	*versicolor*	Влучність, %
virginica	**43**	0	25	63
setosa	0	**49**	1	98
versicolor	7	1	**24**	75
Покриття, %	86	98	48	Точність: 77,3

Можна побачити, що алгоритм *k*–середніх майже ідеально сформував кластер для *Iris setosa*, що не дивно, враховуючи його відокремленість від інших видів. Для двох інших видів результати значно гірше, що теж можна було очікувати, бо за наявних даних більшість точок даних всередині діаграми розсіяння є спірними. Загальна точність алгоритму складає 77,3%, в той час як для МНЛ–класифікатора вона складала 82% (приклад 8.4). При цьому алгоритм ніяк не використовує інформацію про справжній біологічний вид об'єкта. ■

Алгоритм *k*–середніх не має чітко визначеного авторства. Він був незалежно запропонований різними авторами в різних контекстах протягом 1950х–1960х років. Широку відомість алгоритм отримав після публікації роботи Джеймса Мак-Квіна [34]. За даними [56], алгоритм *k*–середніх є другим за популярністю серед усіх методів ІАД.

10.3 Ієрархічні методи кластеризації

Метою ієрархічної кластеризації є створення послідовності вкладених групувань, які можна візуально відобразити у формі бінарного дерева. Таке представлення називають *дендрограмою* (англ. *dendrogram*). Кластери в ієрархії коливаються від дрібних до великих. Найнижчий рівень дерева (листя) складається з кожної точки в окремому кластері, тоді як найвищий рівень (корінь) – це кластер, який об'єднує всі точки. Кластеризація на цих екстремальних рівнях є тривіальною, а всередині дерева знаходяться значущі кластери. Якщо бажана кількість кластерів k задана наперед, то для знаходження шуканого розбиття треба обрати рівень, який містить k кластерів.

Більш формально, розбиття $A = \{A_1, ..., A_p\}$ називається *вкладеним* (англ. *nested*) у розбиття $B = \{B_1, ..., B_q\}$, якщо $p > q$ і для будь-якого кластеру $A_i \in A$ існує кластер $B_j \in B$ такий, що $A_i \subseteq B_j$. Результатом ієрархічної кластеризації є послідовність вкладених розбиттів $C_1 \subset C_2 \subset ... \subset C_n$, де $C_1 = \{\{\mathbf{x}_1\}, ..., \{\mathbf{x}_n\}\}$, а $C_n = \{\{\mathbf{x}_1, ..., \mathbf{x}_n\}\}$. Цю структуру відображає кластерна дендрограма (рис. 10.5).

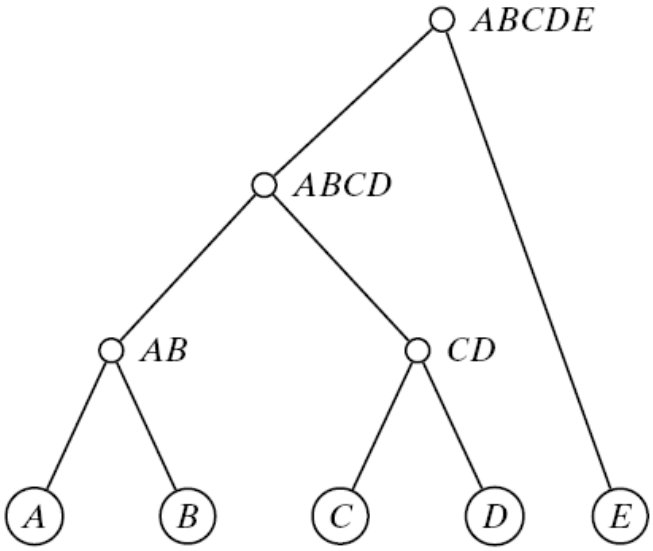


Рис. 10.5. Приклад кластерної дендрограми

Розглянемо об'єднувальну стратегію побудови кластерів у напрямку знизу нагору. Загальна схема цього методу доволі проста. Алгоритм починає роботу з розбиття C_1 з кожним об'єктом у власному кластері. Далі формується матриця відстаней між кластерами і обирається пара кластерів, які розташовані найближче один до одного. Ці кластери об'єднуються між собою, і процес повторюється, доки всі об'єкти не будуть об'єднані у кластер C_n, корінь дендрограми. Оскільки на кожному кроці кількість кластерів зменшується на одиницю, алгоритм триває точно n ітерацій. За бажанням алгоритм можна зупинити, коли залишиться потрібна кількість кластерів k.

Основним кроком алгоритму є визначення найближчої пари кластерів. Будемо вважати, що задана метрика $d(\mathbf{x},\mathbf{y})$ для відстані між об'єктами **x** та **y** (зазвичай евклідова відстань). Оскільки кожний кластер може містити декілька об'єктів, то нема однозначного способу вимірювання відстані між кластерами. Можна скористатись відстанню між центроїдами кластерів, як у методі *k*–середніх. Така метрика називається *середньою відстанню* (англ. *mean distance*). Однак було запропоновано багато інших способів. Розглянемо деякі з них докладніше.

1. Метод *найближчого сусіда* або *одиночного зв'язку* (англ. *single link*). Відстань між двома кластерами C_i, C_j визначається відстанню між двома найближчими об'єктами (найближчими сусідами) у цих кластерах:

$$d(C_i, C_j) = \min\{d(\mathbf{x},\mathbf{y}) \mid \mathbf{x} \in C_i, \mathbf{y} \in C_j\}. \tag{10.7}$$

Назва «одиночний зв'язок» мотивована тим спостереженням, що якщо з'єднати найближчі точки, то між кластерами буде існувати лише одна ланка, оскільки всі інші пари точок будуть більш віддалені між собою. Цей метод дозволяє виділяти кластери як завгодно складної форми, якщо різні частини таких кластерів з'єднані ланцюжками близьких один до одного елементів.

2. Метод *найдальшого сусіда* або *повного зв'язку* (англ. *complete link*). Тут відстань між кластерами C_i, C_j визначається найбільшою відстанню між об'єктами цих кластерів (тобто найбільш віддаленими сусідами):

$$d(C_i, C_j) = \max\{d(\mathbf{x},\mathbf{y}) \mid \mathbf{x} \in C_i, \mathbf{y} \in C_j\}. \tag{10.8}$$

Термін «повний зв'язок» передає той факт, що якщо з'єднати всі пари точок двох кластерів із відстанню не більше $d(C_i, C_j)$, то будуть з'єднані всі можливі пари.

3. Метод *групового середнього* (англ. *group average*). Відстань між двома кластерами визначається як середня відстань між усіма парами об'єктів у них:

$$d(C_i, C_j) = \frac{\sum_{\mathbf{x}\in C_i}\sum_{\mathbf{y}\in C_j} d(\mathbf{x},\mathbf{y})}{n\{C_i\}n\{C_j\}}. \tag{10.9}$$

4. Метод мінімальної дисперсії, або *метод Варда* (англ. *Ward's method*). Як відстань між кластерами береться приріст суми квадратів помилок (10.3) після об'єднання кластерів. Цей спосіб мотивований методами дисперсійного аналізу (п. 4.3). Сума квадратів помилок для кластера C_i дорівнює:

$$\begin{aligned} SSE(C_i) &= \sum_{\mathbf{x}\in C_i} \left\|\mathbf{x} - \bar{\mathbf{x}}_i\right\|^2 = \sum_{\mathbf{x}\in C_i} \mathbf{x}^T\mathbf{x} - 2\sum_{\mathbf{x}\in C_i} \mathbf{x}^T\bar{\mathbf{x}}_i + \sum_{\mathbf{x}\in C_i} \bar{\mathbf{x}}_i{}^t\bar{\mathbf{x}}_i = \\ &= \sum_{\mathbf{x}\in C_i} \mathbf{x}^T\mathbf{x} - n\{C_i\}\sum_{\mathbf{x}\in C_i} \bar{\mathbf{x}}_i{}^t\bar{\mathbf{x}}_i, \end{aligned} \tag{10.10}$$

а для всього розбиття $C = \{C_1, C_2, ..., C_m\}$ –

$$SSE(C) = \sum_{i=1}^{m} \sum_{\mathbf{x} \in C_i} \left\| \mathbf{x} - \overline{\mathbf{x}}_i \right\|^2 . \qquad (10.11)$$

Метрика Варда визначає відстань між кластерами C_i та C_j як зміну у значенні *SSE*(*C*) після об'єднання цих кластерів у $C_{ij} = C_i \bigcup C_j$:

$$d(C_i, C_j) = \Delta SSE_{ij} = SSE_{ij} - SSE_i - SSE_j . \qquad (10.12)$$

Якщо скористатись рівнянням (10.10) і врахувати, що центроїд об'єднаного кластеру $\overline{\mathbf{x}}_{ij} = \dfrac{n\{C_i\}\overline{\mathbf{x}}_i + n\{C_j\}\overline{\mathbf{x}}_j}{n\{C_i\} + n\{C_j\}}$, то після серії перетворень формулу (10.12) можна спростити до:

$$d(C_i, C_j) = \frac{n\{C_i\} n\{C_j\}}{n\{C_i\} + n\{C_j\}} \left\| \overline{\mathbf{x}}_i - \overline{\mathbf{x}}_j \right\| \qquad (10.13)$$

(виведення можна знайти в [57, с. 368–369]). Таким чином, відстань Варда є зваженою середньою відстанню, де ваговий коефіцієнт залежить від розмірів кластерів. Цей метод надає перевагу приєднанню малих кластерів до великих.

10.4 Алгоритм максимізації очікувань (EM–алгоритм)

Методи кластеризації, як і методи класифікації, можна поділити на ординарні, які «жорстко» прив'язують об'єкти до певного кластера, і імовірнісні, які оцінюють імовірність приналежності об'єкта кожному кластеру. Такі методи природно виходять для класу моделей суміші розподілів. В них вважається, що спостереження є реалізацією випадкової величини із деяким параметричним розподілом $F(\mathbf{x}, \boldsymbol{\theta})$, де $\boldsymbol{\theta}$, у свою чергу, є дискретною випадковою величиною, значення якої відповідають кластерам. Найбільш поширеним представником цього класу є модель *гаусової суміші* (англ. *Gaussian mixture*):

$$f(\mathbf{x}) = \sum_{i=1}^{k} f_i(\mathbf{x}) P(C_i) = \sum_{i=1}^{k} f(\mathbf{x} \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) P(C_i), \qquad (10.14)$$

де $f_i(\mathbf{x})$ – щільність нормального розподілу, пов'язаного з *i*–м кластером;

$\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i$ – математичне сподівання і коваріаційна матриця цього розподілу;

$P(C_i)$ – апріорні ймовірності приналежності точки **x** до *i*–го кластеру, які називають *параметрами суміші* (англ. *mixture parameters*). Вони мають задовольняти умову $\sum_{i=1}^{k} P(C_i) = 1$.

Множина параметрів цієї моделі $\boldsymbol{\theta} = \{\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1, P(C_1), ..., \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, P(C_1)\}$ має велику розмірність і їх оцінка стандартними методами математичної статистики є дуже складним завданням з обчислювальної точки зору. Замість цього Артуром Демпстером із колегами був запропонований ітеративний алгоритм *максимізації очікувань* (англ. *expectation-maximization algorithm*, або скорочено

EM– algorithm) [23]. Для спрощення викладок розглянемо цей алгоритм для одновимірного випадку $\mathbf{x} \in R$, коли спостереження обмежені одним атрибутом. Тоді $\mathbf{x} = x$, $\boldsymbol{\mu}_i = \mu_i$, $\boldsymbol{\Sigma}_i = \sigma_i^2$ і

$$f_i(x) = f(x \mid \mu_i, \sigma_i^2) = \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{(x-\mu_i)^2}{2\sigma_i^2} \right). \qquad (10.15)$$

На першому етапі роботи алгоритму обираються початкові значення параметрів. Зазвичай кластери вважаються рівноймовірними: $P(C_i) = 1/k$, μ_i обираються за допомогою генератора псевдовипадкових чисел із діапазону варіації *x*, а дисперсії приймаються рівними одиниці. Далі алгоритм ітеративно повторює два кроки: очікування і максимізацію.

1. Крок *очікування* (англ. *expectation*) полягає в тому, що за допомогою теореми Байєса (2.8) розраховуються апостеріорні ймовірності для кожного кластеру і для кожної точки даних:

$$P(C_i \mid x_j) = \frac{P(x_j \mid C_i)P(C_i)}{\sum_{g=1}^{k} P(x_j \mid C_g)P(C_g)} = \frac{f(x_j \mid \mu_i, \sigma_i^2)P(C_i)}{\sum_{g=1}^{k} f(x_j \mid \mu_g, \sigma_g^2)P(C_g)}. \qquad (10.16)$$

Позначимо $P(C_i \mid x_j)$ через w_{ij}. Ці числа можна трактувати як вагові коефіцієнти, які надають внесок *j*–го спостереження до *i*–го кластеру.

2. Крок *максимізації*[18] (англ. *maximization*) полягає у перерахунку нових значень внутрішньокластерних середніх, дисперсій і апріорних ймовірностей.

Оновлене значення центру кластера μ_i обчислюється як зважене середнє всіх точок даних:

$$\mu_i = \frac{\sum_{j=1}^{n} w_{ij} x_j}{\sum_{j=1}^{n} w_{ij}}. \qquad (10.17)$$

Аналогічно, оновлене значення дисперсії для кластера C_i обчислюється як зважена дисперсія для всіх точок даних:

$$\sigma_i^2 = \frac{\sum_{j=1}^{n} w_{ij} (x_j - \mu_i)^2}{\sum_{j=1}^{n} w_{ij}}. \qquad (10.18)$$

Нарешті, апріорна ймовірність кластера C_i переоцінюється як частка суми вагових коефіцієнтів, що належить до C_i:

$$P(C_i) = \frac{\sum_{j=1}^{n} w_{ij}}{\sum_{g=1}^{k} \sum_{j=1}^{n} w_{gj}} = \frac{\sum_{j=1}^{n} w_{ij}}{n}, \qquad (10.19)$$

[18] Назва кроку пояснюється тим, що використовувані на цьому кроці формули є оцінками параметрів за методом максимальної правдоподібності з математичної статистики [1]. В узагальненнях ЕМ–алгоритму для суміші інших, негаусових розподілів на цьому кроці відбувається максимізація функції правдоподібності.

де останнє перетворення пояснюється тим, що $\sum_{g=1}^{k} w_{gj} = \sum_{g=1}^{k} P(C_g \mid x_j) = 1.$

Ці два кроки повторюються, доки не буде досягнуто деяку умову збіжності; наприклад, доки зміна центрів кластерів між ітераціями $t-1$ та t не стане незначною:

$$\max \left| \mu_i^t - \mu_i^{t-1} \right| \le \varepsilon .$$

Приклад 10.3. Звернемось ще раз до набору даних «іриси Фішера» і скористуємось алгоритмом максимізації очікувань для їх кластеризації за атрибутом «довжина пелюстки». Як і у раніше розглянутих прикладах з цього набору даних, *I. Setosa* виявляється відокремленим від інших видів, в той час як для *I. Versicolor* and *I. Virginica* значення атрибуту перекриваються. Розподіл даних за видами ірисів подано у вигляді скупчення кольорових точок на горизонтальній осі графіків на рис. 10.6. Спробуємо розбити дані на три кластери, які в ідеалі відповідали б зазначеним біологічним видам.

Для ініціалізації алгоритму скористуємось генератором псевдовипадкових величин, щоб обрати значення центрів розподілу в діапазоні варіації атрибуту [1,0; 6,9]. В результаті отримуємо значення $\mu_1 = 3,25;\ \mu_2 = 1,59;\ \mu_3 = 4,52$. Представлення даних рівномірною сумішшю трьох нормальних розподілів з такими середніми і одиничною дисперсією наведено на рис. 10.6*а*.

Виходячи з цих значень, наступним кроком є оцінка ймовірності приналежності кожної із 150 точок даних до кожного з трьох кластерів за формулою (10.16). В результаті отримуємо вагові коефіцієнти $w_{ij}, i = 1,...,150;\quad j = 1,2,3$. За формулами (10.17)–(10.19) розраховуємо оновлені параметри суміші:

$\mu_1 = 3,25;\ \mu_2 = 1,59;\ \mu_3 = 4,52;$

$\sigma_1 = 1,40;\ \sigma_2 = 0,49;\ \sigma_3 = 0,82;$

$P(C_1) = 0,22; P(C_2) = 0,29;\ P(C_3) = 0,49.$

Нова модель даних зображена на рис. 10.6*б*.

Цей процес повторюється, доки середні не перестануть помітно змінюватися. Збіжність при $\varepsilon = 0,01$ досягається на 15–й ітерації. Остаточні параметри:

$\mu_1 = 4,64;\ \mu_2 = 1,46;\ \mu_3 = 4,97;$

$\sigma_1 = 0,71;\ \sigma_2 = 0,17;\ \sigma_3 = 0,84;$

$P(C_1) = 0,13; P(C_2) = 0,33;\ P(C_3) = 0,54.$

Результати кластеризації EM–алгоритмом наведені на рис. 10.6*в*. Границі кластерів показані вертикальними штрих–пунктирними лініями. Матриця помилок для результатів кластеризації наведена в табл. 10.2.

Алгоритм створив відокремлений, компактний кластер для *I. Setosa*. Кластери для двох інших видів ірисів виявилися дуже подібними; проте, *I. Virginica* правильно розпізнається у 98% випадків, а *I. Versicolor* – у 82%. Загальна точність алгоритму 93,3% є досить високою. ■

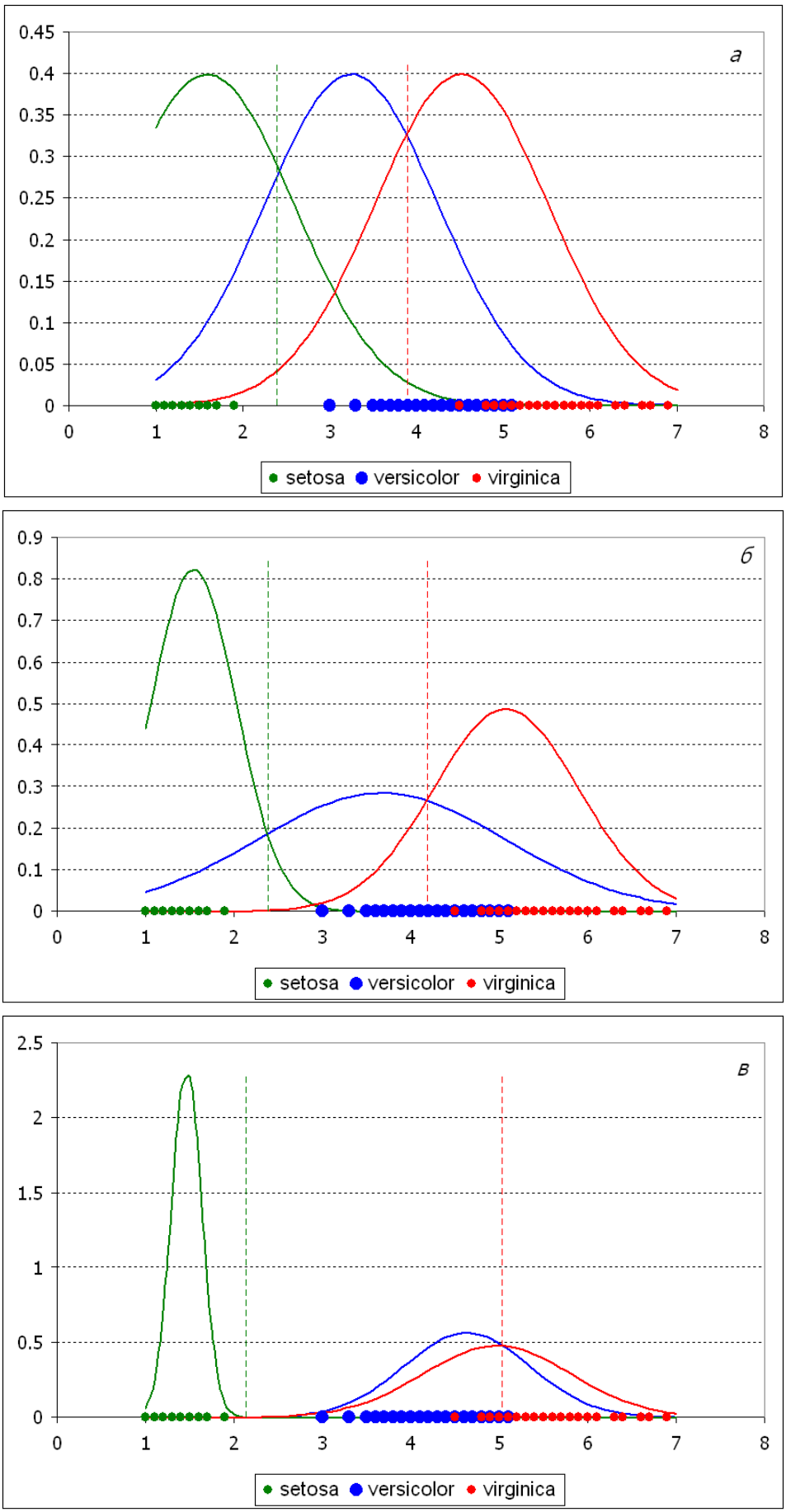


Рис. 10.6. Робота EM–алгоритму для кластеризації ірисів Фішера:
а – початковий розподіл; *б* – після першої ітерації; *в* – фінальний розподіл

Таблиця 10.2 – Матриця помилок для кластеризації ірисів ЕМ–алгоритмом

Справжній вид / Кластер	*virginica*	*setosa*	*versicolor*	Влучність, %
virginica	**49**	0	9	84,5
setosa	0	**50**	0	100
versicolor	1	0	**41**	97,6
Покриття, %	98	100	82	Точність: 93,3

У багатовимірному випадку з l атрибутами даних, формули (10.16), (10.17) та (10.19) залишаються чинними із заміною скалярного значення x_j на вектор $\mathbf{x}_j$. Найбільш суттєвим ускладненням є оцінка коваріаційних матриць $\mathbf{\Sigma}_i$. Їх елементи оцінюються як:

$$\sigma_{ab}^{i} = \frac{\sum_{j=1}^{n} w_{ij}(x_{ja} - \mu_{ia})(x_{jb} - \mu_{ib})}{\sum_{j=1}^{n} w_{ij}}, \; a, b = 1, ..., l \; . \qquad (10.20)$$

Кількість елементів кожної з коваріаційних матриць складає $l(l+1)/2$. Навіть при відносно невеликій розмірності $l = 10$ це вийде 55 параметрів для кожного кластера. Це доволі складно з обчислювальної точки зору, тому при великій кількості атрибутів часто вважають, що компоненти суміші є незалежними один від одного (як у НБК). У такому припущенні для коваріаційних матриць достатньо оцінити l діагональних елементів – дисперсій для кожного з атрибутів. Це значно спрощує задачу.

Рис. 10.7 ілюструє, як виглядає кластеризація двовимірних даних за допомогою алгоритму максимізації очікувань.

ЕМ–алгоритм можна застосовувати і для сумішей розподілів, відмінних від нормального, при відповідній модифікації формул (10.16) – (10.19).

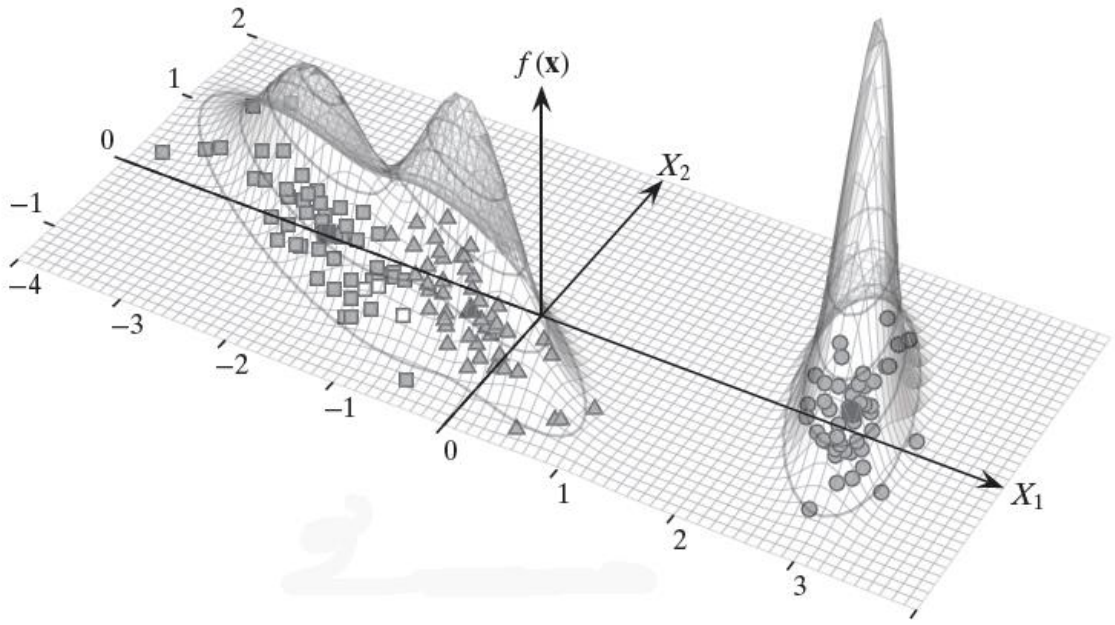


Рис. 10.7. Кластеризація двовимірних даних ЕМ–алгоритмом [57]

Алгоритм *k*–середніх можна розглядати як окремий випадок ЕМ–алгоритму при $P(\mathbf{x}_j \mid C_i) = I\left\{i = \arg\min_{j=1,\ldots,k} \left\|\mathbf{x}_i - \bar{\mathbf{x}}_j\right\|^2\right\}$. Тоді апостеріорна ймовірність $P(C_i \mid \mathbf{x}_j) = 1$, якщо $\mathbf{x}_j \in C_i$ і дорівнює нулю у всіх інших випадках.

10.5 Застосування кластерного аналізу

Кластерний аналіз має різноманітні застосування у різних галузях. За його допомогою можуть вирішуватись такі завдання, як:

– сегментація ринку у маркетингу;

– оцінка роботи персоналу в менеджменті;

– розпізнавання образів у штучному інтелекті;

– аналіз просторових даних в географії, геології, епідеміології тощо;

– аналіз уподобань споживачів у рекомендаційних системах;

– діагностика втручань у захисті комп'ютерних систем;

– класифікацію веб–документів та багато інших.

Кластерний аналіз можна використовувати як окремий інструмент інтелектуального аналізу даних, який допомагає отримати уявлення про розподіл даних, або як етап попередньої обробки даних для подальшого аналізу сформованих кластерів іншими методами ІАД. Деякі алгоритми кластеризації з успіхом застосовуються для виявлення викидів в даних.

Проте, процес кластерного аналізу є досить складним і неоднозначним. Він складається з декількох взаємопов'язаних етапів, і рішення, прийняті на кожному кроці можуть впливати на остаточні результати кластеризації. Важливими рішеннями при проведенні кластерного аналізу є:

– вибір характеристик, на основі яких проводиться кластеризація;

– вибір метрики та методу стандартизації вихідних даних;

– визначення кількості кластерів;

– визначення методу кластеризації та його налаштувань (таких як правила об'єднання кластерів в агломератному алгоритмі);

– оцінка та інтерпретація результатів кластеризації.

Оцінка результатів кластеризації є особливо складним завданням, оскільки «правильне» рішення апріорі невідомо. Методи кластеризації часто тестуються на задачах класифікації (як це було зроблено в прикладах 10.2–10.3), де відомо істинне групування об'єктів. Хороші результати говорять про працездатність алгоритму, але не є гарантією вірного рішення при аналізі нових даних.

Верифікація результатів кластеризації охоплює три основні завдання:

– оцінку якості кластеризації;

– оцінку стабільності кластеризації, тобто чутливості результатів до параметрів алгоритму та/або змін у вихідних даних;

– оцінку тенденції до кластеризації, тобто, чи насправді даним притаманна схильність до групування, яка б вимагала застосування кластерного аналізу.

Для кожного з вищезазначених завдань запропоновано низку показників, які можна поділити на три основні групи.

1. Зовнішні показники використовують інформацію, що відсутня в наборі даних. Вона може бути подана у формі апріорних уявлень, експертних оцінок чи будь-яких інших способів визначення «справжніх» кластерів. Це дає можливість використання всіх тих показників, які оцінюють якість класифікації (див. п. 6.2).

2. Внутрішні показники використовують критерії, які випливають із самих даних. Наприклад, внутрішньокластерні та міжкластерні відстані дозволяють судити про компактність кластерів (наскільки схожі точки в одному кластері) і їх розділення (наскільки далеко одна від одної знаходяться точки в різних кластерах). Так, *індекс Данна* (англ. *Dunn index*) визначається як відношення мінімальної відстані між парами точок з різних кластерів до максимальної відстані між парами точок одного кластеру [57]:

$$D = \frac{\min\limits_{1\le i\le j\le k}\{d(a,b)\,|\,x_a \in C_i, x_b \in C_j\}}{\max\limits_{i=1,\ldots,k}\{d(a,b)\,|\,x_a, x_b \in C_i\}}. \qquad (10.21)$$

Чим більший індекс Данна, тим краще кластеризація, оскільки найближча відстань між точками різних кластерів набагато перевищує найдальшу відстань між точками одного кластеру.

Іншим популярним показником є *силуетний коефіцієнт* (англ. *silhouette coefficient*). Для кожної точки даних можна визначити її «силует» як:

$$s_i = \frac{d_{out}(\mathbf{x}_i) - d_{in}(\mathbf{x}_i)}{\max\{d_{out}(\mathbf{x}_i), d_{in}(\mathbf{x}_i)\}}, \qquad (10.22)$$

де $d_{in}(\mathbf{x}_i)$ – середня відстань від $\mathbf{x}_i$ до точок в її власному кластері, а $d_{out}(\mathbf{x}_i)$ – середня відстань від $\mathbf{x}_i$ до точок у найближчому кластері. Значення цього показника лежить в інтервалі [−1,+1]. Значення, близьке до +1, означає, що $\mathbf{x}_i$ знаходиться набагато ближче до точок у власному кластері ніж до точок інших кластерів. Значення, близьке до нуля, означає, що $\mathbf{x}_i$ знаходиться близько до межі між двома кластерами. Нарешті, значення, близьке до −1, вказує на те, що $\mathbf{x}_i$ ближче до іншого кластера, ніж до свого власного, і тому точка може бути неправильно згрупована.

Силуетний коефіцієнт розраховується далі як середнє значення s_i за усіма точками даних:

$$SC = \frac{1}{n}\sum_{i=1}^{n} s_i\,. \qquad (10.23)$$

Значення, близьке до +1, є ознакою хорошої кластеризації.

3. Відносні показники спрямовані на пряме порівняння різних угрупувань, які виникають при зміні налаштувань алгоритму. Зокрема, відносні показники використовуються для обрання кількості кластерів *k*. Один із підходів до вирішення цієї задачі базується на силуетних коефіцієнтах. Для цього розраховуються значення загального силуетного коефіцієнта (10.23) та внутрішньокластерних силуетних коефіцієнтів $SC_i = \frac{1}{n\{C_i\}} \sum_{j \in C_i} s_j$ для різних значень *k*. Далі обирається значення *k*, яке призводить до найкращого групування із високими значеннями *SC* та SC_i ($i = 1,\ldots,k$).

Іншим можливим варіантом перевірки якості та робастності кластеризації є використання декількох методів та порівняння отриманих результатів. Відсутність подібності не обов'язково є свідченням некоректності результатів, але схожість результатів є ознакою хорошої кластеризації.

Контрольні запитання

1. В чому полягає задача кластеризації?
2. Що обумовлює велику кількість алгоритмів кластеризації?
3. Наведіть основні групи методів кластеризації.
4. Яким вимогам має задовольняти функція відстані між об'єктами?
5. В чому полягає критерій оптимальності в алгоритмі *k*–середніх?
6. Наведіть схему алгоритму *k*–середніх.
7. Як визначаються границі кластерів у методі *k*–середніх?
8. В чому полягає задача ієрархічної кластеризації?
9. Які підходи існують до вирішення задачі ієрархічної кластеризації?
10. Наведіть схему алгоритму ієрархічної кластеризації.
11. Назвіть основні різновиди способів визначення відстані між кластерами.
12. Як пов'язані між собою відстань Варда та евклідова відстань?
13. Як інтерпретується переконливість асоціативного правила?
14. Що мається на увазі під моделлю гаусової суміші?
15. Наведіть загальну схему EM–алгоритму.
16. Як визначаються апостеріорні ймовірності кластерів у методі максимізації очікувань?
17. В чому полягає етап максимізації в EM–алгоритмі?
18. Який зв'язок існує між EM–алгоритмом та методом *k*–середніх?
19. Які рішення слід прийняти під час здійснення кластерного аналізу?
20. В чому полягають основні завдання верифікації результатів кластеризації?
21. Назвіть групи показників, які використовуються при верифікації результатів кластеризації.

22. Що характеризує і як розраховується індекс Данна?

23. Як визначається і які значення приймає силуетний коефіцієнт для окремої точки даних?

24. Яким чином розраховується загальний силуетний коефіцієнт і що він говорить про якість кластеризації?

25. Як використовується силуетний коефіцієнт для визначення кількості кластерів?

Завдання для самостійної роботи

10.1. Наведіть два приклади практичних задач, вирішення яких вимагає кластеризації даних.

10.2. В файлі `gdplife.csv` стовпець GDP містить дані про валовий внутрішній продукт на душу населення в 2021 р. для 180 країн світу (в доларах США).

а) Впорядкуйте країни за зменшенням GDP. Знайдіть квартилі розподілу.

б) Визначте арифметичні середні для кожної з отриманих чотирьох груп.

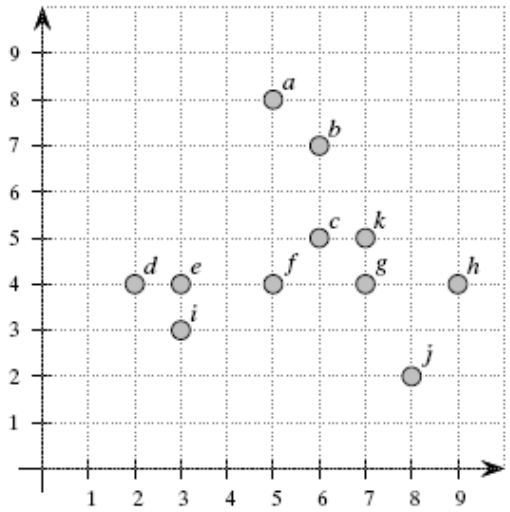
в) Розбийте країни на чотири кластери за рівнем доходів на душу населення за допомогою методу k–середніх. В якості початкових значень центроїдів використовуйте результати з попереднього пункту.

г) Порівняйте фінальну кластеризацію з початковою і дайте змістовну інтерпретацію отриманим результатам.

д) Відкиньте спостереження для чотирьох найбагатших країн світу і заново сформуйте кластери для модифікованого набору даних. Порівняйте результати кластеризації з тими, що були отримані у попередньому пункті. Зробіть висновки про робастність алгоритму k–середніх.

10.3*. Доведіть формулу (10.13).

10.4. За допомогою агломератного методу побудуйте дендрограму для набору зображених на рисунку нижче даних.



Налаштування:

– для визначення відстані між елементами використовуйте манхеттенську відстань;

– для визначення відстані між кластерами використовуйте метод найближчого сусіда;.

– при рівній відстані першими об’єднуйте ті кластери, що ближче до початку координат.

Зобразіть порядок об’єднання кластерів у вигляді дерева, доки не буде сформовано чотири кластери. Наведіть матрицю відстаней для кожного кроку алгоритму.

10.5. Покажіть, що при об’єднанні двох кластерів C_i та C_j в кластер C_{ij} за методом найближчого сусіда дистанція між новим кластером та усіма іншими буде задаватись формулою:

$$d(C_{ij},C_k)=\frac{1}{2}\Big(d(C_i,C_k)+d(C_j,C_k)\Big)-\frac{1}{2}\Big|d(C_i,C_k)-d(C_j,C_k)\Big|.$$

10.6. Покажіть, що при об’єднанні двох кластерів C_i та C_j в кластер C_{ij} за методом групового середнього дистанція між новим кластером та усіма іншими буде задаватись формулою:

$$d(C_{ij},C_k)=\frac{n_i}{n_i+n_j}d(C_i,C_k)+\frac{n_j}{n_i+n_j}d(C_j,C_k),$$

де n_i, n_j – кількість елементів в кластерах C_i та C_j.

10.7. Виконайте пункти в)–д) задачі 10.2 із використанням алгоритму максимізації очікувань. Для початкових значень внутрішньо–кластерних дисперсій використовуйте оцінки дисперсій для кожної з чотирьох груп, отриманих в п. 10.2а.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Грін В. Економетричний аналіз. (Green W. Econometric Analysis. New York: Macmillan, 2000) / Переклад і наукова редакція О.В.Комашка, передмова О.І.Черняка. Київ : Основи, 2005.

2. Марченко О. О., Россада Т. В. Актуальні проблеми Data Mining: Навчальний посібник для студентів факультету комп'ютерних наук та кібернетики. Київ : КНУ ім. Тараса Шевченка, 2017. 150 с.

3. Математична статистика : навч. посіб. [Електронне видання] / С. М. Григулич та ін. Київ : КНЕУ, 2015. 203 с. URL: http://ir.kneu.edu.ua/bitstream/handle/2010/17627/mat_stat.pdf.

4. Огірко О. І., Галайко Н. В. Теорія ймовірностей та математична статистика: навчальий посібник. Львів : ЛьвДУВС, 2017. 292 с.

5. Олійник А. О. Субботін С. О., Олійник О. О. Інтелектуальний аналіз даних : навчальний посібник. Запоріжжя : ЗНТУ, 2012. 278 с.

6. Ординська З. П., Орловський І. В., Руновська М. К. Конспект лекцій з аналітичної геометрії та лінійної алгебри. Київ : НТУ «КПІ», 2014. 176 с.

7. Ситник В. Ф., Краснюк М. Т. Інтелектуальний аналіз даних (дейтамайнінг) : Навч. посібник. Київ : КНЕУ, 2007. 376 с.

8. Слюсарчук П. В. Теорія ймовірностей та математична статистика. Ужгород : Вид-во «Карпати», 2005. 178 с.

9. Теорія ймовірностей. [Електронний ресурс]: навч. посіб. для студ. спеціальності 121 «Інженерія програмного забезпечення» / КПІ ім. Ігоря Сікорського; уклад.: Барабаш О. В., Мусієнко А. П., Свинчук О. В. Електронні текстові дані (1 файл: 3705 Кбайт). Київ: КПІ ім. Ігоря Сікорського, 2021. 193 с.

10. Черняк О. І., Захарченко П. В. Інтелектуальний аналіз даних: Підручник. Київ : Знання, 2014. 599с.

11. Agrawal, R., Imielinski, T., Swami, A. Mining association rules between sets of items in large databases. 1993. Proceedings of the ACM SIGMOD International Conference on Management of Data. ACM, 207–216.

12. Agrawal, R., Srikant, R. Fast algorithms for mining association rules. 1994. Proceedings of the 20th Int. Conference on Very Large Data Bases, 487–499.

13. AgriMetSoft. Multinomial Logistic Regression Calculator. 2019. URL: https://agrimetsoft.com/regressions/Multinomial-Logistic.

14. Altman, N. S. An introduction to kernel and nearest-neighbor nonparametric regression. 1992. *The American Statistician* 46 (3), 175–185. DOI: https://doi.org/10.1080/00031305.1992.10475879.

15. Aurenhammer F, Klein R. Voronoi Diagrams. *Handbook of computational geometry*, 2000. Ch.5(10), 201–290.

16. Azevedo, P. J., Jorge, A. M. Comparing Rule Measures for Predictive Association Rules. Machine Learning: ECML 2007. Lecture Notes in Computer

Science, vol. 4701. Berlin, Heidelberg: Springer, 2007. DOI: https://doi.org/10.1007/978-3-540-74958-5_47.

17. Boser, B. E.; Guyon, I. M., Vapnik, V. N. A training algorithm for optimal margin classifiers. 1992. Proceedings of the fifth annual workshop on Computational learning theory – COLT '92. DOI: https://doi.org/10.1145/130385.130401.

18. Brin, S. et al. Dynamic itemset counting and implication rules for market basket data. 1997. Proceedings of the 1997 ACM SIGMOD International Conference on Management of Data, 255–264.

19. Charles, V., Gherman, T., Paliza, J.C. The Gini Index: A Modern Measure of Inequality. In: Charles, V., Emrouznejad, A. (eds) Modern Indices for International Economic Diplomacy. Palgrave Macmillan, 2022. DOI. https://doi.org/10.1007/978-3-030-84535-3_3.

20. Chicco, D., Jurman, G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. 2020. *BMC Genomics* 21 (6), 1–13. DOI: https://doi.org/10.1186/s12864-019-6413-7.

21. Classification and Regression Trees / Breiman, L., Friedman, J., Olshen R., Stone, C. Belmont, CA : Wadsworth, 1984.

22. Cortes, C., Vapnik, V. Support-vector networks. 1995. *Machine Learning* 20 (3), 273–297. DOI: https://doi.org/10.1007/BF00994018.

23. Dempster, A. P., Laird, N. M., Rubin, D. B. Maximum likelihood from incomplete data via the EM algorithm. 1997. *Journal of the Royal Statistical Society*, Series B, 39 (1), 1–38.

24. Diapers, Beer, and data science in retail. URL: https://canworksmart.com/diapers-beer-retail-predictive-analytics (дата звернення 22.07.2023).

25. Greene, William H. Econometric Analysis. 8th edition. London: Pearson, 2017. 1176 p.

26. Hájek, Alan. Interpretations of Probability. *The Stanford Encyclopedia of Philosophy* (2019 Edition), Edward N. Zalta (ed.). URL: https://plato.stanford.edu/archives/fall2019/entries/probability-interpret.

27. Han, J., Kamber, M. Data Mining: Concepts and Techniques. Elsevier, 2006. 743 p.

28. Han, J., Pei, J., Yin, Y. Mining frequent patterns without candidate generation. 2000. Proceedings of ACM SIGMOD international conference on management of data, 1–12.

29. Hand, D. J., Yu, K. Idiot's Bayes - not so stupid after all? 2001. *International statistical review* 69(3), 385–398.

30. Hovland, C. I. Computer simulation of thinking. 1960. *American Psychologist*, 15(11), 687–693.

31. Hunt, E. B., Marin, J., Stone, P. J. Experiments in Induction. New York : Academic Press, 1966. ISBN 978-0-12-362350-8.

32. Lever, J., Krzywinski, M., Altman, N. Classification evaluation. 2016. *Nat. Methods* 13, 541–542.

33. Li, J., Tong, X. Statistical Hypothesis Testing versus Machine Learning Binary Classification: Distinctions and Guidelines. 2020. *Patterns* 1. 100115. P. 1–10. DOI: https://doi.org/10.1016/j.patter.2020.100115.

34. MacQueen, J. B. Some Methods for Classification and Analysis of Multivariate Observations. 1967. Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability. Vol. 1., 281–297.

35. McFadden, D. L. Conditional Logit Analysis of Qualititative Choice Behavior. Frontiers in Econometrics. New York : Academic Press, 1974. Pp. 105–142.

36. Melnikov, O. Demand for Differentiated Durable Products: The Case of the U.S. Computer Printer Market. 2013. *Economic Inquiry* 51(2), 1277–1298.

37. Mitchell, Tom M. Machine Learning. McGraw Hill, 1997. 432p.

38. Nocedal, J., Wright, S. J. Numerical Optimization. Springer–Verlag, 1999.

39. Numerical recipes in C: the art of scientific computing / Press, W. H. et al. Cambridge University Press, 1992. 994 pp.

40. Ozdemir, S., Kakade, S., Tibaldeshi, M. Principles of Data Science. 2nd edition. Birmingham-Mumbai : Packt Publishing, 2018. 420 p.

41. Phillips, A. W. H. The Relation Between Unemployment and the Rate of Change of Money Wage Rates in the United Kingdom, 1861–1957. 1958. *Economica* 25 (2), 283–299.

42. Quinlan, J. R. Induction of decision trees. 1986. *Machine Learning* 1(1), 81–106.

43. Quinlan, J. R. C4.5: Programs for Machine learning. Morgan Kaufmann Publishers, 1993.

44. Quinlan, J. R. Improved use of continuous attributes in c4.5. 1996. *Journal of Artificial Intelligence Research* 4, 77–90.

45. Ross, Sheldon M. Applied probability models with optimization applications. New York: Dover Publications, 1992. 198 pp.

46. Sachs, L. Applied Statistics: A Handbook of Techniques. Springer Series in Statistics, 1984. 707 p.

47. Scott, D. Box-Cox Transformations. URL: https://onlinestatbook.com/2/transformations/box-cox.html (дата звернення 22.07. 2023).

48. Sopranzetti, B. J. Hedonic regression models. 2015. Handbook of financial econometrics and statistics, 2119–2134.

49. Srikant, R. and Agrawal, R. Mining generalized association rules. 1995. Proceedings of the 21st VLDB conference. 407–419.

50. Srikant, R., Agrawal, R. Mining quantitative association rules in large relational tables. ACM SIGMOD Record, 1996, Vol. 25(2), 1–12. DOI: https://doi.org/10.1145/235968.233311

51. Stevens, S. S. On the Theory of Scales of Measurement. 1946. *Science* 103 (2684), 677–680.

52. Swersky, K. Support Vector Machines vs Logistic Regression. URL: https://www.cs.toronto.edu/~kswersky/wp-content/uploads/svm_vs_lr.pdf (дата звернення: 2.09.2023).

53. Tharwat, A. Classification assessment methods. 2018. *Applied Computing and Informatics*, 168–191. DOI: https://doi.org/10.1016/j.aci.2018.08.003.

54. Vuleta, B. How Much Data Is Created Every Day? SeedScientific, 2021. URL: https://seedscientific.com/how-much-data-is-created-every-day (дата звернення: 11.09.2023).

55. Weisstein, E. W. Stirling Number of the Second Kind. URL: https:// mathworld.wolfram.com/StirlingNumberoftheSecondKind.html (дата звернення: 30.08.2023).

56. Wu, X. et al. Top 10 algorithms in data mining. 2008. *Knowl. Inf. Syst.* 14, 1–37. DOI: https://doi.org/10.1007/s10115-007-0114-2

57. Zaki M. J., Meira W. Jr. Data Mining and Analysis: Fundamental Concepts and Algorithms. Cambridge University Press, 2014. 595 p.

58. Zaki, M. J. et al. New algorithms for fast discovery of association rules. 1997. Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining, 283–286.

СПИСОК ДЖЕРЕЛ ДАНИХ

Д1. Державна служба статистики України. URL: https://ukrstat.gov.ua.

Д2. Протокол Центральної виборчої комісії про результати повторного голосування з виборів Президента України 21.04.2019 р. URL: https://www.cvk.gov.ua/wp-content/uploads/2019/11/vpu_2019_protokol_cvk_30042019.pdf (дата звернення: 12.01.2023).

Д3. Які імена найпопулярніші в Україні. ТОП-60 найпоширеніших імен. 12.02.22. URL: https://www.kf-ks.info/yaki-imena-naypopulyarnishi-v-ukrayini-top-60-nayposhirenishih-imen (дата звернення: 23.01.2023).

Д4. Errera, R. The World Of Printers – An Overview of The Global Printer Market. Posted 01/17/2023. URL: https://www.tonerbuzz.com/blog/printer-market. (дата звернення: 24.07.2023).

Д5. Federal Reserve Economic Data. Consumer Price Index for All Urban Consumers: URL: https://fred.stlouisfed.org/series/CPIAUCSL (дата звернення: 21.06.2023).

Д6. International Monetary Fund. World Economic Outlook Database, April 2023. https://www.imf.org/en/Publications/WEO/weo-database/2023/April. (дата звернення: 14.04.2023).

Д7. Iris flower data set. URL: https://en.wikipedia.org/wiki/Iris_flower_data_set (дата звернення: 27.02.2023).

Д8. Seattle Central College. Quantitative Environmental Learning Project. Data Set #057. URL: https://seattlecentral.edu/qelp/sets/057/057.html (дата звернення: 12.07.2023).

Д9. Sokal, R.R., Hunter, P.E. A morphometric analysis of DDT-resistant and non-resistant housefly strains. *Ann. Entomol. Soc. Amer.* 1955, 48, 499–507.

Д10. The World Bank Group. Life expectancy at birth, total. URL: https://data.worldbank.org/indicator/SP.DYN.LE00.IN (дата звернення: 6.07.2023).

Д11. Trading Economics. URL: https://tradingeconomics.com (дата звернення: 25.06.2023).

Д12. World Health Organization. Global status report on alcohol and health. URL: https://apps.who.int/iris/bitstream/handle/10665/274603/9789241565639-eng.pdf (дата звернення: 26.07.2023).

ПРЕДМЕТНИЙ ПОКАЖЧИК

Навчальне видання

МЕЛЬНИКОВ Олег Станіславович

ІНТЕЛЕКТУАЛЬНИЙ АНАЛІЗ ДАНИХ

Навчально-методичний посібник
для студентів другого (магістерського) рівня підготовки
спеціальностей 122 – Комп’ютерні науки, 124 – Системний аналіз

Відповідальний за випуск проф. Дорофєєв Ю. І.
Роботу до видання рекомендував проф. Безменов М. І.
В авторській редакції

План 2023 р., поз. 129

Підп. до друку 5.12.2023 р. Гарнітура Times New Roman.
Ум. друк. арк. 13,72.

Видавничий центр НТУ «ХПІ».
Свідоцтво про державну реєстрацію ДК № 5478 від 21.08.2017 р.
61002, Харків, вул. Кирпичова, 2

Електронне видання