1. Коротко про основи

Аби краще зрозуміти, як працюють алгоритми науки про дані, необхідно розпочати з основ. Саме тому це найдовший розділ у книжці. Він удвічі довший за інші, які безпосередньо присвячені алгоритмам. Проте завдяки цьому вступу ви отримаєте ґрунтовне уявлення про фундаментальні кроки, які присутні майже в усіх дослідженнях у галузі науки про дані. Ці базові процеси допоможуть вам оцінити контекст, а також обмеження, які виникають у процесі вибору відповідних алгоритмів для використання в дослідженнях.

Існують чотири ключові етапи дослідження в галузі науки про дані. Передусім дані мають бути оброблені та підготовлені до аналізу. Потім на основі вимог нашого дослідження складається короткий список відповідних алгоритмів. Після цього параметри алгоритмів мають бути налаштовані для оптимізації результатів. Нарешті, усе це завершується побудовою моделей, які згодом порівнюють для вибору найкращої.

1.1. Підготовка даних

Наука про дані пов'язана з даними. Якщо їхня якість низька, навіть найскладніший аналіз дасть лише непереконливі

результати. У цьому розділі ми розглянемо основні формати даних, які зазвичай використовуються в аналізі, а також методи оброблення даних для поліпшення результатів.

Формат даних

Аби представити дані для аналізу, найчастіше використовують табличну форму (табл. 1). Кожен рядок є *елементом даних*, що описує окреме спостереження, а кожен стовпчик — змінною, що описує елемент даних. Змінні також відомі під назвами *атрибути, ознаки* або *розмірності*.

Змінні →

Елементи даних ↓

Номер покупки	Біологічний вид покупця	Дата	Фрукти	Риба	Загальна сума покупок, $
1	пінгвін	1 січня	1	так	5,3
2	ведмідь	1 січня	4	так	9,7
3	кролик	1 січня	6	ні	6,5
4	кінь	2 січня	6	ні	5,5
5	пінгвін	2 січня	2	так	6
6	жирафа	3 січня	5	ні	4,8
7	кролик	3 січня	8	ні	7,6
8	кіт	3 січня	?	так	7,4

Таблиця 1. Уявний набір даних про купівлю тваринами продуктів у супермаркеті. Кожен рядок — це купівля, а кожен стовпчик містить інформацію про неї

Маючи різну мету, ми можемо змінювати тип спостережень, представлених у кожному рядку. Наприклад, вибірка в таблиці 1 дає нам змогу вивчати закономірності в ряді трансак-

цій. Проте якщо ми захочемо вивчити закономірності купівлі залежно від дня, нам потрібно представити кожен рядок як сукупність трансакцій за день. Для більш комплексного аналізу ми також можемо додати нові змінні, як-от погоду (табл. 2).

Змінні

Дата	Виторг, $	Кількість покупців	Погода	Вихідні
1 січня	21,5	3	сонячна	так
2 січня	11,5	2	дощить	ні
3 січня	19,8	3	сонячна	ні

Таблиця 2. Переформатований набір даних, що показує загальні щоденні трансакції, з додатковими змінними

Типи змінних

Існують чотири основні типи змінних. Важливо розрізняти їх, аби бути впевненими, що вони придатні для вибраних нами алгоритмів.

- **Двійкова.** Це найпростіший тип змінних, який має лише два можливих значення. У таблиці 1 двійкова змінна використовується для того, аби вказати, чи купували покупці рибу.
- **Категоріальна.** Коли існує понад два варіанти, інформацію можна представити за допомогою категоріальної змінної. У таблиці 1 категоріальна змінна використовується для опису типів покупців.
- **Цілочислова.** Застосовується, коли інформацію можна представити цілим числом. У таблиці 1 така змінна використовується для позначення кількості фруктів, придбаних кожним покупцем.
- **Неперервна.** Це найбільш детальна змінна, що представляє числа з десятковими знаками. У таблиці 1 неперервна змінна використовується для позначення суми, витраченої кожним покупцем.

Вибір змінних

Хоча на першому етапі нам можуть надати набір даних, який містить багато змінних, занадто велика кількість змінних в алгоритмі здатна призвести до повільних обчислень або хибних прогнозів через надмірний інформаційний шум. Отже, нам потрібно скласти короткий список важливих змінних.

Вибір змінних часто є процесом спроб і помилок, коли змінні додають і прибирають з огляду на отримані результати. Для початку ми можемо використовувати прості графіки для вивчення кореляції (див. підрозділ 6.5) між змінними, добираючи найбільш перспективні з них для подальшого аналізу.

Функціональна інженерія

Проте іноді найкращі змінні потрібно сконструювати. Наприклад, маючи бажання передбачити, які споживачі з таблиці 1 не купуватимуть риби, ми могли би подивитися на змінну типу споживача, аби визначити, що кролики, коні та жирафи цього не робитимуть. Проте якби ми згрупували типи споживачів у ширші категорії травоїдних, м'ясоїдних і всеїдних, то мали б змогу дійти більш узагальненого висновку: травоїдні тварини не купують риби.

Крім переформатування однієї змінної, ми також могли б об'єднати кілька змінних за допомогою методу, відомого як зменшення розмірності. Його ми розглянемо в Розділі 3. Зменшення розмірності можна використати для виокремлення найбільш корисної інформації та зведення її в новий, але менший набір змінних для аналізу.

Неповні дані

Не завжди вдається зібрати повні дані. У таблиці 1, наприклад, не було зафіксовано кількості фруктів, придбаних під час останньої купівлі. Відсутність даних здатна завадити аналізу, тому їх слід, за можливості, компенсувати одним із наведених нижче способів.

- **Наближення.** Якщо відсутнє значення належить до двійкового чи категоріального типу змінних, його можна замінити модою (тобто найпоширенішим значенням) цієї змінної. Для цілочислових і неперервних значень можна використати медіану. Застосувавши цей метод до таблиці 1, ми матимемо змогу передбачити, що кіт придбав 5 фруктів, оскільки згідно з іншими 7 записами саме такою є середня кількість куплених фруктів.
- **Обчислення.** Відсутні значення також можна обчислити за допомогою більш досконалих алгоритмів під час навчання з учителем (про це йтиметься в наступному підрозділі). Хоча такі обчислення потребують більше часу, зазвичай вони є точнішими, оскільки алгоритми оцінюють відсутні значення на основі подібних закупівель — на відміну від методу апроксимації, який перевіряє кожний випадок купівлі. З таблиці 1 ми бачимо, що клієнти, які купували рибу, зазвичай придбавали менше фруктів, а отже, за нашими оцінками, кіт мав купити лише два чи три фрукти.
- **Видалення.** У крайньому разі рядки з відсутніми значеннями можна видалити. Проте зазвичай такого уникають, оскільки це зменшує кількість даних, доступних для аналізу. Крім того, видалення елементів даних може призвести до того, що отримана вибірка даних буде зміщена в той чи інший бік відносно певних груп. Наприклад, коти можуть менш охоче розголошувати кількість фруктів, які вони купують, і якщо ми видалимо клієнтів із незареєстрованими трансакціями щодо фруктів, коти будуть недостатньо представлені в нашій остаточній вибірці.

Після того як набір даних буде оброблено, настане час його проаналізувати.

1.2. Відбір алгоритму

У цій книжці ми розглянемо понад десять різних алгоритмів, які можна використовувати для аналізу даних. Вибір алгоритму залежить від типу завдання, яке ми хочемо виконати. Існують лише три основні категорії алгоритмів. У таблиці 3 перераховано алгоритми, які буде розглянуто в цій книжці, а також пов'язані з ними категорії.

Навчання без учителя

Завдання: *Знайти закономірності, що існують у даних.*

Коли ми бажаємо знайти приховані закономірності в нашому наборі даних, то можемо використовувати алгоритми

навчання без учителя. Ці алгоритми є неконтрольованими, оскільки ми не знаємо, на які закономірності слід звертати увагу, тож залишаємо їх на розгляд алгоритму.

	Алгоритм
Навчання без учителя	Кластеризація методом *k*-середніх Метод головних компонент Асоціативні правила Аналіз соціальних мереж
Навчання з учителем	Регресійний аналіз Метод *k*-найближчих сусідів Метод опорних векторів Дерево ухвалення рішень Метод випадкових лісів Нейронні мережі
Навчання з підкріпленням	Багаторукі бандити

Таблиця 3. Алгоритми та відповідні їм категорії

У таблиці 1 таку модель можна застосувати, щоби дізнатися, які товари часто купують разом (використовуючи асоціативні правила, описані в Розділі 4), або згрупувати клієнтів за їхніми покупками (описано в Розділі 2).

Ми маємо змогу перевірити результати моделі, побудованої за допомогою навчання без учителя, непрямими методами, наприклад, з'ясувавши, чи відповідають створені кластери покупців знайомим категоріям (травоїдні та м'ясоїдні тварини).

Навчання з учителем

Завдання: *Використати закономірності в даних для прогнозування.*

Коли ми хочемо зробити прогноз, можна використовувати алгоритми навчання з учителем. Ці алгоритми є керованими, оскільки ми хочемо, щоби вони базували свої прогнози на вже наявних закономірностях.

У таблиці 1 така модель здатна навчитися прогнозувати кількість фруктів, які придбає покупець *(передбачення)*, на основі його типу й того, чи купував він рибу *(змінні-предиктори)*.

Можна безпосередньо оцінити точність цієї моделі, якщо ввести значення для типів майбутніх покупців та їхньої схильності купувати рибу, а потім перевірити, наскільки близькими є передбачення моделі до фактичної кількості куплених фруктів.

Коли ми прогнозуємо цілочислові чи неперервні значення, як-от кількість куплених фруктів, то розв'язуємо проблему *регресії* (див. рис. 1а, с. 24), а коли двійкове чи категоріальне значення, наприклад, чи піде дощ,— проблему класифікації (див. рис. 1б, с. 24). Проте більшість алгоритмів класифікації також здатні генерувати прогнози у формі неперервного значення ймовірності, як-от у твердженнях на кшталт «імовірність дощу становить 75 %», що дає змогу робити прогнози з більшою точністю.