Опануй числа! Наука про дані для нефахівців

Анналін Нг

ОЗНАЙОМЧИЙ ФРАГМЕНТ

Переклад — Олег Буйвол

Чому наука про дані?

Уявіть себе молодим лікарем.

До вас звертається пацієнт зі скаргами на задишку, біль у грудях і періодичну печію. Ви його обстежуєте. Виявляється, його кров'яний тиск і частота серцевих скорочень у нормі, а ще в нього не було жодних попередніх захворювань.

Під час огляду ви виявляєте, що пацієнт повненький. Оскільки його симптоми характерні для людей з надмірною вагою, ви запевняєте його, що все під контролем, і рекомендуєте знайти час для фізичних вправ.

Подібна ситуація занадто часто призводить до встановлення хибного діагнозу, оскільки пацієнти із серцево-судинними захворюваннями мають симптоми, схожі на ознаки звичайного ожиріння. Лікарі часто не проводять подальшого обстеження, яке могло би виявити більш серйозне захворювання.

Судження кожної людини обмежені її суб'єктивним досвідом і неповнотою знань. Це погіршує процес ухвалення рішень і може, як у випадку з недосвідченим лікарем, завадити подальшому обстеженню, яке здатне допомогти дійти більш точних висновків.

Саме тут стане в пригоді наука про дані.

Замість покладатися на судження однієї людини, методи цієї науки дають нам змогу використовувати інформацію з більшої кількості джерел даних для ухвалення кращих рішень. Наприклад, ми в змозі перевірити історії хвороби пацієнтів зі схожими симптомами, аби на основі цієї статистики виявити можливі діагнози, на які раніше не звертали уваги.

Завдяки сучасним обчислювальним технологіям і передовим алгоритмам ми здатні:

• виявляти приховані тенденції у великих масивах даних;

• використовувати тенденції для прогнозування;

• обчислювати ймовірність кожного можливого результату;

• швидко отримувати точні результати.

Цю книжку написано простою мовою (без складної математики) як легкий вступ до науки про дані та її алгоритмів. Аби допомогти вам зрозуміти ключові поняття, ми використовуємо інтуїтивно зрозумілі пояснення та багато візуальних зображень.

Кожному алгоритму присвячено окремий розділ, у якому наведено приклад з реального життя, що пояснює, як працює алгоритм. Дані, використані для цих прикладів, доступні в інтернеті, а їхні джерела вказано наприкінці книжки.

Якщо вам потрібно повторити вивчений матеріал, зверніться до висновків після кожного розділу. Наприкінці книжки ви також знайдете зручні довідкові таблиці, у яких порівнюються переваги та недоліки кожного алгоритму. Також видання містить глосарій загальновживаних термінів.

За допомогою цієї книжки ми сподіваємося дати вам практичне розуміння науки про дані, аби ви могли скористатися її перевагами для ухвалення кращих рішень.

Тож почнімо.

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.

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 трансакцій. 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).

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.

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. 1, а), а коли двійкове чи категоріальне значення, наприклад, чи піде дощ,— проблему класифікації (див. рис. 1, б). 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.

Рис. 1. Регресія передбачає отримання лінії тренду, тоді як класифікація — розподіл елементів даних на групи. Зауважте, що в обох завданнях можна очікувати на помилки. У разі регресії елементи даних відхилятимуться від лінії тренду, тоді як у випадку класифікації вони потраплятимуть до неправильних категорій

Навчання з підкріпленням

Завдання: Використовувати закономірності в даних для прогнозування та поліпшення цих прогнозів відповідно до надходження нових результатів.

На відміну від навчання з / без учителя, де моделі вивчають, а потім застосовують без подальших змін, модель навчання з підкріпленням постійно вдосконалюють, використовуючи інформацію щодо вже отриманих результатів.

Перейдімо від таблиці 1 до реального прикладу: уявіть, що ми порівнюємо ефективність двох рекламних оголошень в інтернеті. Спочатку ми могли би показувати кожне оголошення однаково часто, оцінюючи кількість людей, які натиснули на кожне з них. Модель навчання з підкріпленням сприйматиме це число як відгук про популярність реклами, використовуючи його для налаштування показу на користь більш популярного оголошення. Завдяки цьому ітеративному процесу модель здатна зрештою навчитися показувати лише найкращі оголошення.

Інші фактори

Крім основних завдань, які вони виконують, алгоритми також відрізняються іншими аспектами, як-от здатність аналізувати різні типи даних або характер згенерованих результатів. Ці аспекти розглядаються в наступних розділах, присвячених кожному алгоритму, а також наведені у зведених таблицях Додатка А (навчання без учителя) і Додатка Б (навчання з учителем).

1.3. Налаштування параметрів

Безумовно, численні алгоритми, доступні в науці про дані, призводять до виникнення величезної кількості потенційних моделей, які ми можемо побудувати. Але навіть один алгоритм здатний генерувати різні результати — залежно від того, які параметри встановлено.

Параметри — це опції, які використовуються для налаштування алгоритму. Це подібно до налаштування радіо на потрібний частотний канал. Різні алгоритми мають різні параметри, доступні для налаштування. Загальні параметри, пов’язані з алгоритмами, що розглядаються в цій книжці, наведено в Додатку В.

Зрозуміло, що точність моделі страждає, коли її параметри не налаштовано належно. Погляньте на рис. 2, щоби побачити, як один алгоритм класифікації може генерувати кілька меж для розділення сірих і чорних точок.

Рис. 2. Порівняння результатів прогнозування за тим самим алгоритмом із різними параметрами

На рис. 2, а алгоритм виявився надто чутливим і сприйняв випадкові варіації в даних як стійкі закономірності. Ця проблема відома як перенавчання. Надмірно підігнана модель даватиме високоточні прогнози для поточних даних, але буде менш придатною для узагальнення майбутніх даних.

Натомість на рис. 2, в алгоритм був занадто нечутливим і не помітив основних закономірностей. Ця проблема відома як

недонавчання. Недостатньо пристосована модель, найвірогідніше, нехтуватиме значущими тенденціями, що призведе до менш точних прогнозів як для поточних, так і для майбутніх даних.

• • •

Далі — у повній книжці «Опануй числа! Наука про дані для нефахівців».
nashaknyga.com.ua