



Огляд задач статистичного навчання

Статистичне навчання охоплює широкий спектр інструментів, призначених для розуміння даних. Їх можна розділити на дві групи: *контрольоване навчання (навчання з вчителем)* та *неконтрольоване навчання (навчання без вчителя)*. У широкому сенсі, контрольоване навчання передбачає побудову статистичної моделі для прогнозування або оцінки *вихідної змінної* на основі однієї або декількох *вхідних змінних*. Такі задачі виникають у різноманітних сферах, зокрема в бізнесі, медицині, астрофізиці та державній політиці тощо. У разі неконтрольованого навчання є початкові дані, але немає контрольованого виходу даних; проте можна вивчати їхні взаємозв'язки та структуру на основі відсутності результативної змінної. Для кращого розуміння сфер застосування статистичного навчання розглянемо три реальних набори даних, які будуть використані в цій книжці.

Дані про заробітну плату

У цьому прикладі (набір даних **Wage**) ми досліджуємо низку факторів, які впливають на заробітну плату різних груп чоловіків із Середньоатлантичного регіону США. Зокрема, ми хочемо виявити зв'язки між заробітною платою працівника (змінна **wage**) та його віком (**age**), рівнем освіти (**education**), а також календарним роком (**year**). Подивіться, наприклад, на графік (рис. 1.1 ліворуч), де показано зв'язок між заробітною платою і віком працівників із цього набору даних. Видно, що заробітна плата (**wage**) збільшується при зростанні віку (**age**), але потім знову знижується приблизно після 60 років. Синя лінія відображає середній

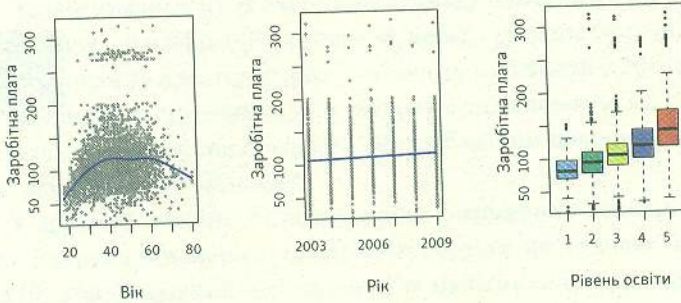


Рисунок 1.1. Інформація про доходи чоловіків із Середньоатлантичного регіону США згідно з набором даних *Wage*. Ліворуч: *wage* як функція від *age*. З віком середня заробітна плата зростає приблизно до 60 років, після чого починає знижуватися. У центрі: *wage* як функція від *year*. З 2003 по 2009 рік спостерігається повільне, але стабільне щорічне зростання середньої заробітної плати приблизно на 10 000\$. Праворуч: діаграма розмаху *wage* як функція від *education*, де 1 означає найнижчий рівень освіти (відсутність диплома про середню освіту), а 5 — найвищий рівень (науковий ступінь). У середньому заробітна плата (*wage*) зростає з підвищенням рівня освіти (*education*).

рівень заробітної плати (*wage*) для певного віку, дає змогу побачити цей тренд чіткіше.

Знаючи вік працівника, можна *спрогнозувати* його заробітну плату за цією кривою. Однак на рис. 1.1 також добре видно значний розкид цього показника, із чого випливає, що сама лише змінна *age* навряд чи дасть змогу з високою точністю передбачити *wage* для конкретної людини.

У нас також є інформація щодо рівня освіти кожного працівника та його заробітної плати *wage* за кожен календарний рік *year*. Графіки в центрі та праворуч на рис. 1.1 відображають залежності змінної *wage* від предикторів *year* та *education* відповідно, які свідчать про їхній зв'язок із *wage*. У період з 2003 по 2009 рік зарплата щороку лінійно зростала приблизно на 10 000\$, хоча це зростання дуже слабке порівняно з коливаннями даних. Окрім того, зарплата більша в людей з вищим рівнем освіти: працівники з найнижчим рівнем освіти (1) загалом заробляють значно менше, ніж працівники з найвищим рівнем (5). Зрозуміло, що найточніший прогноз *wage* для конкретної людини можна отримати, об'єднавши інформацію про її вік (*age*), рівень освіти (*education*) та календарний рік (*year*). У розділі 3 ми розглянемо лінійну регресію, яку можна застосувати для прогнозування *wage* за цими даними. В ідеалі нам потрібно врахувати нелінійний зв'язок між змінними *wage* та *age*. У розділі 7 ми обговорюємо клас методів, призначених для вирішення цієї проблеми.

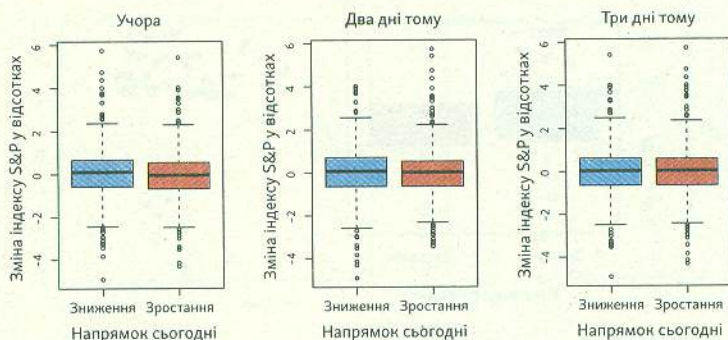


Рисунок 1.2. Ліворуч: коробчасті діаграми зміни індексу S&P за попередній день у відсотках, для днів зростання та падіння ринку (за даними **Smarket**). По центру та праворуч: те саме, що й ліворуч, але наведено відсоткові зміни за 2 та 3 попередніх дні відповідно.

Дані про ринок акцій

За набором даних **Wage** прогнозується неперервне або кількісне вихідне значення. Часто таку ситуацію називають задачею *регресії*. Однак у деяких випадках можна зіткнутися з необхідністю передбачити нечислове значення, тобто *категоріальну*, або *якісну*, результативну змінну. Так, у розділі 4 використовується набір даних про ринок акцій, у якому зібрані щоденні зміни індексу Standard & Poor's 500 (S&P) протягом 5-річного періоду (з 2001 по 2005 рік). Ми будемо посилалися на нього як на **Smarket**. Завдання — спрогнозувати *зростання* або *зниження* індексу на основі зміни його відсоткового значення за останні 5 днів. Тут проблема статистичного навчання не передбачає прогнозування числового значення. Натомість передбачається зростання (**Up**) або зниження (**Down**) показника ринку акцій для того чи іншого дня. Така проблема відома як задача *класифікації*. Модель, яка здатна з високою точністю прогнозувати напрямок руху ринку, була б дуже корисною!

На рис. 1.2 ліворуч наведено дві коробчасті діаграми розмаху відсоткових змін S&P порівняно з попереднім днем: перша для 648 днів, коли індекс зростав наступного дня, та для 602 днів — коли він падав. Ці дві діаграми майже ідентичні, що свідчить про те, що не існує простої стратегії для прогнозування поточного стану індексу S&P на основі його попередніх значень. Решта графіків, на яких наведені подібні діаграми відсоткових змін за попередні 2 та 3 дні, також вказують на відсутність вираженого зв'язку між минулим і поточним станами індексу. Безумовно, така відсутність зв'язку очікувана: за наявності тісних кореляцій

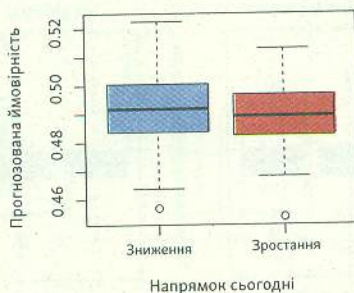


Рисунок 1.3. Ми застосували модель квадратичного дискримінантного аналізу для підмножини даних з набору Smarket за період з 2001 по 2004 рік і спрогнозували ймовірність падіння фондового ринку для даних на 2005 рік. У середньому прогнозна ймовірність зниження ринку вища для тих днів, коли падіння насправді мало місце. На основі цих результатів можна правильно передбачити напрямок руху ринку в 60 % випадків.

між послідовними днями можна було б застосувати просту торгову стратегію для отримання прибутку. Проте в розділі 4 ми детально досліджуємо ці дані за допомогою інших методів статистичного навчання. Виявляється, у цих даних присутні певні слабкі тенденції, які дозволяють спрогнозувати напрям руху ціни на ринку з точністю 60 % випадків (принаймні для цього 5-річного періоду; рис. 1.3).

Дані про рівень експресії генів

Два попередніх приклади ілюструють дані, які мають і вхідні, і вихідні змінні. Однак існує ще один важливий клас проблем, який стосується випадків, коли ми маємо лише вхідні змінні, але без залежної (вихідної) змінної. Наприклад, виконуючи маркетингові дослідження, ми маємо демографічні дані певної кількості наявних або потенційних клієнтів. Тоді може виникнути потреба з'ясувати, а які клієнти між собою схожі, чи можна їх згрупувати за спільними характеристиками. Такий підхід відомий як задача *кластеризації*. На відміну від попередніх прикладів, тут не постає завдання передбачити будь-яку результативну змінну.

Ми присвятили розділ 12 обговоренню статистичних методів навчання для проблем, де відсутня жодна реальна вихідна змінна. Зокрема, ми розглядаємо набір даних **NCI60**, який складається з 6830 значень рівня експресії генів для кожної з 64 клітинних ліній ракових пухлин. Нас не цікавить прогнозування конкретної вихідної змінної, нас цікавить можливість виявлення груп або кластерів за типами раку в клітинах на основі замірів експресії їхніх генів. Це складне питання, частково через те,