Лекція 3

Pandas. Лінійні моделі

§11 Знайомство з Pandas

Див. jupyter notebook "11ЗнайомствоPandas.ipynb"

§12 Знайомство з машинним навчанням

Розглянемо типи задач із класу задач **навчання із учителем**. Навчання з учителем дуже схожа на задачу інтерполяції. Інтерполяція - задача відновлення функції по декількох точках, у яких відомі її значення.

Навчання із учителем - теж відновлення загальної закономірності за скінченним числом прикладів.

Приклад: чи сподобається фільм користувачеві

Нехай є деякий сайт, присвячений кіно

Необхідно зрозуміти, чи сподобається користувачеві фільм

Існує велика кількість прикладів – ситуацій, коли інші користувачі заходили на сторінки фільмів, вирішували подивитися фільм і далі ставили оцінку, за якою можна зрозуміти, сподобався їм чи фільм ні. ***Задача машинного навчання складається у знаходженні загальної закономірності з такої інформації***.

Основні позначення

x — об'єкт, $\boldsymbol{X}$ — простір об’єктів, $y = y(x)$ — відповідь на об'єкті x, $\boldsymbol{Y}$ — простір відповідей.

Об'єктом називається сутність, яка має ознаки (характеристики, фічі). У даному прикладі об'єктом є пара **користувач-фільм**.

Простір об'єктів - це множина всіх можливих об'єктів, для яких може знадобитися робити передбачення. У цьому прикладі це **множина всіх можливих пар** користувач-фільм.

Відповіддю буде називатися те, що потрібно передбачити. У цьому випадку відповідь - **сподобається користувачеві фільм чи ні**.

Простір відповідей, тобто множина всіх можливих відповідей, складається із двох можливих елементів: -1 (користувачеві фільм не сподобався) та +1 (сподобався).

Ознака - це число, що характеризує об'єкт.

Ознаковим описом об'єкта називається сукупність всіх ознак:

$$x = (x^1, \ldots x^d).$$

Вибірка, алгоритм навчання

Навчальна вибірка $X = (x_i, y_i)_{i=1}^{l}$ – приклади, на основі яких буде будуватися загальна закономірність. У вищезгаданому випадку y_i - це оцінка відповідного фільму відповідного користувача.

Прогноз буде виконуватись на основі деякої **моделі** (**алгоритму**) $a(x)$, тобто деякої функції із простору $\boldsymbol{X}$ у простір $\boldsymbol{Y}$. Ця функція повинна бути легко реалізована на комп'ютері, щоб її можна було використовувати в системах машинного навчання. Прикладом такої моделі є лінійний алгоритм:

$$a(x) = \text{sing}(\text{w}_0 + w_1 x^1 + \ldots + w_d x^d).$$

Для оцінки якості роботи алгоритму вводиться функціонал похибки $Q(a, X)$ — **похибка алгоритму** на на вибірці X. Наприклад, функціонал похибки може бути часткою неправильних відповідей.

Завдання навчання полягає в підборі такого алгоритму a, для якого досягається мінімум функціонала похибки. Кращий у цьому сенсі алгоритм вибирається з деякого сімейства A алгоритмів.

Однорівневе дерево рішень

Приклад сімейства алгоритмів є однорівневе дерево рішень:

$$A = \{[x^j > t] \mid \forall j, t\}.$$

Тут квадратні дужки відповідають так званій **нотації Айверсона**. Якщо логічний вираз усередині цих дужок — ***істина***, то значення дужок дорівнює **1**, в випадку ***хибності*** — **0**.

Алгоритм працює в такий спосіб. Якщо значення певної ознаки x^j менше деякого граничного значення t, то даний алгоритм повертає відповідь 0 (фільм не сподобається), в іншому випадку — +1 (користувачеві фільм сподобається).

Однорівневе дерево рішень може бути використано для побудови складних композицій алгоритмів.

§13 Навчання на розмічених даних

Постановка задачі

Для навчальний вибірки

$\boldsymbol{X} = (x_i, y_i)_{i=1}^{l}$ (розмічені дані)

потрібно знайти такий алгоритм $a \in \boldsymbol{A}$, на якому буде досягатися мінімум функціонала похибки

$Q(a, X) \to \min_{a \in \boldsymbol{A}}$ – **це є задача навчання із учителем**

Залежно від множини можливих відповідей $\boldsymbol{Y}$, задачі діляться на кілька типів.

Задача бінарної класифікації

Простір відповідей складається із двох відповідей $Y = \{0,1\}$.

Множина об'єктів, які мають однакову відповідь, називається **класом**.

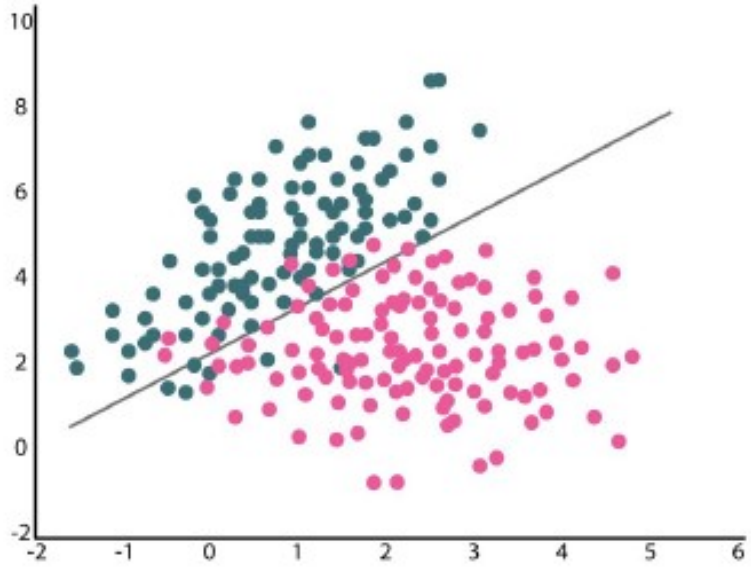


Рис. 1.1: Задача бінарної класифікації

Приклади задач бінарної класифікації:

- Чи сподобається користувачеві фільм?
- Чи поверне клієнт кредит?

Задача багатокласової класифікації

Класів може бути більше, ніж два.

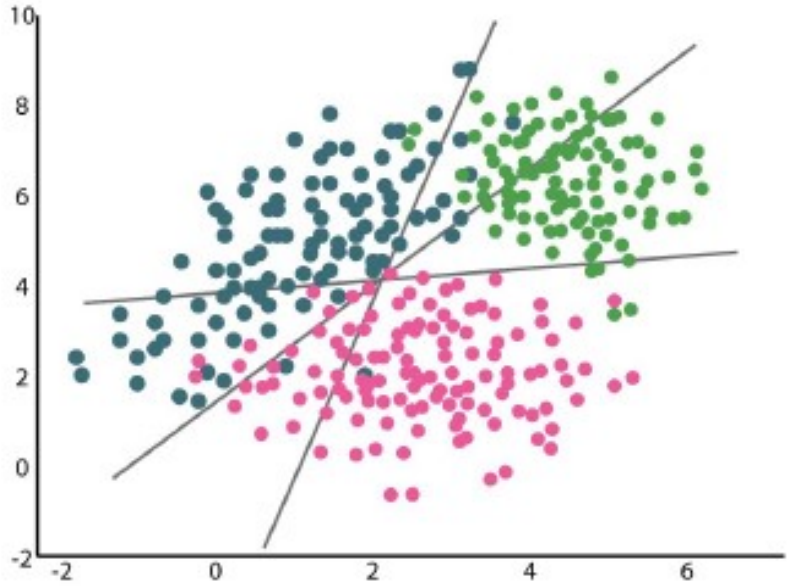


Рис. 1.2: Задача багатокласової класифікації

Приклади задач багатокласової класифікації:

- З якого сорту винограду зроблене вино?
- Яка тема статті?
- Машина якого типу зображена на фотографії: мотоцикл, легкова або вантажна машина?

Задача регресії

Коли y є дійсною змінною, то говорять про задачу регресії.

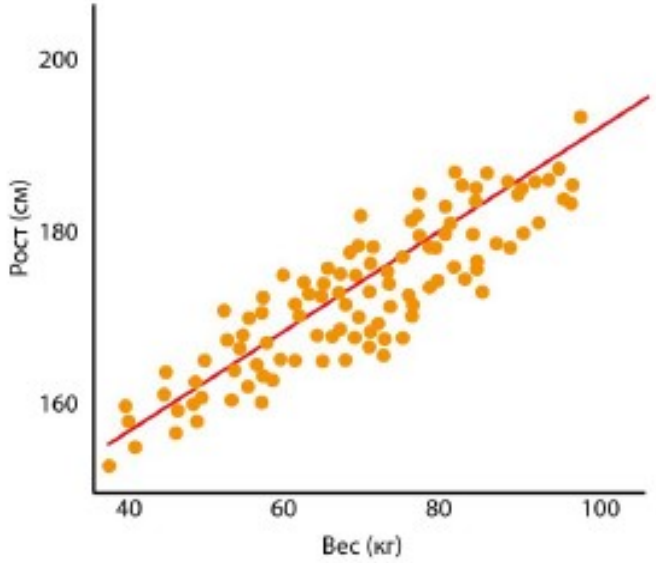


Рис. 1.3: Задача регресії

Приклади задач регресії:

- Прогноз температури на завтра.
- Прогнозування виторгу магазина за рік.
- Оцінка віку людини за його фото.

Задача ранжирування

Із задачою ранжування зіштовхуємося з нею щодня, коли шукаємо що-небудь в Інтернеті. Після того, як ми ввели запит, відбувається ранжирування сторінок за релевантністю (ступеню відповідності) їх запиту. Тобто для кожної сторінки оцінюється її релевантність у вигляді числа, а потім сторінки сортуються за зменшенням релевантності.

Завдання задачі ранжування полягає в передбаченні релевантності для пари (запит, сторінка).

§14 Навчання без учителя

Види навчання

Задача навчання без учителя **-** це така задача, у якій є тільки об'єкти, а відповідей немає.

Також бувають «проміжні» постановки.

У випадку ***часткового навчання*** є об'єкти, деякі з яких з відповідями.

У випадку ***активного навчання*** (проведення експерименту) одержання відповіді звичайно дуже дорого, тому алгоритм повинен спочатку вирішити, для яких об'єктів потрібно отримати відповідь, щоб найкраще навчитися.

Задача кластеризації

Задано множину об'єктів. Необхідно знайти групи схожих об'єктів. Є дві основні проблеми: не відома кількість кластерів і не відомі істині кластери, які потрібно виділяти. Тому задача вирішується дуже важко - тут неможливо оцінити якість рішення.

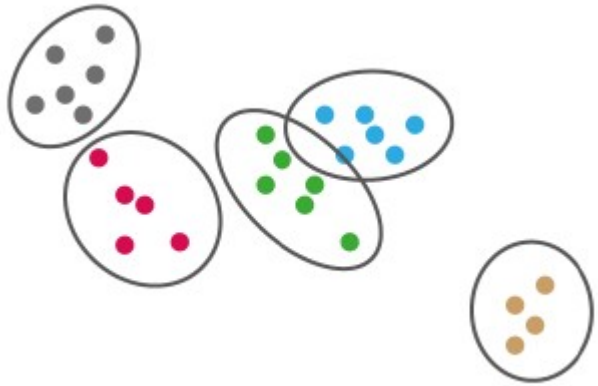


Рис. 1.4: Задача кластеризації

Приклади задач кластеризації:

- Сегментація користувачів (інтернет-магазину або оператора зв'язку)
- Пошук схожих користувачів у соціальних мережах
- Пошук генів зі схожими профілями експресії

Задача візуалізації

Задача візуалізації – необхідно зобразити багатовимірну (*d-вимірну*) вибірку так, щоб зображення наочно показувало структуру об'єктів.

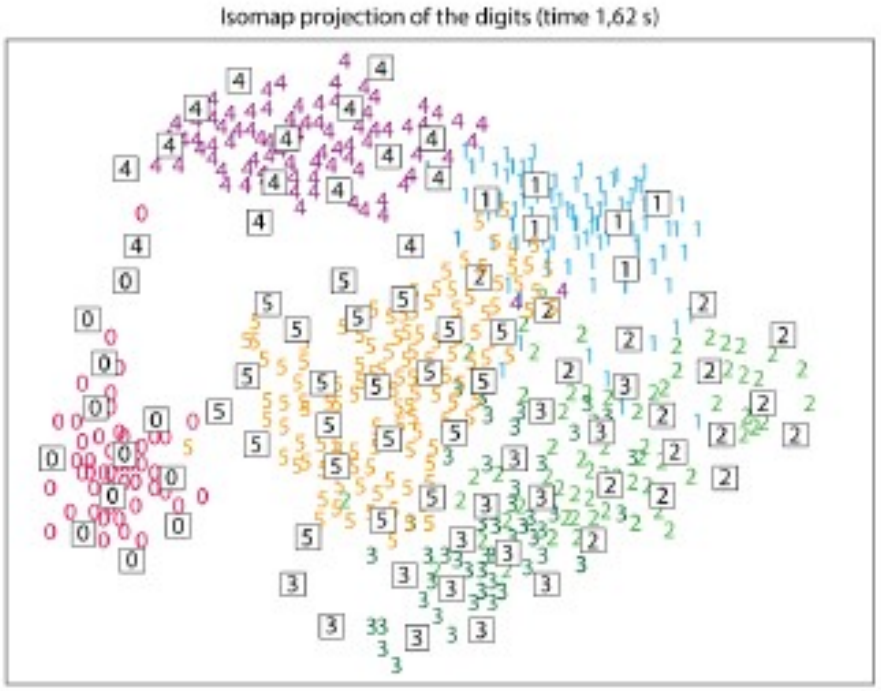


Рис. 1.5: Задача візуалізації

Зображено набору даних MNIST (відцифровка рукописних креслень цифр, вектор ознак - яскравість окремих пікселів)

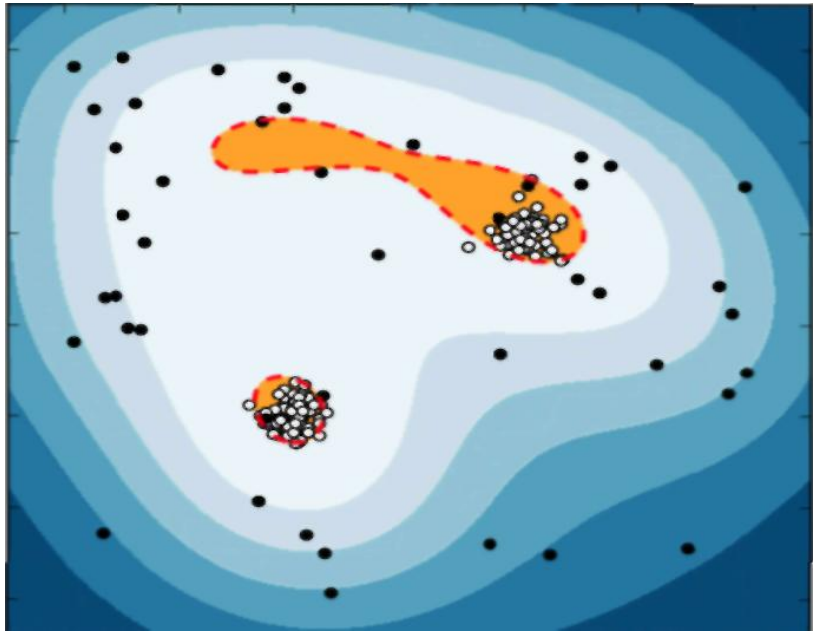
Пошук аномалій

Пошук аномалій: необхідно виявити, що даний об'єкт не схожий на всі інші, тобто є аномальним.

Сладна задача: приклади аномальних об'єктів відсутні.

Приклади такого типу задач:

- Визначення поломки в системах літака (за показниками сотень датчиків)
- Визначення поломки інтернет-сайту
- Виявлення проблем у моделі машинного навчання



§15 Ознаки в машинному навчанні

Існує кілька класів, або типів ознак. І у всіх свої особливості – їх потрібно по-різному обробляти й по-різному враховувати в алгоритмах машинного навчання.

Множина значень *j*-ої ознаки позначаємо D_j.

Бінарні ознаки

Бінарні ознаки приймають два значення: $D_j = \{0,1\}$.

Приклади:

- Чи вище дохід клієнта середнього доходу у місті?
- Колір фрукта - зелений?

Якщо відповідь на питання так — ознака покладається такою, що дорівнює 1, в іншому випадку —0.

Дійсні ознаки

Для дійсних ознак $D_j = \boldsymbol{R}$.

Приклади:

- Вік
- Площа квартири
- Кількість дзвінків до call-центру

Категоріальні ознаки

Для категоріальних ознак D_j є неупорядкованою множиною (неможливість порівняння «більше-менше»)

Приклади:

- Колір очей
- Місто
- Освіта (у деяких задачах може бути уведений порядок)

З категоріальними ознаками досить важко працювати.

Порядкові ознаки

Частинним випадком категоріальних ознак є порядкові ознаки. У цьому випадку D_j — упорядкована множина.

Приклади:

- Роль у фільмі (Перший план, другий план, масовка)
- Тип населеного пункту (упорядковані за населеністю)
- Освіта

Вони відрізняються тим, що у випадку порядкових ознак «відстань» між двома значеннями ознаки не має змісту. Наприклад, відмінність значення 3 від значення 2 може бути не такою ж істотною, як відмінність 1 від 0.

Багатозначні ознаки

Багатозначна ознака — це така ознака, значенням якого на об'єкті є підмножина деякої множини. Приклад:

- Які фільми подивився користувач
- Які слова входять у текст

Розподіл ознак

Проблеми, з якими можна зіштовхнутися під час роботи з ознаками.

Існування викидів. Викидом називається такий об'єкт, значення ознаки на якому відрізняється від значення ознаки на більшості об'єктів.

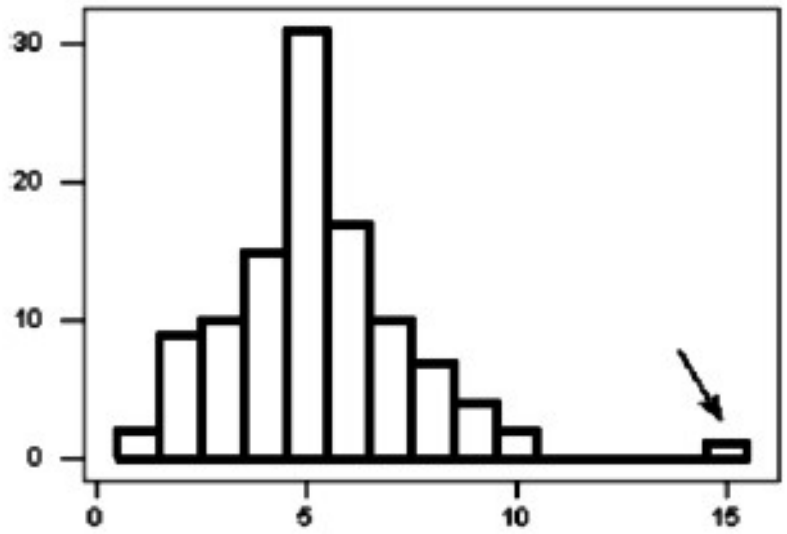


Рис. 1.6: Приклад викиду

Наявність викидів суттєво ускладнює алгоритми машинного навчання, які будуть намагатися врахувати і їх теж. *Викиди як правило виключають* із даних, щоб не заважати алгоритму машинного навчання шукати закономірності в даних.

Розподіл ознак. Не завжди ознака має такий розподіл, що дозволяє відповістити на необхідне запитання. Наприклад, може бути занадто мало даних про клієнтів з невеликого міста, тому що зібрати достатню статистику не вдалося.