ОСНОВИ СИСТЕМ ШТУЧНОГО ІНТЕЛЕКТУ, НЕЙРОННИХ МЕРЕЖ та ГЛИБОКОГО НАВЧАННЯ

Модуль 2. Навчання з вчителем

Лекція 2.7.
Класифікація. Метод kNN.

Класифікація

Формально:

Маємо множину $\mathbb{O}$ об'єктів $o^{(j)},\ j = 1,2,\dots,M$

Кожен об'єкт $o^{(i)}$ має сукупність характеристик - ознак $x_i^{(j)}, i = 1,2,\dots,N$ з множини $\mathbb{X}$.

Маємо множину $\mathbb{C}$ класів $\mathrm{c}^{(\mathrm{k})},\ k = 2,\dots,\mathrm{K}$

Існує невідома залежність (правило) $\mathbb{F}$, яка на підставі пар $\langle o^{(j)}, \mathrm{c}^{(\mathrm{k})} \rangle$ визначає, чи належить об'єкт $o^{(j)}$ до класу $\mathrm{c}^{(\mathrm{k})}$.

Завдання: знайти правило $\tilde{\mathbb{F}}$, максимально наближене до $\mathbb{F}$. Тобто, знайти вирішальне правило, що дозволяє класифікувати довільний об'єкт o за його ознаками.

Методи Класифікації

- Регресійні методи, логістична регресія
- ***Метод k-найближчих сусідів (kNN)***
- Метод опорних векторів (SVM)
- Наївний Байєс (ймовірнісний класифікатор)
- Дерева рішень
- ….

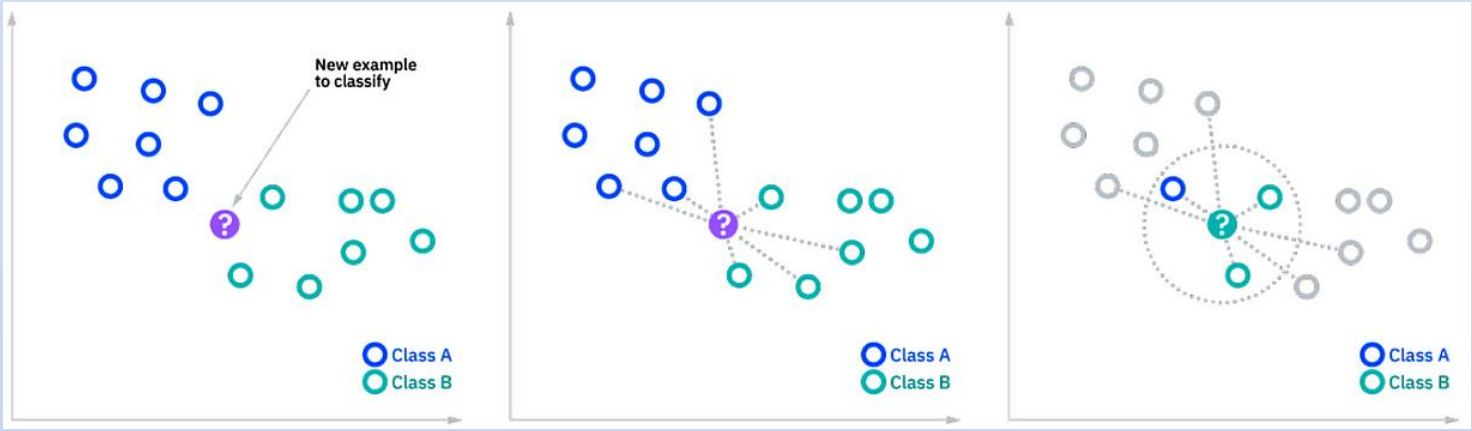
k-найближчих сусідів (kNN)

KNN (*k-nearest neighbor method*) — метод навчання з вчителем. Метод використовується як для класифікації, так і для регресії. В обох випадках вхідні дані складаються з ***k*** найближчих навчальних прикладів у наборі даних.
При **класифікації** результатом **k-NN** є належність класу. Об'єкт класифікується за допомогою множини голосів його сусідів, при цьому об'єкт належить до класу найбільш поширеного серед його ***k*** найближчих сусідів.
При **регресії** результатом є числове значення властивості об'єкту, що є середнім з ***k*** найближчих сусідів.

k-найближчих сусідів (kNN)

Спрощено ***KNN*** працює за простим принципом: об’єкт, швидше за все, буде схожий на своїх «ближчих» сусідів.

Алгоритм розглядає «k» найбільш схожих точок даних (сусідів) і приймає більшість голосів у разі класифікації або середнє значення у разі регресії.



Відстань між точками даних обчислюється за допомогою різних метрик, але найчастіше використовуються Евклідова відстань, відстань Мінковського,

Відстань

«Манхетенська» відстань (метрика Мінковского, L1)

$$dist(X,Y) = \sum_{i=1}^{N} |x_i - y_i|$$

Евклідова відстань (метрика L2)

$$dist(X,Y) = \sqrt{\sum_{i=1}^{N} (x_i - y_i)^2}$$

Косинусна метрика – кут між векторами

$$dist(X,Y) = 1 - \frac{\sum_{i=1}^{N} x_i y_i}{\sum_{i=1}^{N} (x_i)^2 \sum_{i=1}^{N} (y_i)^2}$$

k-найближчих сусідів (kNN)

1. **Попередня обробка даних:** Масштабувати функції. Різні масштаби між об’єктами можуть спотворити відстані та привести до неточних прогнозів. Масштабування включають min-max scaling та нормалізацію Z-оцінки.
2. **Вибір показника відстані:** Обирається показник відстані, який відповідає характеру даних.
3. **Визначення значення «k»**: визначення відповідної кількості сусідів є фундаментальним кроком.
4. **Навчання моделі:** KNN «ледачий» учень, тобто модель явно не будується під час «навчання». Зберігається весь набір даних у пам’яті.
5. **Передбачення:** Для нової точки даних обчислюється відстань між нею та **всіма** іншими точками в наборі даних. Далі обирається «***k***» найближчих і приймається більшість голосів (для класифікації) або середнє значення (для регресії).
6. **Оцінка моделі:** після прогнозів оцінюється якість точність моделі за допомогою відповідних показників.

k-найближчих сусідів (kNN)

Наданий алгоритм **kNN** називається **Brute-Force**, оскільки використовується метод повного перебору для пошуку найближчих сусідів, що робить його простим у реалізації, але вельми повільним при роботі з великим обсягом даних.
Для вирішення цієї проблеми в реалізації ***scikit-learn*** передбачені більш просунуті методи, що базуються на бінарних деревах, що дозволяє отримати значний приріст у продуктивності.

Дивись:

- Приклад 1
- Приклад 2
- Scikit-learn KNeighborsRegressor — scikit-learn 1.5.1 documentation

Контрольні запитання

- **Надайте загальну постановку задачі класифікації.**
- **Пояснить сутність методу kNN вирішення задачі класифікації.**
- **Пояснить процес пошуку параметру *k* при вирішенні задачі класифікації методом kNN .**

Рекомендована ЛІТЕРАТУРА

- **Глибинне навчання:** Навчальний посібник / Уклад.: В.В. Литвин, Р.М. Пелещак, В.А. Висоцька В.А. – Львів: Видавництво Львівської політехніки, 2021. – 264 с.
- Тимощук П. В., Лобур М. В. **Principles of Artificial Neural Networks and Their Applications: Принципи штучних нейронних мереж та їх застосування:** Навчальний посібник. – Львів : Видавництво Львівської політехніки, 2020. – 292 с.
- Morales M. **Grokking Deep Reinforcement Learning. –** Manning, 2020. – 907 с.
- Trask Andrew W. **Grokking Deep Learning.** – Manning, 2019. – 336 с.

Корисні та цікави посилання

- **Машинне навчання**

https://uk.wikipedia.org/wiki/машинне_навчання

- **Львівська політехніка**

http://www.mmf.lnu.edu.ua/ar/1739

http://www.mmf.lnu.edu.ua/ar/1743

Приклади дивись

Lec_02_07_Exmpl_1.md

Lec_02_07_Exmpl_2.md

The END
Модуль 2. Лекція 2.7.