ОСНОВИ СИСТЕМ ШТУЧНОГО ІНТЕЛЕКТУ, НЕЙРОННИХ МЕРЕЖ та ГЛИБОКОГО НАВЧАННЯ

Модуль 3. Навчання без вчителя

Лекція 3.3.
Кластеризація. Метод k-means.

Класичний AI / Класичний ML

Навчання без вчителя: Маємо великий набір даних. В цих даних є приховані закономірності.

Задача – знайти закономірності, наприклад, розбивши дані на певні групи чи кластери.

Кластеризація

Формально:

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

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

Передбачається, що є множина $\mathbb{C}$ класів (кластерів) $c^{(k)}$, $k = 1,2,\dots,K < M$ (іноді K відомо, іноді – невідомо).

Але (на відміну від класифікації)!

належність об’єкту $o^{(j)}$ до класу $c^{(k)}$ - невідома.

Кластеризація

Визначена деяка метрика $d(o^{(j)}, o^{(i)})$ – відстань від між об'єктом $o^{(j)}$ та об'єктом $o^{(i)}$.

Завдання: розбити вибірку $o^{(j)}$, $j = 1,2, \dots, M$ на непересічні підмножини – кластери так, щоб кожен кластер складався з об'єктів, близьких по метриці $d(.,\ .)$, а об'єкти різних кластерів істотно відрізнялися. При цьому кожному об'єкту $o^{(j)}$ приписується відповідний кластер – клас $c^{(k)}$.

Загальний підхід до кластеризації

Типова послідовність вирішення задачі

- **Відбір сукупності (вибірки) об'єктів для кластеризації.**
- **Визначення характеристик, по яких об'єкти оцінюються.**
- **Обчислення міри (відстань, схожість) між об'єктами.**
- **Застосування конкретного методу кластерного аналізу для створення груп схожих об'єктів.**
- **Перевірка достовірності результатів кластеризації.**
- **Якщо необхідно, коректування вибірки об'єктів.**

K-means

Алгоритм ***k-means*** розбиває набір на ***K*** наборів $\mathbb{C} = \{c^{(1)} \dots c^{(K)}\}$, таким чином, щоб мінімізувати суму квадратів відстаней від кожного об'єкту кластера до його центру (центру мас кластера !). Тобто потрібно знайти

$$\arg\min_{C} \sum_{k=1}^{K} \sum_{o^{(j)} \in C^{(k)}} d(o^{(j)}, \mu^{(k)})^2$$

Де $\mu^{(k)}$ центр ***k***-го кластеру $C^{(k)}$

$d(o^{(j)}, \mu^{(k)})$ функція відстані між $o^{(j)}$ та $\mu^{(k)}$ (типова Евклід):

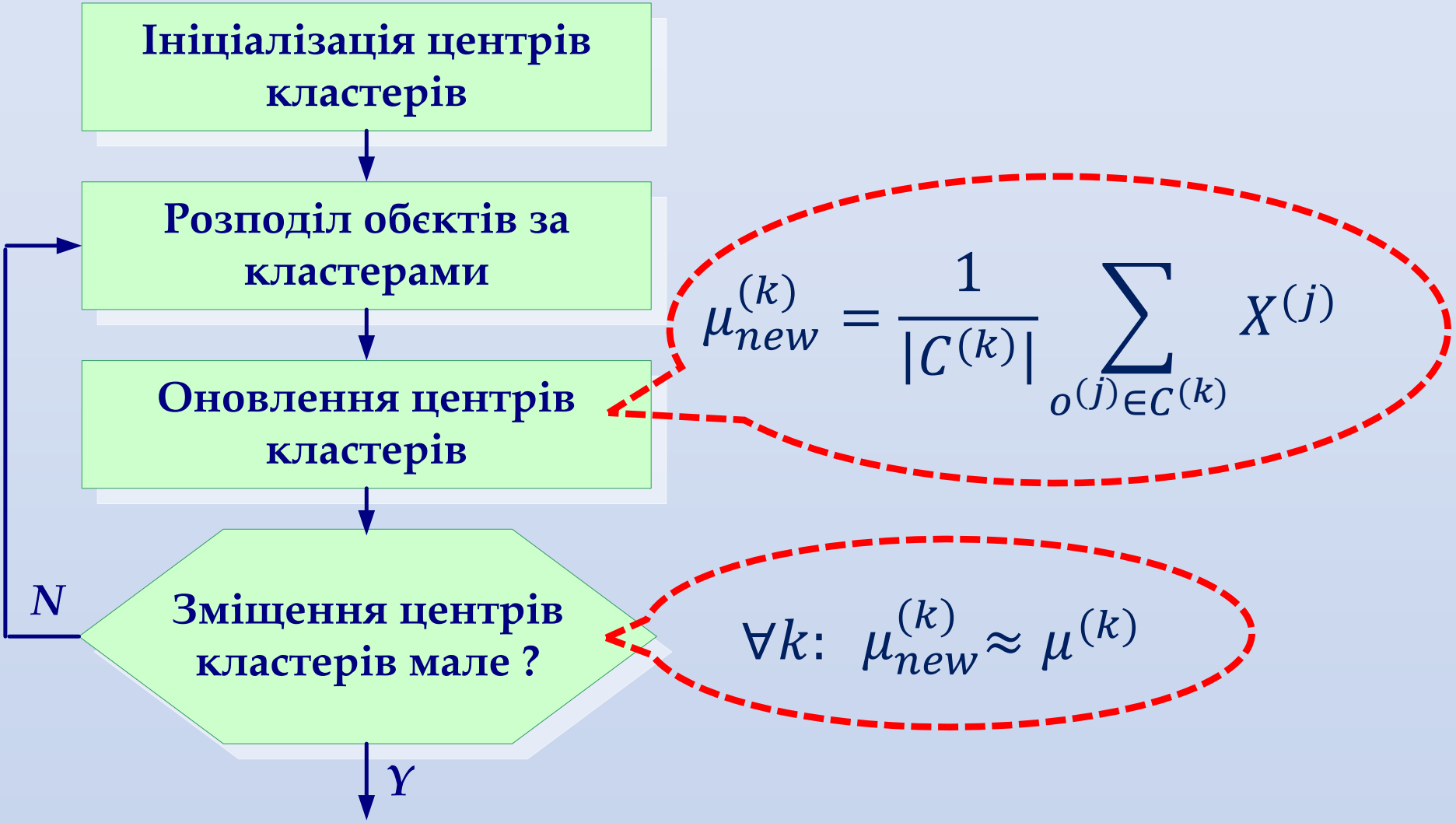
$$d(x, \mu_i) = \sqrt{\sum_{i=1}^{N} (o_i^{(j)} - \mu_i^{(k)})^2}$$

K-means

Початкова ініціалізація $\mu^{(k)}$ - випадкові вектори

$$C^{(s)} = \{o^{(j)} : d(o^{(j)}, \mu^{(s)})^2 \leq d(o^{(j)}, \mu^{(k)})^2\}$$
$$\forall k = 1, \ldots, K$$
$$\forall j = 1, \ldots, M$$

K-means


Ініціалізація центрів кластерів
Розподіл обєктів за кластерами
Оновлення центрів кластерів
Зміщення центрів кластерів мале ?
N
Y
$\mu_{new}^{(k)} = \frac{1}{|C^{(k)}|} \sum_{o^{(j)} \in C^{(k)}} X^{(j)}$
$\forall k: \ \mu_{new}^{(k)} \approx \mu^{(k)}$

K-means

Сильні сторони :

- Порівняно висока ефективність при простоті реалізації
- Висока якість кластеризації
- Можливість паралельного виконання
- Існування безлічі модифікацій
- У деяких випадках модель може не досягти збіжності.

Недоліки алгоритму:

- Кількість кластерів є параметром алгоритму
- Чутливість до початкових умов, викидів та шумів

K-means

Недоліки алгоритму:

- Викиди, далекі від центрів реальних кластерів, однаково враховуються під час обчислення їх центрів.
- Можливість збіжності до локального оптимуму
- Ітеративний підхід не дає гарантії збіжності до раціонального рішення.
- Алгоритм не застосовується до даних, котрим не визначено поняття " середнього " , наприклад, категоріальним даним.

Модифікації K-means

K-medians:
для обчислення центроїдів використовується не середнє, а медіана, що робить алгоритм більш стійким до аномальних значень у даних.

C-means:
визначає ймовірність того, що об'єкт належить до того чи іншого кластера.

Fuzzy C-means:
Дозволяється нечітке кластерне призначення.

Контрольні запитання

- **Надайте загальну постановку задачі кластеризації.**
- **Пояснить сутність алгоритму k-means для вирішення задачі кластеризації**
- **Опишіть метод «силуету» для оцінки якості вирішення задачі кластеризації**

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 посилання

Cluster Analysis

https://en.wikipedia.org/wiki/Cluster_analysis

K-means

https://en.wikipedia.org/wiki/K-means_clustering

Sklearn clustering

https://scikit-learn.org/stable/modules/clustering.html#silhouette-coefficient

Silhouette (clustering)

https://en.wikipedia.org/wiki/Silhouette_(clustering)

Calinski–Harabasz index

https://en.wikipedia.org/wiki/Calinski%E2%80%93Harabasz_index

The END
Модуль 3. Лекція 03.