ОСНОВИ СИСТЕМ ШТУЧНОГО ІНТЕЛЕКТУ, НЕЙРОННИХ МЕРЕЖ та ГЛИБОКОГО НАВЧАННЯ

Модуль 3. Навчання без вчителя

Лекція 3.2.
Огляд методів кластеризації. Оцінка якості кластеризації.

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.

Кластерний аналіз. Кластеризація

Кластерний аналіз (data clustering, cluster analysis, data clustering, clustering)
– процес розбиття заданої вибірки об'єктів (ситуацій) на підмножини, які називаються кластерами, так, щоб кожен кластер складався з схожих об'єктів, а об'єкти різних кластерів істотно відрізнялися.
Завдання кластеризації належить до статистичної обробки, а також до широкого класу завдань некерованого навчання (без вчителя).

Основна мета → знаходженні «схожих» об'єктів у виборці.
Головна проблема → що таке схожість , скільки кластерів ?

Кластерний аналіз – сукупність суттєво різних методів та алгоритмів розбиття об'єктів.

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 $c^{(k)}$.

Методи кластеризації

- **Центроідні методи (метод k-середніх, k-means)**
- Моделі зв'язності (ієрархічна кластеризація)
- Статистичні моделі (багатовимірний нормальний розподіл за EM-алгоритмом)
- Графові методи …
- Групові моделі …Регресійні методи, логістична регресія
- Нейронні мережі (нейронна мережа Кохонена)
- ….

Загальні методи кластеризації

На основі центроїд
Centroid Based:
Kmeans, Kmeans+, KMedods

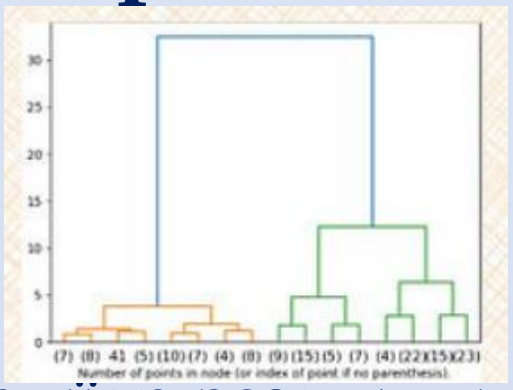
Плюси:

- Простота: легко реалізувати та інтерпретувати.
- Ефективність: добре масштабуються великі набори даних.
- Універсальність: можуть використовуватися для різних типів даних та завдань..

Мінуси:

- Чутливі до шуму та викидів у даних.
- Нездатні виявити кластери довільної форми (придатні для кластерів сферичної / еліптичної форми.
- Немає автоматичного способу визначення оптимального числа кластерів.

Загальні методи кластеризації



Ієрархічні методи

Connectivity-based:

Hierarchical Clustering

Плюси:

- Простота. Відносно прості у реалізації та розумінні.
- Наочність. Результат можна надати у вигляді дендрограми, яка наочно демонструє ієрархію кластерів та його взаємозв'язку.
- Універсальність. Задачі з довільним числом кластерів, задачі з кластерами складної форми.
- Нечутливість до викидів.
- Не потрібно попередньо визначати кількість кластерів.

Мінуси:

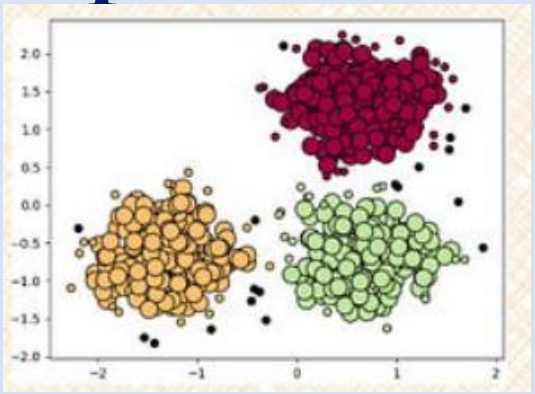
- Не підходить для багатокласових завдань.
- Чутлива до мультиколлінеарності.
- Не може обробляти нелінійні залежності.

Загальні методи кластеризації

На основі щільності

Density-based:

DBSCAN, HDBSCAN, OPTICS



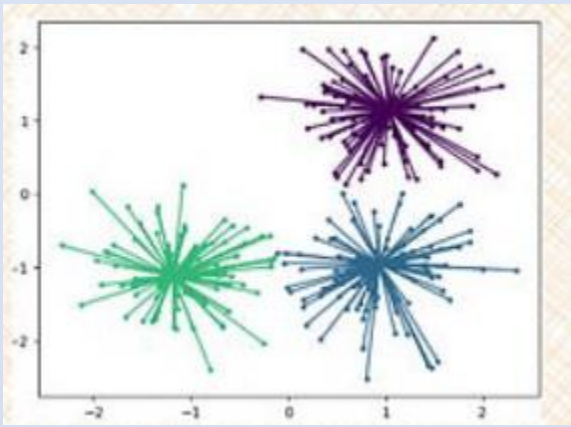
Плюси:

- Не вимагає попередньої вказівки числа кластерів.
- Виявлення кластерів довільної форми.
- Виділення викидів.
- Стійкість до шуму.
- Відносно невелика кількість параметрів.

Мінуси:

- Чутливість до параметрів.
- Складність у високорозмірних просторах.
- Проблеми з кластерами різної щільності.
- Неефективність великих даних.

Загальні методи кластеризації



Базовані на графах

Graph-based:

Affinity Propagation, Spectral Clustering

Плюси:

- Природне подання даних (граф !!!.
- Універсальність: категоріальні дані, текстові дані та мультимедійні дані.
- Здатність виявляти кластери складної форм.
- Стійкість до шуму.
- Легка візуалізація та інтерпретація результатів.

Мінуси:

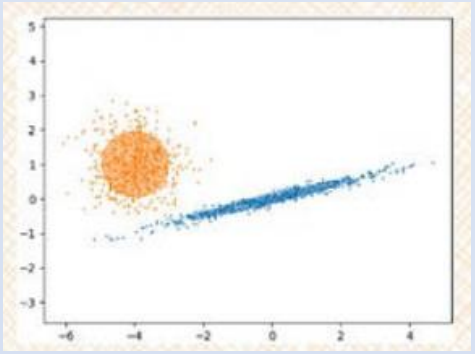
- Можуть бути складнішими у реалізації.
- Вибір міри подібності робер графа може бути непростим.
- Визначення оптимальної кількості кластерів.

Загальні методи кластеризації

Базовані на розподілі

Distribution-based:

Gaussian Mixture Models(GMM)



Плюси:

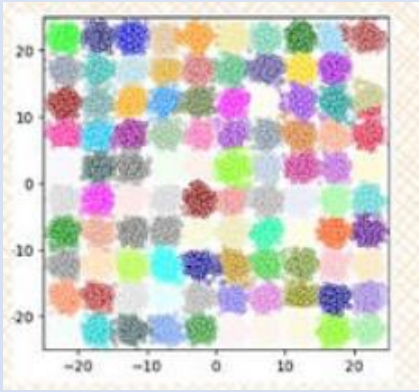
- Можуть моделювати складні структури даних, де точки в кластерах можуть мати різну форму та розміри.
- Надають ймовірнісну інтерпретацію належності точки до кластера.
- Більш стійкі до шуму та викидів.
- Автоматичне визначення числа кластерів:

Мінуси:

- Чутливість до вибору ініціалізації.
- Обчислювальна складність.
- Необхідність попередньої обробки даних.
- Інтерпретація результатів може бути складною.

Загальні методи кластеризації

На базі стиску даних
Compression-based:
BIRCH



Плюси:

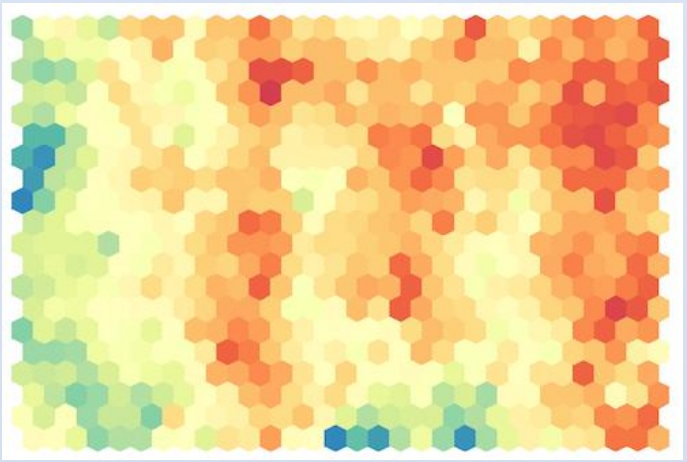
- Можуть швидко обробляти великі набори даних.
- Добре масштабуються на набори даних з високою розмірністю.
- Стійкість до шуму.
- Результати легко візуалізувати та інтерпретувати.

Мінуси:

- Чутливість до вибору алгоритму стиснення.
- Втрата інформації в процесі стиснення.
- Складність інтерпретації.

Загальні методи кластеризації

На основі моделі
Model-based:
Self-Organizing Map (SOM)



Плюси:

- Дозволяють візуалізувати подібність між об'єктами в низькорозмірному просторі.
- Можуть виявляти нелінійні взаємозв'язки даних.
- Стійкість до шуму та викидів.

Мінуси:

- Чутливість до вибору параметрів
- Складність інтерпретації - може бути складно інтерпретувати, чому певні об'єкти потрапляють в той самий кластер.

Якість кластеризації

Універсальна метрики для оцінки якості кластеризації відсутня. Суттєво залежить від конкретної задачі.

Деякі підходи, що враховують середню відстань від об'єктів до центрів кластерів:

- Silhouette score (Силует).
- Calinski-Harabasz index.
- Dunn index.

Silhouette score

Завдання кластеризації вирішено.
Визначена кількість кластерів (класів) ***K.*** Для кожного об’єкту$o^{(j)} \in \mathbb{O}$ визначено до якого кластеру (класу) він належить $o^{(j)} \in \mathrm{C}^{(\mathrm{k})}, k = 1, \dots, K$.

1. **Для кожного** $o^{(j)} \in \mathrm{c}^{(\mathrm{k})}$ обчислюється

$$a^{(j)} = \frac{1}{|C^{(k)}| - 1} \sum_{o^{(i)} \in C^{(k)}, i \neq j} d(o^{(j)}, o^{(i)}$$

Тобто середня відстань між j-м об'єктом та всіма об'єктами, що належать до того ж самого кластеру

Silhouette score

2. Для кожного $o^{(j)} \in C^{(\mathrm{k})}$ обчислюється середня відстань до об'єктів інших класів (для кожного іншого класу окремо - $C^{(\mathrm{k})} \neq C^{(\mathrm{i})}$) та визначається **середня різниця**

$$b^{(j)} = \min_{k \neq i} \frac{1}{|C^{(k)}|} \sum_{o^{(j)} \in C^{(k)}, o^{(i)} \notin C^{(k)}} d(o^{(j)}, o^{(i)})$$

Це найменша середня відстань до всіх точок у будь-якому іншому кластері, з яких не є $C^{(\mathrm{k})}$. Кластер з цією найменшою середньою відмінністю називається «**сусіднім кластером**» Це «наступний кластер», що найбільш підходить для об'єкту кластера k.

Silhouette score

3. Силует $s^{(j)}$ кожного об'єкту $o^{(j)} \in C^{(\mathrm{k})}$ визначається як

$$s^{(j)} = \frac{b^{(j)} - a^{(j)}}{\max\{a^{(j)}, b^{(j)}\}}, \qquad s^{(j)} = 0, if \left|C^{(k)}\right| = 1$$

$$s^{(j)} = \begin{cases} 1 - {}^{a}/_{b}, & if\ a < b \\ 0, & if\ a = b \\ {}^{b}/_{a} - 1, & if\ a > b \end{cases}$$

$$-1 \le s^{(j)} \le 1$$

Середнє $s^{(j)} \approx 1$ → об'єкти в кластері $C^{(\mathrm{k})}$ щільно згруповані.

Silhouette score

Коефіцієнт силуету – максимальне значення середнього $s^{(j)}$ по всім об'єктам всіх кластерів

$$SC = \max_k \{mean(s^{(j)})\}$$

Sklearn clustering

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

Контрольні запитання

- **Надайте огляд методів 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

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 02.