ОСНОВИ СИСТЕМ ШТУЧНОГО ІНТЕЛЕКТУ, НЕЙРОННИХ МЕРЕЖ та ГЛИБОКОГО НАВЧАННЯ

Модуль 3. Навчання без вчителя

Лекція 3.4.
Кластеризація. Метод DBSCAN.

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 $c^{(k)}$.

Надалі: об'єкт $o^{(j)}$, $j = 1,2,\dots,M$ будемо розуміти як деяку точку $\boldsymbol{p}$ в $\boldsymbol{N}$-вимірному векторному просторі.

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.**

Методи кластеризації

- Центроідні методи (метод k-середніх, k-means)
- **Методи щільності**
- 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)
- ….

DBSCAN

DBSCAN (**density-based spatial clustering of applications with noise**) алгоритм засновано на щільності даних: для заданої множини точок у деякому просторі він відносить в одну групу точки, які розташовані найбільш щільно (точки з багатьма сусідами) та розмічає точки, які лежать в областях з невеликою щільністю (чиї сусіди розташовані занадто далеко) як **outliners** (викиди, аномалії).

DBSCAN один з найпоширеніших алгоритмів кластеризації.

DBSCAN. Визначення

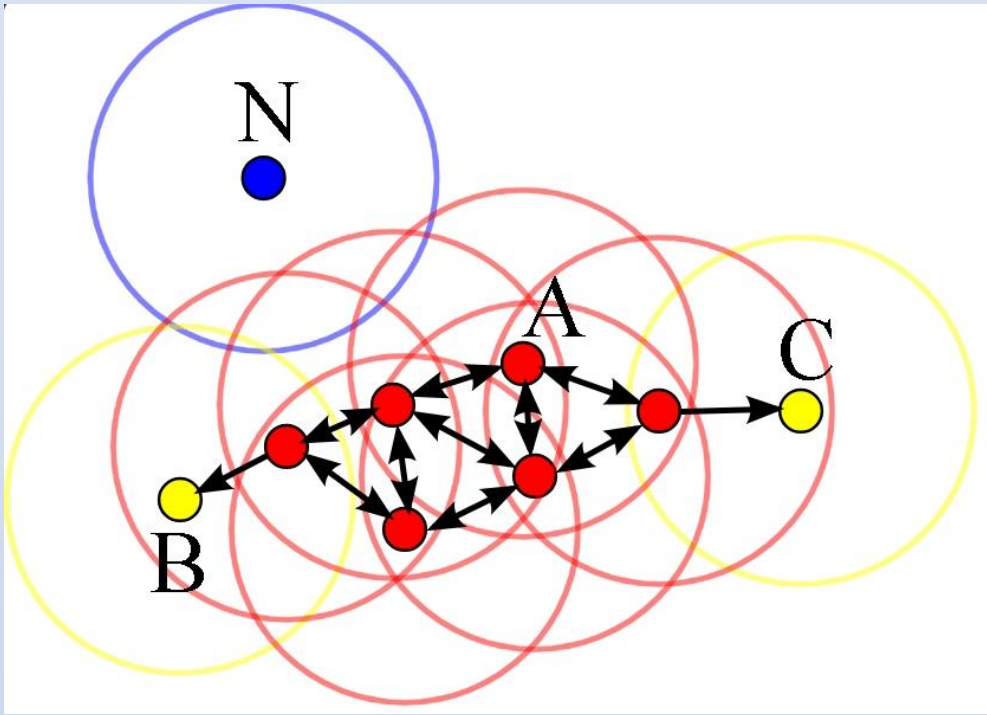
- ***Eps (ε)*** **-** радіус околу точки **p**.
- ***MinPts*** **-** мінімальна кількість точок необхідних для утворення щільної області.
- ***core point - ядрова точка*** **p** - коли не менш ***MinPts*** точок знаходяться на відстані ***ε*** від неї, включно з **p**. Кажуть, що ці точки безпосередньо досяжні з **p**.
- ***directly reachable point –*** точка **q** безпосередньо досяжна з **p**, якщо **q** знаходиться на відстані не більшій ніж **ε** від ядрової точки **p**.
- ***reachable poinr -*** точка **q** є досяжною з **p**, якщо існує шлях p_1, ... p_n з точок $p_1 = p$ та $p_n = q$, де кожна p_{i+1} безпосередньо досяжна з p_i (всі p_i повинні бути ядровими, можливо за виключенням **q**).
- ***outliers or noise points –*** викиди – всі інші.

DBSCAN.

Точки **A** (червоні) є **ядровими (*kernel*)**, бо окіл цих точок радіуса **ε** містить щонайменше 4 точки, включно з нею самою. Всі вони досяжні одна з одної - утворюють спільний кластер.

Точки **B** та **C** не будуть ядровими, але **досяжні** з **A** або з інших ядрових точок і тому належать кластеру.

MinPts = 4.



Точка **N** - шумова, бо не є ні ядровою ні безпосередньо досяжною.

DBSCAN.

Input: *DB*: Database
Input: ε: Radius
Input: *minPts*: Density threshold
Input: *dist*: Distance function
Data: *label*: Point labels, initially *undefined*

```
1  foreach point p in database DB do                                  # Перегляд всіх точок
2      if label(p) ≠ undefined then continue                          # Пропуск розглянутих точок
3      Neighbors N ← RangeQuery(DB, dist, p, ε)                       # Пошук сусідів
4      if |N| < minPts then
5          label(p) ← Noise                                           # Не ядро -> шум
6          continue
7      c ← next cluster label                                         # Новий кластер
8      label(p) ← c
9      Seed set S ← N \ {p}                                           # Розширення сусідів
10     foreach q in S do
11         if label(q) = Noise then label(q) ← c
12         if label(q) ≠ undefined then continue
13         Neighbors N ← RangeQuery(DB, dist, q, ε)
14         label(q) ← c
15         if |N| < minPts then continue                              # Перевірка корневої точки
16         S ← S ∪ N
```

DBSCAN.

Проблеми:

- Як оцінити якість кластеризації ?
- Як обрати ***Eps (ε), MinPts*** ?

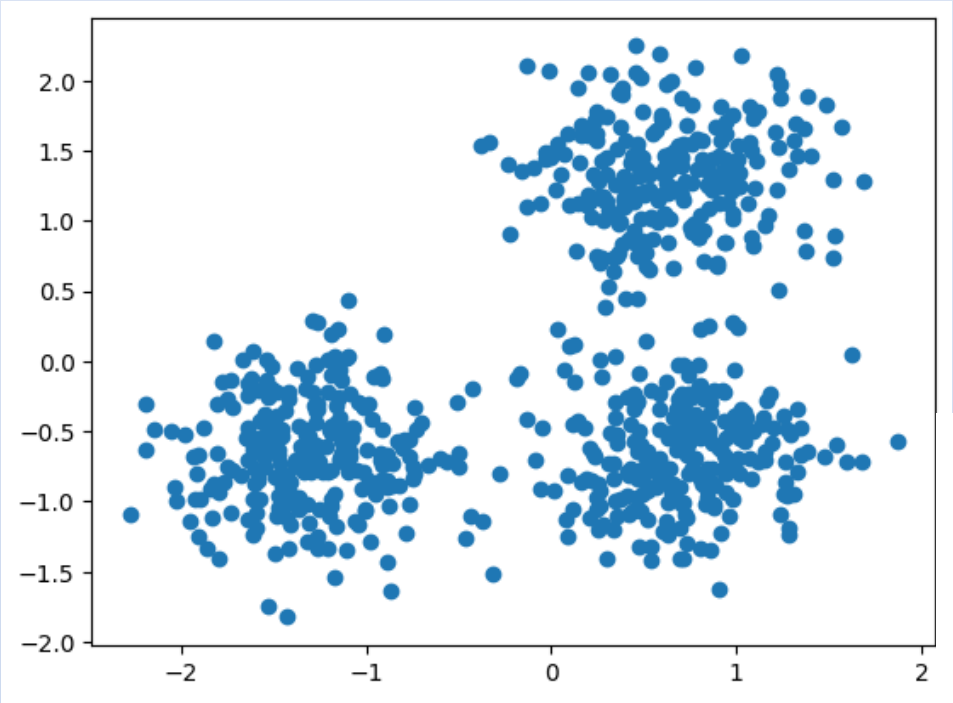
Якість кластеризації

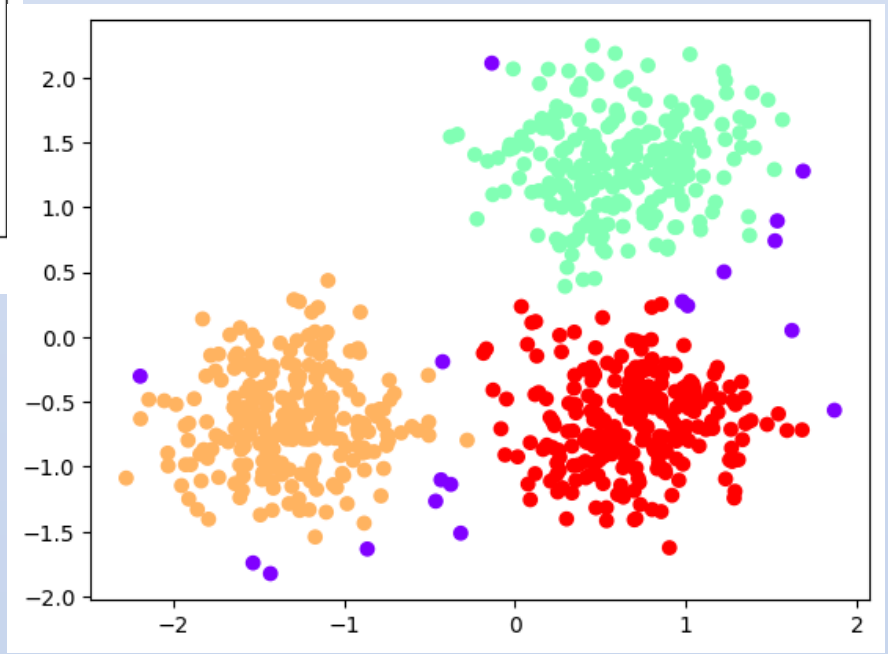
Універсальна метрики для оцінки якості кластеризації відсутня. Залежить від конкретної задачі.

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

- Silhouette score (Силует)
- Calinski-Harabasz index
- Dunn index
- **Elbow Method**

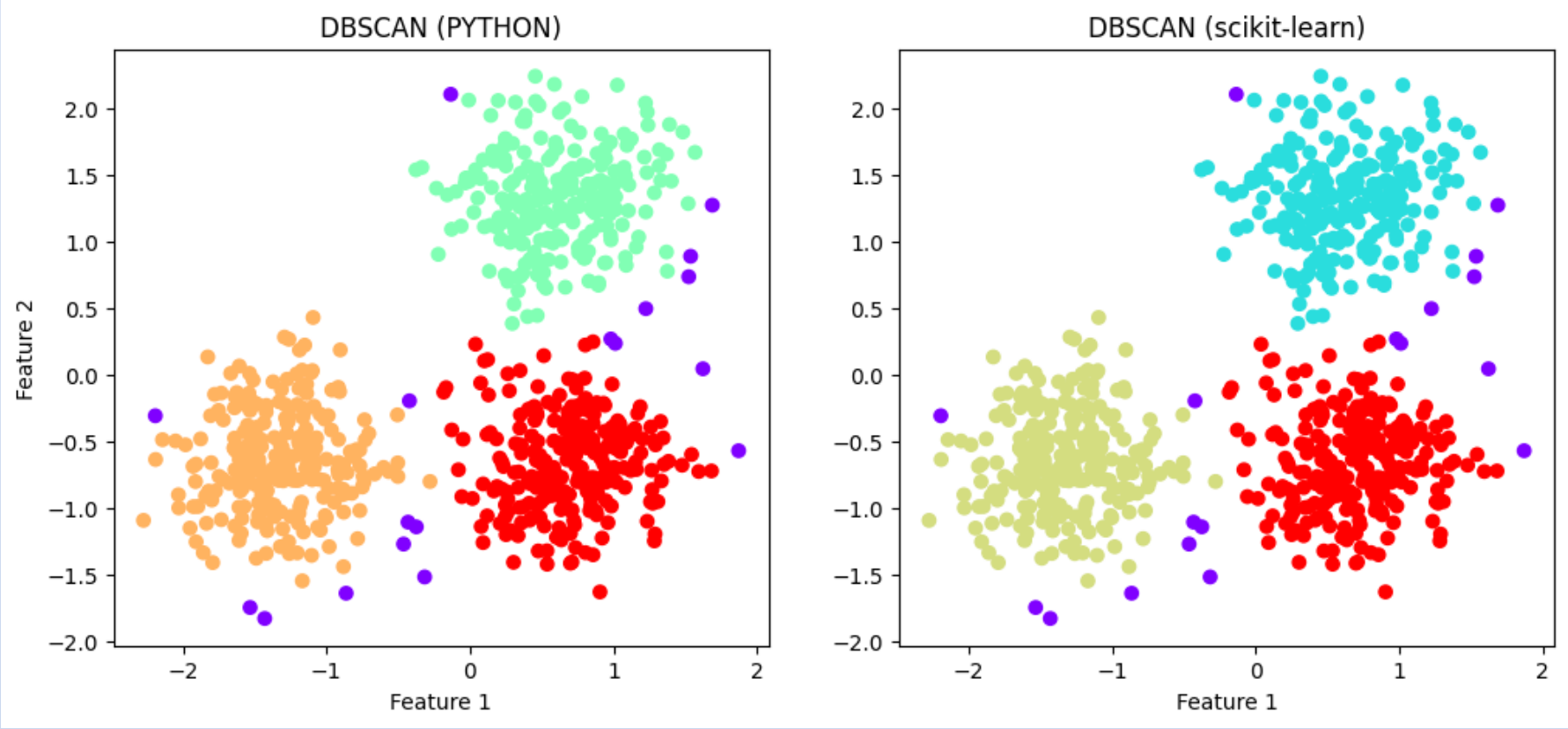
Приклад





Дивись lec_03_04_Exmpl_1.md

Приклад



Дивись lec_03_04_Exmpl_1.md

DBSCAN

Сильні сторони :

- Не потрібно вказувати апріорну кількість ***k*** кластерів, як це потрібно для ***k-means***.
- Знаходить кластери довільної форми. За допомогою параметру ***MinPts*** можна позбутись ефекту одного зв'язку, коли різні кластери зв'язані тонкою лінією.
- Має поняття шуму, і є надійним для виявлення аномалій.
- Потребує всього два параметри ***Eps***, ***MinPts*** і здебільшого нечутливий до впорядкування точок.
- Параметри ***Eps***, ***MinPts*** можуть бути визначені експертно, якщо дані зрозумілі.

DBSCAN

Недоліки алгоритму:

- DBSCAN не є детерміністичним: точки на межі, які досяжні з декількох кластерів, можуть належати одному або іншому кластеру в залежності від порядку обробки даних..
- Якість кластеризації залежить від функції відстані. Зазвичай евклідова норма. Ця метрика може стати негодящою через так зване прокляття розмірності, що ускладнює пошук придатного значення для ***Eps***.
- Не може кластеризувати набори даних з великим перепадом щільностей, оскільки неможливо підібрати поєднання значень ***Eps***, ***MinPts***, яке б відповідало різним кластерам.

Модифікації DBSCAN

OPTICS (*Ordering points to identify the clustering structure*) — вирішує проблему визначення значущих кластерів в наборах даних різної щільності.

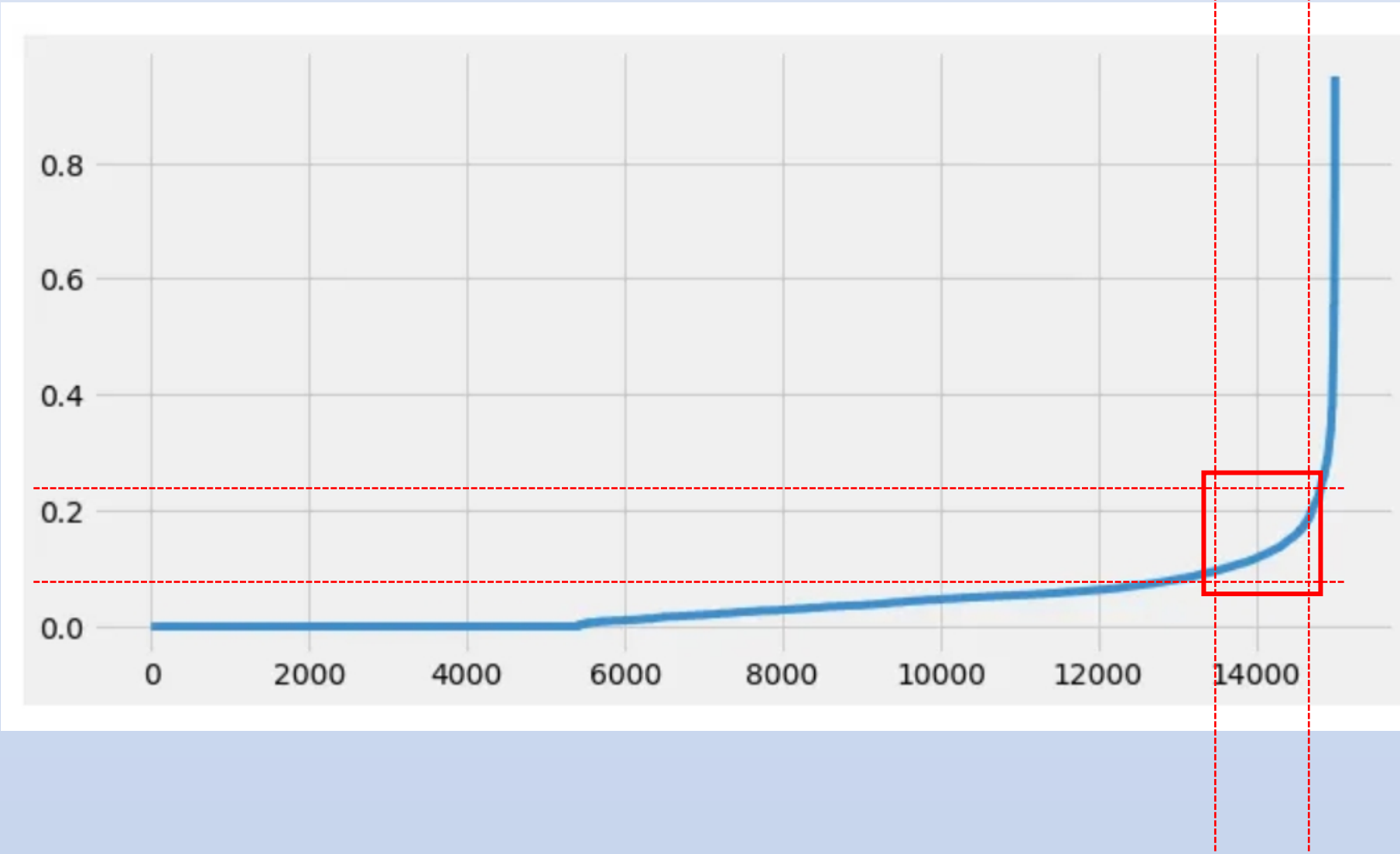
HDBSCAN: ієрархічний варіант DBSCAN, швидший за OPTICS, в якому можна отримати розбиття на найбільші кластери з ієрархії.

Elbow Method \ Метод ліктя (коліна)

Метод ліктя — це евристика, яка використовується для визначення кількості кластерів у наборі даних. Метод складається з побудови відносини середньої відстані в кластерах до середньої відстані між кластерами як функції кількості кластерів і подальшого вибору ліктя (коліна) кривої як кількості кластерів для використання.

Можна використовувати для вибору кількості параметрів в інших керованих даними моделях, таких як ***eps*** у DBSCAN.

Elbow Method \ Метод ліктя (коліна)



Дивись lec_03_04_Exmpl_2.md

Контрольні запитання

- **Надайте загальну постановку задачі кластеризації.**
- **Пояснить сутність алгоритму DBSCAN для вирішення задачі кластеризації**
- **Опишіть метод «локтя» для оцінки якості вирішення задачі кластеризації та визначення *Eps* для алгоритму DBSCAN**

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

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

Sklearn clustering
https://scikit-learn.org/stable/modules/clustering.html#dbscan

Determining the number of clusters in a data set
https://en.wikipedia.org/wiki/Determining_the_number_of_clusters_in_a_data_set

Elbow Method
https://en.wikipedia.org/wiki/Elbow_method_(clustering)

The END
Модуль 3. Лекція 04.