

ОСНОВИ СИСТЕМ ШТУЧНОГО ІНТЕЛЕКТУ, НЕЙРОННИХ МЕРЕЖ та ГЛИБОКОГО НАВЧАННЯ

Модуль 3. Навчання без вчителя

Лекція 3.1.

Кластеризація. Загальні визначення.

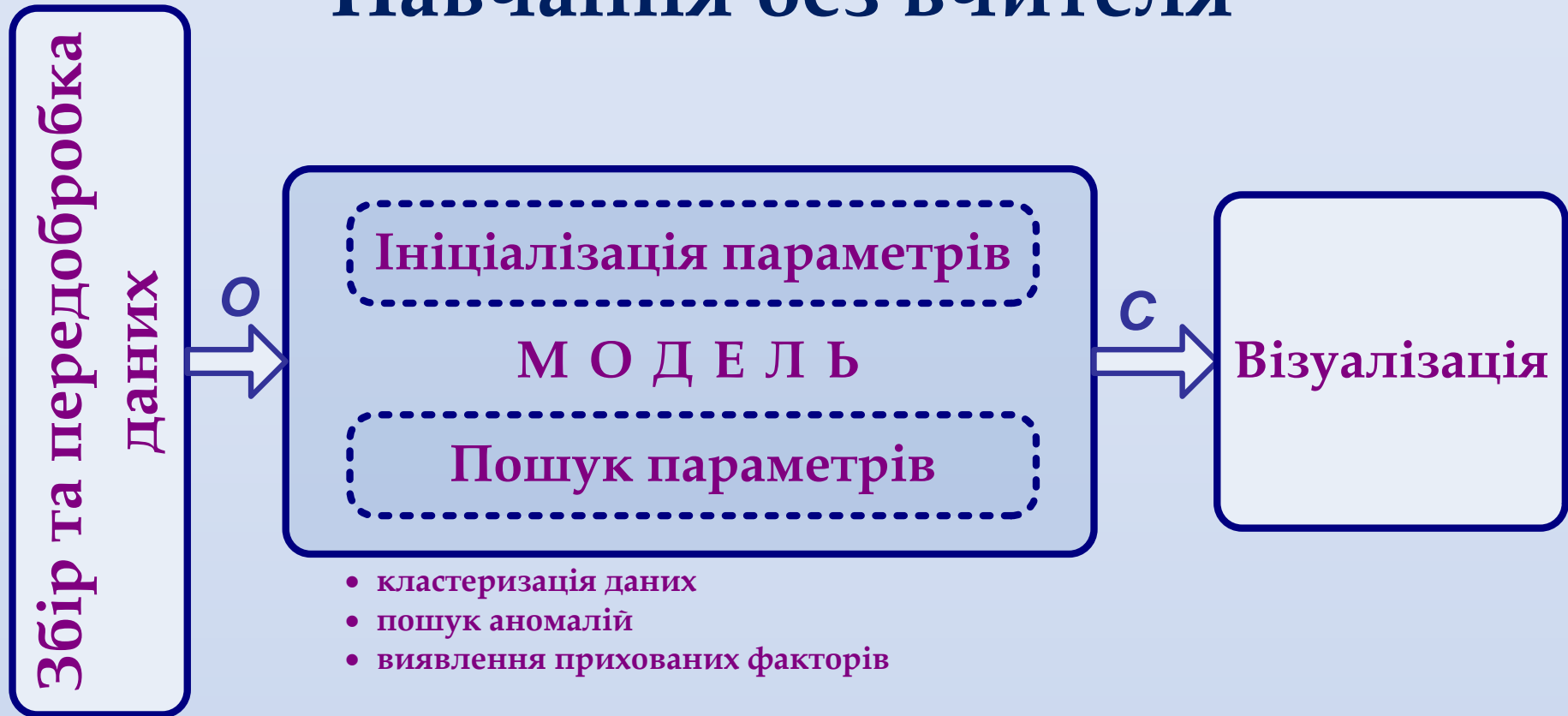
Класичний AI / Класичний ML



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

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

Навчання без вчителя



!!! Ніяких міток (Data are not Labeled)

Класичний AI / Класичний ML



Завдання кластеризації полягає в розбитті безлічі об'єктів на групи (кластери), так щоб об'єкти всередині одного кластера були більш схожі один з одним, ніж об'єкти із різних кластерів.

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

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 множину об'єктів $o^{(j)}$, $j = 1, 2, \dots, M$ 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 \rightarrow клас $c^{(k)}$.

Приклади завдань кластеризації

Медична діагностика: угруповання пацієнтів за схожими признаками

Сегментація ринку: поділ клієнтів на групи за схожими характеристиками

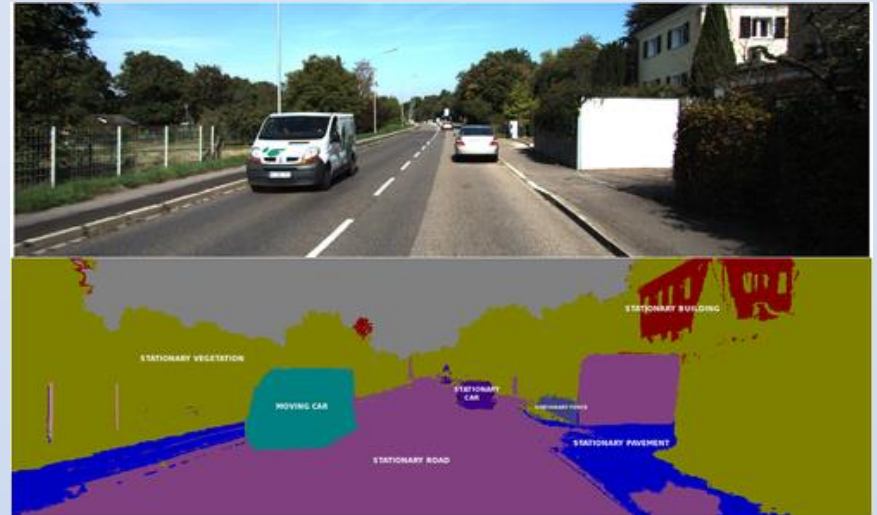
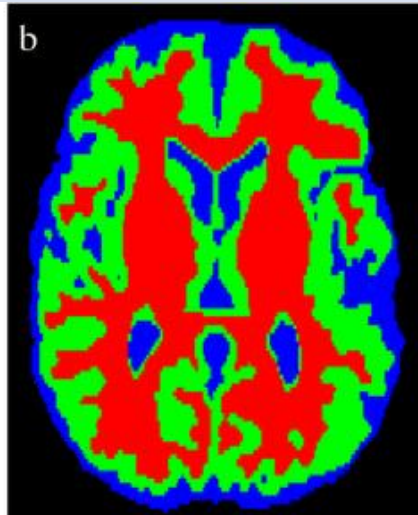
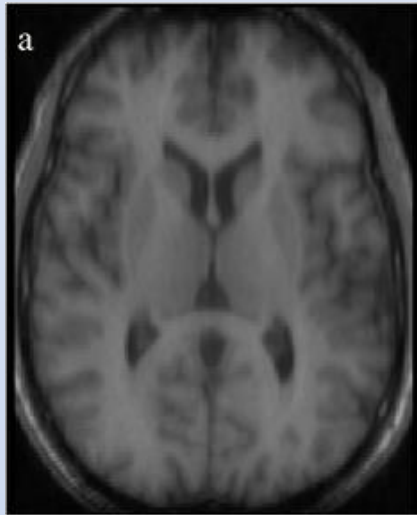
Аналіз соціальних мереж: виявлення спільнот за схожими інтересами (музика, спорт, політика ...)

Обробка зображень: групування зображень за кольором, текстурою, формами

Обробка текстових документів: групування за тематикою, стилем, ...

Приклад: СЕГМЕНТАЦІЯ

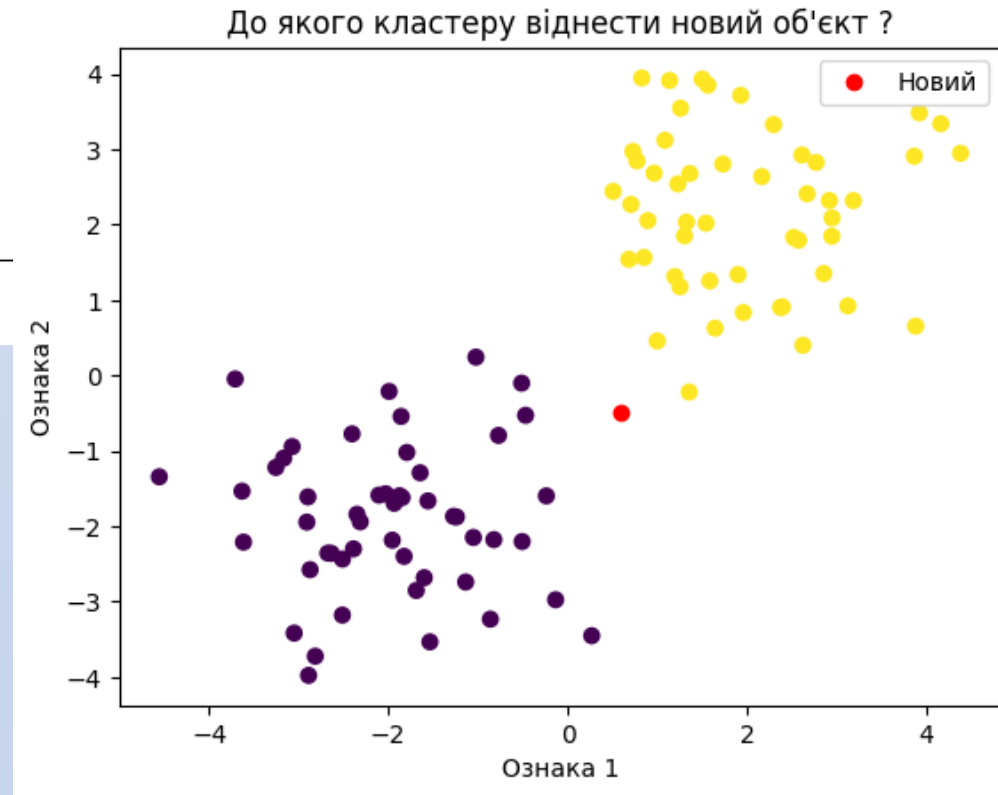
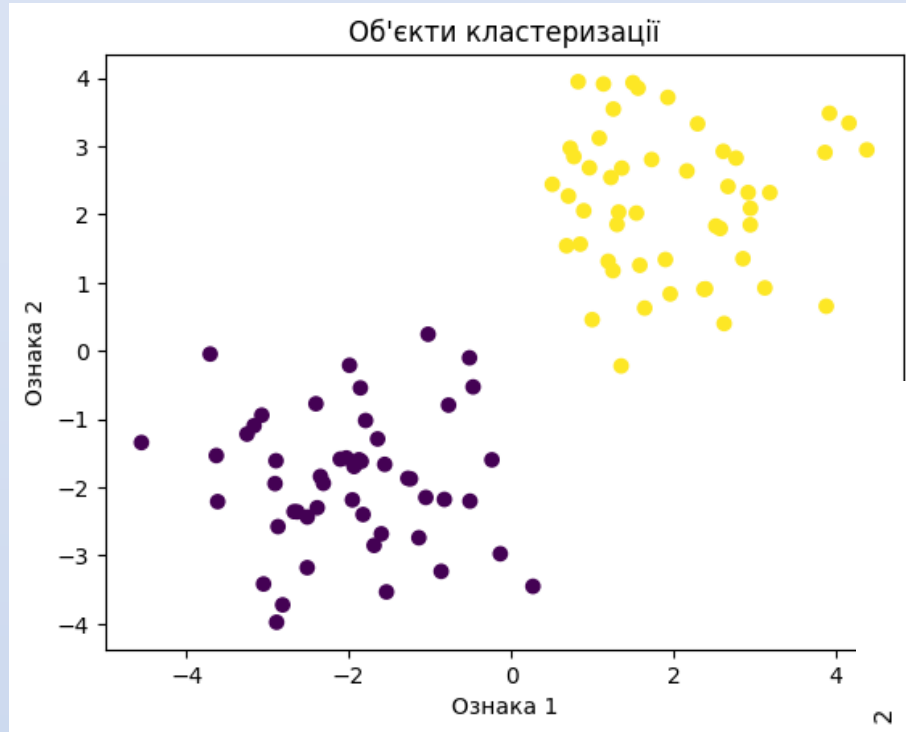
Сегментація: поділ зображення на об'єкти (предмети, фонові області)



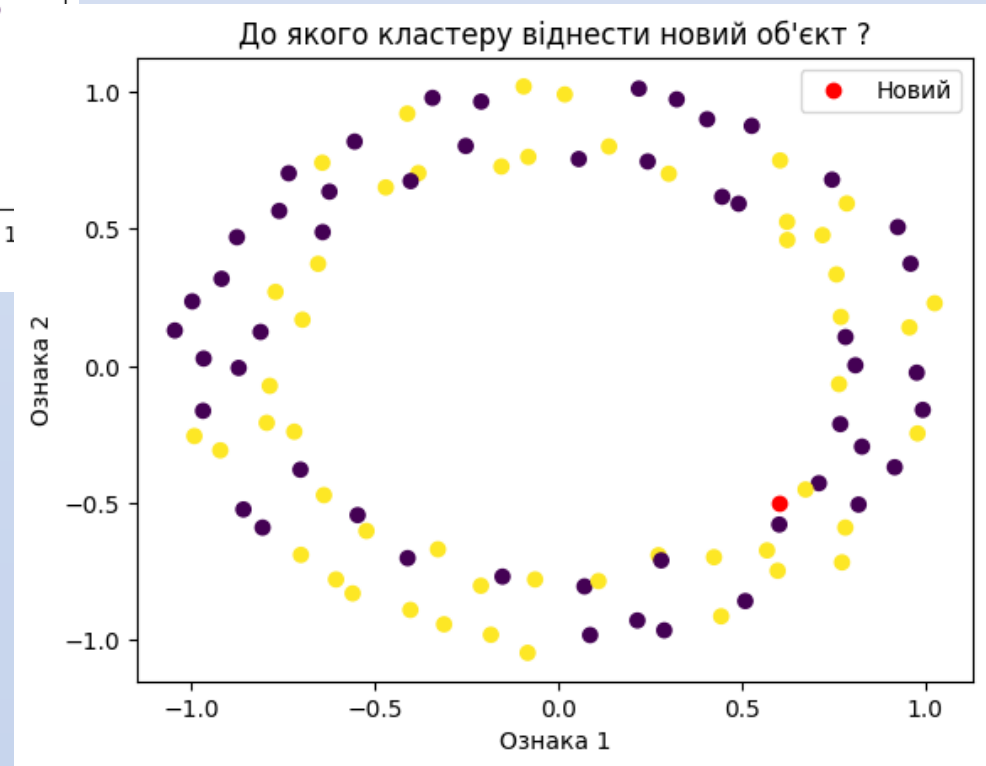
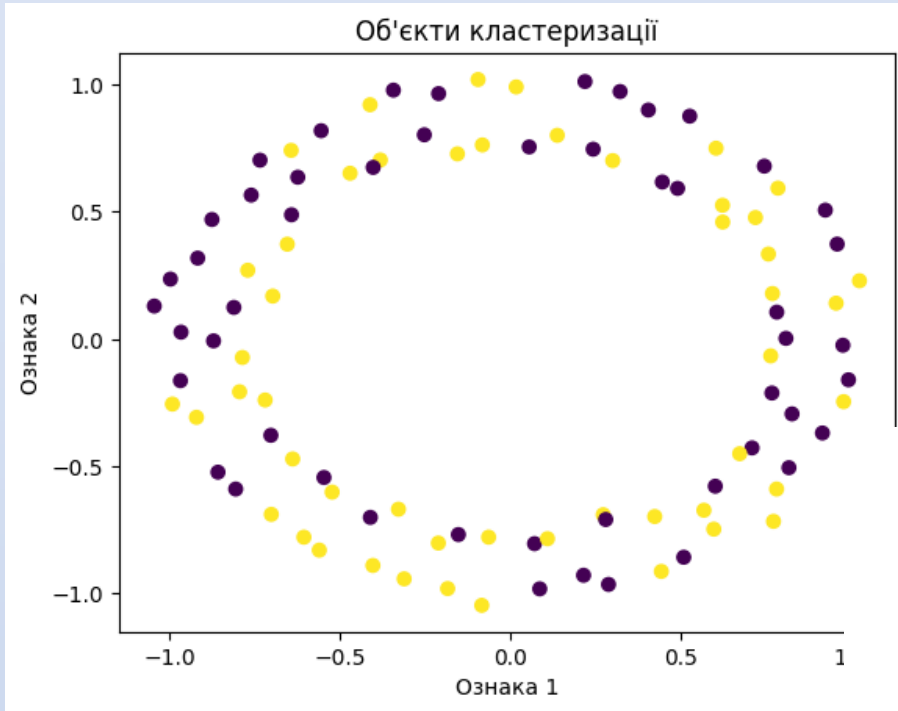
Сегментація медичних зображень для виявлення пухлин та інших аномалій.

Сегментація в автономних автомобілях для виявлення пішоходів та інших об'єктів на дорозі

Проблеми кластеризації



Проблеми кластеризації



Проблеми кластеризації

1. Відсутність єдиного "правильного" рішення:
Не існує універсального алгоритму кластеризації, який би підходив для всіх задач.
2. Визначення оптимальної кількості кластерів.
3. Існує багато критеріїв якості кластеризації.
4. Чутливість до шуму та викидів.
5. Інтерпретація результатів.
6. Прокляття розмірності.

Контрольні запитання

- Пояснить сутність машинного навчання без вчителя.
- Надайте загальну постановку задачі кластеризації.
- Надайте прикладі практичної задачі кластеризації.
- Визначте проблеми вирішення задач 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_\(clustering\)](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 01.