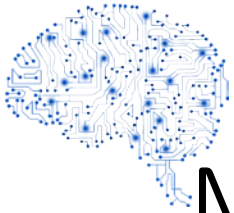
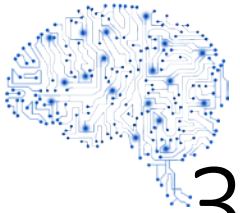
Кластеризація

Професор, д.е.н. Ставицький А.В.



Мета кластеризації

- Пошук існуючих структур.
- Кластеризація є описовою процедурою, вона не робить жодних статистичних висновків, але надає можливість провести розвідницький аналіз та вивчити структуру даних.
- Кластерний аналіз призначений для розбиття множини об’єктів на визначене чи невідоме число класів (= кластерів) згідно з певним критерієм якості класифікації
- Кластер – це група об’єктів зі спільними властивостями.
- Характеристики кластера:
  - внутрішня однорідність;
  - зовнішня ізольованість



Завдання кластерного аналізу:

- розробка типології або класифікації
- дослідження корисних концептуальних схем групування об’єктів
- представлення гіпотез на основі дослідження даних
- перевірка гіпотез або досліджень для визначення, чи дійсно типи (групи), виділені тим або іншим способом, присутні у наявних даних

Як правило, при практичному використанні кластерного аналізу одночасно вирішується декілька завдань.

Неформальні вимоги

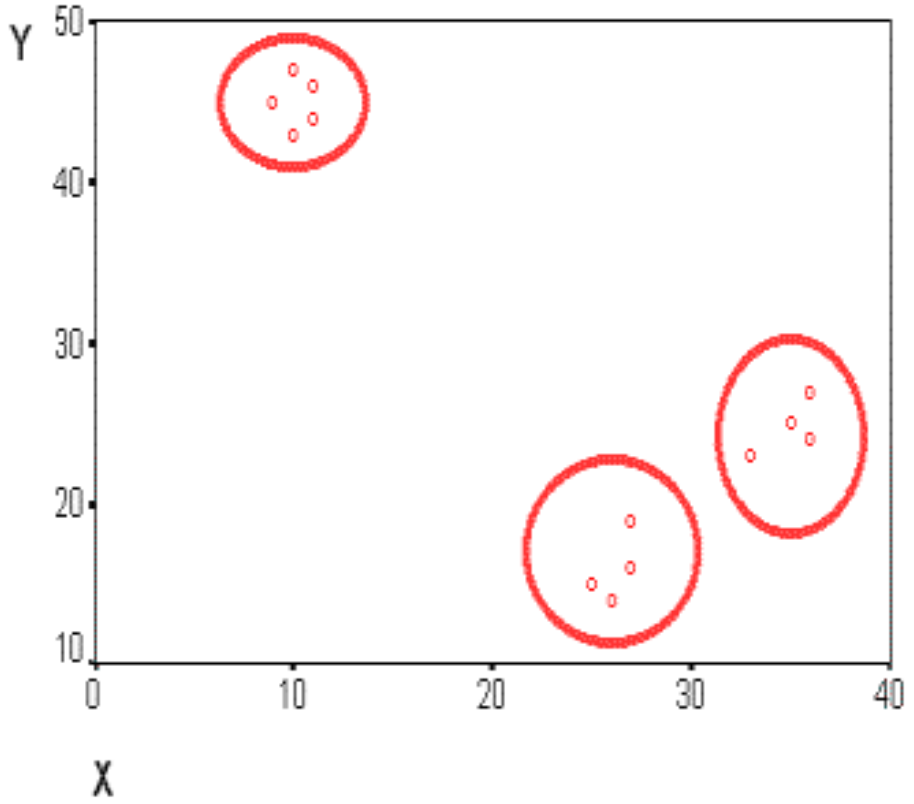
- всередині кластерів об’єкти повинні бути тісно пов’язані між собою
- об’єкти різних кластерів повинні бути далекими одне від одного
- за інших рівних умов розподіли об’єктів по кластерам повинні бути рівномірними

Приклади конкретних задач кластеризації

- Сегментація цільової аудиторії сайту
- Ідентифікація груп сімей — споживачів певного товару для розробки стратегії позиціонування бренду
- Тематичне моделювання електронних листів
- Кластеризація символів в незалежності від їх шрифту, розміру тощо (для подальшого розпізнавання)

Процедура кластерного аналізу

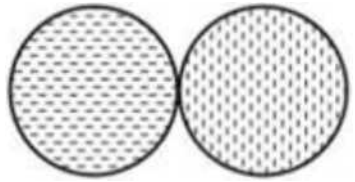
Номер за порядком	Ознака X	Ознака Y
1	27	19
2	11	46
3	25	15
4	36	27
5	35	25
6	10	43
7	11	44
8	36	24
9	26	14
10	26	14
11	9	45
12	33	23
13	27	16
14	10	47



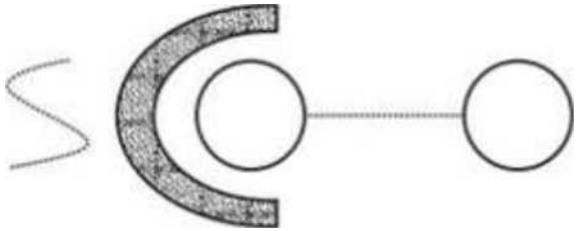
Діаграма розсіювання змінних X та Y

Типи кластерів

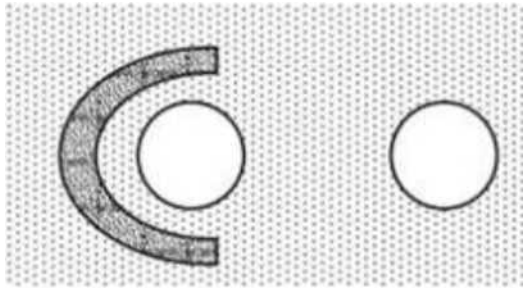
(a) Кластери відокремлені. Кожна точка знаходиться ближче до всіх точок свого кластера, ніж до будь-якої точки іншого кластера.



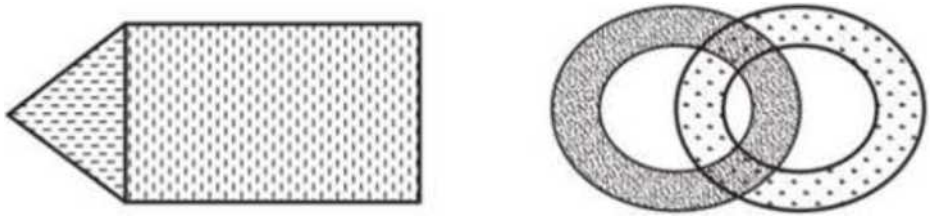
(b) Кластери на основі центрів. Точка знаходиться ближче до центру свого скупчення, ніж до центру будь-якого іншого скупчення.



(c) Кластери з сусідством. Кожна точка ближча принаймні до однієї точки в своєму скупченні, ніж до будь-якої точки в іншому скупченні.

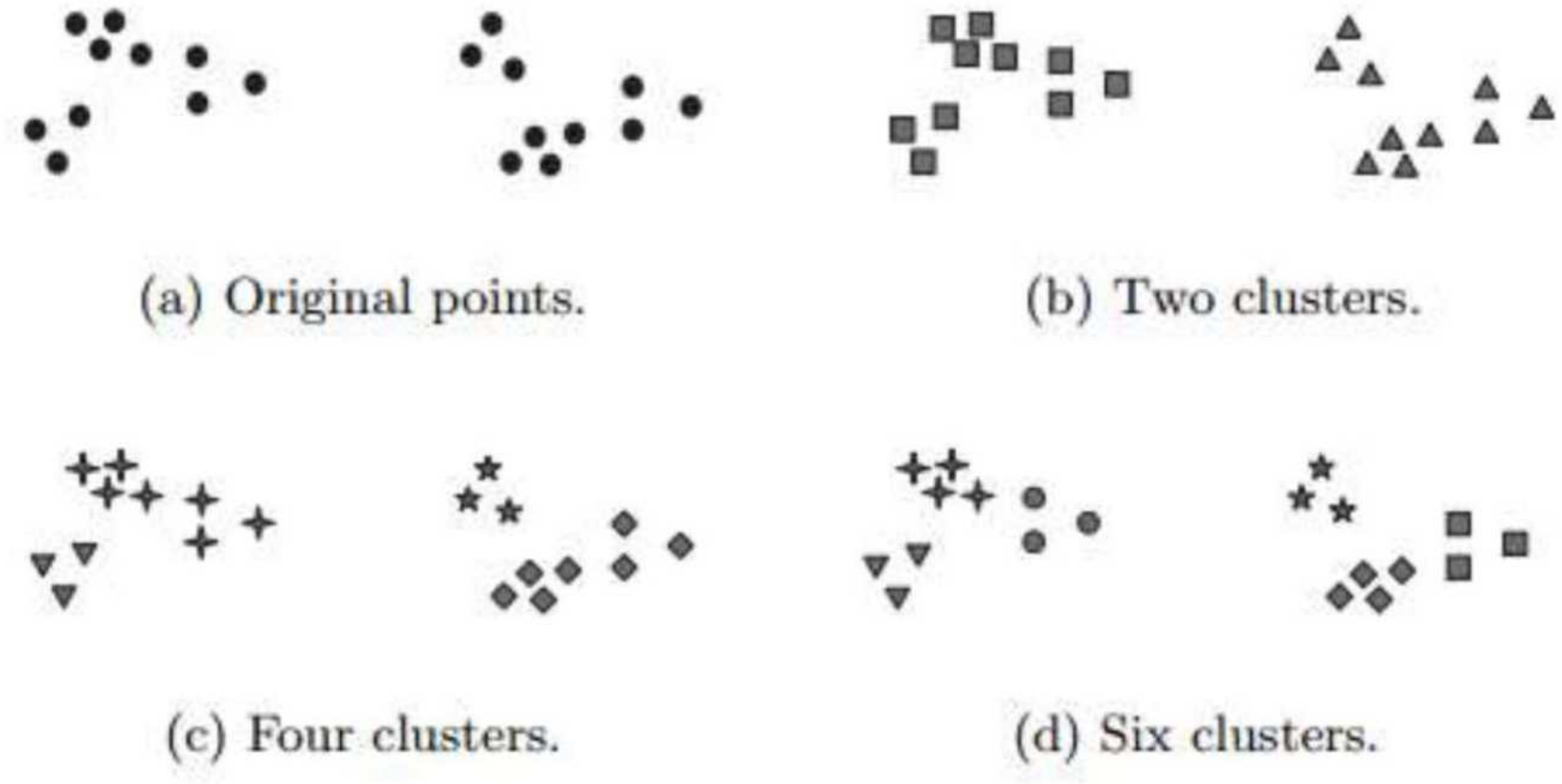


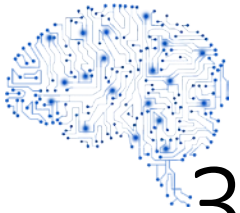
(d) Кластери на основі щільності. Кластери - це регіони високої щільності «сепаратистів» за регіонами низької щільності.



(e) Концептуальні кластери. Точки в кластері мають спільну властивість, яка походить від усього набору точок. (Точки в перетині кіл належать обом.)

20 точок, але скільки кластерів?





Задача кластеризації ставиться нечітко

- Невідомі властивості кластерів
- Невідома їх кількість
- Невідомо, чи є вони взагалі?
- Відсутня навчальна вибірка
- Відсутні очевидні критерії (метрики) якості

- Але на відміну від класифікації не потрібні мітки (не потрібен учитель). Тому можна обробляти суттєво більшу кількість об’єктів!

Питання

- Скільки кластерів?
- Як рахувати відстань між ними (чи їх близькість)?
- За якими правилами об’єднувати елементи в кластери?
- Чи відомі приблизна форма і розмір кластерів?
- Чи можуть вони бути вкладеними?
- Чи можуть об’єкти належати одночасно декільком кластерам?

Результат кластеризації

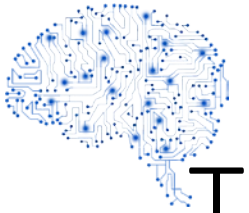
- Розбиття об’єктів на групи
- Знаходження типових точкових представників класів
- Знаходження нетипових представників класів (викидів)
- Побудова повної ієрархії груп об’єктів (таксономія)
- Стиснення даних
- Розуміння даних

Процедури оцінки якості кластеризації:

- ручна перевірка
- встановлення контрольних точок та перевірка на отриманих кластерах
- визначення стабільності кластеризації шляхом додавання до моделі нових змінних
- створення та порівняння кластерів з використання різних методів

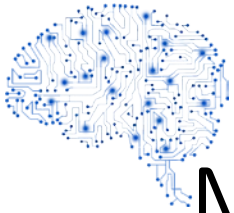
Підходи до кластеризації

- засновані на розділенні даних (Partitioning algorithms), в тому числі ітеративні:
  - розділення об’єктів на k кластерів;
  - ітеративний перерозподіл об’єктів для покращення кластеризації
- ієрархічні алгоритми (Hierarchy algorithms):
  - агломерація
- методи, засновані на концентрації об’єктів (Density-based methods):
  - засновані на можливості з’єднання об’єктів
  - ігнорують шуми, знаходження кластерів довільної форми
- грід-методи (Grid-based methods):
  - квантування об’єктів у грід-структури
- модельні методи (Model-based):
  - використання моделі для знаходження кластерів, що найбільше відповідають даним



Типи кластерних алгоритмів

- Пласкі алгоритми:
  - починають роботу розділенням елементів по групах випадковим чином
  - ітеративно покращують результат
  - головний алгоритм: A-середніх (A-means)
- Жорстка кластеризация:
  - кожен елемент належить строго одному кластеру
- Алгоритми ієрархічної кластеризації:
  - створюють ієрархію
  - знизу-вгору (агломеративні)
  - зверху-вниз (розділяючі)
- М'яка кластеризация:
  - елемент може належати кільком кластерам
  - головний алгоритм: C-середніх (C-means)



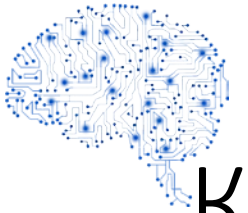
Математичні характеристики кластера:

- Центр – середнє геометричне місце точок у просторі змінних;
- Діаметр: максимальна відстань між будь-якими двома точками в кластері
- Радіус: максимальна відстань від якогось «центру» до будь-якої з точок кластера
- Щільність: кількість точок в кластері, поділена на «обсяг», тобто на радіус в якійсь степені
- Міжкластерна відстань: відстань між центрами, між найближчими точками, середня відстань між усіма парами
- Середньоквадратичне відхилення
- Розмір кластера

- Спірний об’єкт – це об’єкт, який за мірою подібності може бути віднесений до декількох кластерів. Об’єкт відноситься до кластера, якщо відстань від об’єкта до центра кластера менша за радіус кластера. Якщо ця умова не виконується для двох та більше кластерів, об’єкт є спірним. Неоднозначність даного завдання може бути усунена експертом або аналітиком.

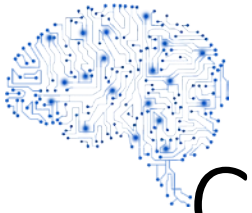
Центри кластерів

- В евклідовому просторі є центр ваги (центроїд) — середнє арифметичне координат точок кластера
- У неевклідовому просторі (наприклад, в просторі слів) центроїда немає. Центром (кластроїдом) вибирається одна з точок кластера, що мінімізує
  - максимальну відстань до інших точок
  - суму відстаней
  - суму квадратів відстаней



Критерії зупинки кластеризації

- Побудована потрібна кількість кластерів
- Характеристики кластерів (діаметр, щільність, ...) досягли граничних значень
- Кластери не змінилися під час останньої ітерації



Способи нормування вихідних даних:

- Z-шкали (Z-Scores) – із значень змінних віднімається їх середнє, й ці значення діляться на стандартне відхилення;
- максимум 1 (значення змінних діляться на їх максимум);
- середнє 1 (значення змінних діляться на їх середнє);
- розкид від -1 до 1 (лінійним перетворенням змінних домагаються розкиду значень від -1 до 1);
- розкид від 0 до 1 (лінійним перетворенням змінних домагаються розкиду значень від 0 до 1).

Метод К-середніх

- Кожен кластер визначається своїм центроїдом
- Критерій кластеризації: мінімізувати усереднену квадратичну відстань від центроїда
- Визначення центроїда:

$$\vec{\mu}(\omega) = \frac{1}{|\omega|} \sum_{\vec{x} \in \omega} \vec{x}$$

де w позначає кластер

- Ітеративно застосовуємо два кроки алгоритму:
  - перерозподіл: зараховуємо кожен об’єкт до найближчого центроїду
  - перерахунок: заново розраховуємо кожен центроїд як середнє об’єктів, віднесених до кластеру на попередньому кроці

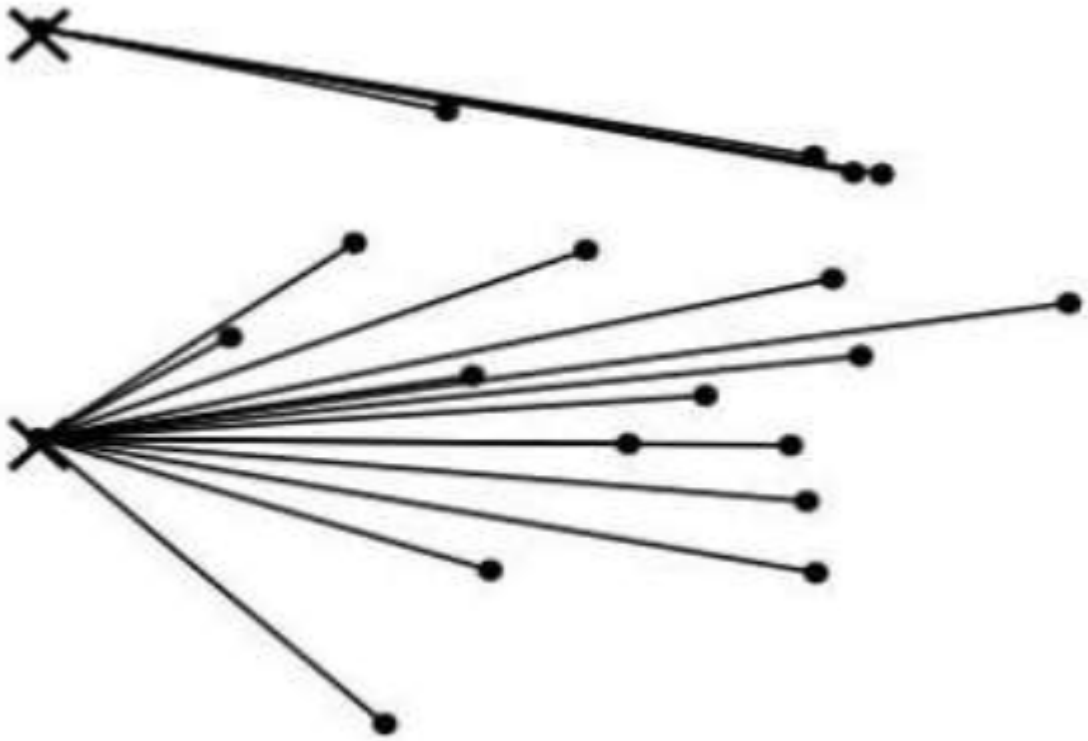
Приклад: кластеризация набору даних методом К-середніх – 1

Приклад: кластеризация набору даних методом К-середніх – 2

- Вибираємо випадковим чином два центроїда

Приклад: кластеризация набору даних методом К-середніх – 3

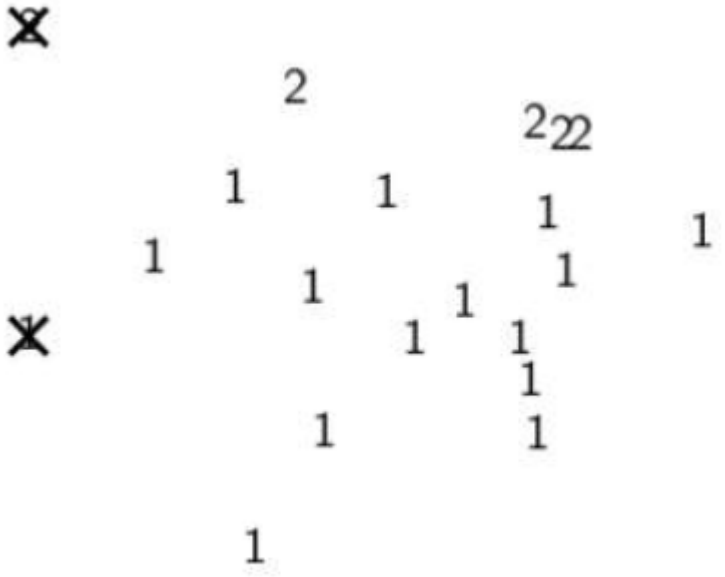
- Ітерація 1. Розподіляємо кожну точку до найближчого центроїда



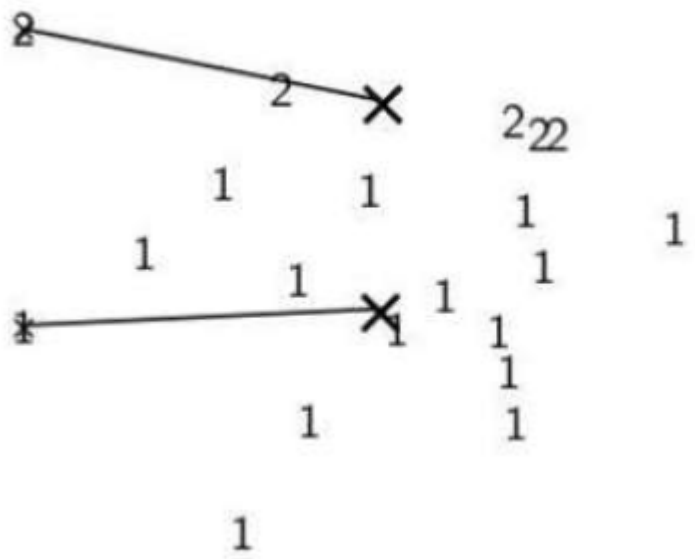
Приклад: кластеризация набору даних методом К-середніх – 4

- Ітерація 1.

Результат розподілу

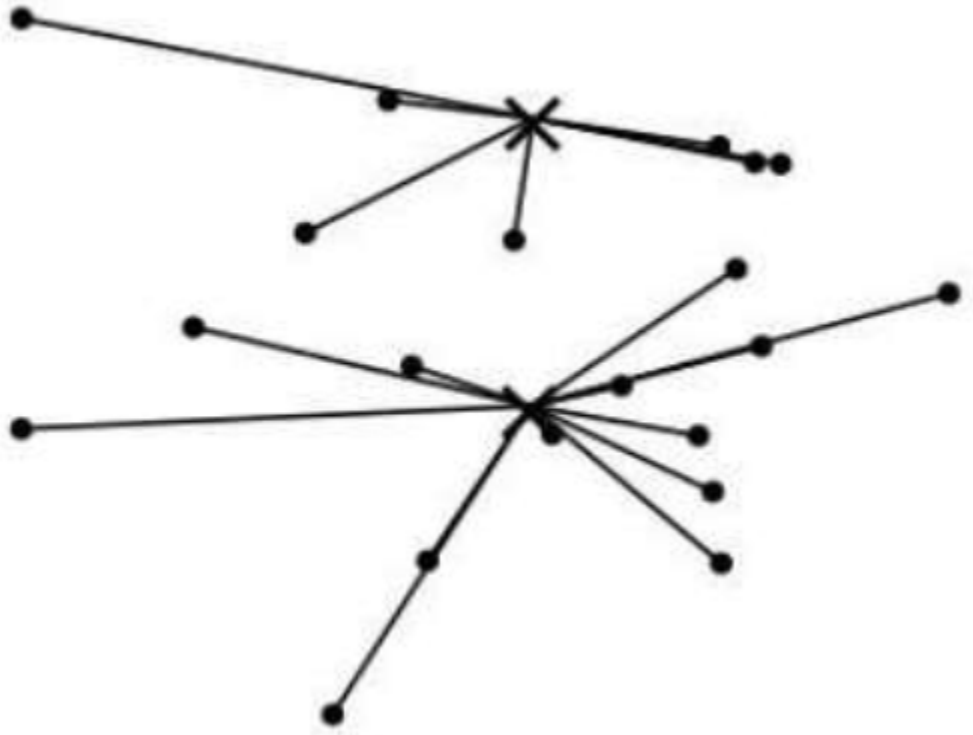


Перераховуємо центроїди кластерів



Приклад: кластеризация набору даних методом К-середніх – 5

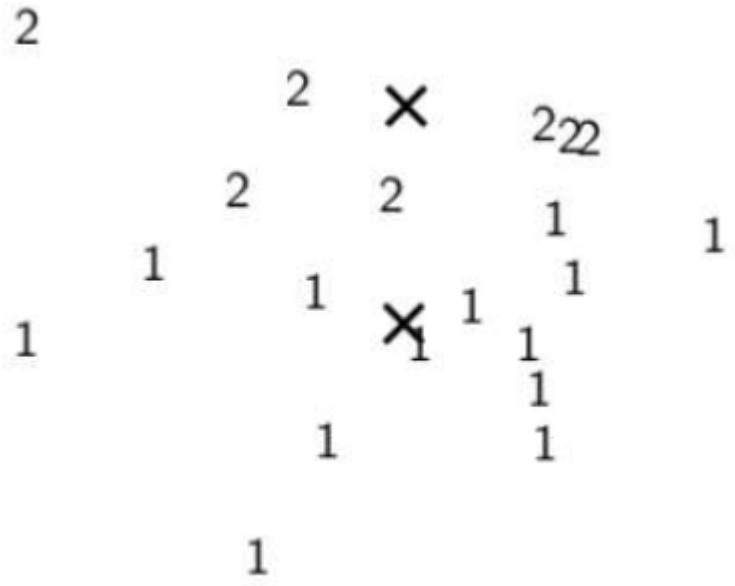
- Ітерація 2. Розподіляємо кожну точку до найближчого центроїда



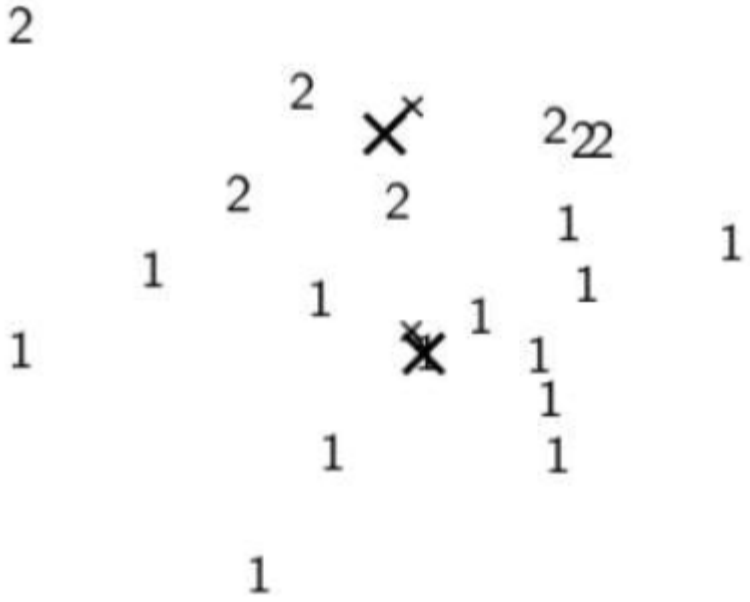
Приклад: кластеризация набору даних методом К-середніх – 6

- Ітерація 2.

Результат розподілу

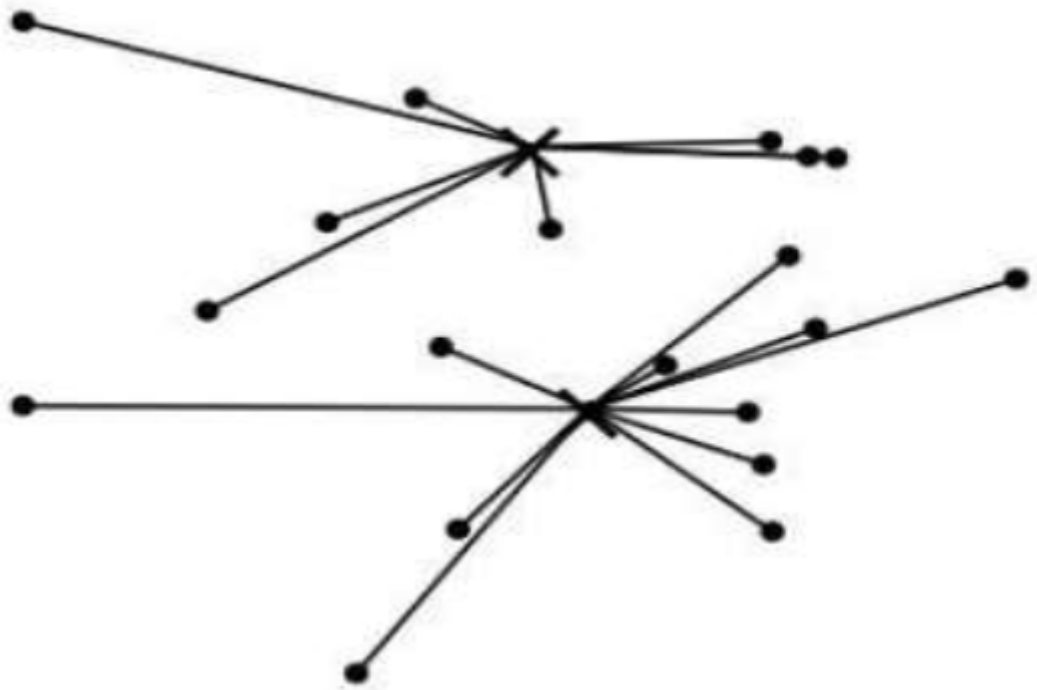


Перераховуємо центроїди кластерів



Приклад: кластеризация набору даних методом К-середніх – 7

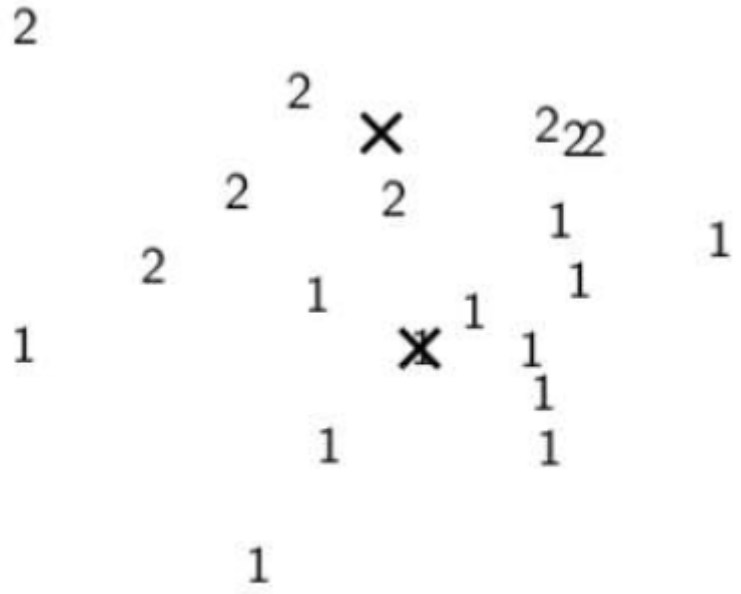
- Ітерація 3. Розподіляємо кожну точку до найближчого центроїда



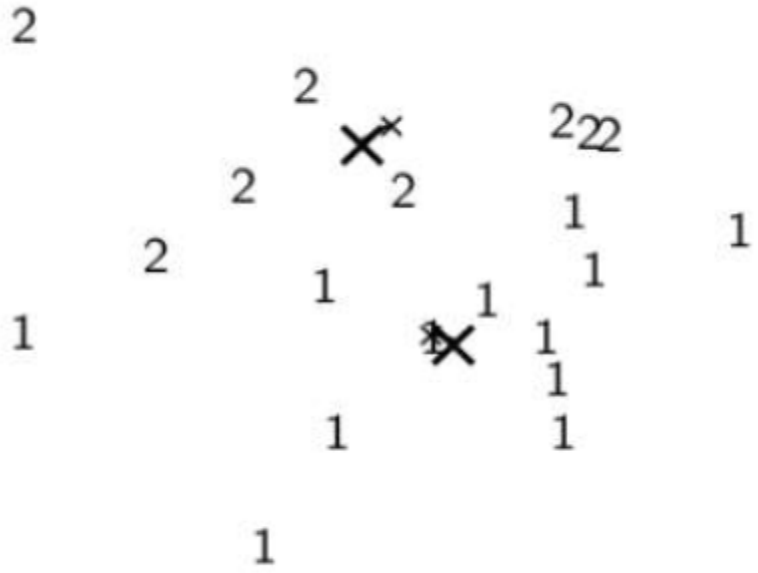
Приклад: кластеризация набору даних методом К-середніх – 8

- Ітерація 3.

Результат розподілу

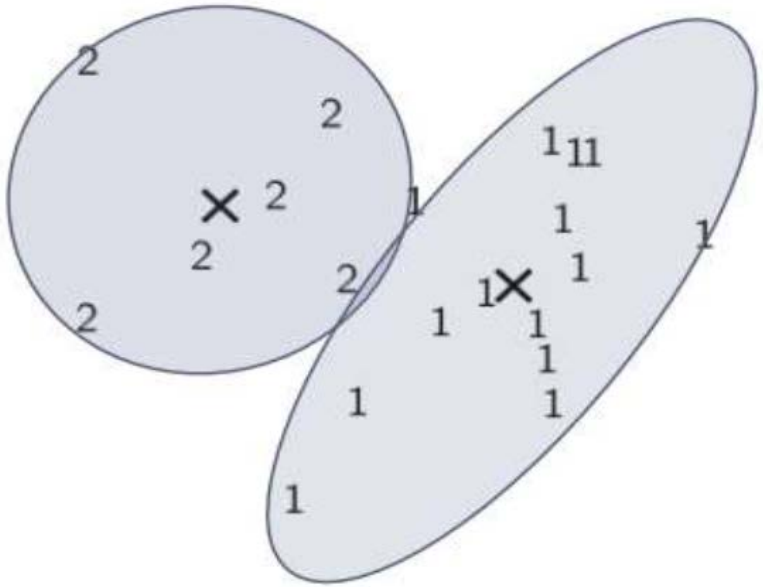


Перераховуємо центроїди кластерів

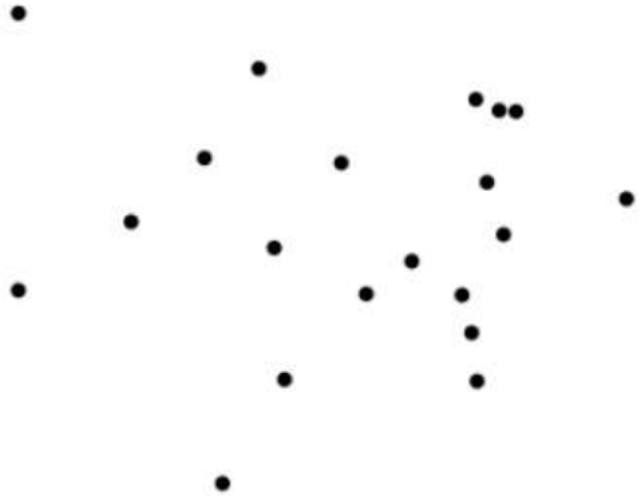


Приклад: кластеризация набору даних методом К-середніх – 9

- Після 8-ї ітерації



Початок був такий

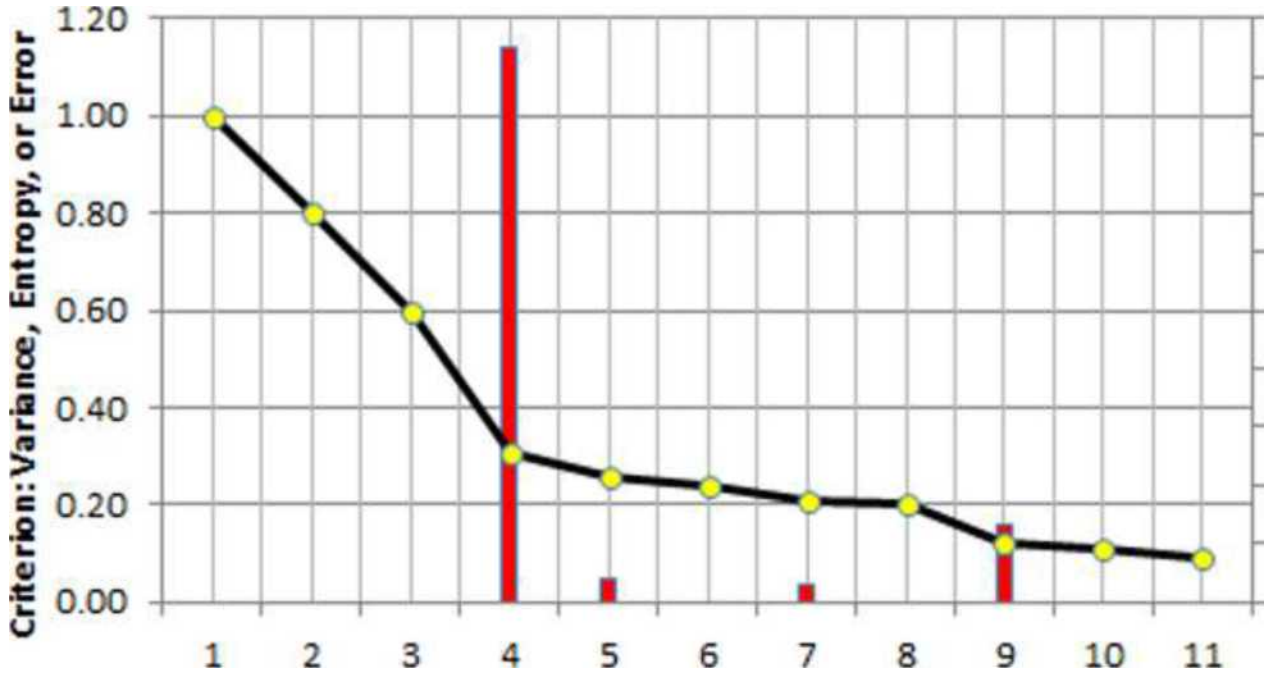


Як визначити кількість кластерів?

- Число кластерів K має бути визначено заздалегідь
  - Евристика: наприклад, знаючи характер об’єктів, припустимо «прийнятне» число кластерів
- Проста цільова функція для пошуку K
  - Починаємо з одного кластера (K = 1)
  - Продовжуємо додавати кластери (= збільшуємо K)
  - Нараховуємо штраф за кожен новий кластер
  - Балансуємо штрафи за нові кластери і вигоду від меншої середньої дистанції від центроїду
  - Вибираємо K з найкращим балансом
  - Для даної кластеризації, визначте вартість штрафу для об’єкта як квадрат відстані до центроїда
  - Загальний штраф для кластера розрахуйте як суму штрафів всіх об’єктів у кластері RSS (K) (Residual Sum of Squares)
  - Кожен кластер додатково штрафується фіксованим параметром λ
  - Цільова функція: мінімізувати RSS (K) + K λ
  - Залишається проблемою як знайти оптимальне значення λ

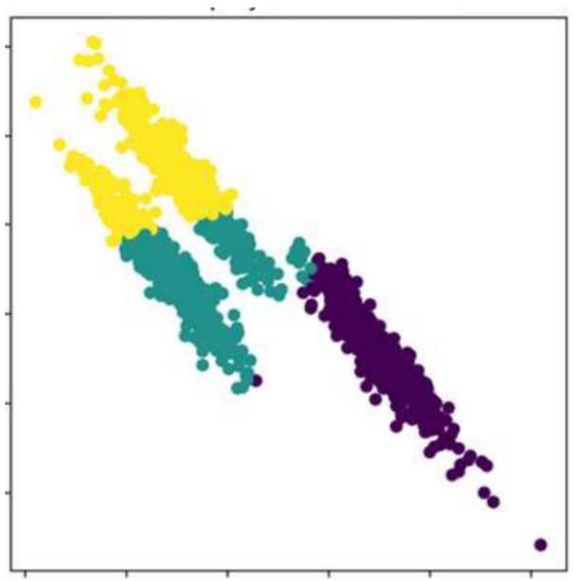
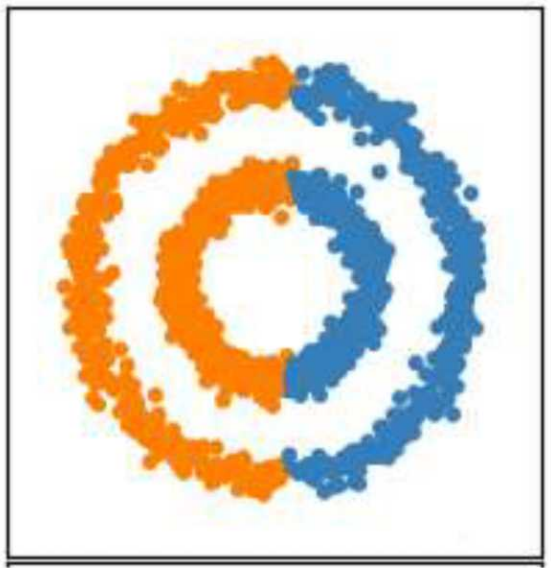
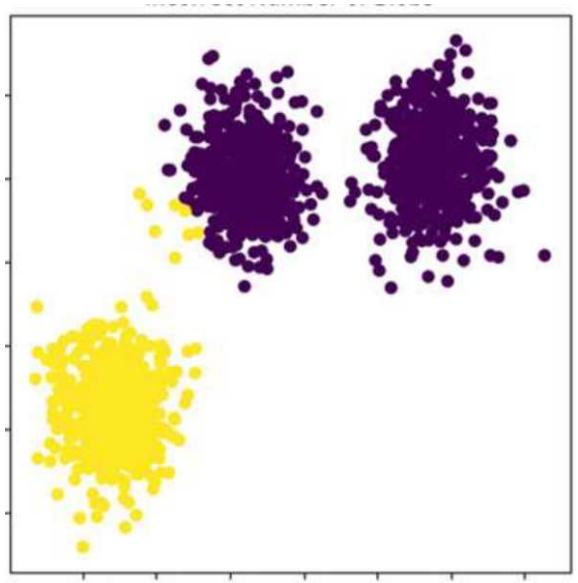
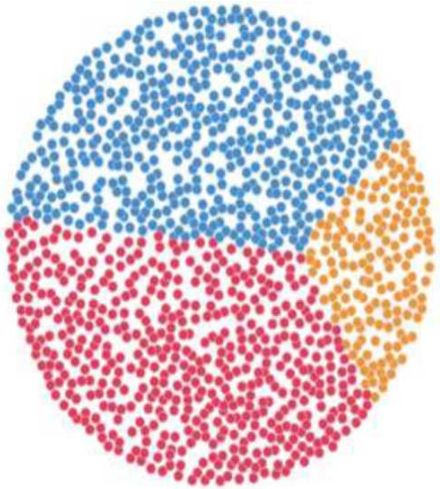
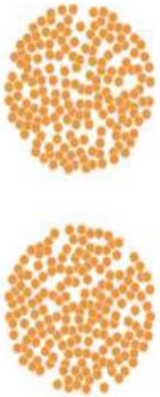
Пошук «ліктя» («elbow») на кривій

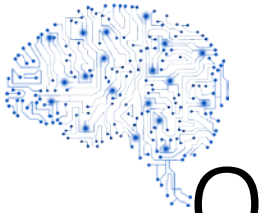
- Виберіть кількість кластерів, при якій крива стає більш «пласкою» - як згин у лікті
- У даному випадку: 4



Проблеми К-середніх

- Ілюстрація ситуацій, коли k-середні засоби створюватимуть неінтуїтивні та, можливо, несподівані кластери. Вхідні дані не відповідають якомусь неявному припущенню, і в результаті виникають небажані кластери



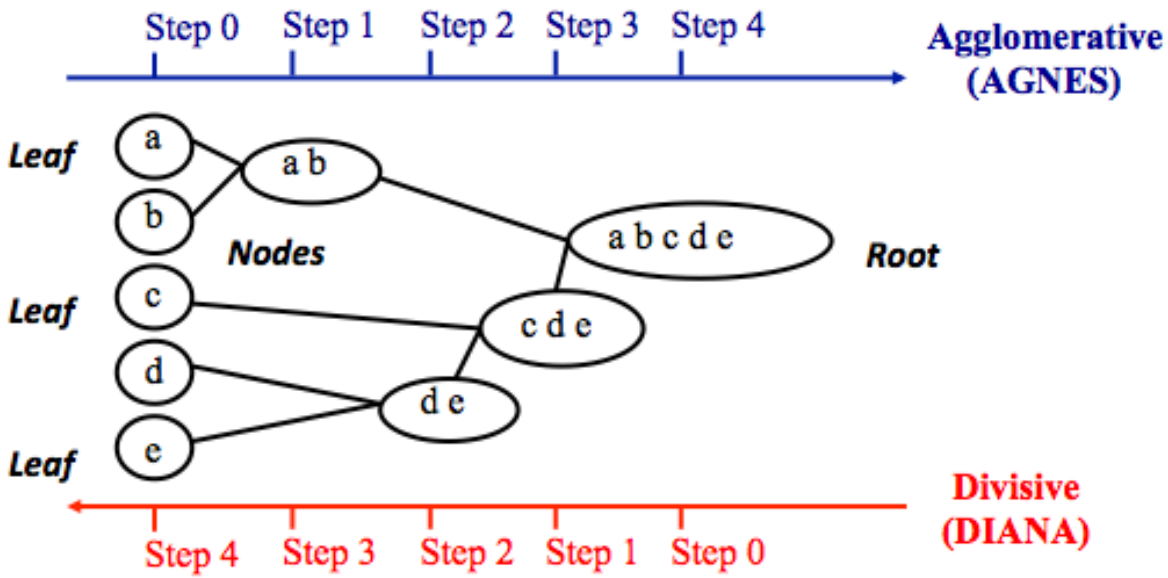
Особливості методу К-середніх

- Кластеризація може завершитись на локальному оптимумі, тому для високоякісного результату необхідна початкова ініціалізація
- Необхідно заздалегідь визначити кількість кластерів
- Чутливість до шумів та викидів
- Можливе застосування тільки для числових даних
- Неможливо будувати кластери неопуклої форми

- Загальна складність алгоритму: O(IKNM). Зазвичай, кількість кластерів (K), розмірність об’єктів (M) і число ітерацій (I) набагато менші кількості об’єктів (N), тому метод є ефективним.

Ієрархічні методи кластерного аналізу

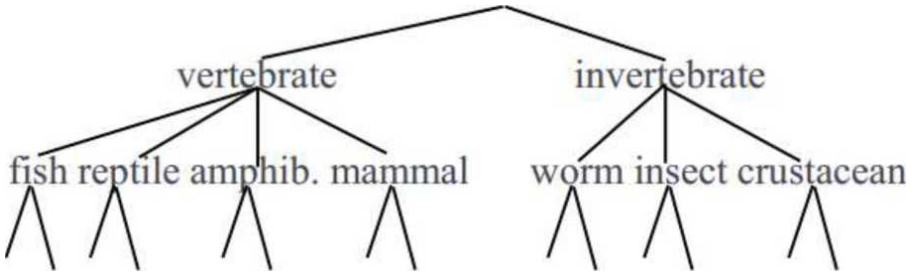
- ієрархічні агломеративні методи (Agglomerative Nesting, AGNES)
- ієрархічні дівізійні (подільні) методи (DIvisive ANAlysis, DIANA)



- Неієрархічні (ітеративні)

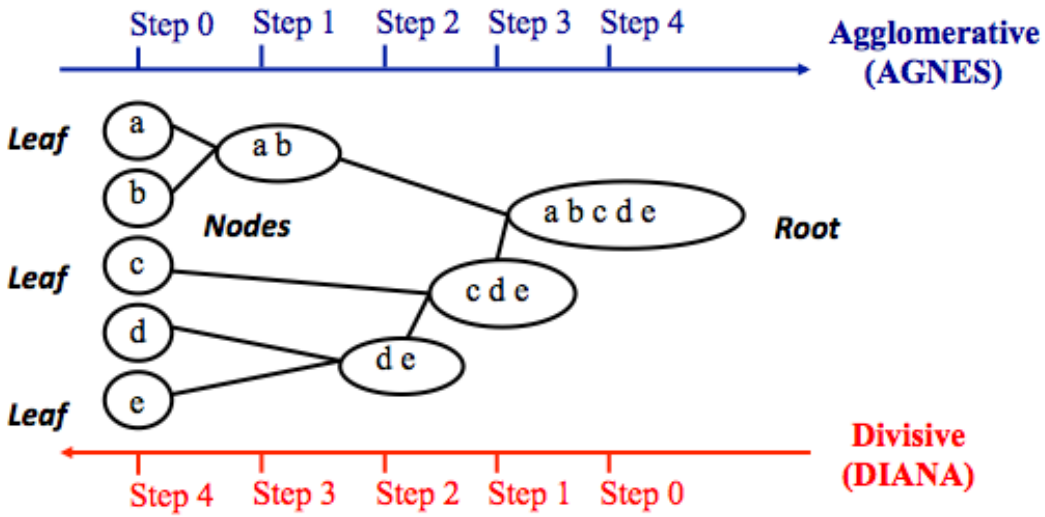
Ієрархічна кластеризация

- Завдання ієрархічної кластеризації — побудувати ієрархію кластерів
- Ієрархія будується автоматично:
  - або згори-вниз (агломеративні алгоритми) - AGNES (AGglomerative NESting): ROCK, CURE, CHAMELEON
  - або знизу-вгору (алгоритми розділення) - DIANA (DIvisive ANAlysis): BIRCH, MST
- Як в агломеративній, так і в роздільній ієрархічній кластеризації, користувачам потрібно вказати бажану кількість кластерів та умову завершення



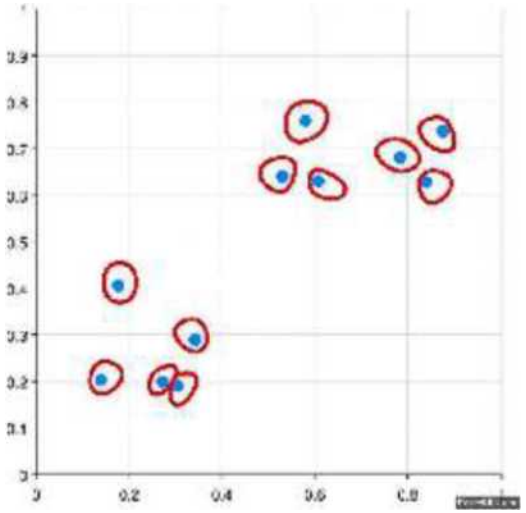
Ієрархічна кластеризация

- На початку роботи агломеративного алгоритму кожна точка розглядається як кластер, потім алгоритм намагається об’єднати найближчі сусідні точки в один більший кластер і так далі, щоб зрештою об’єднати всі кластери в один великий кластер
- Алгоритм розділення спочатку розглядає всі точки множини як один кластер; на подальших кроках деякі кластери вищого рівня рекурсивно розщеплюються для побудови діаграми
- Ці підходи протилежні один одному



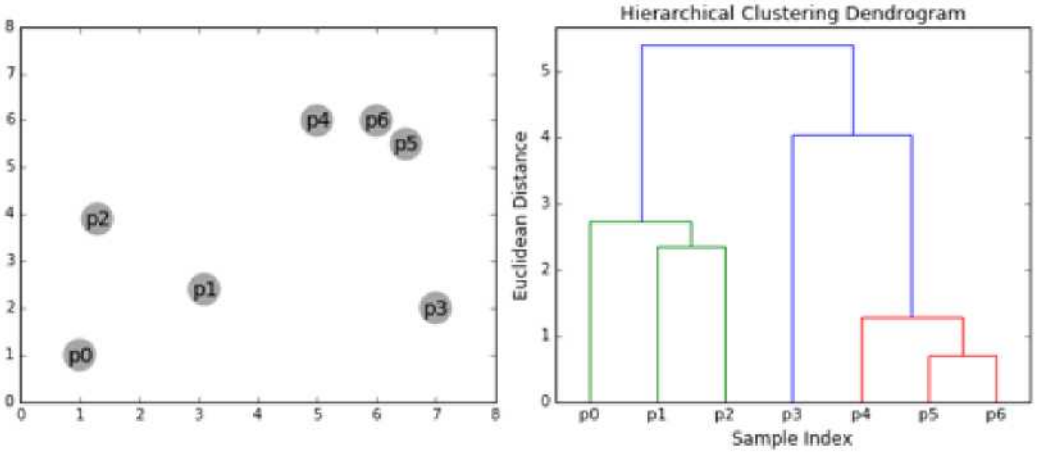
Ієрархічна агломеративна кластеризация

- Найвідоміший метод побудови знизу-вгору: ієрархічна агломеративна кластеризація
- Будує ієрархію у вигляді двійкового дерева
- Використовує міру близькості для визначення подібності двох кластерів
- Алгоритм:
  - Спочатку кожен об’єкт розглядається як окремий кластер
  - По черзі об'єднуємо два найбільш схожих кластера
  - До тих пір поки не залишиться один кластер
  - Історія об'єднань формує дерево ієрархії
  - Така історія зображується дендограмою

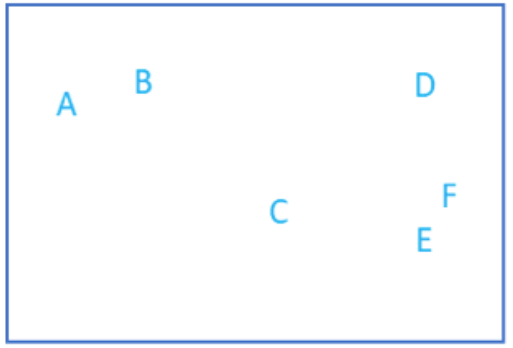


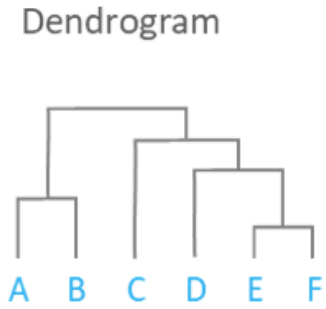
Дендограма

- Дендрограма - це тип деревної діаграми, що показує ієрархічні взаємозв'язки між різними наборами даних
- Дендрограма містить пам'ять ієрархічного алгоритму кластеризації, тому, просто переглянувши дендрограмму, ви можете сказати, як формується кластер



- Відстань між точками даних означає несхожість
- Висота блоків представляє відстань між кластерами

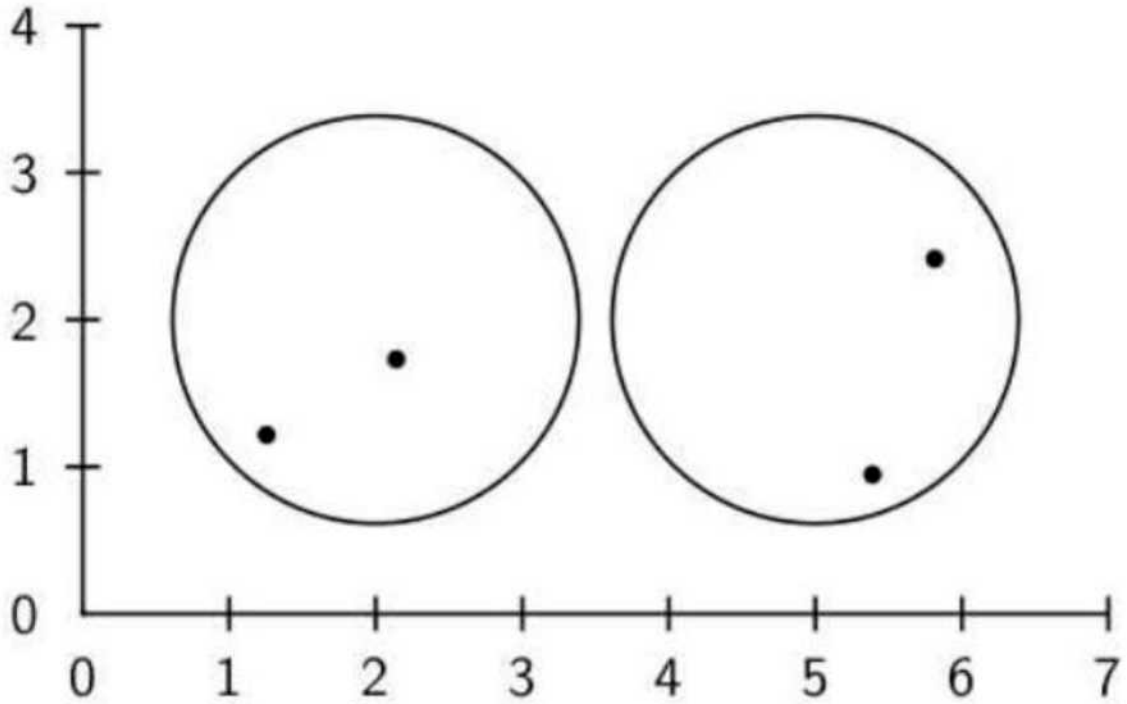




Як обчислити близькість кластерів?

- Попарна близькість
  - Одиночний зв'язок (Single-linkage):
    - максимальна близькість будь-яких двох об’єктів
  - Повний зв'язок (Complete-linkage):
    - мінімальна близькість будь-яких двох об’єктів
- Центроїдна близькість
  - Центроїдний зв'язок (Centroid-linkage):
    - середня близькість всіх пар об’єктів (не виключаючи пари об’єктів всередині кластерів)
    - рівносильно близькості центроїдів
  - Групове-середнє (Average-linkage):
    - середня близькість всіх пар об’єктів, включаючи пари всередині кластерів

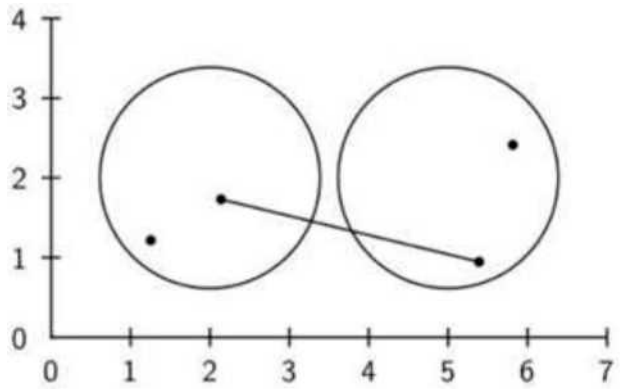
Приклад – 1



Приклад – 2

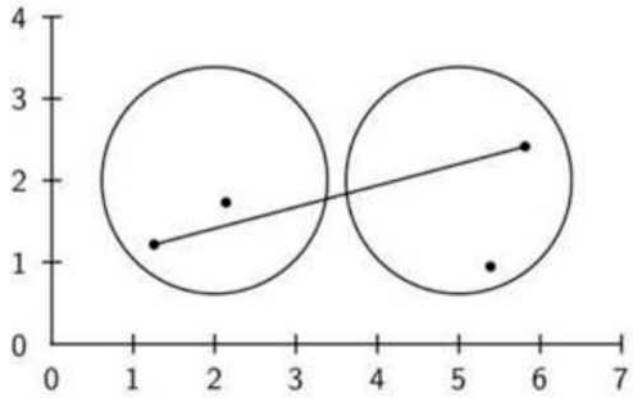
Одиночний зв'язок: максимальна близькість

- подібність двох кластерів - це подібність між їх найбільш подібними членами (найближчий сусід)
- приділяється увага найближчим точкам, ігнорується структура кластера
- можливість будувати кластери неправильної форми
- такий вид зв’язку чутливий до даних з шумами та викидами



Повний зв'язок: мінімальна близькість

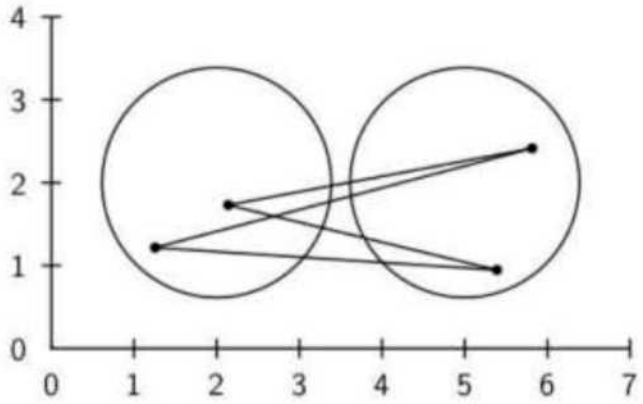
- подібність двох кластерів рахується як подібність їх найменш подібних членів
- два кластери, об’єднуючись, формують кластер з найменшим діаметром
- на виході - кластери компактної форми
- чутливий до викидів



Приклад – 3

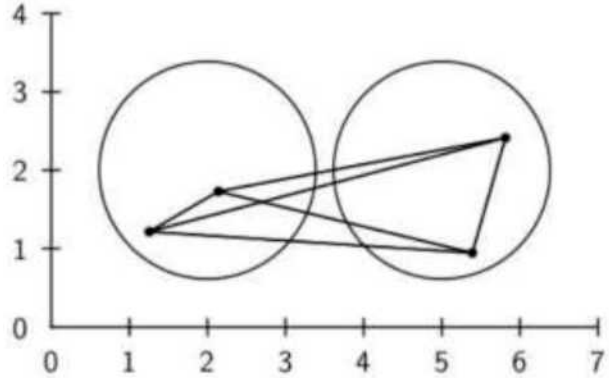
Середня міжкластерна близькість (міжцентроїдна близькість)

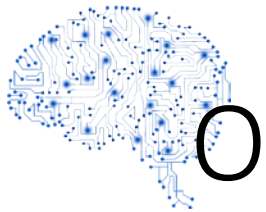
- багато обчислень
- дозволяють об’єднувати в кластери дані без істотних змін через викиди та шуми



Групове-середнє (середня внутрішньокластерна близькість)

- багато обчислень
- дозволяють об’єднувати в кластери дані без істотних змін через викиди та шуми





Обчислювальна складність ієрархічної кластеризації

- Обчислюємо близькість всіх NxN пар об’єктів
- Потім, на кожній ітерації:
- Скануємо O(NxN) близькостей для знаходження максимальної
- Об'єднуємо два кластери
- Обчислюємо близькість між створеним кластером і всіма рештою
- Всього O(N) ітерацій, кожна вимагає O(NxN) сканувань
- Загальна складність: $O(N^3)$
- Існує більш раціональна модифікація алгоритму зі складністю $O(N^2)$

Пласка чи ієрархічна кластеризація?

- Пласка кластеризація значно швидше (O(IKNM) і O(N^2)), добре підходить для великих обсягів даних
- Для стабільного передбачуваного результату використовують ієрархічну кластеризацію
- Ієрархічну кластеризацію також застосовують тоді, коли потрібна структура кластерів
- Іноді ієрархічна кластеризація використовується для визначення числа K, а в подальшому використовується пласка кластеризація

Рішення, що приймаються на етапах кластерного аналізу:

- прийняття рішення, чи використовувати всі спостереження або виключити деякі дані чи вибірки з набору даних;
- вибір метрики та методу стандартизації вихідних даних;
- визначення кількості кластерів (для ітеративного кластерного аналізу);
- визначення методу кластеризації - вибір методу кластеризації є вирішальним при визначенні форми та специфіки кластерів;
- аналіз результатів кластеризації: перевірка випадковості розбиття на кластери, надійності та стабільності розбиття на підвибірках даних, інтерпретація результатів кластеризації тощо;
- перевірка результатів кластеризації здійснюється формальними та неформальними методами.

Навчання під наглядом

- На сучасному етапі кластеризація часто виступає першим кроком при аналізі даних
- Після виділення схожих груп застосовуються інші методи, для кожної групи будується окрема модель (Semi-Supervised Learning)

Приклади конкретних задач кластеризації

- Сегментація цільової аудиторії сайту
- Ідентифікація груп сімей — споживачів певного товару для розробки стратегії позиціонування бренду
- Тематичне моделювання електронних листів
- Кластеризація символів в незалежності від їх шрифту, розміру тощо (для подальшого розпізнавання)
- Кластеризация для підвищення продажів
  - Підвищити продажі в магазині
  - В додачу до основного товару пропонувати супутні
  - Кожен покупець — унікальний, але потрібно розділити на групи (кластери)
  - Кожен товар має свій час піку продаж.
  - Є час, коли краще продаються дорогі товари, час — коли дешевші

Кластеризация для підвищення продажів - 1

- Підвищити продажі в магазині
- В додачу до основного товару пропонувати супутні
- Кожен покупець — унікальний, але потрібно розділити на групи (кластери)
- Кожен товар має свій час піку продаж.
- Є час, коли краще продаються дорогі товари, час — коли дешевші

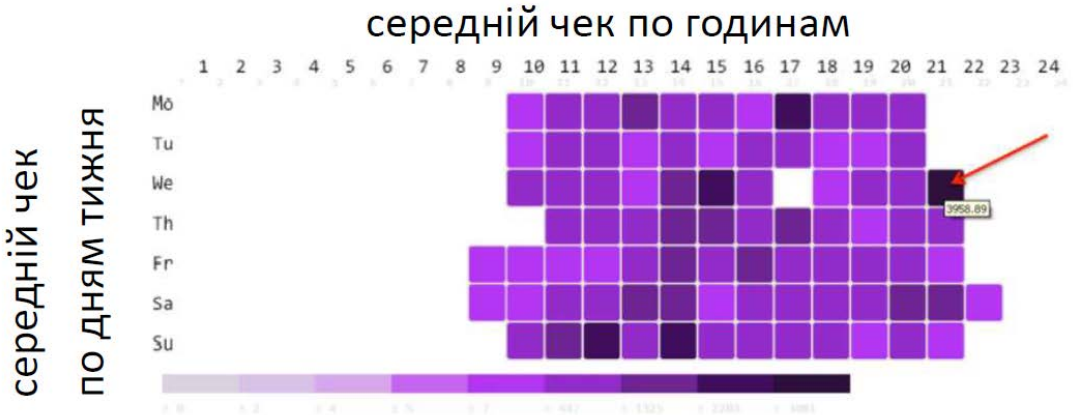
- Покупці цієї категорії (=з цього кластеру):
  - з ймовірністю 27% будуть купляти карти пам’яті;
  - з ймовірністю 13% - навушники тощо

Назва категорії	Ймовірність покупки, %
Карта пам’яті Micro SD	27
Навушники	13
Стартовий пакет	7
Чохли	3
Кабелі аудіо-відео	1
Реле напруги	1

Кластеризация для підвищення продажів - 2

Показник	Значення
Кількість клієнтів в кластері	228
Загальний обіг кластера	377 607 грн
Загальна кількість чеків	364
Середній чек по кластеру	1 037 грн

В цей час обіг складає – 6 146 грн, а середній чек – 439 грн



Це найкращий час для пропозиції дорогих телефонів і аксесуарів!

- Для клієнтів з цього кластера основним товаром є мобільні телефони
- З ним потрібно пропонувати навушники, стартові пакети, чохли, карти пам'яті, оскільки ці товари теж увійшли до кластеру з високою ймовірністю покупки. Консультанти в середу о 15:00 і 19:00, а також у понеділок о 17:00 повинні рекомендувати найдорожчу продукцію цього кластера та розширити викладку дорогих моделей і аксесуарів на вітрині
- У суботу о 15:00 досить низькі показники по обігу при великій кількості чеків
- Необхідно попрацювати над збільшенням середнього чека:
  - збільшити кількість консультантів у відділі «Мобільні телефони»
  - провести акцію або зробити невеликі знижки на супутній товар

Питання?