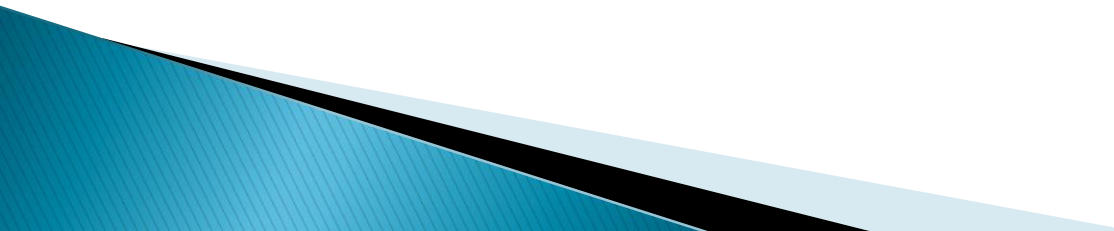


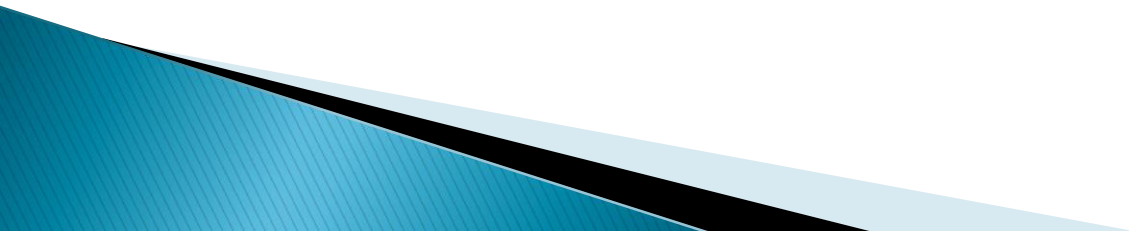
Кластерний аналіз

А.В. Корнилюк

План

1. Характеристика кластерного аналізу
 2. Кластерний аналіз методом K-mean, K-mode, K-medoids
 3. Ієрархічний кластерний аналіз
 4. Кластерний аналіз в R
- 

1. Характеристика кластерного аналізу



Характеристика кластерного аналізу

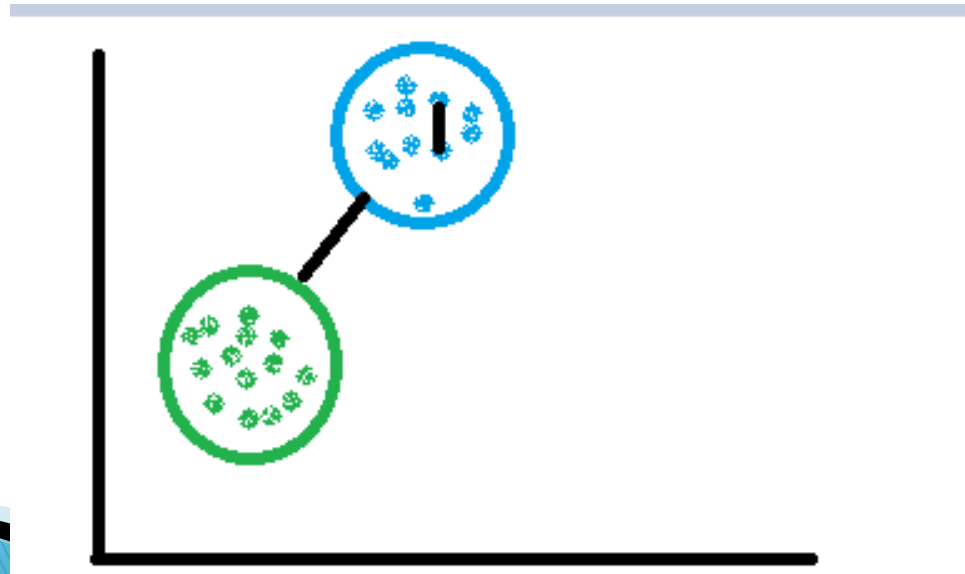
Кластерний аналіз – завдання розбиття певної множини об'єктів на групи, які називаються кластерами так, щоб кожен кластер складався з схожих об'єктів, а об'єкти різних кластерів істотно відрізнялися

Кластерний аналіз застосовується:

- Як окремий метод дослідження
 - Як допоміжний етап при застосуванні інших методів дослідження (н-д, регресійного аналізу)
- 

Характеристика кластерного аналізу

- Значна схожість між об'єктами всередині кластеру
- Низька схожість між кластерами

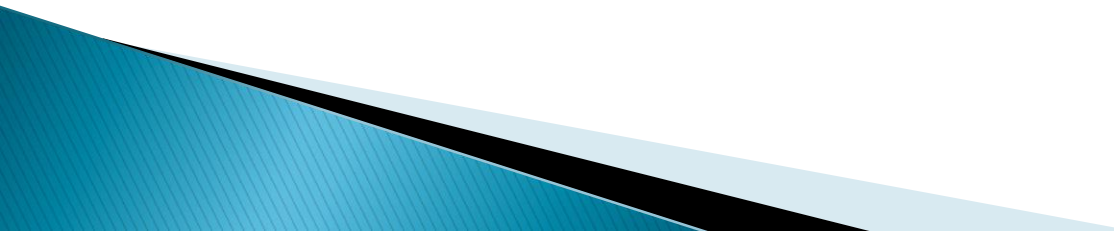


Характеристика кластерного аналізу

- ✓ Відсутність цільової змінної
 - ✓ Розкриває структуру даних без надання пояснень чи інтерпретації
 - ✓ Отримані кластери неінформативні без надання пояснень
 - ✓ Суб'єктивний метод дослідження через складність побудови “ідеальних” кластерів
- 

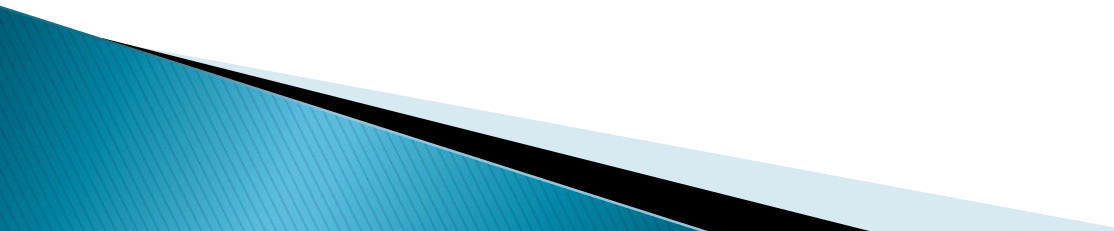
Характеристика кластерного аналізу

Застосування кластерів у фінансовій науці:

- ✓ Вибір\оптимізація портфелю ЦП
 - ✓ Групування банків (компаній, клієнтів) зі схожим рівнем ризику
 - ✓ Групування власників страхових полісів зі схожими характеристиками (страховою сумою, страховими виплатами)
- 

Характеристика кластерного аналізу

Алгоритм проведення кластерного аналізу:

1. Формування вибірки та визначення критеріїв кластеризації
 2. Вибір методу визначення відстані між кластерами
 3. Стандартизація даних
 4. Вибір методу кластерного аналізу. Формування кластерів
 5. Інтерпретація результатів. Формування висновків
- 

Характеристика кластерного аналізу

Методи кластерного аналізу:

- ▶ Ієрархічні: агломеративні (Agglomerative Nesting, AGNES); розділюючі (DIANA)
- ▶ Неієрархічні: н-д, методи секціонування (K-means, K-medoids) тощо

Характеристика кластерного аналізу

- ▶ **Методи секціонування** – методи, що на через оптимізацію цільової функції розподіляють вибірку об'єктів на певну кількість частин
- ▶ Найчастіше цільовою функцією виступає сума квадратів помилок (SSE)

Характеристика кластерного аналізу

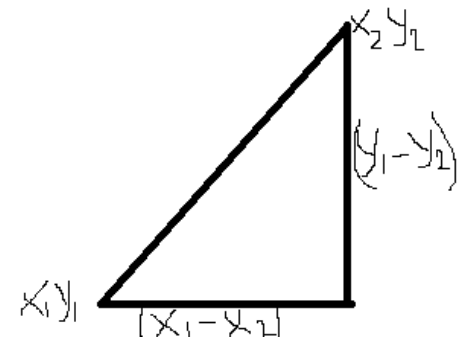
Вибір методу розрахунку відстані

- ▶ Евклідова відстань (найбільш розповсюджений метод)

$$\sqrt{(X_1 - X_2)^2 + (Y_1 - Y_2)^2}$$

У загальному вигляді:

$$\sqrt{(A_1 - A_2)^2 + (B_1 - B_2)^2 + \dots + (Z_1 - Z_2)^2}$$



Характеристика кластерного аналізу

Вибір методу розрахунку відстані:

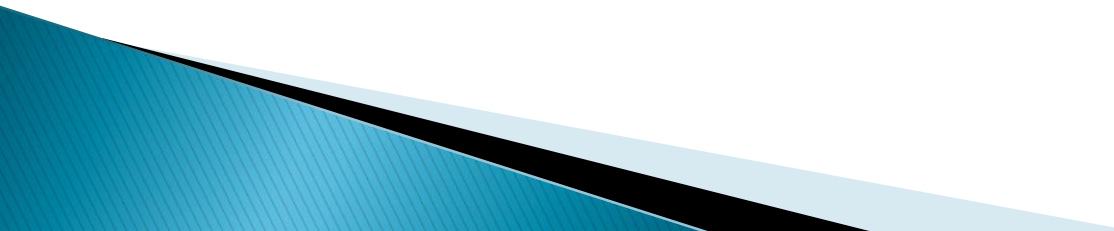
- ▶ Відстань по Манхеттену (зменшується вплив крайніх значень)

$$|A_1 - A_2| + |B_1 - B_2| + \dots + |Z_1 - Z_2|$$



Характеристика кластерного аналізу

Методи розрахунку відстані:

- ▶ Нормалізована евклідова відстань
 - ▶ Метрика Мінковського
 - ▶ Косинусна відстань
 - ▶ Кореляційна відстань
 - ▶ Відстань Хемінга
 - ▶ інші
- 

Характеристика кластерного аналізу

Стандартизація даних – здійснюється з метою нормалізації даних та приведення їх в одну “вагову категорію”

$$x_{ij}^H = \frac{x_{ij} - \bar{x}_j}{\sigma_j}$$

2. Кластерний аналіз методом K-mean, K-mode, K-medoids

K-meap кластеринг

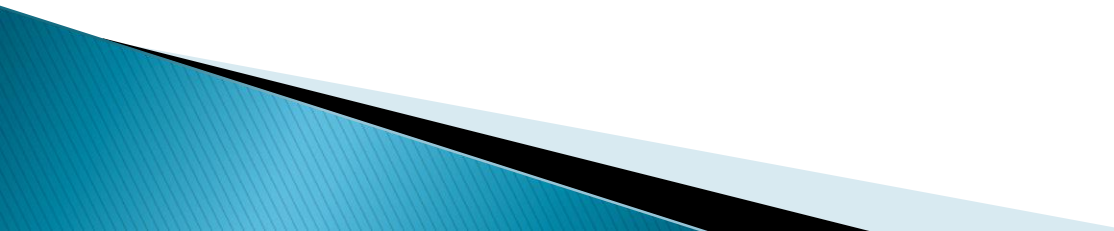
- ▶ Розподіл вибірки даних D із n об'єктів на K кластерів, оптимізуючи цільову функцію (кожен об'єкт відноситься до кластера з найближчим до нього середнім значенням)

Завдання: Мінімізація суми квадратів помилок, де c_k – центроїд кластеру C_k

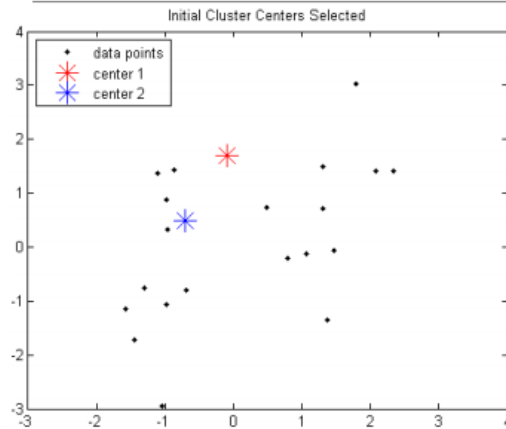
$$SSE(C) = \sum_{k=1}^K \sum_{x_i \in C_k} \|x_i - c_k\|^2$$

K-means кластеринг

Алгоритм:

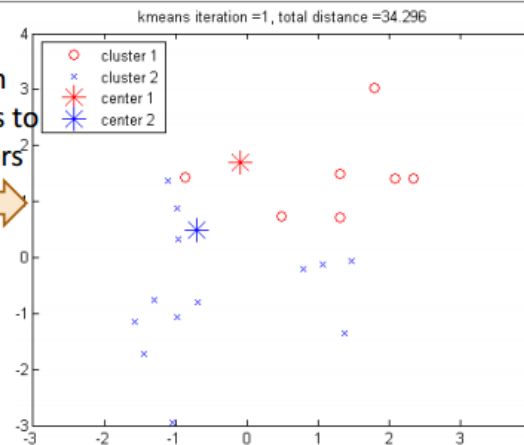
1. Визначити кількість кластерів
 2. Обрати центроїди для кожного кластеру
 3. Згрупувати спостереження навколо найближчого центроїду
 4. Обрати нові центроїди
- 

K-mean кластеринг

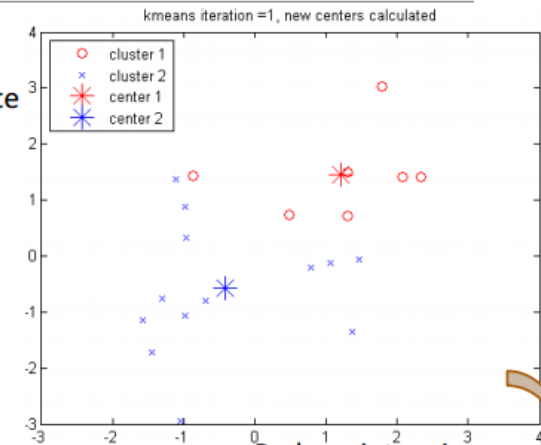


The original data points & randomly select $K = 2$ centroids

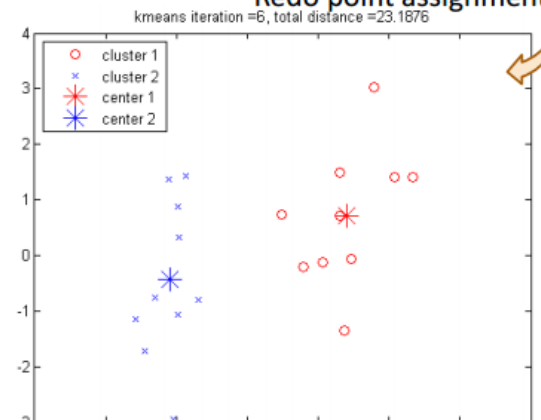
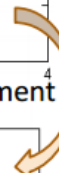
Assign
points to
clusters



Recompute
cluster
centers



Redo point assignment



Execution of the K-Means Clustering Algorithm

Select K points as initial centroids

Repeat

- Form K clusters by assigning each point to its closest centroid
- Re-compute the centroids (i.e., *mean point*) of each cluster

Until convergence criterion is satisfied

K-meap кластеринг

Визначення кількості кластерів:

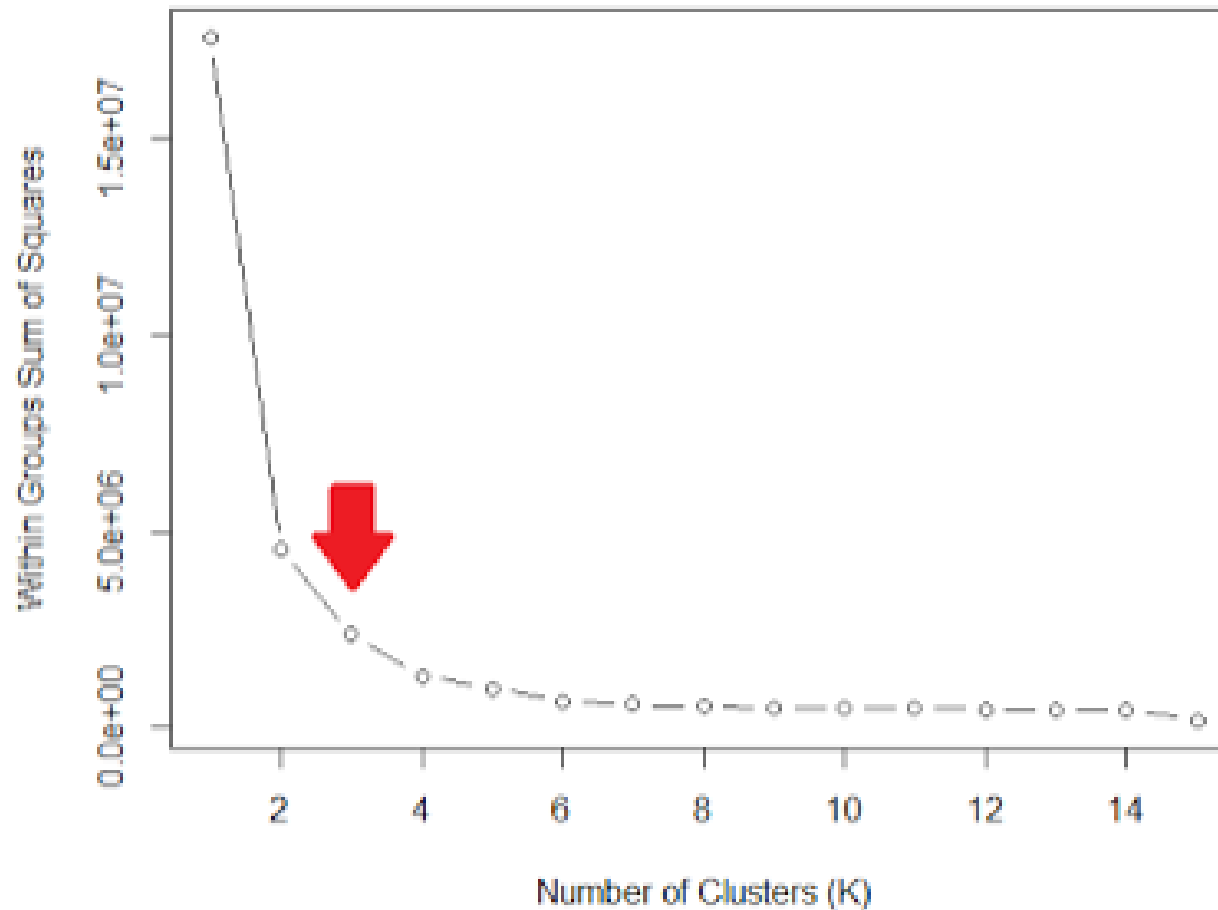
$$k = \sqrt{n}$$

Якщо вибірка велика:

- ▶ метод “ліктя”
- ▶ інформаційні критерії AIC, BIC, DIC
- ▶ суб’єктивний підхід

K-mean кластеринг

Optimal K With WSS



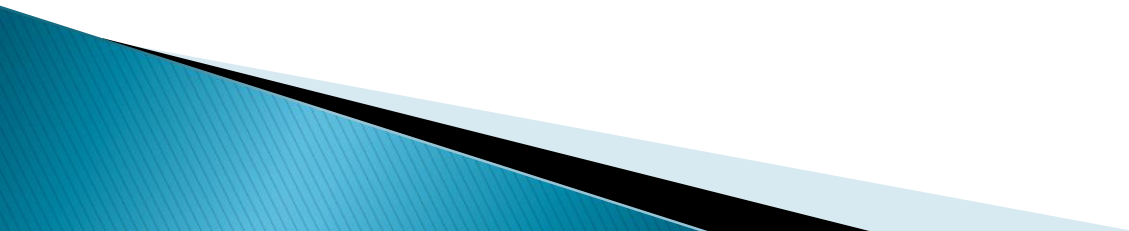
Within Group Sum of Squares by Cluster Number

K-means кластеринг

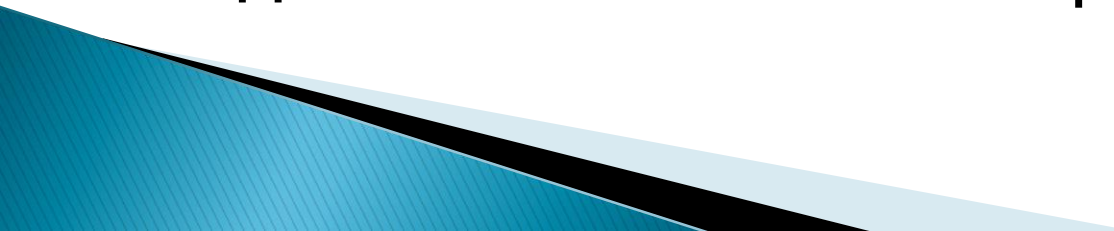
Результат аналізу методом K-means залежить від:

- заданої кількості кластерів
- заданої кількості ітерацій

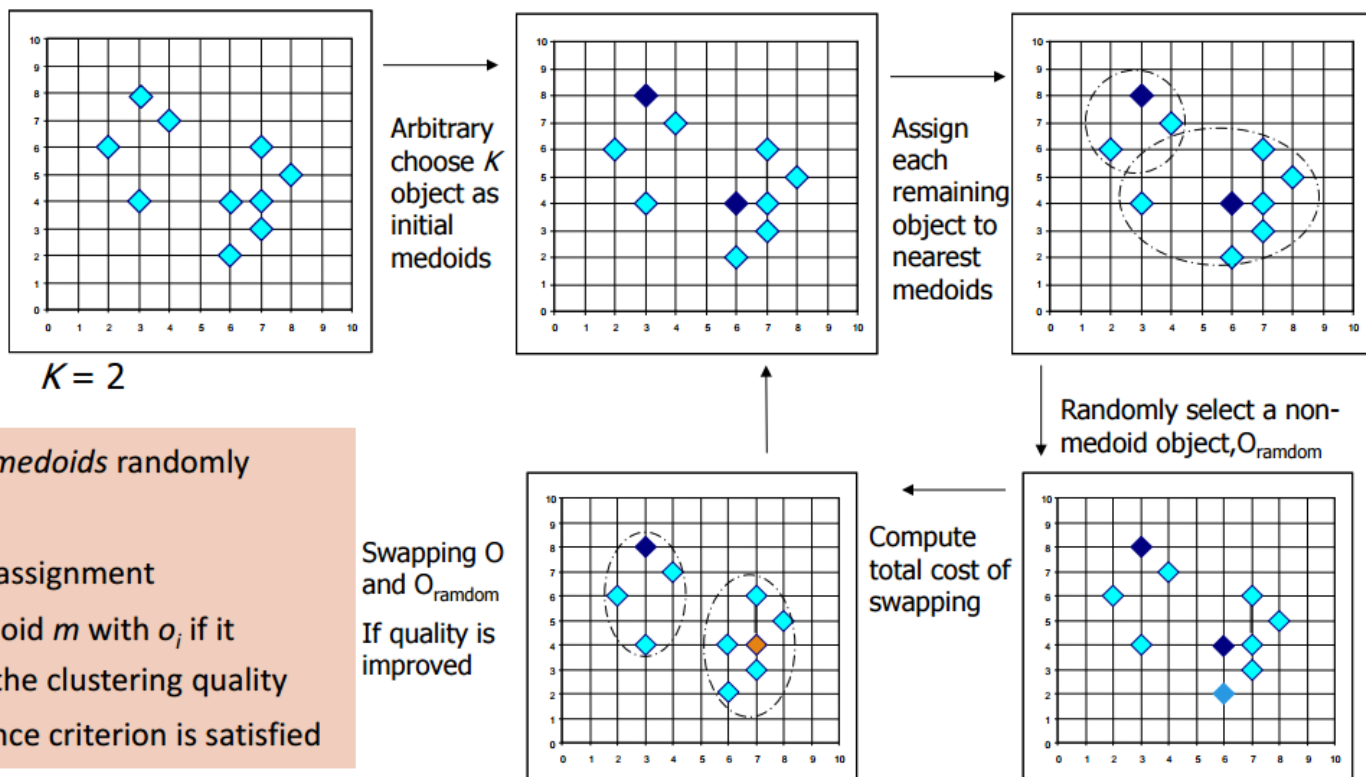
метод K-means чутливий до екстремальних значень!



K-medoid кластеринг

- ▶ Використання K-medoids дозволяє скоротити вплив екстремальних значень у вибірці
 - ▶ Медоїд – такий об'єкт, що належить кластеру та найменше відрізняється від інших об'єктів кластеру
 - ▶ Медоїд (на відміну від центроїду) має бути одним із об'єктів спостереження
- 

K-medoid кластеринг



K-medians кластеринг

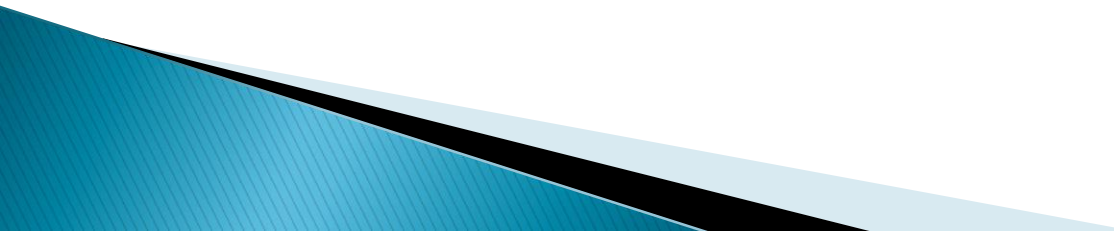
K-medians – менш чутливий до крайніх значень

- ▶ Замість середнього для розрахунку центроїду використовується медіана

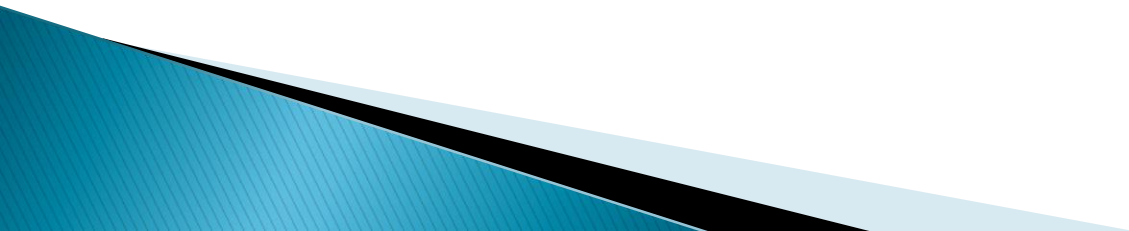
- ▶ Цільова функція:

$$S = \sum_{k=1}^K \sum_{x_{ij} \in C_k} |x_{ij} - med_{kj}|$$

Інші методи кластерингу

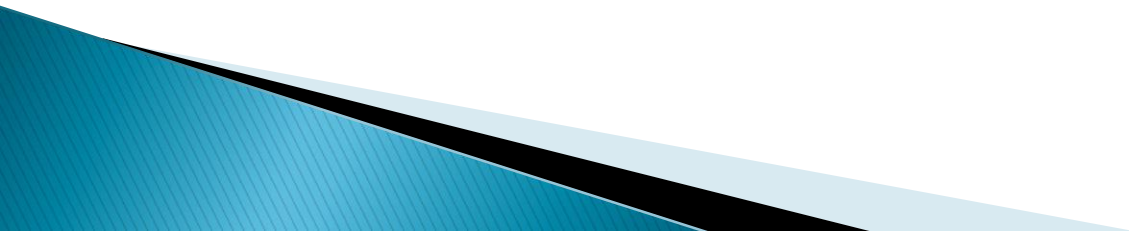
- ▶ **K-modes** – використовується для кластеризації категоріальних даних
 - ▶ **K-prototype** – використовується для кластеризації змішаних даних
- 

3. Ієрархічний кластерний аналіз



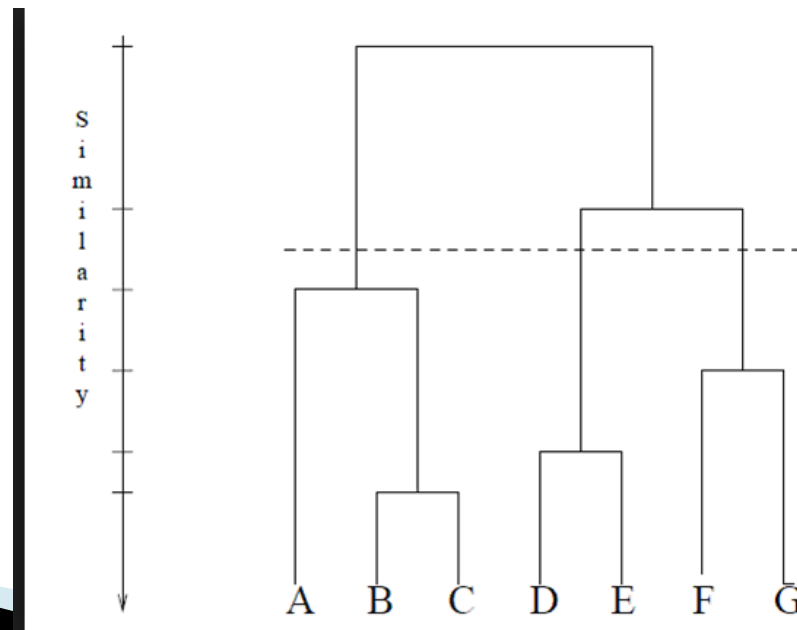
Ієрархічний кластерний аналіз

Ієрархічний кластерний аналіз – сукупність алгоритмів впорядкування даних, візуалізація яких забезпечується за допомогою графів



Ієрархічний кластерний аналіз

Не потребує визначення кількості кластерів наперед (необхідна кількість кластерів може бути досягнута “відсіканням” дендрограми)



Ієрархічний кластерний аналіз

Методи побудови кластерів:

- ▶ **Агломеративний** – усі спостереження розглядаються як окремі кластери, поступово укрупнюючись
- ▶ **Розподільчий** – уся сукупність спостережень розглядається як один кластер, поступово розподіляючись на менші кластери

Ієрархічний кластерний аналіз

Алгоритм агломеративного ієрархічного кластерингу:

1. Знайти два найближчі спостереження
2. Об'єднати ці спостереження
3. Повторити пп 1 і 2

3. Кластерний аналіз в R

