Лабораторна робота №5

**НЕІЄРАРХІЧНІ МЕТОДИ КЛАСТЕРНОГО АНАЛІЗУ**

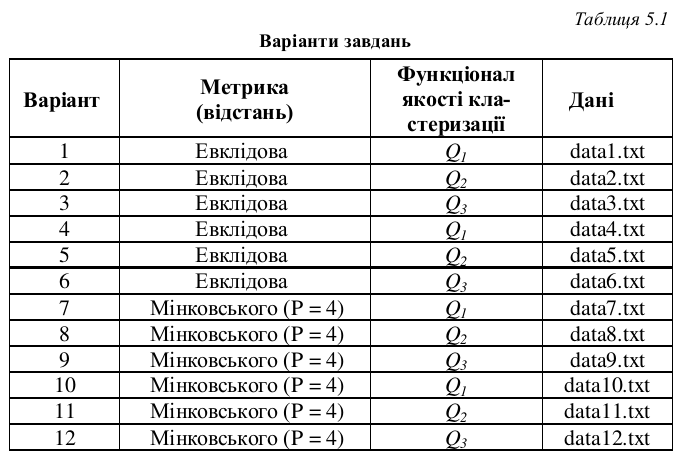
Мета роботи: Практичне засвоєння неієрархічного кластерного аналізу багатовимірних даних на приклады методу *k*-середніх.

### Завдання до роботи

1. Ознайомитися з конспектом лекцій та рекомендованою літературою, а також додатком Е, що містить короткі теоретичні відомості про неієрархічні методи кластерного аналізу та особливості їх застосування в Matlab.
2. Вивчити функції MATLAB: scatter, gscatter, min, pdist, hist, std.
3. Завантажити дані відповідно до вашого варіанту (табл. 5.1). Побудувати графічне зображення експериментальних даних (діаграму розсіювання). Візуально оцінити кількість кластерів *k* за побудованим зображенням.
4. Розробити алгоритм кластеризації *k*-середніх і програмно його реалізувати в середовищі MATLAB.
5. Виконати кластерний аналіз висхідних даних методом *k*-середніх (параметри методу див. в табл. Е.1). Визначити найбільш оптимальну кількість кластерів *k*.
6. Розрахувати центри отриманих кластерів. Відобразити графічно знайдені кластери (використати діаграму розсіювання у кольорі).
7. Оформити звіт з роботи.
8. Відповісти на контрольні питання.

### Зміст звіту

1. Тема та мета роботи.
2. Короткі теоретичні відомості.
3. Номер варіанту та набір даних для обробки (якщо він великий – навести фрагмент). Графічне зображення експериментальних даних.
4. Опис процесу використання програмного забезпечення для обробки набору даних, що має бути ілюстрований зображеннями екранних форм.
5. Тексти програм.
6. Результати роботи програмного забезпечення, що відображують результати неієрархічної кластеризації ал-горитмом *k*-середніх.
7. Висновки, що містять відповіді на контрольні запитання, а також відображують результати виконання роботи та їх критичний аналіз



### Контрольні запитання

1. В чому полягає задача неієрархічного кластерного аналізу?
2. Для яких задач обробки експериментальних даних використовуються методи неієрархічного кластерного аналізу?
3. В чому сутність алгоритму *k*-середніх?
4. Наведіть основні етапи ієрархічного кластерного аналізу за методом *k*-середніх.
5. З якою метою в алгоритмі *k*-середніх вводиться матриця розбиттів *U*?
6. Які критерії зупинки автоматичної кластеризації використовуються на практиці?
7. Яким чином оцінити кількість кластерів в алгоритмі *k*-середніх?
8. З якою метою використовуються функції MATLAB: scatter, gscatter, min, pdist, hist, std?

**Додаток E**

**НЕІЄРАРХІЧНІ МЕТОДИ КЛАСТЕРНОГО АНАЛІЗУ У ПАКЕТІ MATLAB**

Задача кластерного анализу полягає у організації даних, що спостерігаються у наочні структури - кластери. Вихідні дані можуть бути представлені у вигляді матриці X, строки якої відповідають об'єктам (спостереженням), а стовбці їх ознакам (див. табл. Е.1). Потрібно розбити задану множину об'єктів на кластери.

*Таблиця Е.1*

**

За великої кількості об'єктів (N > 200) ієрархічні методи кластерного аналізу непридатні до застосування, через великі обчислювальні витрати і складність інтерпретації дерева кластерів. В таких випадках можуть бути використані неієрархічні методи кластерного аналізу, одним з яких є метод *k*-середніх. Сутність цього методу полягає в наступному. Припустимо, що вже маємо гіпотези відносно кількості кластерів - тобто заздалегідь визначена кількість кластерів *k*, на які необхідно розбити наявні об'єкти. Серед множини об'єктів обирають *k* об'єктів в якості початкових центрів кластерів. Для кожного об'єкту розраховують відстані до центрів кластерів, і даний об'єкт відноситься до того кластера, відстань до якого виявилася мінімальною. Після чого для цього кластеру (в якому змінилася кількість об'єктів) розраховується нове положення центру кластеру (як середнє за кожною ознакою *Xi*) за всіма залученими до кластеру об'єктами. В загальному випадку, в результаті застосування методу *k*-середніх вихідна множина об'єктів розділюється рівно на *k* різних кластерів, розташованих на якнайбільших відстанях один від одного.

Відстань між об'єктами *ni* та *nj* може бути обчислена, наприклад, за такими формулами:

1. Евклідова відстань



2. відстань Мінковського



Для зберігання інформації про приналежність об'єкту до деякого кластеру в методі *k*-середніх вводиться матриця *U =* {*uij*}*, i =* 1*,…, N* та *j =* 1, 2. В першому стовбці матриці *U* містяться індекси кластерів, до яких відносяться об'єкти даних в другому стовбці – відстані від об'єктів до відповідних центрів кластерів.

Алгоритм *k*-середніх є ітераційним. Блок-схема алгоритму неієрархічної кластеризації по методу *k*-середніх наведена на рис. Е.1.

*Алгоритм методу k-середніх*

*Крок 1.* Ініциалізація початкових параметрів методу (блок 1, рис. Е.1).

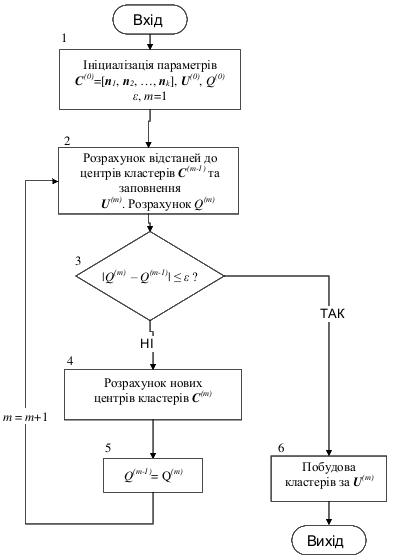
Задати: k - кількість передбачуваних кластерів, матрицю координат центрів кластерів *С(0)* = {*сl(0)*}, *l* = 1,2,…, *k* (наприклад, обрати *k* об'єктів із файлу висхідних даних), матрицю *U(0)*, початкове значення функціоналу якості кластеризації *Q(0)* (деяке велике число) і точність його обчислення (для зупинки алгоритму). Встановити номер ітерації *m* = 1.

*Крок 2.* Розрахувати відстані від об'єктів *n1, n2, …, nN* до центрів кла-стерів *С(m-1)*, визначених на попередній ітерації. Заповнити матрицю *U(m)*, виходячи з розташування центрів *С(m-1)* обчислених на попередній ітерації (блок 2). Розрахувати значення функціоналу якості кластеризації *Q(m)* (з урахуванням *С(m-1)*).

*Крок 3.* Перевірити умову зупинки алгоритму | *Q(m) – Q(m-1)* | (блок 3). При цьому оцінюється, чи призвело нове об'єднання об'єктів в кластери до суттєвого покращення якості кластеризації. Якщо умова виконується, то завершити процес кластерізації (блок 6). Інакше перейти на крок 4.

*Крок 4.* Розрахувати нове положення центрів кластерів *С(m)* як середнє аріфметичне за координатами об'єктів, що входять у відповідні кластери (блок 4).

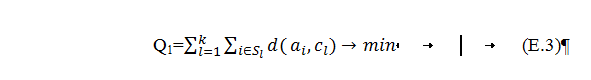
*Крок 5*. Встановити *Q(m-1) = Q(m)* і перейти на крок 2 (нової ітерації) з *m = m* + 1 (блок 5).



*Рис. Е.1.* Блок-схема алгоритму неієрархічної кластеризації за методом k-середніх

Автоматичне групування об'єктів в кластери припиняється при виконанні одного з критеріїв зупинки автоматичного групування. На практиці використовуються такі види критеріїв якості автоматичного групування (функціонали якості кластеризації):

1. Сума відстаней до центрів кластерів:



*l* – номер кластеру, *l* = 1, 2, …, *k, ni* – вектор ознак *i*-го об'єкту в *l*-ом кластері, *Sl* – множина об'єктів в *l*-ому кластері.

1. Сума квадратів відстаней до центрів кластерів:



*l* – номер кластеру, *l* = 1, 2, …, *k, ni* – вектор ознак *i*-го об'єкту в *l*-ому кластері, *Sl* – множина об'єктів в *l*-ому кластері.

1. Сума внутрішньоикластерних відстаней між об'єктами:



Головним недоліком алгоритму *k*-середніх є те, що здійснюється локальна, а не глобальна мінімізація функціоналу *Q*.