

НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ

«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ   
ІМЕНІ ІГОРЯ СІКОРСЬКОГО»

ФАКУЛЬТЕТ БІОМЕДИЧНОЇ ІНЖЕНЕРІЇ

КАФЕДРА БІОМЕДИЧНОЇ КІБЕРНЕТИКИ

**Практична робота №1**

## з дисципліни «Вступ до інтелектуального аналізу даних»

# на тему: «Дослідження застосування різних форм критерію

# роздільності для визначення оптимальної кількості кластерів в

# задачах аналізу функціонального стану серцево-судинної системи людини»

**Виконав:**

студент гр. БС-03

Затуловський Г.А.

**Перевірив:**

доц. Павлов В.А.

Зараховано від \_\_\_.\_\_\_.\_\_\_\_\_\_\_

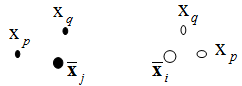
\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

(підпис викладача)

Київ-2022

**Теоритичні відомості:**

***Дослідження застосування різних форм критерію роздільності для визначення оптимальної кількості кластерів в задачах аналізу функціонального стану серцево-судинної системи людини***

Варіанти крітерію роздільності:

1. **Внутрішньокласова дисперсія**

min Dw  (внутрішньокласова відстань min dw)

1. **Міжкласова дисперсія**

max DB (міжкласова відстань max dB)

1. Нормована внутрішньокласова дисперсія

min Dw = Dw/ (Dw + DB) (нормована внутрішньокласова відстань min dw = dw/ (dw + dB) )

– тут нормуємо ділячи на повну дисперсію/відстань кластерів

1. Нормована міжкласова дисперсія

max DB = DB/ (Dw + DB) (нормована міжкласова відстань max dB = dB/ (dw + dB) )

1. **Власне критерій роздільності:**

на основі дисперсій min Dr= Dw / DB та

на основі відстані dr= dw / dB

Нормовані форми критерію 3,4 застосовують замість первинних 1,2 коли порівнюються варіанти розділення на суттєво відмінних (порядок) кількостей кластерів (3 та 33). При застосуванні форми 5, очевидно нема необхідності у нормуванні.

Розглянемо далі ***приклад застосування*** таких критерії для оцінки якості розбиття даних методом *k-середніх.*

Було здійснено вимірювання у студентів даних тиск-пульс при виконанні проби

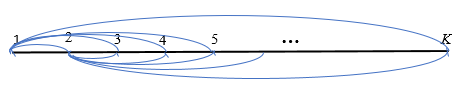
Мартине (до навантаження, зразу після виконання 20 присідань, та після періодів відпочінку). У файлі даних представлено результати розбиття даних методом *k-середніх* у просторі 3\*6=18 координат: «систолічний тиск - діастолічний тиск – пульс» у 6 станах людини: спокій, навантаження, та через 4 періоди відпочинку (період 1 хвилина).

*Одержано 3 варіанти*  розбиття при різних варіантах кількості кластерів 8-12-14 (реально 8-10-13, інші кластери трактуємо, як псевдокластери – одержано з 1 об’єктом у кластері)

Кожне розбиття представлено переліком об’єктів що попали у кластери, об’єкти представлені у матриці даних відповідними строками значень 18-ти координат-ознак. Строки даних послідовно зверху вниз представляють відповідні кластери з номерами (1-8) чи (1-12) чи (1-14).

Розглянемо далі вирази для критеріїв 1,2,5 детальніше

1. На основі ***середньоквадратичного відхилення:***

 (1) 

У чисельнику– ***внутришньокласова*** дисперсія, у знаменнику- ***міжклассова дисп***

2. На основі ***відстані*** (Евклідова норма) теж відношення

 (2)

Для більш детального представлення співвідношень скоректуємо позначення:

Поточний номер кластеру (тут *j* у чисельнику, *j,i* у знаменнику, далі *q* ) перенесемо наверх, тоді розуміючи під нормою - Евклідову відстань між ***i*-тою** точкою ***q*-того** кластеру та його центром

можемо записати:

, чи 

Тоді (1) прийме вид (1\*)

 (1\*)

А (2) - вид (2\*)

 (2\*\*)

де *xip,* *i*=*1,...,nq* , *p=1,...,18* це *р -****та координата*** *i-****тої точки*** *(*їх всього *nq* одиниць *) в q-***тому кластері** (всього кластерів *K* штук) а ,,- **середні точки** *q*-того, *i*-того и *j*-того кластерів по р-тій координаті.

Необхідно застосувавши дані по m змінним (в прикладі m=18 ) у K кластерах (в прикладі K=8 K=12 (10) и K=14(13) кластерів ) и по nq q=1,...K точкам в кoжному кластері.

Визначити, для якого варианту розбиття на кластери (К=8, К=12(10), К=14(13)) критерії приймають найкращі значення.

**Показники для кожного розбиття розрахувати для чисельника, знаменника, Ч/З для двох варіантів міри – всього 6 варіантів** (зробити таблицю)

**Завдання:**

Використувавши дані функціонального стану серцево-судинної системи людини - вимірювання тиск-пульс модифікованої проби Мартине, задайте в алгоритмі к-середніх різні значення кількості кластерів: одержіть розбиття на 8, 12(10) та 14(13) кластерів.

*В режимі ВП виконуємо тільки 2-гу частину, Розбиття на кластери надається у файлі додатку*

***Частина 2***. Дослідження застосування різних форм критерію роздільності для визначення оптимальної кількості кластерів функціонального стану серцево-судинної системи людини

***Завдання.*** Розрахувати значення критериїв роздільності відповідно до додатку 1 (теоретчні відомості до лаборатоної роботи 1). Здійснити дослідження розбиття на кластери. Зробити висновки що до якості розбиття та визначити найкращий варіант кластеризації.

Набір даних, що додається (файл додається) складається з вимірів тиск-пульс у 6-ти режимах: спокій, зразу після навантаження, та 4 виміри послідовно через хвилину відпочинку.

Задачу виконати на двох наборах даних:

1. Веріфувати розрахунок на повному наборі ознак.

Додатковет завдання

1. Індивідуальний варіант розрахунку визначити наступним чином: у множину ознак включаються режим спокою та режим зразу після навантаження. Крім того випадковим чином долучаються дані 2-х чи 3-х режимів (з чотирьох, що залишилися).

Результати розрахунку представити у вигляді таблиці значень критеріїв для кожного варіанту розбиття на кластери.

*У режимі ВП задачу виконати тільки для п.1 Додаткове завдання виконується за бажанням для отримання оцінки «respect»*

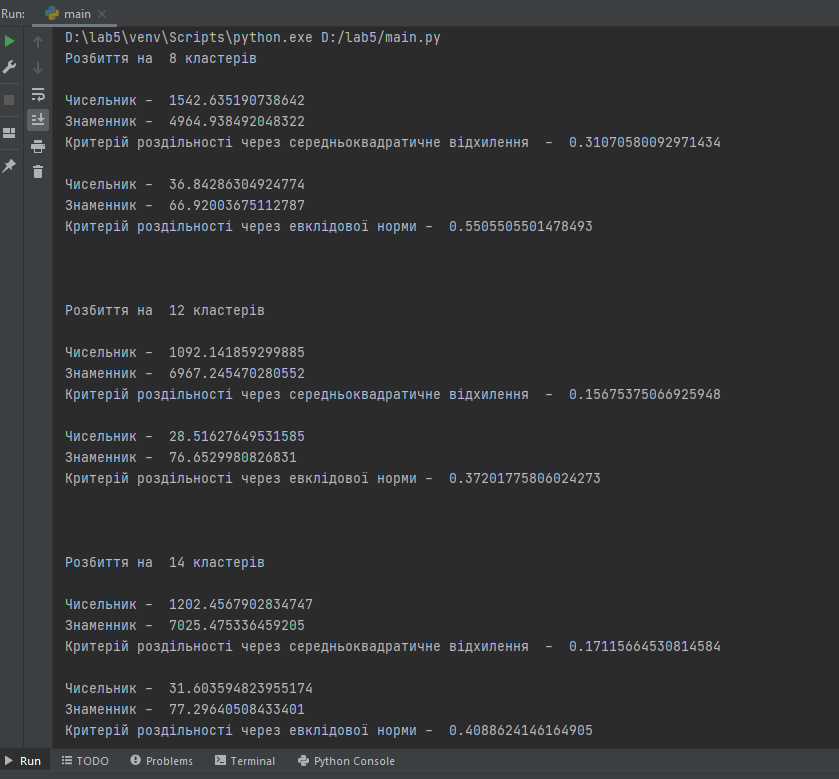
**Виконання:**

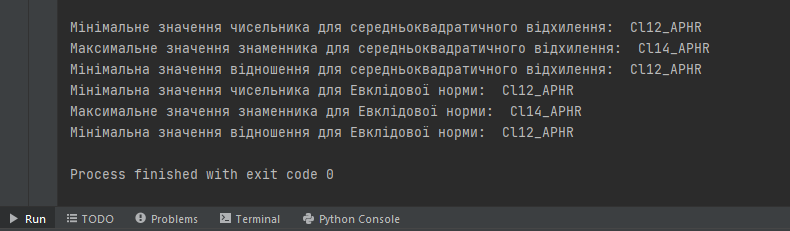
Дані які, використовують в цій лабораторній роботі мітять інформацію для визначення оптимальної кількості кластерів функціонального стану серцево-судинної системи людини

**Лістінг:**

import pandas as pd  
import numpy as np  
from math import comb  
def task(cluster\_data, k):  
 cluster\_mean = [cluster\_data[i].mean(axis=0) for i in range(len(cluster\_data))]  
 cluster\_dev = [0] \* len(cluster\_data)  
 ans = [0, 0]  
 print("Розбиття на ",k, "кластерів\n")  
 l = 1  
 x = 0  
 for q in range(1,3):  
 for i, cluster in enumerate(cluster\_data):  
 for j in cluster:  
 x += np.sum((j - cluster\_mean[i]) \*\*2) \*\* (1/q)  
 cluster\_dev[i] = x / len(cluster)  
 x = 0  
 cluster\_dev\_mean = np.sum(cluster\_dev)/len(cluster\_data)  
  
 sum = 0  
 for i in range(len(cluster\_data)):  
 for j in range(i + 1, len(cluster\_data)):  
 sum = np.sum((cluster\_mean[i] - cluster\_mean[j])\*\*2 )\*\*(1/q) + sum  
 clusters = sum / comb(len(cluster\_data),2)  
  
 ans[q-1] = np.array([cluster\_dev\_mean,clusters,cluster\_dev\_mean/clusters])  
  
 if l == 1:  
 print("Чисельник - ", cluster\_dev\_mean, '\nЗнаменник - ', clusters,'\nКритерій роздільності через середньоквадратичне відхилення - ', cluster\_dev\_mean / clusters,"\n")  
 l = l - 1  
 else:  
 print("Чисельник - ",cluster\_dev\_mean,'\nЗнаменник - ',clusters,'\nКритерій роздільності через евклідової норми - ',cluster\_dev\_mean/clusters,"\n")  
 l = l + 1  
  
 print("\n")  
  
 return ans  
  
df = pd.read\_excel('lab\_1.xlsx', engine='openpyxl')  
  
  
df\_8 = [df[df["Cl8\_APHR"] == i + 1].drop(columns=['Cl8\_APHR', 'Cl12\_APHR', 'Cl14\_APHR']).values for i in range(8)]  
df\_12 = [df[df["Cl12\_APHR"] == i + 1].drop(columns=['Cl8\_APHR', 'Cl12\_APHR', 'Cl14\_APHR']).values for i in range(12)]  
df\_14 = [df[df["Cl14\_APHR"] == i + 1].drop(columns=['Cl8\_APHR', 'Cl12\_APHR', 'Cl14\_APHR']).values for i in range(14)]  
  
result\_1 = [0,0,0]  
result\_2 = [0,0,0]  
var = ['Cl8\_APHR', 'Cl12\_APHR', 'Cl14\_APHR']  
result\_1[0],result\_2[0] =task(df\_8, 8)  
  
result\_1[1],result\_2[1] =task(df\_12, 12)  
  
result\_1[2],result\_2[2] =task(df\_14, 14)  
  
print("Мінімальне значення чисельника для середньоквадратичного відхилення: ",var[np.argmin(np.array(result\_2), axis=0)[0]])  
print("Максимальне значення знаменника для середньоквадратичного відхилення: ",var[np.argmax(np.array(result\_2), axis=0)[1]])  
print("Мінімальна значення відношення для середньоквадратичного відхилення: ",var[np.argmin(np.array(result\_2), axis=0)[2]])  
print("Мінімальна значення чисельника для Евклідової норми: ",var[np.argmin(np.array(result\_1), axis=0)[0]])  
print("Максимальне значення знаменника для Евклідової норми: ",var[np.argmax(np.array(result\_1), axis=0)[1]])  
print("Мінімальна значення відношення для Евклідової норми: ",var[np.argmin(np.array(result\_1), axis=0)[2]])

**Результат роботи:**

****

****

**Рішення по мінімальному значенню чисельника:**

середньоквадратичного відхилення - 12 кластерів

eвклідової норми - 12 кластерів

**Рішення по максимальному значенню знаменника:**

середньоквадратичного відхилення - 14 кластерів

eвклідової норми - 14 кластерів

**Рішення по мінімальному значенню відношення:**

середньоквадратичного відхилення – 12 кластерів

eвклідової норми – 12 кластерів

Отже, можно зробити висновок, що для внутрішньокластерної дисперсії, ми намагаємось мінімізувати значення, а для міжкластерної дисперсії навпаки максимізувати. Таким чином ми отримаємо якісну кластеризацію. Кластери повині бути угрупованні, таким чином, щоб можно розрізняти їх. Розглянувши результатами розбиття на 12 кластерів є найкращім, для двох виразів для критеріїв було отримано найкраще значення для критерію роздільності (його критерій роздільності є найменшим – 0.15675, через середньоквадратичне

відхилення та 0.37202) і внутрішньокласової дисперсії ( для середньоквадратичного відхилення – 1092.1419 та евклідової норми – 28.5163. Після чого вже за якістю де розбиття на 14 кластерів котра має кращі результати міжкласової дисперсії (для середньоквадратичного відхилення –7025.4743 та евклідової норми – 77.2964). Розділення на 8 кластерів не має найкращих результатів, тому можно вважати його найгіршим.

**Висновок:** Ми дослідили застосування різних форм критерію роздільності для визначення оптимальної кількості кластерів в задачі аналізу функціонального стану серцево-судинної системи людини