Звіт
до лабораторної роботи №5:
«Метод k-середніх»

студента 1-го курсу магістратури
факультету комп'ютерних наук та кібернетики
Кравця Олексія

Зміст

1 Постановка задачі

Використовуючи метод k-середніх необхідно провести кластерізацію точок на площині. Візуалізувати результати використовуючи діаграму Вороного.

2 Метод k-середніх

Метод k-середніх – популярний метод кластерізації. Алгоритм намагається мінімізувати сумарне квадратичне відхилення точок кластерів від центрів цих кластерів:

$$V = \sum_{i=1}^{k} \sum_{x \in S_i} (x - \mu_i)^2$$

де k – число кластерів, S_i – отримані кластери, $i = 1, 2, \ldots, j$, а μ_i – центри мас всіх векторів x з кластера S_i.

Головна ідея алгоритму полягає у перерахуванні на кожній ітерації центрів мас всіх кластерів, що були отримані на минулій ітерації. Після цього вектори знову розбиваються на кластери відповідно до отриманих центрів мас.

Алгоритм завершується, коли на деякій ітерації не змінюються центри кластерів. Це досягається за скінченну кількість кроків.

3 Результати

Сгенеруємо набір точок на площині, що після цього будемо кластерізувати на 4 класи. Сгенеруємо точки з такими розподілами:

- Нормальний розподіл. По осі x маємо $\mu = 3, \sigma = 0.2$, по осі y маємо $\mu = 5, \sigma = 0.2$
- Нормальний розподіл. По осі x маємо $\mu = 1, \sigma = 0.1$, по осі y маємо $\mu = 1, \sigma = 0.7$
- Нормальний розподіл. По осі x маємо $\mu = 5, \sigma = 1$, по осі y маємо $\mu = 2, \sigma = 0.01$
- Нормальний розподіл. По осі x маємо $\mu = 5, \sigma = 0.01$, по осі y маємо $\mu = 1, \sigma = 0.01$

Оберемо початкові центри кластерів:

$$\mu_0 = (3, 4) \quad \mu_1 = (2, 1) \mu_2 = (5, 4) \quad \mu_3 = (4, 1)$$

На рисунку 1 можна побачити точки, що будемо кластерізувати (чорні точки), початкове наближення для центрів кластерів. Також на рисунку 1 зображено діаграму Вороного.

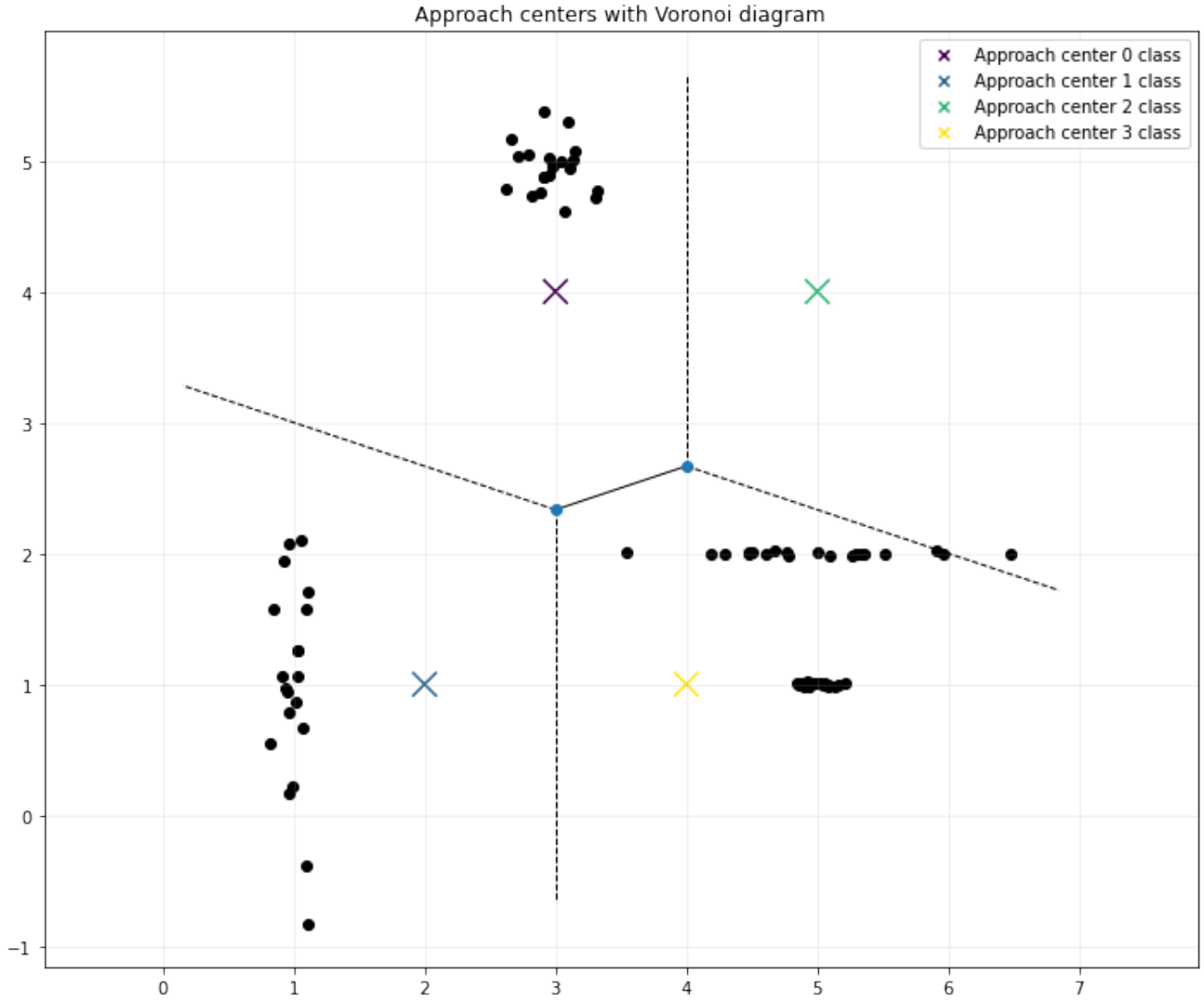

Рис. 1: Точки для кластерізації

Давайте поглянемо на вигляд методу після однієї ітерації. На рисунку 2 зображено результат **методу k-середніх** після 1 ітерації. Розглянемо відстань на яку змістилися центри кластерів:

- Центр кластеру 0 змістився на 0.95
- Центр кластеру 1 змістився на 1
- Центр кластеру 2 змістився на 2.49
- Центр кластеру 3 змістився на 1.06

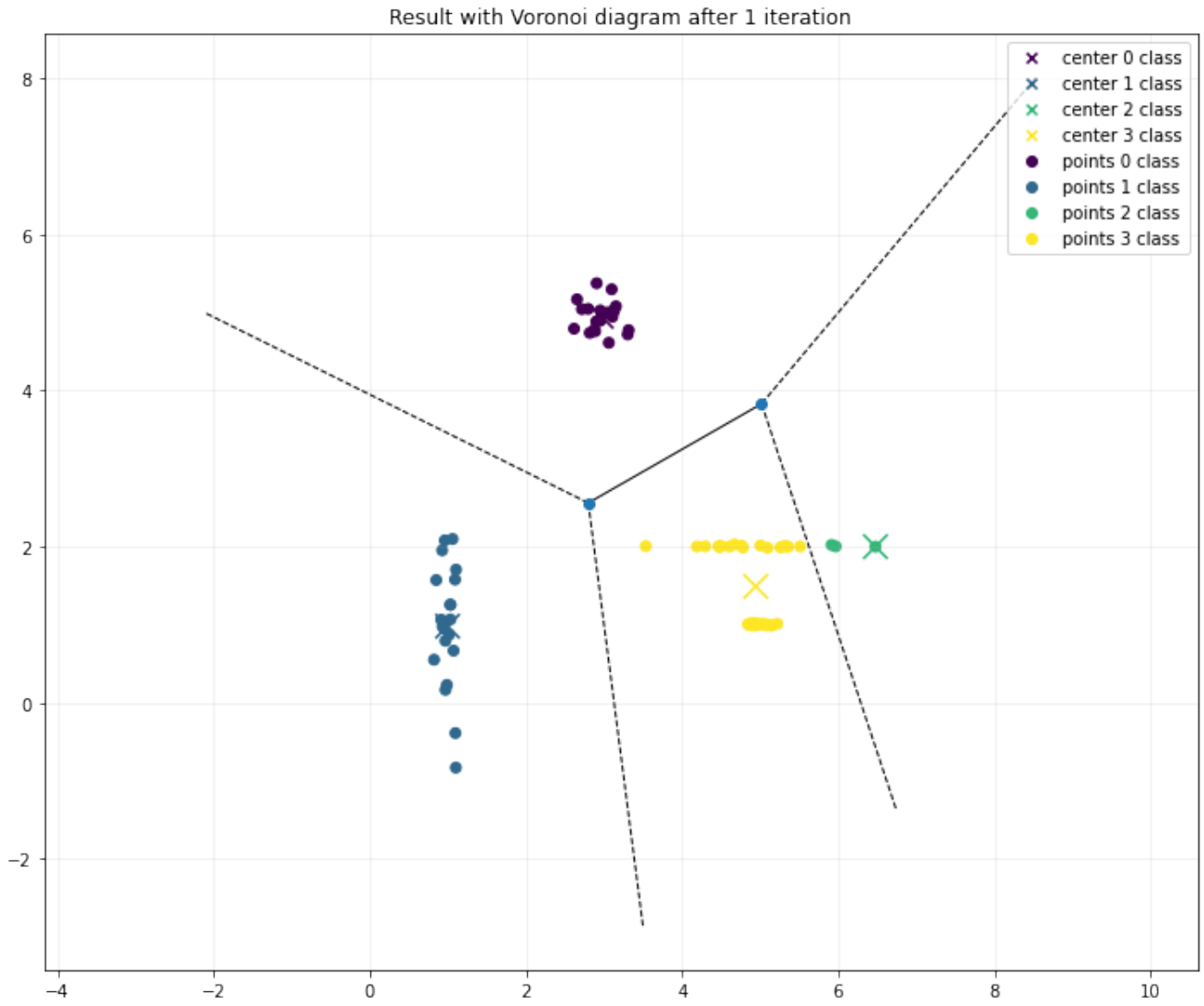

Рис. 2: Після 1 ітерації

Тепер запустимо алгоритм, доки не отримаємо ітерацію для якої, максимальне зміщення центрів кластерів буде менше 10^{-5}. Виведемо результат на рисунку 3.

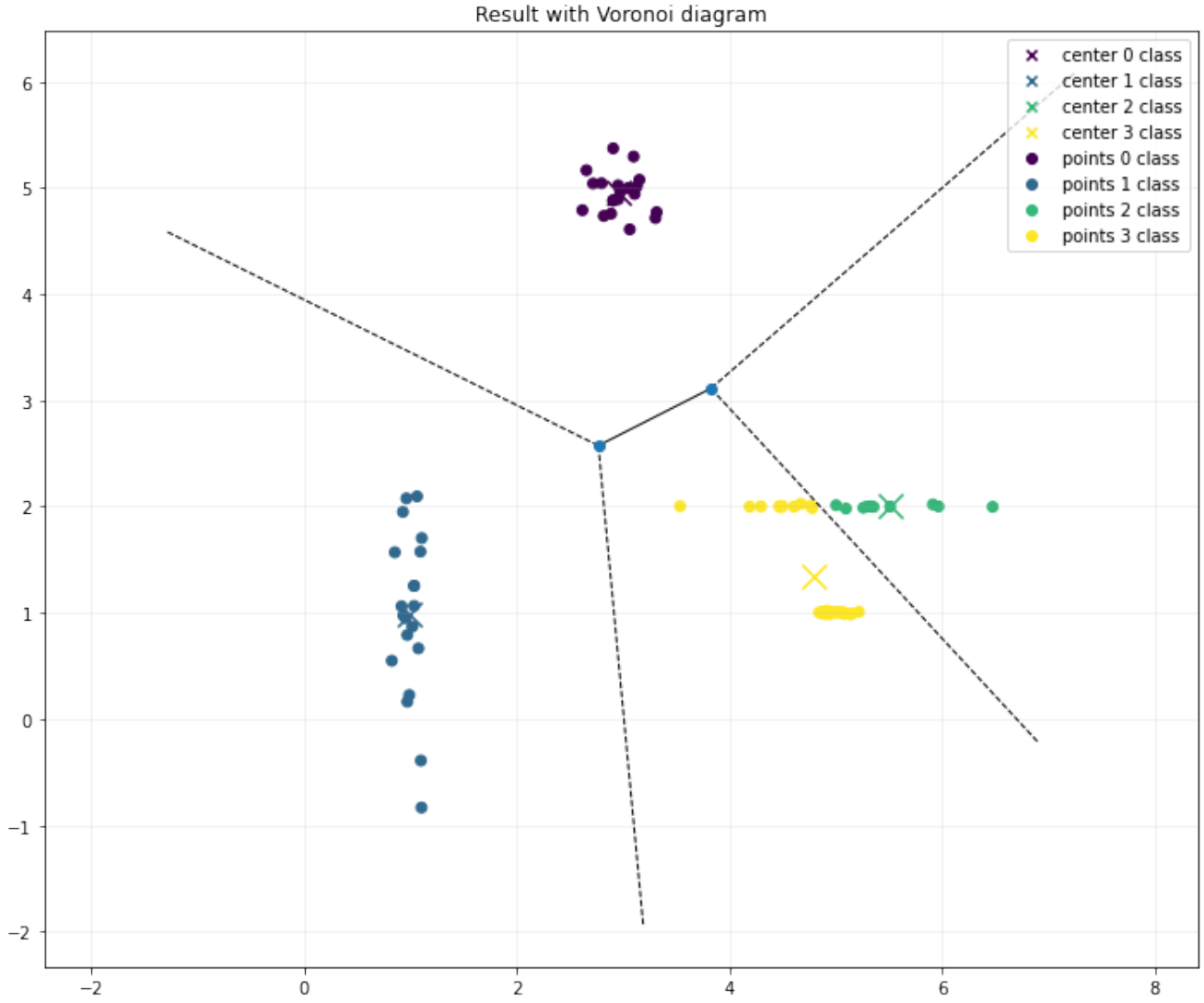

Рис. 3: Результат

4 Висновки

Метод k-середніх спрацював на тестових даних коректно та розділив точки на 4 класи. Але метод не зміг знайти кластери, що відповідали б заданим нормальним розподілам точок. Про це свідчить неточне розділення точок на **2** та **3** кластери. Можливо більш точний підбір початкових центрів кластерів може вирішити цю проблему.