**Звіт**
**до лабораторної роботи №4:**
**«Багатовимірна класифікація»**

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

# Зміст

# 1 Постановка задачі

Маємо навчальну вибірку у вигляді векторів багатовимірного простору. Вибірка розподілена на 3 класи, що не перетинаються.

Необхідно, на базі навчальної вибірки, провести класифікацію тестової вибірки трьома різними методами.

# 2 Опис методів

Для класифікації використовуються центри мас класів

$$M_k = \frac{\sum_{j=1}^{L} \mathbf{y}_j^k}{L}$$

де $M_k$ – центр мас $k$-го класу, $L$ – кількість елементів (векторів) у класі, $y_j^k$ – елемент (вектор) $k$-го класу.

Нехай $M^1, \ldots, M^m$ – центи мас класів. $X = (x_1, \ldots, x_n)$ – точка, яку необхідно класифікувати.

## 2.1 Метод 1

Будемо класифікувати точки за найбільш суттєвою різницею до центрів мас класів. Тобто для кожного виміру $i \in \{1, 2, \ldots, n\}$ будемо рахувати відстань від точки до центрів мас класів та рахувати наступний вираз для кожного класу

$$\frac{|M_i^k - x_i|}{\left(\sum_{j \neq k} |M_i^j - x_i|\right) / (m-1)} \tag{1}$$

ділимо відстань до поточного класу на середню відстань до інших класів. Або

$$\frac{|M_i^k - x_i|}{\min_{j \neq k} |M_i^j - x_i|} \tag{2}$$

ділимо відстань до поточного класу на мінімальну відстань до інших класів.

В результаті отримуємо $m \cdot n$ значень (для кожного виміру та кожного класу). Вибираємо мінімальне значення, і клас, що йому відповідає, буде результатом.

## 2.2 Метод 2

Для класифікації точки обираємо той клас, що має найменшу відстань о точки по найбільшій кількості вимірів. Отже, рахуємо наступний вираз

$$\left( \underset{j \in \{1, \ldots, m\}}{\arg\min} |M_1^j - x_1|, \ldots, \underset{j \in \{1, \ldots, m\}}{\arg\min} |M_n^j - x_n| \right) \tag{3}$$

І вибираємо той клас, що зустрівся більше всіх.

## 2.3 Метод 3

Порахуємо суму відстаней від точки до центрів мас класів по кожному виміру. Оберемо той клас, у якого сумма найменша. Отже, обираємо клас за наступним правилом

$$\underset{j \in \{1, \ldots, m\}}{\arg\min} \left( \sum_{i=1}^{n} |M_i^j - x_i| \right) \tag{4}$$

# 3 Результати

## 3.1 Тести на площині

Для наочності провели класифікацію точок на площині. У **Методі 1** було використано рівняння (2).

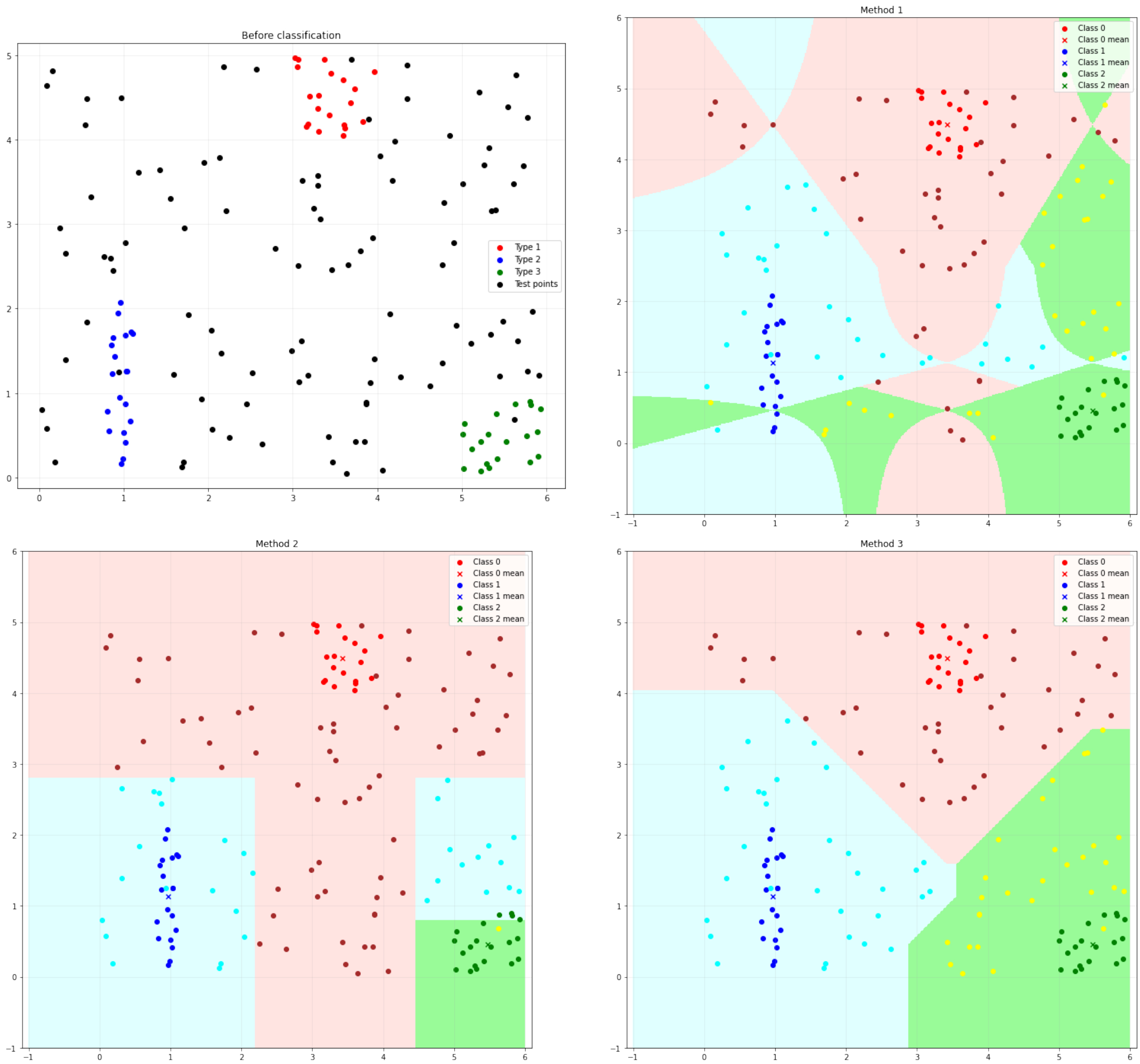


Рис. 1:

Якщо розглянути результати більш детально, то можна помітити, що **Метод 2** не може точно класифікувати багато точок. Розглянемо перші 10 тестових точок

```
Dot: [5.65745822 1.61601466] has class: [1, 2]
Dot: [3.11274373 3.51509479] has class: [0]
Dot: [2.18177761 4.85891041] has class: [0, 1]
Dot: [5.77468377 1.25891148] has class: [1, 2]
Dot: [2.98349104 1.50439155] has class: [0, 1]
Dot: [1.70904297 0.18443474] has class: [1, 2]
Dot: [3.657386   2.51339512] has class: [0, 1]
Dot: [0.30887251 1.39323232] has class: [1]
Dot: [5.44959532 1.19780945] has class: [1, 2]
```

```
Dot: [0.86936923 2.4472638 ] has class: [1]
```

Легко помітити, що **Метод 2** не може класифікувати деякі точки.

## 3.2 Багатовимірна класифікація

Для тестування багатовимірної класифікації я обрав набір даних «Іриси Фішера». Кожний елемент набору даних це – чотиривимірний вектор, що розподілений в один з трьох класів. Точність класифікації я оцінюю так

$$\text{точність} = \frac{\text{кількість правильних класифікацій}}{\text{загальна кількість класифікацій}}$$

Також будемо проводити крос-валідацію. Розділимо набір даних на 5 частин.

Виведемо таблицю з отриманою точністю.

| Метод | *fold 1* | *fold 2* | *fold 3* | *fold 4* | *fold 5* | mean |
|---|---|---|---|---|---|---|
| **Метод 1** *min* | 0.867 | 0.867 | 0.733 | 0.9 | 0.8 | 0.833 |
| **Метод 1** *mean* | 0.9 | 0.867 | 0.767 | 0.933 | 0.833 | 0.86 |
| **Метод 2** | 0.933 | 0.9 | 0.933 | 0.967 | 0.833 | 0.913 |
| **Метод 3** | 0.967 | 0.9 | 0.8 | 0.967 | 0.8 | 0.887 |

# 4 Висновок

**Метод 2** показав найкращі результати на Ірисах Фішера, однак для класифікації точок на площині цей метод показує себе погано. Він не може класифікувати деякі точки.