Міністерство Освіти і Науки України

Київський Національний Університет імені Тараса Шевченка
Факультет Інформаційних Технологій
Кафедра Інформаційних систем та технологій

Звіт з лабораторної роботи № 3
з дисципліни «Data Science та машинне навчання»
Тема: **« МЕТОДИ РОЗДІЛЯЮЧИХ ФУНКЦІЙ »**

Виконав студент 1-го курсу
магістратури
групи ІРма-12
Гаврасієнко Є.О.

Київ – 2025

Мета роботи:

1. Опанувати непараметричні методи «навчання з учителем», засновані на

лінійних розділяючих функціях і методику побудови шматково-лінійних вирішальних правил;

2. Отримати навички статистичного оцінювання показників якості класифікації з використанням системи MathCAD для моделювання та подання об'єктів у вигляді даних спостережень.

Завдання:

1. За заданими (згідно з варіантом) двовимірними даними спостережень $\xi i = (xi, yi)$ двох класів об'єктів a1 і a2 за правилом найближчого сусіда провести межі між класами:

a. за вибірковими значеннями – межу $g1(x, y) = 0$;

b. за вибірковим середнім – межу $g2(x, y) = 0$.

2. Побудувати вирішальні правила g1 та g2.

3. Згенерувати масиви N даних спостережень (N = 100) класів a1 і a2 у припущенні, що спостерігається двовимірний випадковий вектор, компоненти якого – некорельовані нормально розподілені величини. В якості параметрів розподілу класів відповідно) взяти їх статистичні оцінки, отримані за заданими вихідними даними.

4. Змоделювати процеси розпізнавання спостережень згідно вирішуючих правил g1 і g2 і порівняти ефективності класифікаторів за емпіричними оцінками ймовірностей правильних рішень.

5. Оформити звіт про лабораторну роботу , який повинен містити короткі теоретичні відомості, алгоритми моделювання даних та прийняття рішень, графічні подання реалізацій спостережень та меж між класами, висновки.

Хід роботи

Варіант для виконання згідно таблиці - 2

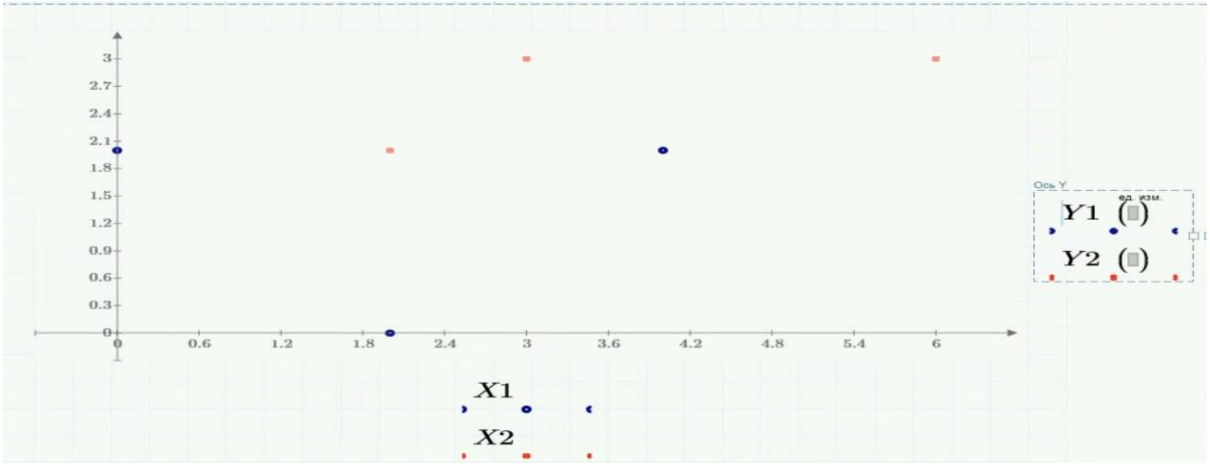
Параметри для варіанту:

Вхідні дані

Клас a_1		Клас a_2	
$x_{11} := 0$	$y_{11} := 2$	$x_{24} := 2$	$y_{24} := 2$
$x_{12} := 2$	$y_{12} := 0$	$x_{25} := 3$	$y_{25} := 3$
$x_{13} := 4$	$y_{13} := 2$	$x_{26} := 6$	$y_{26} := 3$

Завдання 1:

Нанесемо точки на площину для більш чіткого розуміння їх положення:



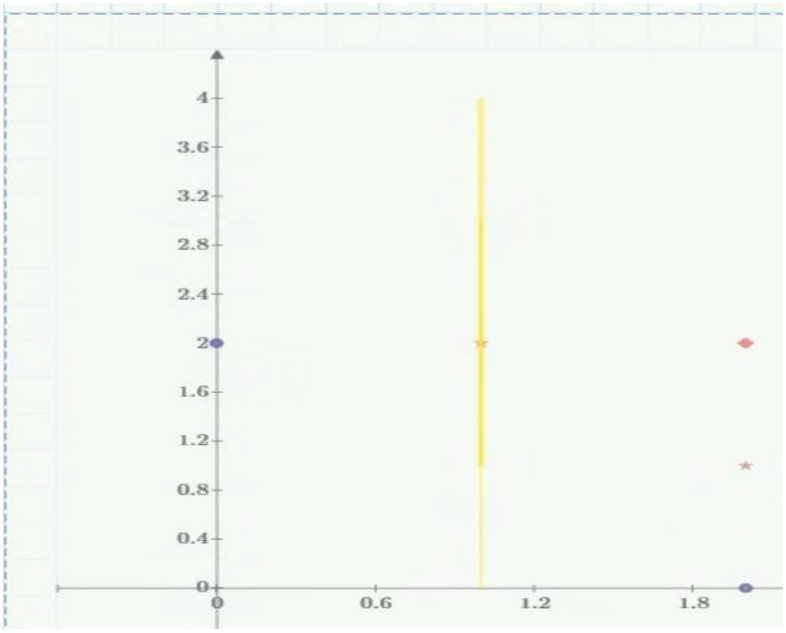
Далі будемо шукати координати середини відрізків, що сполучають об’єкти різних класів для проведення нормалі (перпендикулярних ліній) до цих відрізків, які і будуть слугувати межами між об’єктами цих класів. Шукати ці точки будемо за правилом **найближчого сусіда**.

Спочатку знайдемо координати середини відрізку між точками $\xi 1 \in a1$ та

$\xi 4 \in a2$ та розрахуємо функцію нормалі для цієї пари точок:

$$norm3(n) := \frac{(y_{13}+y_{25})}{2} + \left(\frac{(y_{25}-y_{13})}{-(x_{25}-x_{13})}\right)\cdot\left(n - \frac{(x_{13}+x_{25})}{2}\right)$$

Зобразимо нормаль на площині:



Схожим чином виконаємо розрахунок нормалей для інших точок і нанесемо їх на площину

$$norm1 := 1..4 \qquad norm5 := 0, 2.5..2.6$$

$$norm3(n) := \frac{(y_{13}+y_{25})}{2} + \left(\frac{(y_{25}-y_{13})}{-(x_{25}-x_{13})}\right)\cdot\left(n - \frac{(x_{13}+x_{25})}{2}\right)$$

$$norm4(n) := \frac{(y_{13}+y_{26})}{2} + \left(\frac{(y_{26}-y_{13})}{-(x_{26}-x_{13})}\right)\cdot\left(n - \frac{(x_{13}+x_{26})}{2}\right)$$

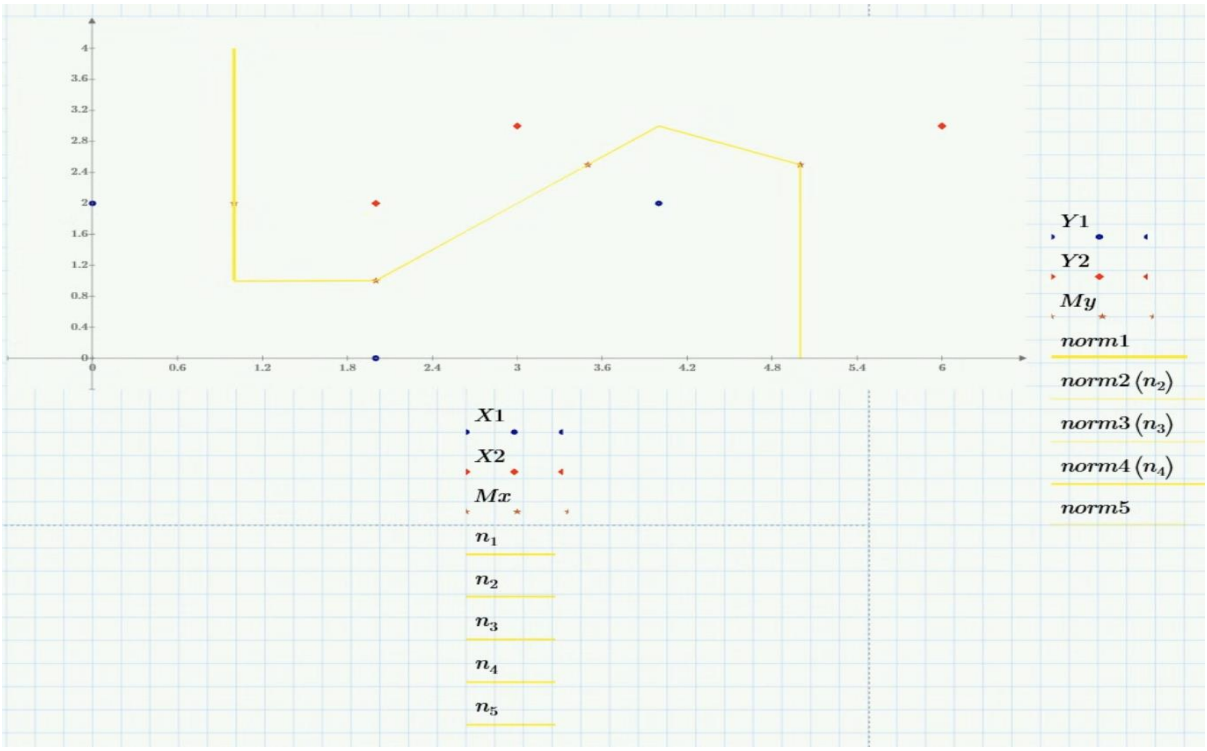
$$norm2(n) := \frac{(y_{12}+y_{24})}{2} + \left(\frac{(y_{24}-y_{12})}{300}\right)\cdot\left(n - \frac{(x_{12}+x_{24})}{2}\right)$$

Прорахуємо середні значення для всіх комбінацій

$$Mxn := \left[\frac{(x_{11} + x_{24})}{2} \right] \qquad Myn := \left[\frac{(y_{11} + y_{24})}{2} \right]$$

$$Mx := \begin{bmatrix} \frac{(x_{11} + x_{24})}{2} \\ \frac{(x_{12} + x_{24})}{2} \\ \frac{(x_{13} + x_{25})}{2} \\ \frac{(x_{13} + x_{26})}{2} \end{bmatrix} \qquad My := \begin{bmatrix} \frac{(y_{11} + y_{24})}{2} \\ \frac{(y_{12} + y_{24})}{2} \\ \frac{(y_{13} + y_{25})}{2} \\ \frac{(y_{13} + y_{26})}{2} \end{bmatrix}$$

Нанесемо нормалі на графік для зображення в подальшому межі між класами



Проміжки, які ми використали для побудови неперервної ламаної, що буде відділяти класи один від одного

$$n_1 := 1 \qquad n_3 := 2..4 \qquad n_5 := 5$$

$$n_2 := 1..2 \qquad n_4 := 4..5$$

Отже, за отриманою межею, визначимо вирішальне правило (3.3):

$$
\begin{aligned}
a_i := \ & \text{if } x_i \le 1 \\
& \quad \text{if } y_i \ge 2 \\
& \quad\quad 1 \\
& \quad \text{else} \\
& \quad\quad 2 \\
& \text{else if } 1 \le x_i \le 2 \\
& \quad \text{if } y_i \ge 1 \\
& \quad\quad 1 \\
& \quad \text{else} \\
& \quad\quad 2 \\
& \text{else if } 2 \le x_i \le 3.5 \\
& \quad \text{if } y_i \ge 2.5 \\
& \quad\quad 1 \\
& \quad \text{else} \\
& \quad\quad 2 \\
& \text{else if } 3.5 \le x_i \le 5 \\
& \quad \text{if } y_i \ge 2.5 \\
& \quad\quad 1 \\
& \quad \text{else} \\
& \quad\quad 2 \\
& \text{else if } x_i > 5 \\
& \quad 2
\end{aligned}
$$

- Якщо $x < 1$:
  - Якщо $y \geq 2$, то клас **1**, інакше клас **2**.
- Якщо x знаходиться в діапазоні $[1, 2]$:
  - Якщо $y \geq 1$, то клас **1**, інакше клас **2**.
- Якщо x знаходиться в діапазоні $[2, 3.5]$:
  - Якщо $y \geq 2.5$, то клас **1**, інакше клас **2**.
- Якщо x знаходиться в діапазоні $[3.5, 5]$:
  - Якщо $y \geq 2.5$, то клас **1**, інакше клас **2**.
- Якщо $x > 5$, то клас **2**.

Оскільки дані спостережень – випадкові величини ξі, можна провести межу між класами після статистичної обробки даних – знаходження середніх вибіркових значень. Визначимо компоненти векторів - статистичних оцінок математичних очікувань (МО) класів a1 та a2:

$$
x_i := \begin{bmatrix} x_{11} \\ x_{12} \\ x_{13} \\ x_{24} \\ x_{25} \\ x_{26} \end{bmatrix} \qquad y_i := \begin{bmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{24} \\ y_{25} \\ y_{26} \end{bmatrix}
$$

$$m1_0 := \frac{1}{3} \cdot \sum_{i=1}^{3} x_i \qquad m1_1 := \frac{1}{3} \sum_{i=1}^{3} y_i$$

$$m2_0 := \frac{1}{3} \cdot \sum_{i=4}^{6} x_i \qquad m2_1 := \frac{1}{3} \sum_{i=4}^{6} y_i$$

$$m1_0 := \frac{(x_{11} + x_{12} + x_{13})}{2} = 3 \qquad m1_1 := \frac{(y_{11} + y_{12} + y_{13})}{2} = 2$$

$$m2_0 := \frac{(x_{24} + x_{25} + x_{26})}{2} = 5.5 \qquad m2_1 := \frac{(y_{24} + y_{25} + y_{26})}{2} = 4$$

Знайдемо рівняння межі g2 між класами a1 та a2 після усереднення даних спостережень. Ця межа проходить через точку з координатами:

$$My := \begin{bmatrix} m1_1 \\ m2_1 \end{bmatrix} \quad Mx := \begin{bmatrix} m1_0 \\ m2_0 \end{bmatrix}$$

$$M\left(\frac{(m1_0 + m2_0)}{2}, \frac{(m1_1 + m2_1)}{2}\right) \rightarrow M(4.25, 3.0)$$

$$x_0 := 4.25 \qquad y_0 := 3.0$$

$$k := \frac{(m2_1 - m1_1)}{(m2_0 - m1_0)} = 0.8$$

$$kt := \frac{1}{k} = 1.25$$

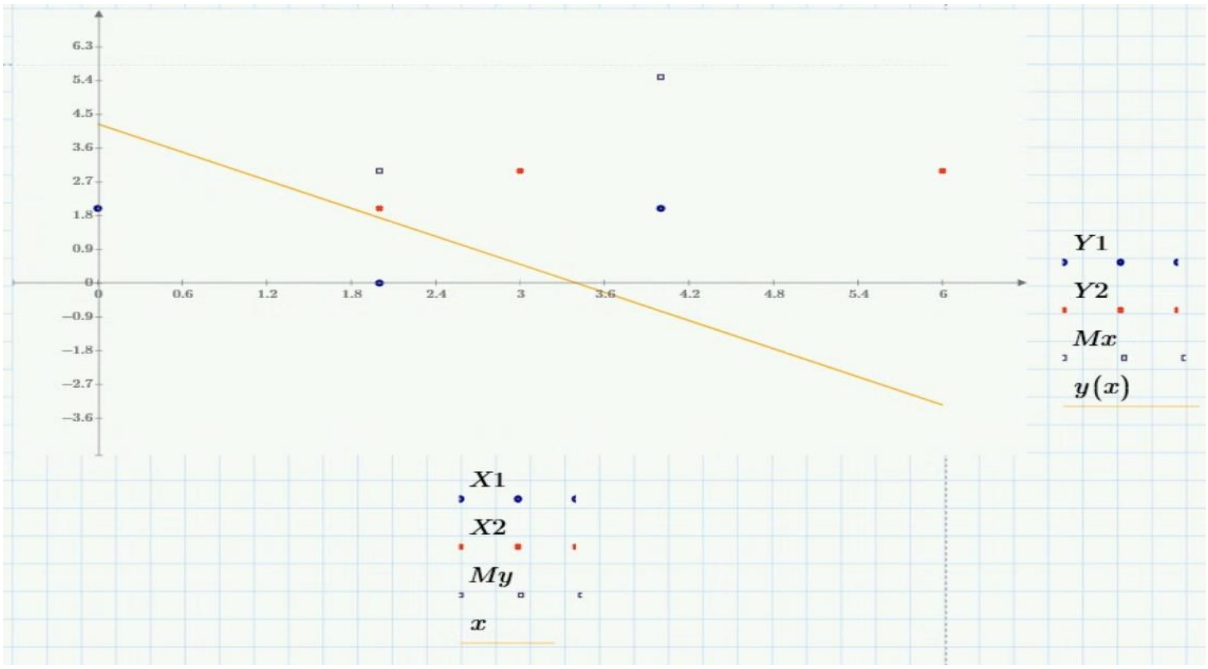
$$g_2(x, y) := -y - (x - x_0) \cdot kt + y_0$$

Отже, за отриманою межею, визначимо вирішальне правило (3.4):

$$y(x) := -kt\ x + x_0 \qquad x := m1_1 \qquad x := 0, 1..6$$

{Якщо y<= y(x) , тоді γ_1, інакше γ_2}

Зобразимо лінію на площині разом з точками:



Завдання 3

Для експериментальної перевірки якості роботи класифікаторів за правилами (3.3) і (3.4) змоделюємо результати спостережень – масиви ({x1i}; {y1i}) та ({x2i}; {y2i}), відповідні класам a1 і a2. Вважаємо, що x1 та y1 – некорельовані компоненти двовимірної випадкової величини, що підпорядковується нормальному закону розподілу з МО і середньоквадратичним відхиленням (СКВ) В якості значень параметрів розподілу приймемо їх статистичні оцінки:

$$i := 1,2..6$$

$$D1_0 := \frac{1}{2} \cdot \sum_{i=1}^{3} (x_i - m1_0)^2 \qquad \sigma1_0 := \sqrt{m1_0} = 1.732$$

$$D1_1 := \frac{1}{2} \cdot \sum_{i=1}^{3} (y_i - m1_1)^2 \qquad \sigma1_1 := \sqrt{m2_0} = 2.345$$

$$D2_0 := \frac{1}{2} \cdot \sum_{i=3}^{6} (x_i - m2_0)^2 \qquad \sigma2_0 := \sqrt{m1_1} = 1.414$$

$$D2_1 := \frac{1}{2} \cdot \sum_{i=3}^{6} (y_i - m2_1)^2 \qquad \sigma2_1 := \sqrt{m2_1} = 2$$

Визначимо функцію користувача, яка здійснює алгоритм генерації масиву реалізацій нормально розподіленої випадкової величини:

$$n := 48 \qquad k := 1..n$$

$$Norm(z, m, \sigma) := \sqrt{\frac{12}{n}} \cdot \sigma \cdot \left(\sum_k \mathrm{rnd}(1) - \frac{n}{2} \right) + m$$

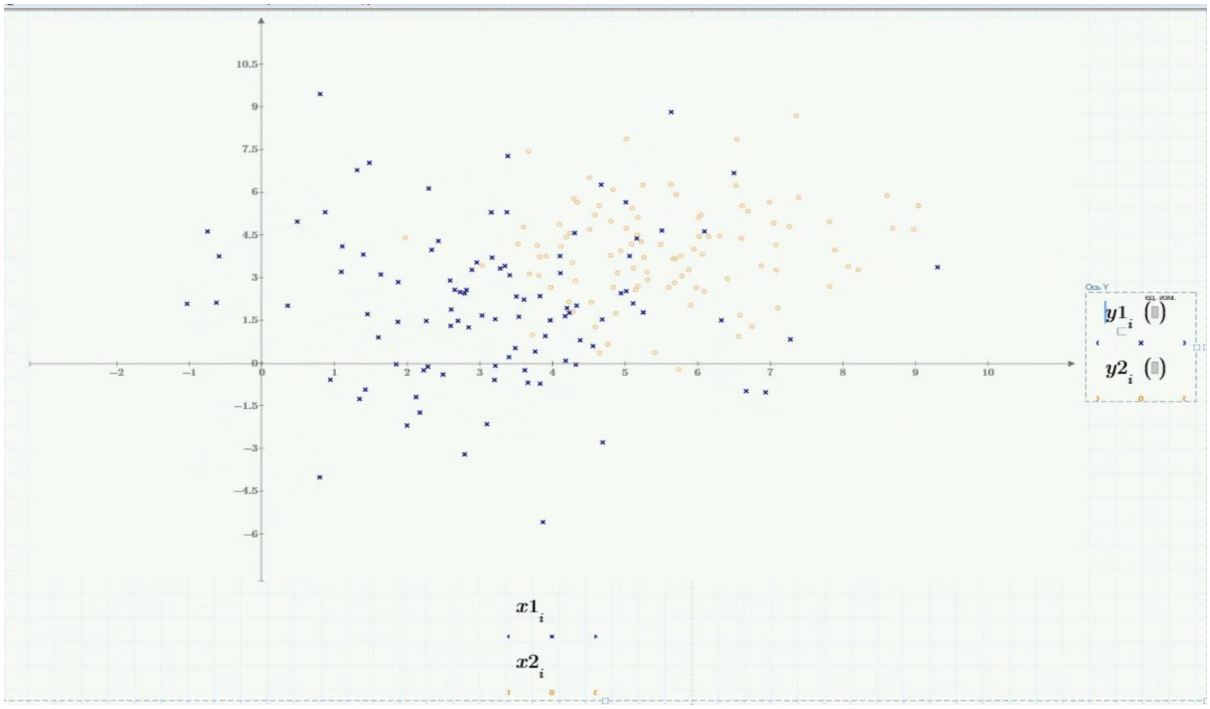
$$N := 100 \qquad i := 0..N-1$$

Формальними аргументами цієї функції є номери елементів масиву реалізацій і параметри нормального розподілу МО m і СКО σ випадкової величини, яка моделюється. Отримаємо 100 даних спостережень класу a1 та a2

$$x1_i := Norm(i, m1_0, \sigma1_0) \qquad y1_i := Norm(i, m1_1, \sigma1_1)$$

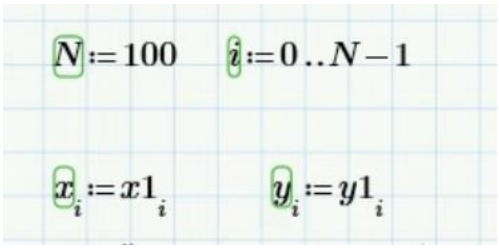
$$x2_i := Norm(i, m2_0, \sigma2_0) \qquad y2_i := Norm(i, m2_1, \sigma2_1)$$

Зобразимо їх на площині:

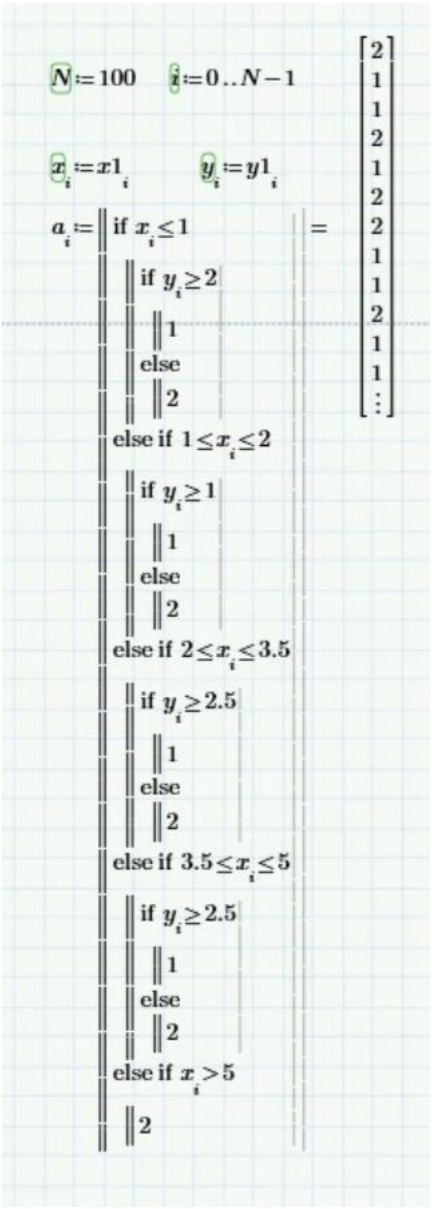


Завдання 4

Виконаємо розпізнавання контрольної вибірки ({xi}; {yi}) за вирішальним правилом (3.3). Нехай контрольна вибірка належить класу a1, тоді



Формалізуємо опис процедури ухвалення рішення (3.3):



Проведемо розпізнавання контрольної вибірки ({xi}; {yi}) за вирішальним правилом (3.4):

$$b_i := \text{if}\left(a_i \le 1, 1, 0\right)$$
$$P22 := \frac{1}{N} \cdot \sum_i b_i = 0.48$$
$$P12 := 1 - P22 = 0.52$$

Проведемо дослідження для правил 3.3 та 3.4 для класу a2:

$$N := 100 \qquad i := 0..N-1$$
$$x_i := x2_i \qquad y_i := y2_i$$

$$b_i := \text{if}\left(a_i \le 1, 1, 0\right)$$
$$P22 := \frac{1}{N} \cdot \sum_i b_i = 0.32$$
$$P12 := 1 - P22 = 0.68$$

За проведеними дослідженнями побудуємо таблицю результатів:

	Правило 3.3	Правило 3.4
P11	0.48	0.46
P21	0.52	0.54
P22	0.32	0.30
P12	0.68	0.70

Висновки

Виходячи з результатів досліджень, побудовані мною вирішальні правила за вибірковими значеннями та за узагальненим значенням дали дуже схожі

результати , а отже побудовано мною межі розділення класів є вірними. При цьому , метод вибіркових значень статистикчно був точніший у задачі правильної класифікації об’єкта і менше припустився помилок. Отже можна зробити висновок, що узагальнене значення може негативно вплинути на вірну класифікацію об’єктів.