Міністерство Освіти і Науки України

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

Звіт з лабораторної роботи № 1
з дисципліни **«Data Science та машинне навчання»**
Тема: **«Задача класичного виявлення. Статистичні критерії прийняття рішення»**

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

Київ – 2024

Мета роботи:

1. Вивчити методику побудови вирішального правила з використанням критеріїв максимальної правдоподібності і максимуму апостеріорної ймовірності
2. Отримати навички оцінювання показників якості двоальтернативного непараметричного розпізнавання

Завдання:

1. Для заданих (згідно з варіантом) значень параметрів нормальних законів розподілу (m1, σ1) і (m2, σ2), які характеризують два класи об'єктів спостереження a1 та a2, визначити умовні за класом щільності ймовірності результатів спостережень
2. Побудувати вирішальне правило за критерієм максимального правдоподібності (1.5).
3. Розрахувати теоретичні величини ймовірностей помилок розпізнавання першого та другого роду за критерієм (1.5).
4. Для заданих (згідно з варіантом) значень апріорнихймовірностей p1 та p2 появи класів a1 та a2 визначити умовні щільності повної ймовірності результатів спостережень та апостеріорні ймовірності класів a1 і a2.
5. Побудувати вирішальне правило за максимальною критерієм апостеріорної ймовірності (1.3).
6. Розрахувати теоретичні величини ймовірностей помилок розпізнавання першого та другого роду за критерієм (1.3).
7. Порівняти ефективність вирішальних правил, побудованих за критеріям максимальної правдоподібності та максимальної апостеріорної ймовірності.
8. Оформити звіт про лабораторну роботу, який повинен мати короткі теоретичні відомості, результати розрахунків, графіки досліджуваних статистичних характеристик та висновки.

Хід роботи

Варіант для виконня згідно таблиці - 3

Параметри для варіанту:
математичне очікування **m** та середньоквадратичне відхилення до класів:
- Клас 1 (a1): m1 = 0, σ1 = 1
- Клас 2 (a2): m2 = 2, σ2 = 0.8

апріорні ймовірності для класів: p1 = 0.1, p2 = 0.9
кількість точок для побудови графіку: N = 200

Завдання 1:

Для початку , необхідно визначити щільності ймовірності результатів спостережень відповідно до класу . З умови маємо, що розподіл ознак об’єктів нормальний, а отже, **функція розподілу** матиме вигляд функції користувача трьох аргументів:

$$f(z, m, \sigma) := \frac{1}{\sigma \cdot \sqrt{2 \cdot \pi}} \cdot \exp\left(\frac{-(z-m)^2}{2 \cdot \sigma^2}\right)$$

Де m - це математичне сподівання, а σ ^ 2 - це дисперсія випадкової величини.
При використанні 200 точок, рівномірно розподілених у масиві для побудови графіка, необхідно визначити мінімальні (xmin) і максимальні (xmax) значення для двох класів.

Границі діапазону встановлюються за правилом «трьох сигм», яке передбачає, що випадкова величина x, розподілена за нормальним законом, знаходиться в інтервалі m ± 3σ з імовірністю 0,997.

Для класу 1 (x ∈ a1) інтервал значень [x1min, x1max] визначається так:

$$x1min := m_1 - 3 \cdot \sigma_1 = 0.2 \qquad x1max := m_1 + 3\ \sigma_1 = 3.8$$

Для класу 2 (x ∈ a2) межі [x2min, x2max] визначаються наступним чином:

$$x2min := m_2 - 3 \cdot \sigma_2 = 0.9 \qquad x2max := m_2 + 3 \cdot \sigma_2 = 5.1$$

Верхня та нижня межа значень параметру x:

$$xmin := min(x1min, x2min) = 0.2$$
$$xmax := \max(x1max, x2max) = 5.1$$

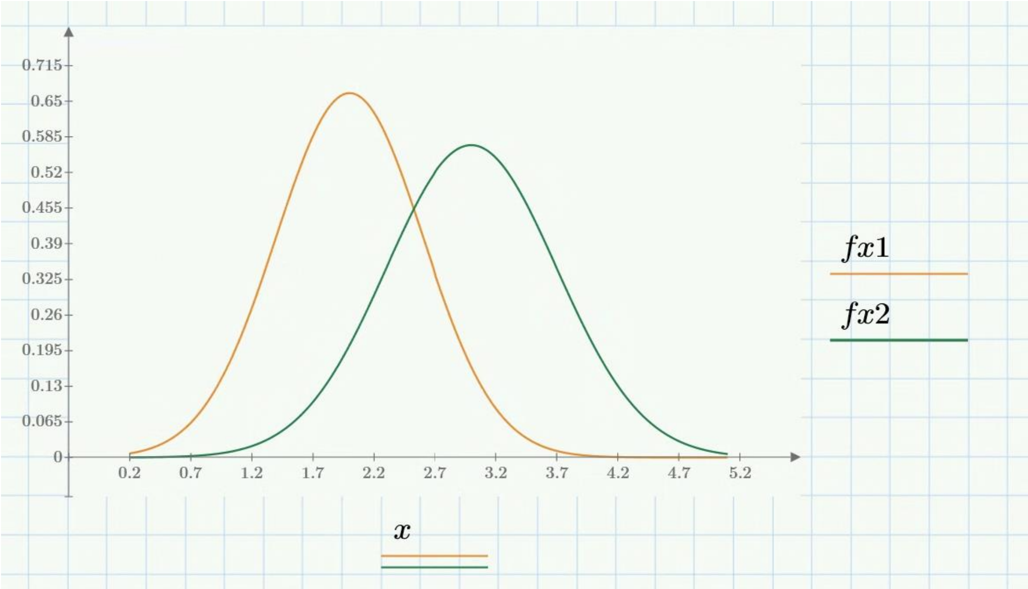
Ділимо цей інтервал на нашу кількість точок (N – 1) та визначаемо координати точок розділення:

$$i := 0..N-1 \quad x_i := xmin + \frac{(xmax - xmin)}{N-1} \cdot i = \begin{bmatrix} 0.2 \\ 0.225 \\ 0.249 \\ 0.274 \\ 0.298 \\ 0.323 \\ 0.348 \\ 0.372 \\ 0.397 \\ 0.422 \\ 0.446 \\ 0.471 \\ \vdots \end{bmatrix}$$

Маючи математичні очікування, стандартне відхилення та масив координат точок розділення по осі OX, ми можемо застосувати функцію нормального розподілу та побудувати графік для ознак класів 1 та 2:

$$fx1_i := f(x_i, m_1, \sigma_1) = \begin{bmatrix} 0.007 \\ 0.008 \\ 0.009 \\ 0.011 \\ 0.012 \\ 0.013 \\ 0.015 \\ 0.017 \\ 0.019 \\ 0.021 \\ 0.023 \\ 0.026 \\ \vdots \end{bmatrix} \quad fx2_i := f(x_i, m_2, \sigma_2) = \begin{bmatrix} 1.912 \cdot 10^{-4} \\ 2.199 \cdot 10^{-4} \\ 2.527 \cdot 10^{-4} \\ 2.9 \cdot 10^{-4} \\ 3.323 \cdot 10^{-4} \\ 3.804 \cdot 10^{-4} \\ 4.349 \cdot 10^{-4} \\ 4.966 \cdot 10^{-4} \\ 5.664 \cdot 10^{-4} \\ 6.451 \cdot 10^{-4} \\ 7.339 \cdot 10^{-4} \\ 8.339 \cdot 10^{-4} \\ \vdots \end{bmatrix}$$

Будуємо графік для умовних за класом щільності ймовірності ознаки x:



Завдання 2:

Побудуємо вирішальне правило максимальної правдоподібності для нормального розподілу ознак двох класів. У математичній статистиці — це метод оцінювання невідомого параметра шляхом максимізації функції правдоподібності. Він ґрунтується на припущенні про те, що вся інформація про статистичну вибірку міститься у цій функції. Метод використовується для створення статистичної моделі на основі даних, і забезпечення оцінки параметрів моделі.

$$x^2 \cdot \left(\sigma_2{}^2 - \sigma_1{}^2\right) + x \cdot \left(2\, m_2 \cdot \sigma_1{}^2 - 2 \cdot m_1 \cdot \sigma_2{}^2\right) + m_1{}^2 \cdot \sigma_2{}^2 - m_2{}^2 \cdot \sigma_1{}^2 - 2 \cdot \sigma_1{}^2 \cdot \sigma_2{}^2 \cdot \ln\left(\frac{\sigma_2}{\sigma_1}\right)$$

Отримуємо таке рівняння.

Виведемо наступні позначення:

$$d_1 := \sigma_1{}^2 = 0.36 \quad d_2 := \sigma_2{}^2 = 0.49 \qquad a := d_2 - d_1 = 0.13$$

$$b := 2 \cdot \left(m_2 \cdot d_1 - m_1 \cdot d_2\right) = 0.2 \qquad c := m_1{}^2 \cdot d_2 - m_2{}^2 \cdot d_1 - 2 \cdot d_1 \cdot d_2 \cdot \ln\left(\frac{\sigma_1}{\sigma_2}\right) = -1.226$$

Виводимо формулу для пошуку максимального та мінімального порогу за цим критерієм:

$$xg_1 := \frac{-b - \sqrt{b^2 - 4 \cdot a \cdot c}}{2 \cdot a} = -3.935 \qquad xg_2 := \frac{-b + \sqrt{b^2 - 4 \cdot a \cdot c}}{2 \cdot a} = 2.396$$

Зобразимо на графіку отримані межі розділу між класами xg1 та xg2. У розрахунках видно, що xg1 не потрапляє у [0.2, 5.1] масив найбільш ймовірних значень параметра x. Тому необхідно перевизначити верхню і нижню межу параметра для включення цих значень:

$$xmin := \mathbf{if}\left(xmin > xg_1, xg_1, xmin\right) = -3.935$$

$$xmax := \mathbf{if}\left(xmax < xg_2, xg_2, xmax\right) = 5.1$$

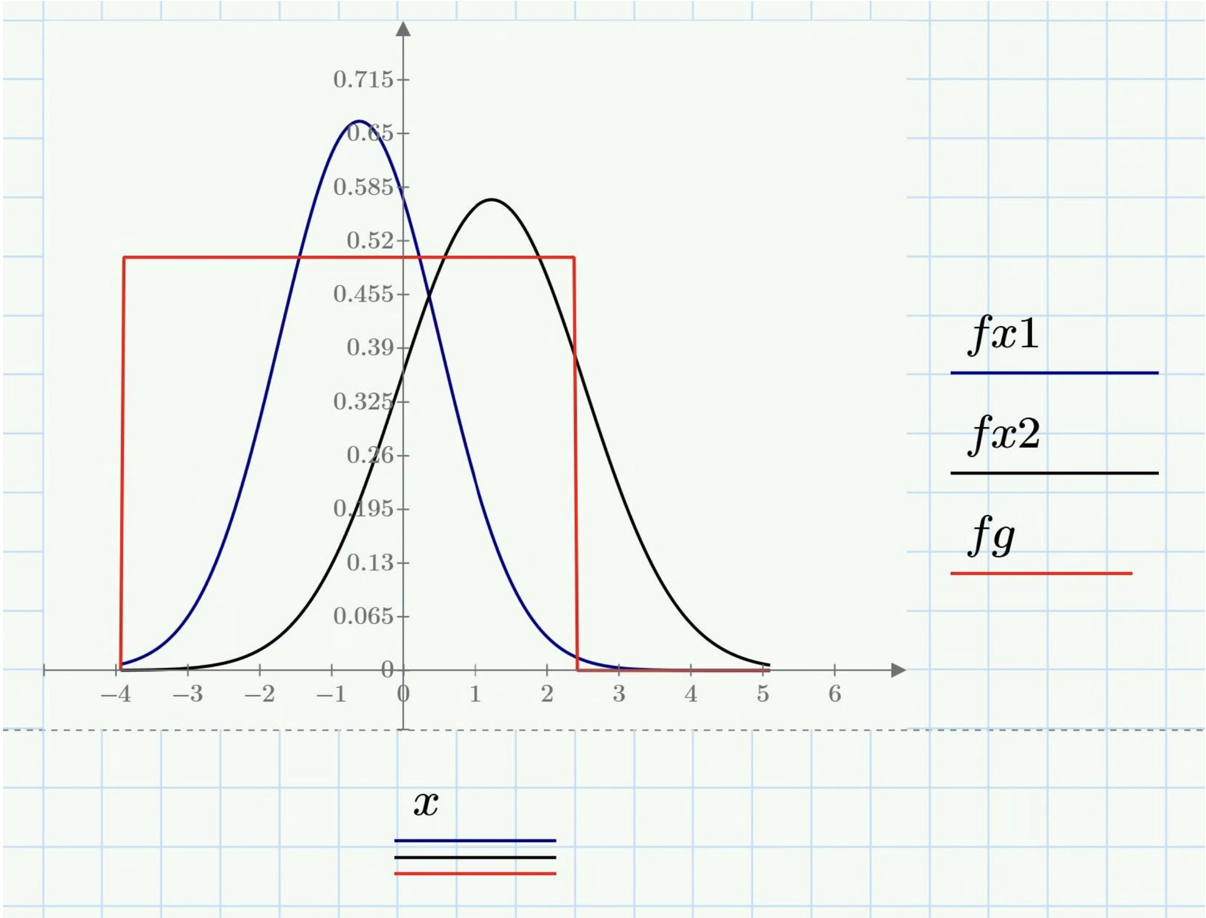
Переписуємо функцію розподілу для класів з новими значеннями та новими координатами точок розподілу:

$$x_i := xmin + \frac{(xmax - xmin)}{N-1} \cdot i = \begin{bmatrix} -3.935 \\ -3.889 \\ -3.844 \\ -3.798 \\ -3.753 \\ -3.708 \\ -3.662 \\ -3.617 \\ -3.571 \\ -3.526 \\ -3.481 \\ -3.435 \\ \vdots \end{bmatrix}$$

Для відображення порогу розділення классів використовуємо квадратну функцію:

$$fg_i := \mathbf{if}\left(xg_1 < x_i < xg_2, 0.5, 0\right)$$

Будуємо графік:



Завдання 3:

Розрахувуєм теоретичні величини ймовірностей помилок розпізнавання першого та другого роду за критерієм “Максимальна правдоподібність”

Помилка першого роду - ймовірність віднести ознаку до класу а1, коли він насправді належить класу а2:

$$P_{21} := \int_{xmin}^{xg_1} f\left(z, m_2, \sigma_2\right) \mathrm{d}z + \int_{xg_1}^{xmax} f\left(z, m_2, \sigma_2\right) \mathrm{d}z = 0.999$$

Помилка другого роду - ймовірність прийняття рішення на користь класу а2, коли в реальності спостерігається клас а1:

$$P_{12} := \int_{xg_1}^{xg_2} f\left(z, m_1, \sigma_1\right) \mathrm{d}z = 0.745$$

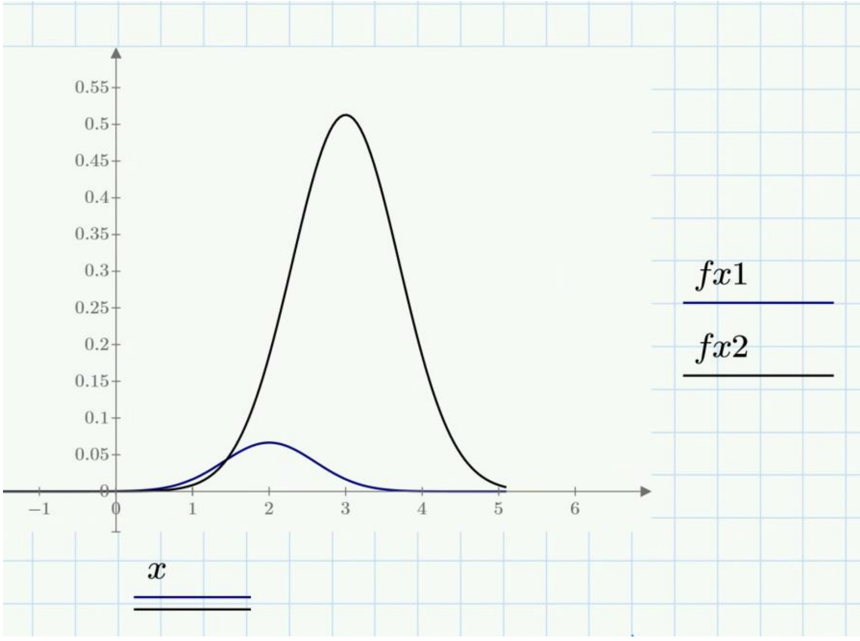
Ймовірність розпізнавання за нашими показниками буде:

$$P := 1 - 0.5 \cdot \left(P_{21} + P_{12}\right) = 0.128$$

Використовуючи значення апріорних ймовірностей з пункту 1, на інтервалі значень [xmin, xmax] побудуємо графік з урахуванням цих значень для визначення умовних щільностей повної ймовірності (задавши відповідні функції):

$$fx1_i := p_1 \cdot f\left(x_i, m_1, \sigma_1\right)$$

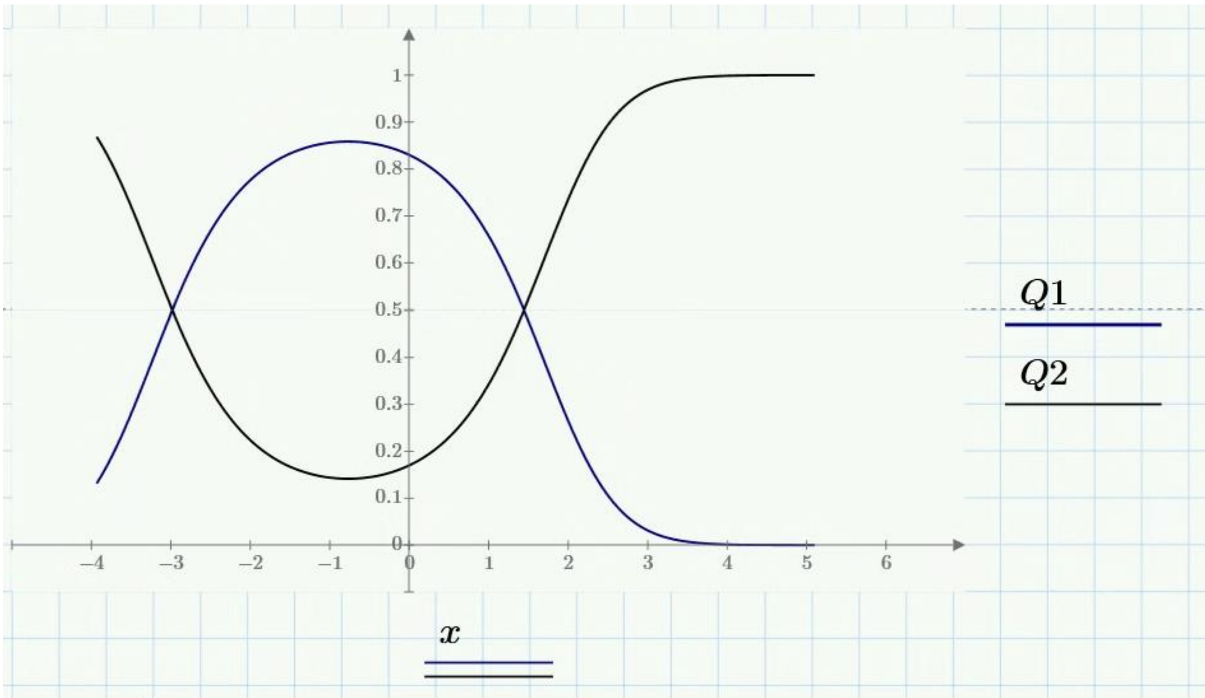
$$fx2_i := p_2 \cdot f\left(x_i, m_2, \sigma_2\right)$$



Також, маючи відомі значення густини ймовірності для функцій обох класів на певному інтервалі, визначимо апостеріорні ймовірності для обох цих класів за формулами:

$$Q1_i := \frac{fx1_i}{fx1_i + fx2_i} \qquad Q2_i := \frac{fx2_i}{fx1_i + fx2_i}$$

Побудуємо графік:



Завдання 5:

Побудуємо вирішальне правило за максимальною критерієм апостеріорної ймовірності (1.3)

Введемо позначення для скорочення обрахунків квадратного рівняння вигляду:

$$d_1 := {\sigma_1}^2 = 0.36$$

$$d_2 := {\sigma_2}^2 = 0.49$$

$$a := d_2 - d_1 = 0.13$$

$$b := 2 \cdot (m_2 \cdot d_1 - m_1 \cdot d_2) = 0.2$$

$$c := {m_1}^2 \cdot d_2 - {m_2}^2 \cdot d_1 - 2 \cdot d_1 \cdot d_2 \cdot \ln\left(\frac{\sigma_1 \cdot p_1}{\sigma_2 \cdot p_2}\right) = -0.45$$

$$x^2(\sigma_2^2 - \sigma_1^2) + x(2m_2\sigma_1^2 - 2m_1\sigma_2^2) + m_1^2\sigma_2^2 - m_2^2\sigma_1^2 - 2\sigma_1^2\sigma_2^2 \ln\left(\frac{\sigma_2 \mathrm{p1}}{\sigma_1 \mathrm{p2}}\right)$$

Знаходимо мінімальний та максимальні пороги xg1, xg2:

$$xg_1 := \frac{-b - \sqrt{b^2 - 4 \cdot a \cdot c}}{2 \cdot a} = -2.783$$

$$xg_2 := \frac{-b + \sqrt{b^2 - 4 \cdot a \cdot c}}{2 \cdot a} = 1.245$$

Для того, щоб пороги потрапили у інтервал функції розподілу , необхідно перевизначити цей інтервал з урахуванням мінімального і максимального порогів:

$$xmin := \mathbf{if}\,(xmin > xg_1, xg_1, xmin) = -3.935$$

$$xmax := \mathbf{if}\,(xmax < xg_2, xg_2, xmax) = 5.1$$

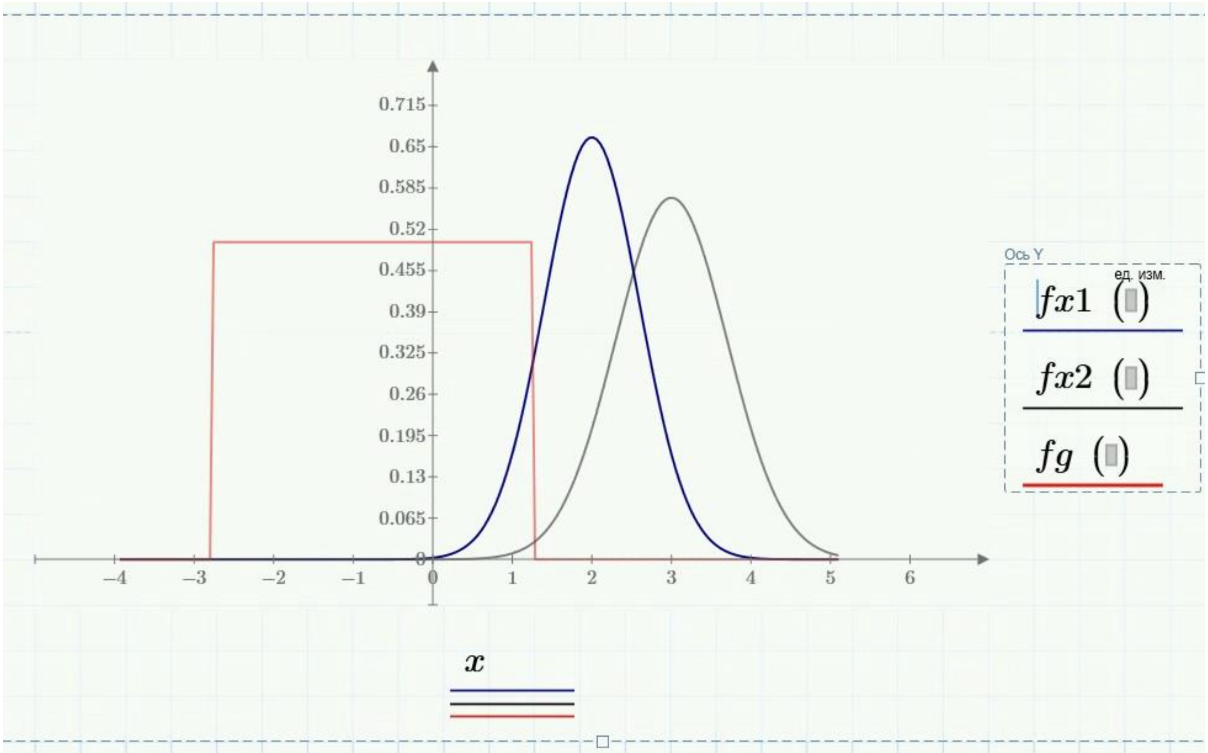
Також, необхідно перевизначити координати точок розділу графіка на N-1 частин, і перемалювати з їх урахуванням графіки розподілу для обох класів. Також, візуалізуємо пороги визначення класів за допомогою квадратної функції:

$$x_i := xmin + \frac{(xmax - xmin)}{N-1} \cdot i$$

$$fx1_i := f\left(x_i, m_1, \sigma_1\right) \qquad fx2_i := f\left(x_i, m_2, \sigma_2\right)$$

$$fg_i := \mathbf{if}\left(xg_1 < x_i < xg_2, 0.5, 0\right)$$

Малюємо графік:



Завдання 6:

Розрахуємо теоретичні величини ймовірностей помилок розпізнавання першого та другого роду за критерієм (1.3). Максимальна апостеріольна ймовірність

Ймовірність помилки першого роду:

$$P_{21} := \int_{xmin}^{xg_1} f\left(z, m_2, \sigma_2\right) \, \mathrm{d}z + \int_{xg_1}^{xmax} f\left(z, m_2, \sigma_2\right) \, \mathrm{d}z = 0.999$$

Ймовірність помилки другого роду:

$$P_{12} := \int_{xg_1}^{xg_2} f(z, m_1, \sigma_1)\, dz = 0.104$$

Ймовірність правильного розпізнавання:

$$P := 1 - (p_2 \cdot P_{21} + p_1 \cdot P_{12}) = 0.091$$

Завдання 7:

Порівняємо ефективність вирішальних правил, побудованих за критеріям максимальної правдоподібності та максимальної апостеріорної ймовірності.

За розрахунками даної лабораторної роботи, вирішальне правило за критерієм апостеріорної ймовірності дало нижчий результат ймовірності правильного розпізнавання об’єкта і співвідношення його з конкретним класом.

Висновки

У ході лабораторної роботи я освоїв методику побудови вирішальних правил, використовуючи максимуму апостеріорної ймовірності та критерії максимальної правдоподібності. Використання цих підходів дозволяє приймати обґрунтовані рішення на основі статистичних даних.