ЛЕКЦІЯ 2
ЯДЕРНІ ФУНКЦІЇ В ЗАДАЧАХ КЛАСИФІКАЦІЇ

Вислів «ядро» (kernel) означає центр, зерно чогось, наприклад, алгоритму, формули, системи тощо. У різних галузях це різні речі. У нашому випадку ми умовимося, що це просто функція подібності, і розглядатимемо її між двома векторами (хоча це не завжди так, може бути і більше, і можна порівнювати з їх допомогою навіть функції).

Візьмемо найпростіший приклад – лінійне ядро, воно ж скалярне множення, воно внутрішній продукт. У нас є два вектори [3, 4] та [5, 6]. Ми їх скалярно перемножили:

$3 * 5 + 4 * 6$ та отримали 39.

Це і є наша міра подібності. У контексті лінійного ядра її складно інтерпретувати, бо це проекція одного вектора в інший, помножена на довжину вектора, який проектується. Але що більше саме число (скаляр), то більше схожі вектора, і навпаки.

Перетворення із підйомом.

Отже, коли ми зрозуміли концепцію ядра, можна перейти до маніпуляцій з даними. Сам термін «перетворення з підйомом» не є загальноприйнятим, але добре допомагає відрізнити цей метод від перетворення з ядром (який ми розглянемо пізніше).

- Задача класифікації. У машинному навчанні у багатьох моделях завдання класифікації зводиться до пошуку гіперплощини, яка максимально точно розділить класи. Але не завжди у вихідному просторі ознак дані лінійно розділяються, тому ми можемо «підняти» дані в більш високий простір ознак, де вони, можливо, будуть розділено лінійно, як показано на малюнках (рис. 1, 2).

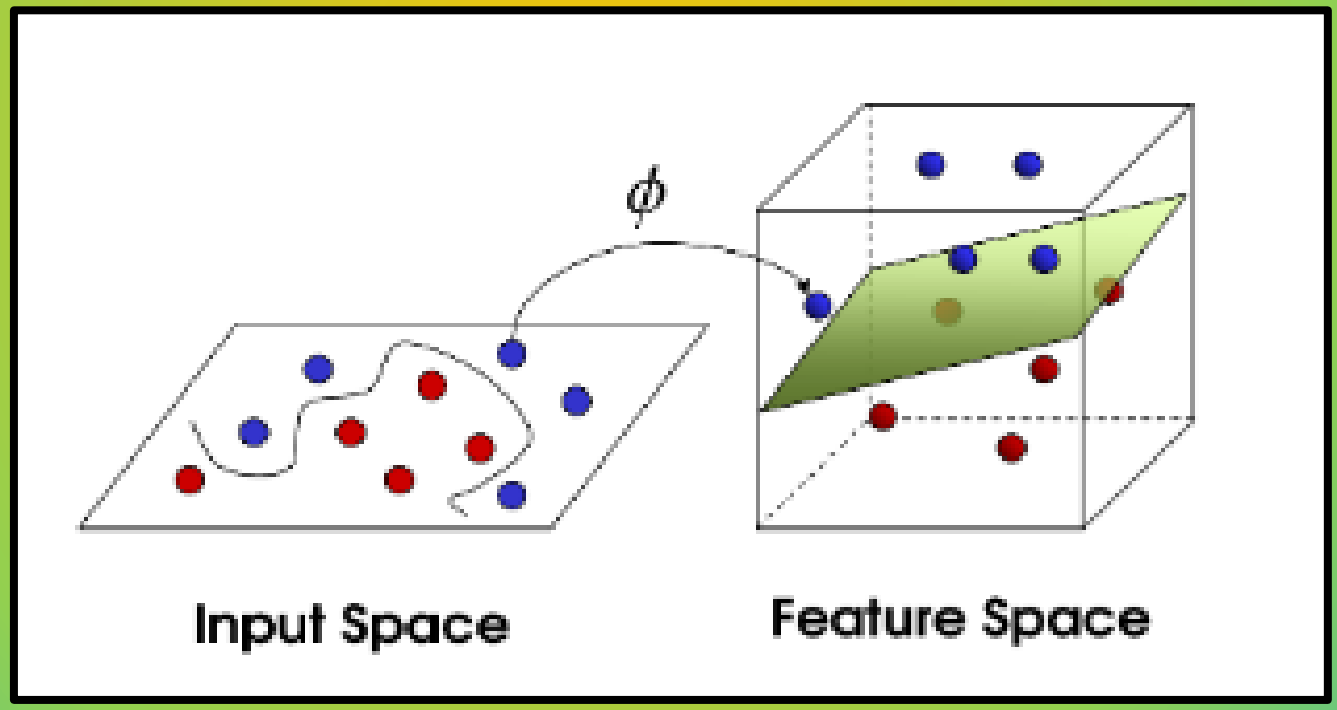


Рис.1. Перетворення простору.

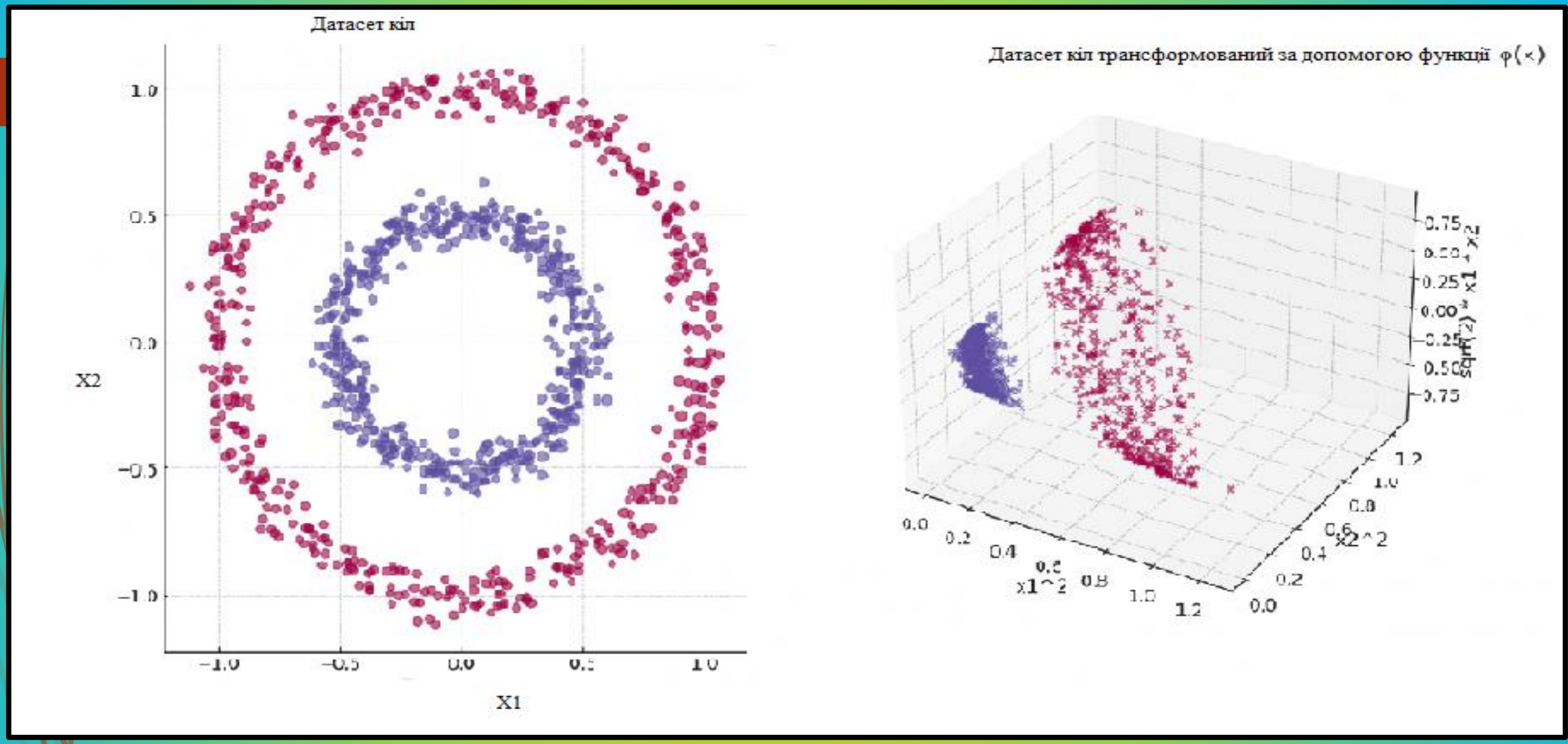


Рис. 2. Трансформація датасету.

Функція $\varphi(\vec{x}) = \begin{bmatrix} x_1^2 \\ x_2^2 \\ \sqrt{2} \cdot x_1 \cdot x_2 \end{bmatrix}$, виконує відображення двовимірного простору в тривимірний.¶

Як бачимо, дані в тривимірному просторі лінійно розділяються. Ми беремо простір R^D, використовуємо перетворення з підйомом у простір R^J, де $D < J$, а потім застосовуємо метод типу логістичної регресії, для лінійної класифікації. Однак це може бути громіздким для функції $\varphi(\vec{x})$. Для N точок даних, «піднятих» у J вимірювань, нам знадобиться O(N × J) операцій тільки для попередньої обробки даних. Але ми можемо уникнути обчислення $\varphi(\vec{x})$ повністю, продовжуючи робити лінійну класифікацію у цьому піднятому просторі, використовуючи певні перетворення. Ці перетворення, називаються **перетворення з ядром**.¶

МАТРИЦЯ-ГРАМА.¶

У лінійній алгебрі матриця Грама для набору векторів $v_1, \ldots, v_n$ у Гільбертовому просторі є ермітовою матрицею скалярних добутків, елементи якої задані скалярними добутками $G_{ij} = \langle v_i, v_j \rangle$.¶

Якщо ми застосуємо скалярне множення (лінійне ядро, «внутрішній продукт») попарно до всіх даних у нашому наборі, то отримаємо матрицю Грама, яка відображає лінійні залежності між нашими даними, формуючи **вихідний** простір ознак. Бо наше ядро $k(x_n, x_m) = \varphi(x_n) \cdot \varphi(x_m)$, де $\varphi(\vec{x}) = \vec{x}$.¶

Приклад:

Візьмемо 4 вектори в 2-х мірному просторі і попарно знайдемо їх подібність, побудувавши матрицю Грама.

$$\mathbf{v}_1 = \begin{bmatrix} 1 \\ 2 \end{bmatrix}, \quad \mathbf{v}_2 = \begin{bmatrix} 3 \\ 6 \end{bmatrix}, \quad \mathbf{v}_3 = \begin{bmatrix} 1 \\ 10 \end{bmatrix}, \quad \mathbf{v}_4 = \begin{bmatrix} -100 \\ -100 \end{bmatrix}$$

$$G = \begin{bmatrix} \mathbf{v}_1 \cdot \mathbf{v}_1 & \mathbf{v}_1 \cdot \mathbf{v}_2 & \mathbf{v}_1 \cdot \mathbf{v}_3 & \mathbf{v}_1 \cdot \mathbf{v}_4 \\ \mathbf{v}_2 \cdot \mathbf{v}_1 & \mathbf{v}_2 \cdot \mathbf{v}_2 & \mathbf{v}_2 \cdot \mathbf{v}_3 & \mathbf{v}_2 \cdot \mathbf{v}_4 \\ \mathbf{v}_3 \cdot \mathbf{v}_1 & \mathbf{v}_3 \cdot \mathbf{v}_2 & \mathbf{v}_3 \cdot \mathbf{v}_3 & \mathbf{v}_3 \cdot \mathbf{v}_4 \\ \mathbf{v}_4 \cdot \mathbf{v}_1 & \mathbf{v}_4 \cdot \mathbf{v}_2 & \mathbf{v}_4 \cdot \mathbf{v}_3 & \mathbf{v}_4 \cdot \mathbf{v}_4 \end{bmatrix}$$

$$\mathbf{v}_1 \cdot \mathbf{v}_1 = 1 \cdot 1 + 2 \cdot 2 = 1 + 4 = 5$$
$$\mathbf{v}_1 \cdot \mathbf{v}_2 = 1 \cdot 3 + 2 \cdot 6 = 3 + 12 = 15$$
$$\mathbf{v}_1 \cdot \mathbf{v}_3 = 1 \cdot 1 + 2 \cdot 10 = 1 + 20 = 21$$
$$\mathbf{v}_1 \cdot \mathbf{v}_4 = 1 \cdot (-100) + 2 \cdot (-100) = -100 - 200 = -300$$
$$\mathbf{v}_2 \cdot \mathbf{v}_2 = 3 \cdot 3 + 6 \cdot 6 = 9 + 36 = 45$$
$$\mathbf{v}_2 \cdot \mathbf{v}_3 = 3 \cdot 1 + 6 \cdot 10 = 3 + 60 = 63$$
$$\mathbf{v}_2 \cdot \mathbf{v}_4 = 3 \cdot (-100) + 6 \cdot (-100) = -300 - 600 = -900$$
$$\mathbf{v}_3 \cdot \mathbf{v}_3 = 1 \cdot 1 + 10 \cdot 10 = 1 + 100 = 101$$
$$\mathbf{v}_3 \cdot \mathbf{v}_4 = 1 \cdot (-100) + 10 \cdot (-100) = -100 - 1000 = -1100$$
$$\mathbf{v}_4 \cdot \mathbf{v}_4 = (-100) \cdot (-100) + (-100) \cdot (-100) = 10000 + 10000 = 20000$$

$$G = \begin{bmatrix} 5 & 15 & 21 & -300 \\ 15 & 45 & 63 & -900 \\ 21 & 63 & 101 & -1100 \\ -300 & -900 & -1100 & 20000 \end{bmatrix}$$

Ця матриця не змінює наш вихідний простір ознак: дані залишаються у 2-мірному просторі, незважаючи на наявність 4 стовпців, оскільки кожен стовпець є елементом, що формує наш вихідний простір ознак.

Повернемося до нашої функції $\varphi(\vec{x})$ і виконаємо наступну операцію: піднімемо наші попередні вектори у 3-мірний простір, попарно обчислимо їхній «внутрішній продукт» і побудуємо матрицю Грама.

Фактично формула кожного елемента матриці виглядатиме так:

$$\varphi(\mathbf{x}_n)^\top \varphi(\mathbf{x}_m) = \begin{bmatrix} x_{n,1}^2 & x_{n,2}^2 & \sqrt{2}x_{n,1}x_{n,2} \end{bmatrix} \cdot \begin{bmatrix} x_{m,1}^2 \\ x_{m,2}^2 \\ \sqrt{2}x_{m,1}x_{m,2} \end{bmatrix}$$

$$= x_{n,1}^2 x_{m,1}^2 + x_{n,2}^2 x_{m,2}^2 + 2x_{n,1}x_{n,2}x_{m,1}x_{m,2}$$

Таким чином, ми отримаємо відображення наших даних у 3-мірному просторі. Ми використали перетворення з підвищенням розмірності, при цьому обчислено ще й скалярний добуток. Розкриємо це питання. Перейдемо до перетворення з ядром.

Перетворення з ядром.

Ми обчислюємо, як у нашому алгоритмі, кожен елемент матриці наступним чином:

$$\varphi(\mathbf{x}_n)^\top \varphi(\mathbf{x}_m) = \begin{bmatrix} x_{n,1}^2 & x_{n,2}^2 & \sqrt{2}x_{n,1}x_{n,2} \end{bmatrix} \cdot \begin{bmatrix} x_{m,1}^2 \\ x_{m,2}^2 \\ \sqrt{2}x_{m,1}x_{m,2} \end{bmatrix} =$$

$$x_{n,1}^2 x_{m,1}^2 + x_{n,2}^2 x_{m,2}^2 + 2x_{n,1}x_{n,2}x_{m,1}x_{m,2}$$

Виконаємо це перетворення за допомогою поліноміального ядра:

$$\left(\mathbf{x}_n^\top \mathbf{x}_m\right)^2 = \left(\begin{bmatrix} x_{n,1} & x_{n,2} \end{bmatrix} \cdot \begin{bmatrix} x_{m,1} \\ x_{m,2} \end{bmatrix} \right)^2 =$$

$$\left(x_{n,1}x_{m,1} + x_{n,2}x_{m,2}\right)^2 =$$

$$\left(x_{n,1}x_{m,1}\right)^2 + \left(x_{n,2}x_{m,2}\right)^2 + 2\left(x_{n,1}x_{m,1}\right)\left(x_{n,2}x_{m,2}\right) = \varphi(\mathbf{x}_n)^\top \varphi(\mathbf{x}_m)$$

Розберемо суть виконаних перетворень:

Замість того, щоб переносити наші дані в 3-мірний простір і обчислювати скалярний добуток, ми тільки-но обчислили внутрішній продукт у двовимірному просторі і потім звели суму в квадрат. Хоча обидва варіанти мають однакову кількість математичних символів, фактична кількість операцій для другого підходу набагато менша. Це відбувається тому, що внутрішній продукт у двовимірному просторі – це два множення та сума. Квадрат це просто квадрат скаляра, тому всього 4 операції. Перший підхід займав 9 операцій.

У цьому полягає суть перетворення з ядром: ми можемо уникнути «дорогих» операцій переведення у велику розмірність, знайшовши відповідну функцію ядра $k(x_n, x_m)$, ***еквівалентну*** **скалярному множенню у просторі вищого виміру. Це дозволяє нам дешево побудувати матрицю Грама, дбаючи лише про її масштабування.**

Отримана матриця Грама моделюватиме лінійні залежності, але вже у вищому просторі. Іншими словами, перетворення з ядром дозволяє «дешево» (з точки зору алгоритмічних етапів) виконати перетворення з підвищенням розмірності

Теорема Мерсера.

Теорема Мерсера у роботах по функціональному аналізу, дає умови формування ядерного перетворення в машинному навчанні.

Джеймс Мерсер довів, що будь-яка позитивно визначена функція $k(x_n, x_m)$ при $(x_n, x_m) \in R^D$ визначає скалярний добуток векторного простору V. Таким чином, якщо ми маємо функцію φ(?), така що $\langle \varphi(x_n), \varphi(x_m) \rangle^V$ є допустимим скалярним добутком в V, ми знаємо, що існує функція ядра, яка може виконати перетворення з підвищенням розмірності «дешево». Як альтернатива, якщо ми маємо позитивно визначене ядро, ми можемо реконструювати його неявну базисну функцію φ(?).

Нехай X буде непустою множиною, іноді званою індексною множиною. Симетрична функція K: X × X > R називається позитивно напіввизначеним ядром на X, якщо:

$$\sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j K(x_i, x_j) \geq 0,$$

Де $x_1,.....x_n \in X$, $n \in N, c_1,.....c_n \in R$.

Тобто, якщо матриця Грама $\boldsymbol{K}$, елементи якої визначено як $K_{ij} = K(x_i, x_j)$, є позитивно напіввизначеною для всіх наборів x_i.

Це рівносильно тому, що власні значення λ_i матриці Грама $\boldsymbol{K}$, задовільняють умови:

$\lambda_i \geq 0$ *для всіх* i.

Таким чином, ядро може виконати перетворення з ядром при умові:

- Ядро є безперервною функцією;
- Ядро є симетричним;
- Власні значення матриці Грама, побудованої за допомогою ядра не від'ємні.

Приклади позитивно визначених ядер:

Деякі приклади позитивно визначених ядер, визначених в евклідовому просторі R^D, включають:

Лінійне ядро:

$$K(x, y) = x^T y, \quad x, y \in R^d;$$

Поліноміальне ядро:

$$K(\mathbf{x}, \mathbf{y}) = (\mathbf{x}^T \mathbf{y} + r)^n, \quad \mathbf{x}, \mathbf{y} \in \mathbb{R}^d, r \geq 0, n \geq 1$$

Гаусове ядро (RBF ядро):

$$K(\mathbf{x}, \mathbf{y}) = e^{-\frac{\|\mathbf{x}-\mathbf{y}\|^2}{2\sigma^2}}, \quad \mathbf{x}, \mathbf{y} \in \mathbb{R}^d, \sigma > 0,$$

или

$$K(\mathbf{x}, \mathbf{y}) = e^{-\lambda\|\mathbf{x}-\mathbf{y}\|^2}, \quad \mathbf{x}, \mathbf{y} \in \mathbb{R}^d, \lambda > 0$$

Де σ та λ – це ширина ядра. У першому випадку чим менше значення σ, тим менше значення під експонентою і тим чутливіше ядро до змін. У другому, навпаки, що більше значення λ, то більш чутливо.

Нескінченномірний простір та Гаусове (RBF) ядро.

Зараз ми зрозуміємо алгебраїчно, в чому суть гаусівського ядра, щодо його здатності повертати скалярний добуток в нескінченномірному просторі.

Вихідний вираз для ядра RBF при $\lambda = 1/2$

$$k_{\mathrm{RBF}}(x, y) = \exp\left(-\frac{1}{2}\|x - y\|^2\right)$$

Норма різниці у вигляді скалярного добутку:

$$\|x - y\|^2 = (x - y) \cdot (x - y)$$

Розкриємо дужки в скалярному добутку:

$$(x - y) \cdot (x - y) = \langle x - y, x - y \rangle$$

Лінійність та комутативність скалярного добутку:

$$\langle x - y, x - y \rangle = \langle x, x \rangle - \langle x, y \rangle - \langle y, x \rangle + \langle y, y \rangle$$

Вираз спрощується:

$$\langle x - y, x - y \rangle = \langle x, x \rangle - 2\langle x, y \rangle + \langle y, y \rangle$$

Підставляємо в експоненту:

$$k_{\text{RBF}}(x, y) = \exp\left(-\frac{1}{2}(\langle x, x\rangle - 2\langle x, y\rangle + \langle y, y\rangle)\right)$$

Розділення експоненти:

$$\exp\left(-\frac{1}{2}(\langle x, x\rangle - 2\langle x, y\rangle + \langle y, y\rangle)\right) =$$

$$\exp\left(-\frac{1}{2}\langle x, x\rangle\right) \cdot \exp\left(-\frac{1}{2}\langle y, y\rangle\right) \cdot \exp(\langle x, y\rangle)$$

Визначення складових:

Представимо перші два множники як ??, тому що в даному випадку вони нас не цікавлять через те, що враховують норму векторів, а не взаємозв'язок між векторами x та y:

$$C = \exp\left(-\frac{1}{2}\|x\|^2\right) \cdot \exp\left(-\frac{1}{2}\|y\|^2\right)$$

Оновлений вираз для ядра RBF:

$$k_{\text{RBF}}(x, y) = C \cdot \exp(\langle x, y\rangle)$$

Розкладемо експоненту в ряд Тейлора:

$$\exp(x) = \sum_{r=0}^{\infty} \frac{x^r}{r!}$$

Застосуємо ряд Тейлора до множника $\exp(\langle x, y \rangle)$:

$$\exp(\langle x, y \rangle) = \sum_{r=0}^{\infty} \frac{\langle x, y \rangle^r}{r!}$$

Поліноміальне ядро:

$$k_{\text{poly}(r)}(x, y) = (\langle x, y \rangle)^r$$

Підстановка поліноміального ядра:

$$\exp(\langle x, y \rangle) = \sum_{r=0}^{\infty} \frac{k_{\text{poly}(r)}(x, y)}{r!}$$

Підсумковий вираз для ядра RBF:

$$k_{\text{RBF}}(x, y) = C \cdot \sum_{r=0}^{\infty} \frac{k_{\text{poly}(r)}(x, y)}{r!}$$

$$C = \exp\left(-\frac{1}{2}\|x\|^2\right) \cdot \exp\left(-\frac{1}{2}\|y\|^2\right)$$

Висновок:

- *Таким чином, ядро RBF можна розглядати як нескінченну суму поліноміальних ядер. Тому матриця Грама, побудована з їх допомогою, моделюватиме наші дані в нескінченномірному просторі.*

Для ілюстрації, візьмемо наш попередній датасет make_circles і навчимо на ньому повнозв'язну нейронну мережу без функцій активації (2 входи, 3 нейрони в прихованому шарі, 2 виходи) і відобразимо дані після виходу із прихованого шару (рис. 1).

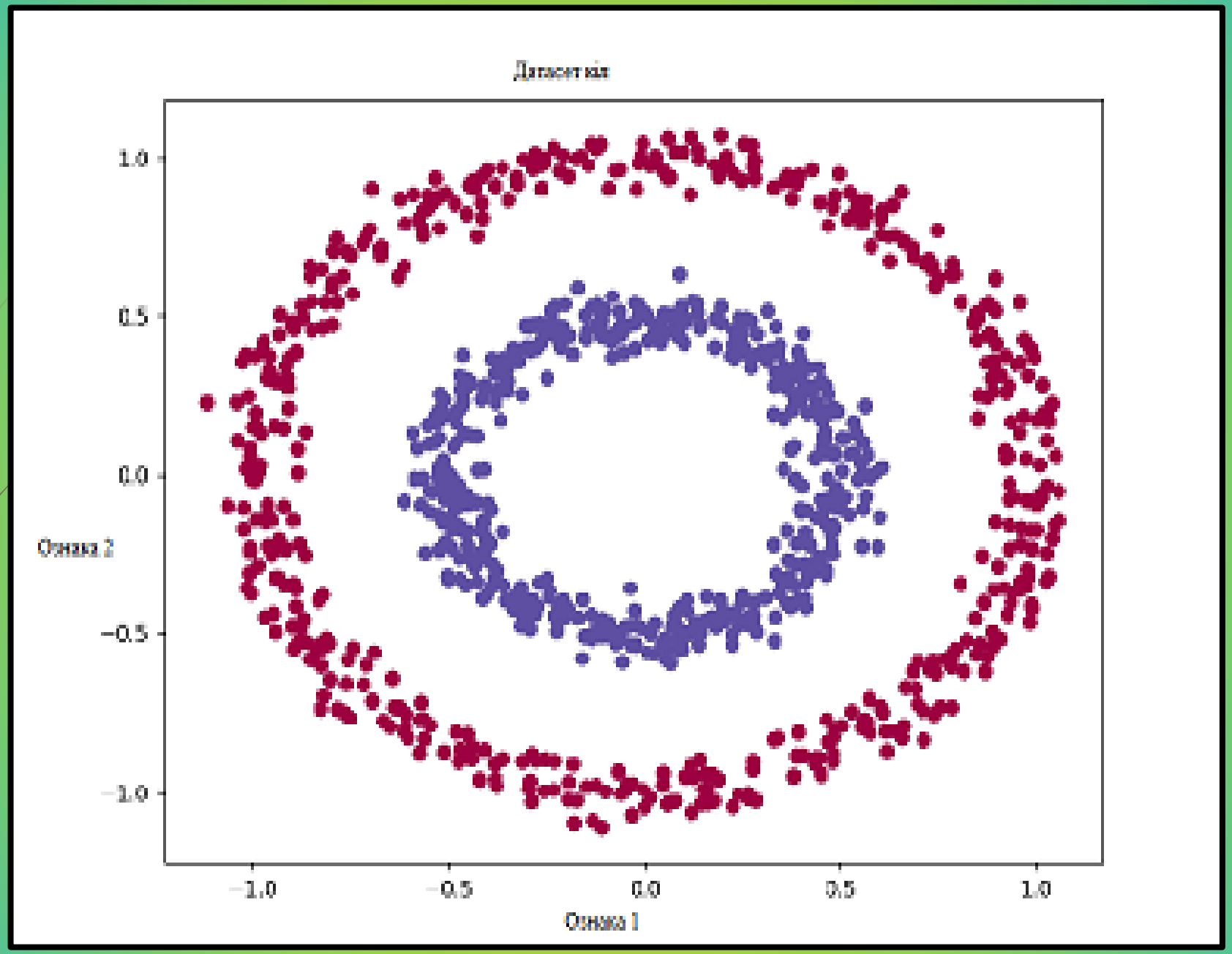


Рис.1

Виходи прихованого шару нейромережі (рис. 2)

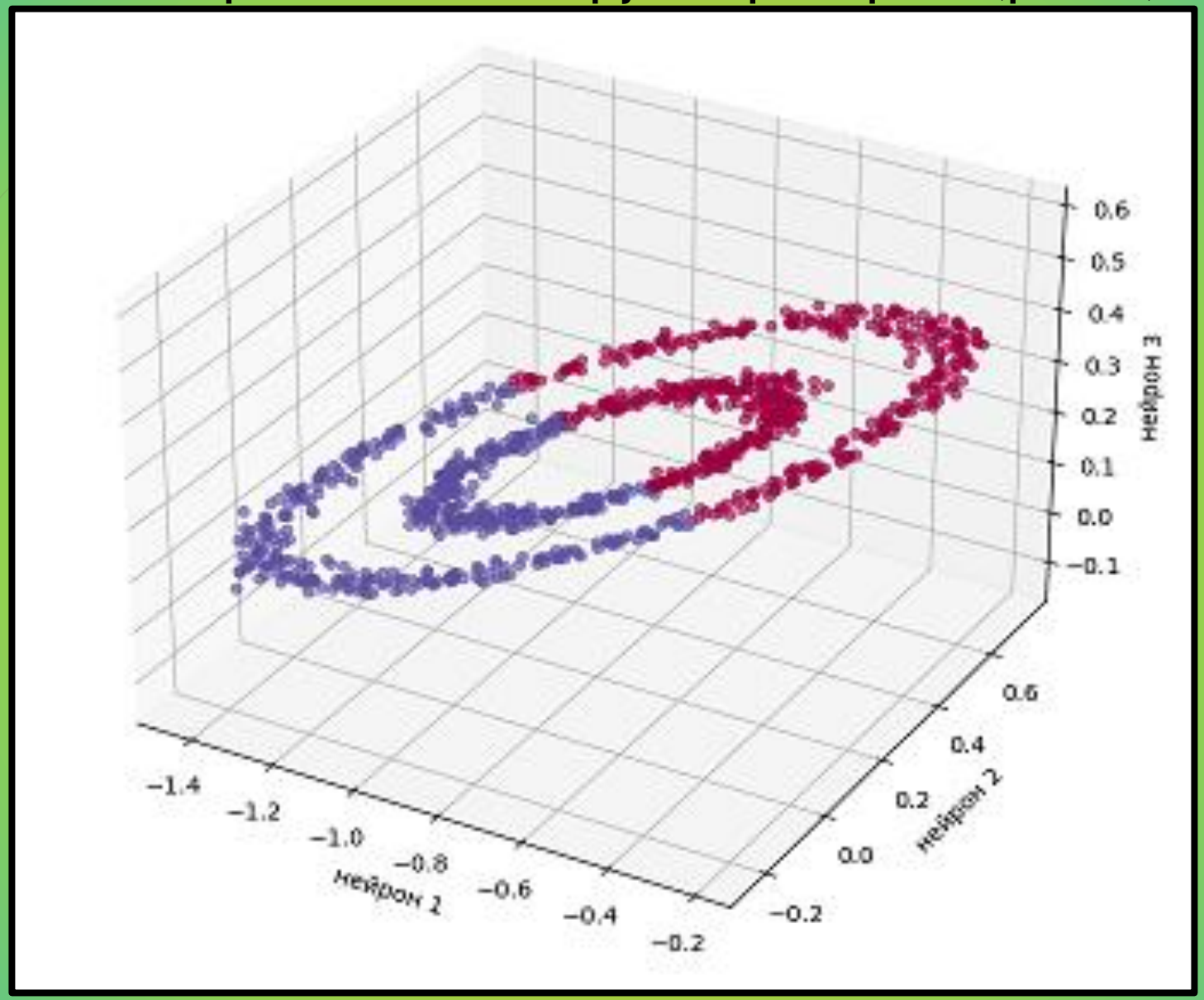


Рис. 2 Виходи прихованого шару нейромережі

- Як можна помітити, без функції активації наша функція ϕ(·), яка представлена прихованим шаром, здатна зробити перетворення з підйомом, моделюючи лінійні залежності між даними. Тому наш класифікатор у вигляді вихідного шару нейромережі не здатний передбачити справжні мітки.
- Тепер розглянемо випадок із функціями активації (sigmoid) (рис. 3). З ними наша нейронна мережа стає складнішою функцією ϕ(·), яка краще навчається змінювати простір ознак так, щоб наприкінці вихідний шар міг правильно класифікувати вектори.

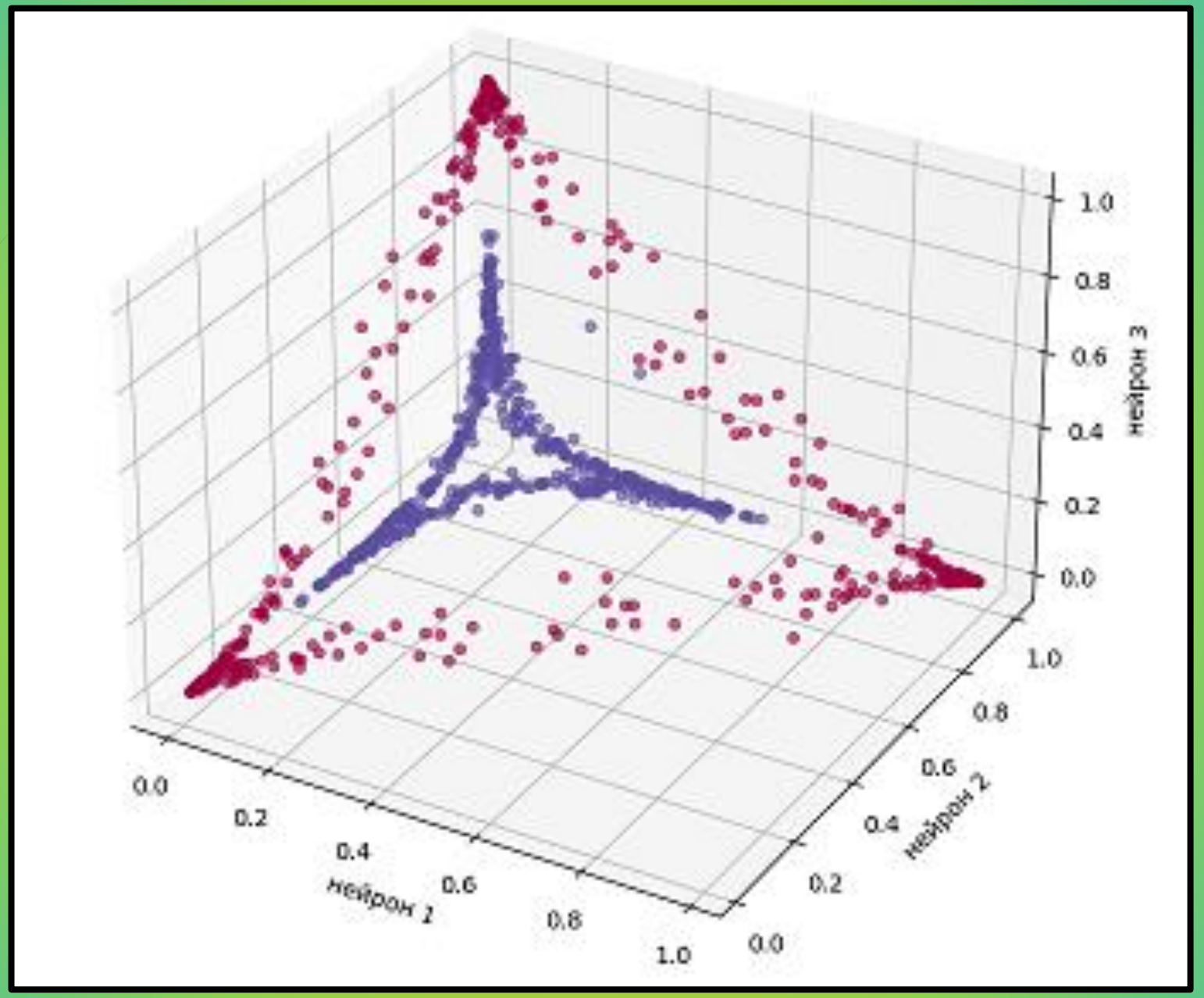


Рис.3 Випадок із функціями активації (sigmoid).

- Повернемося знову до нашої нейромережі без функції активації. Чи обов'язково використовувати функції активації? Можна додати модульність та попередньо обробити дані, використовуючи перетворення з ядром. В даному випадку ми будемо використовувати RBF ядро і представимо наші дані в нескінченномірному просторі, де нейромережа без функції активації зможе їх класифікувати максимально точно (Рис. 4).

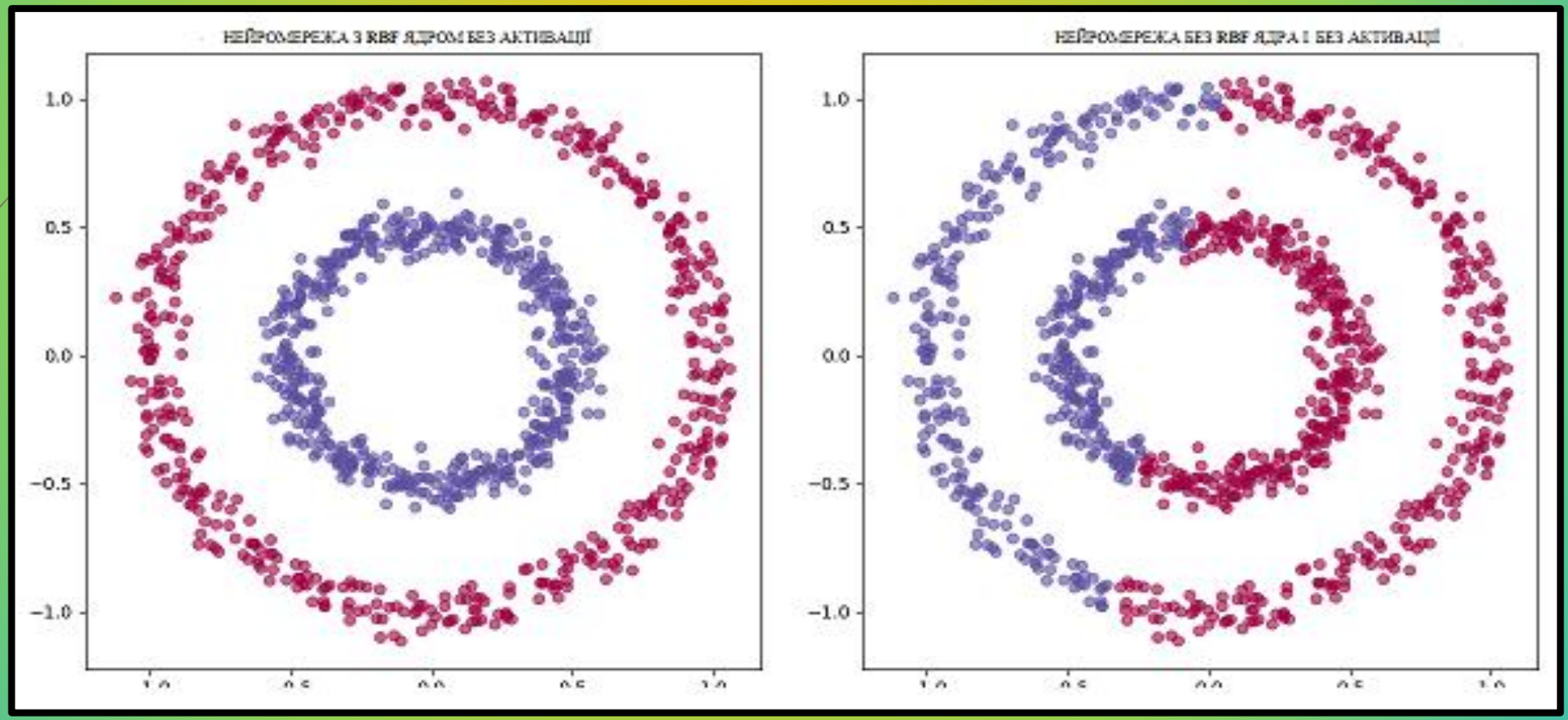


Рис. 4 Нейромережа з ядром RBF і без нього (без активації)

- Розглянемо ще одну відому архітектуру – дерево рішень. Розглянемо його функціонал.
- Саме собою воно ніяк не трансформує простір, але при цьому воно ділить і структурує його. Геометрично дерево рішень розбиває простір ознак на осі, перпендикулярні до осей ознак.
- Розглянемо це на прикладі двовимірного простору ознак з ознаками $x1$ і $x2$:

1. *Початковий стан:*

   Простір ознак є площиною, де кожна точка відповідає певним значенням $x1$ і $x2$.

2. *Перше розбиття:*

   Дерево рішень обирає ознаку $X1$ і граничне значення $x1$. Простір розбивається на дві частини по лінії $X1 = x1$:

   Усі точки, у яких $X1 \leq x1$, знаходяться ліворуч від лінії.

   Усі точки, у яких $X1 > x1$, знаходяться праворуч від лінії.

   Геометрично це означає, що простір ділиться вертикальною лінією.

3. Друге розбиття:

Наразі кожна з двох частин простору розглядається окремо. Припустимо, що у лівій частині ($X1 \leq x1$) вибирається ознака $X2$ і порогове значення $x2$. Простір тепер ділиться горизонтальною лінією $X2 = x2$:

Усі точки, у яких $X2 \leq x2$, знаходяться нижче лінії.

Усі точки, у яких $X2 > x2$, знаходяться вище лінії.

Геометрично це означає, що ліва частина простору тепер ділиться горизонтальною лінією.

Геометричний приклад роботи дерева рішень у двовимірному просторі ознак для класифікації чаю: смачний чи ні, на основі його температури та рівня цукру (рис. 5).

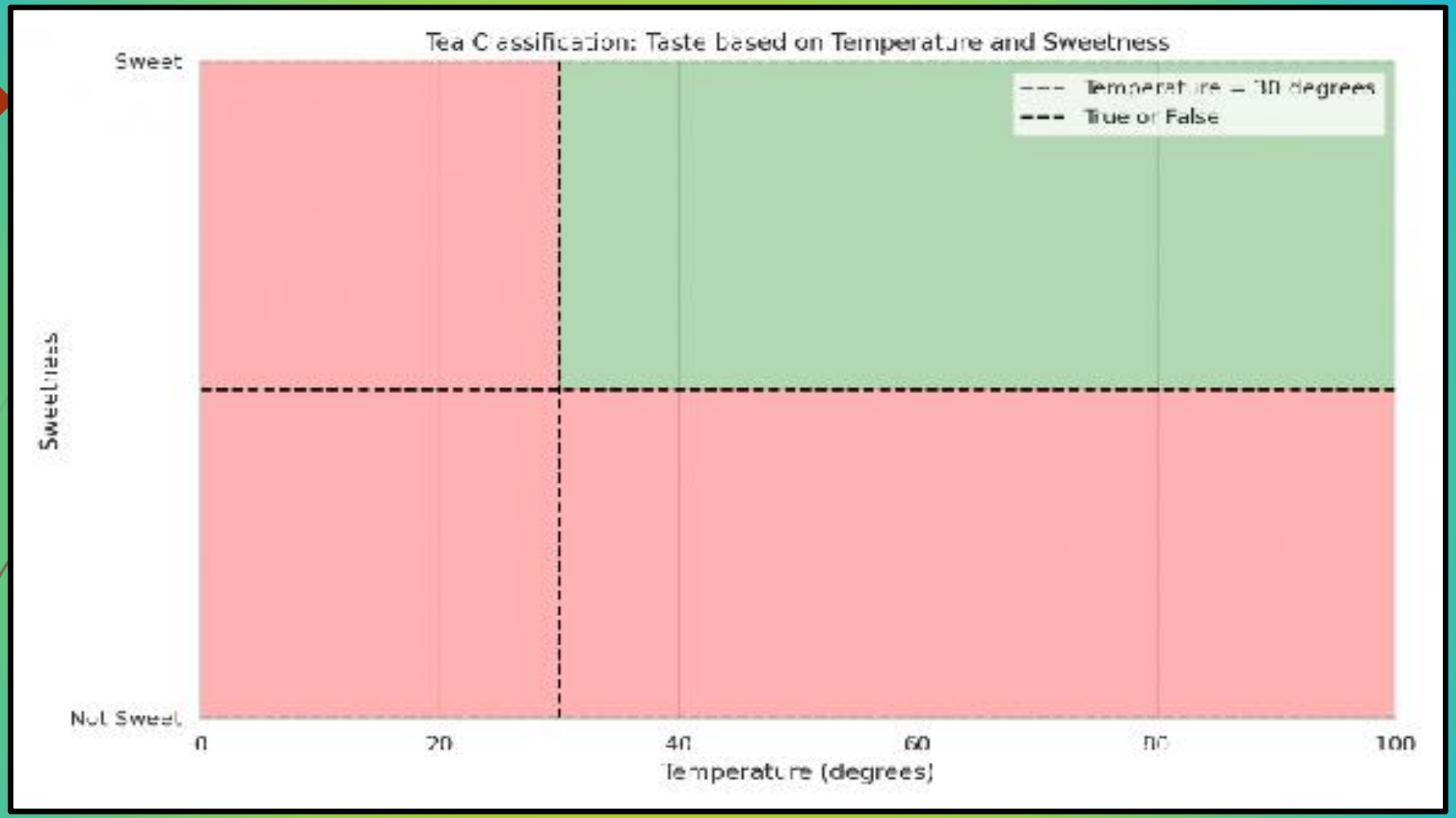


Рис. 5 Геометричний приклад роботи дерева рішень у двовимірному просторі ознак для класифікації чаю: смачний чи ні, на основі його температури та рівня цукру