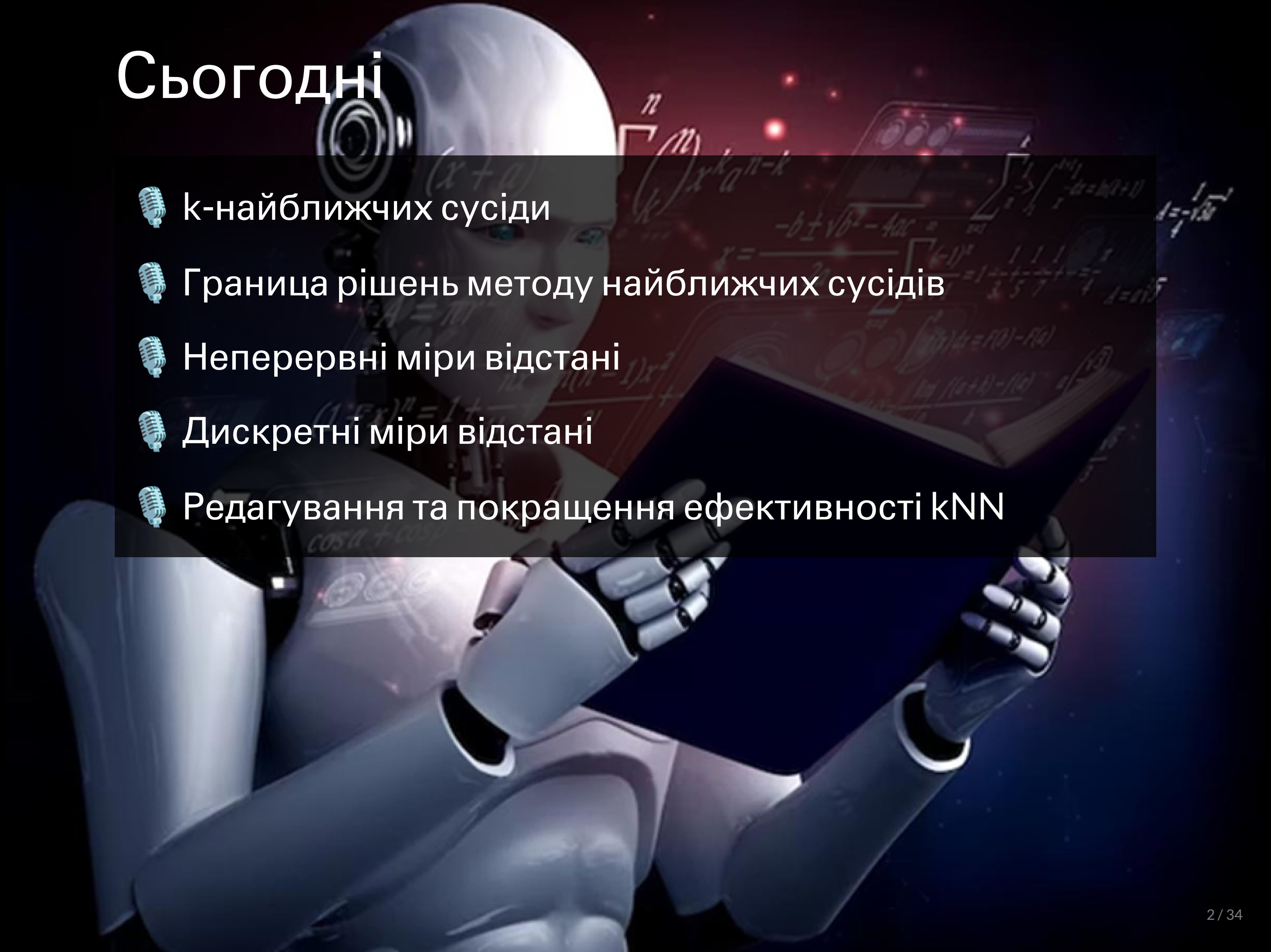
Машинне навчання

Лекція 5: Метод найближчих сусідів

Кочура Юрій Петрович
iuriy.kochura@gmail.com
@y_kochura

Сьогодні

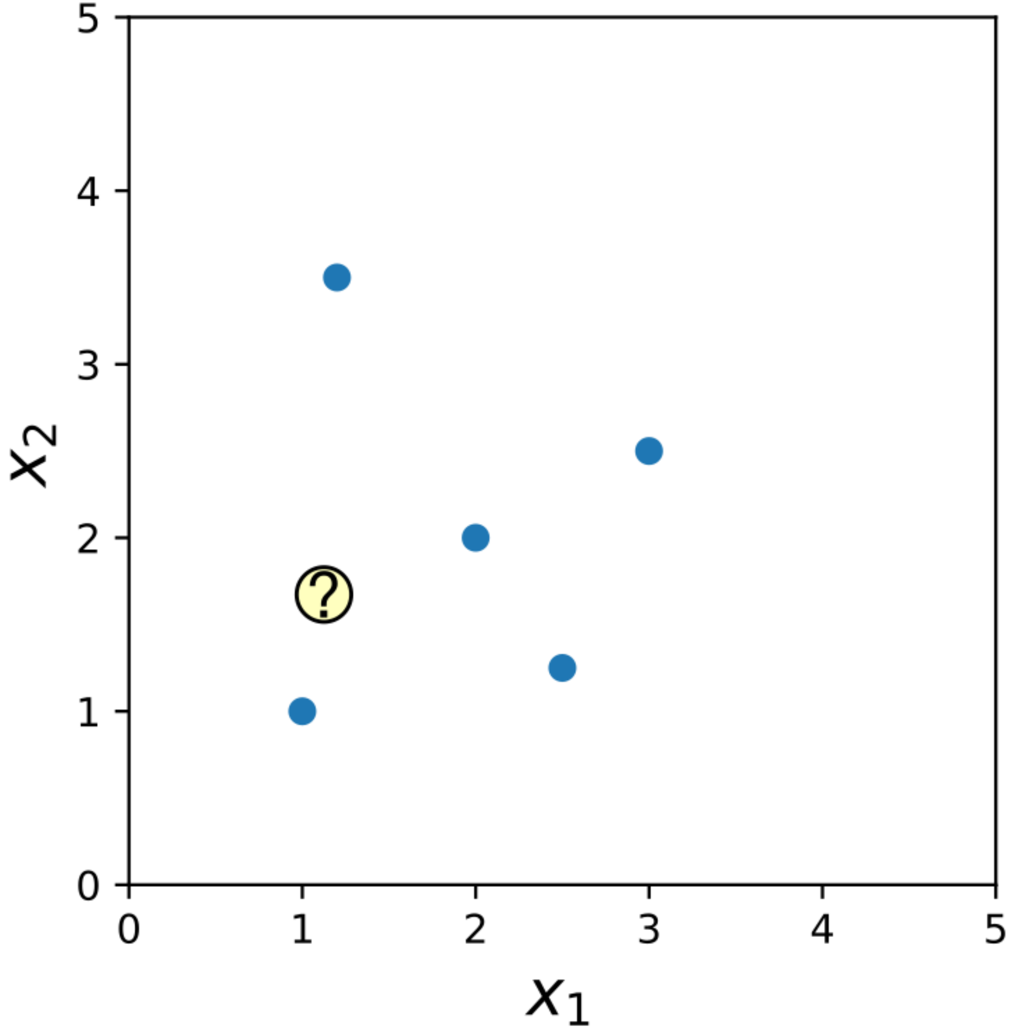
- 🎙️ k-найближчих сусіди
- 🎙️ Границя рішень методу найближчих сусідів
- 🎙️ Неперервні міри відстані
- 🎙️ Дискретні міри відстані
- 🎙️ Редагування та покращення ефективності kNN



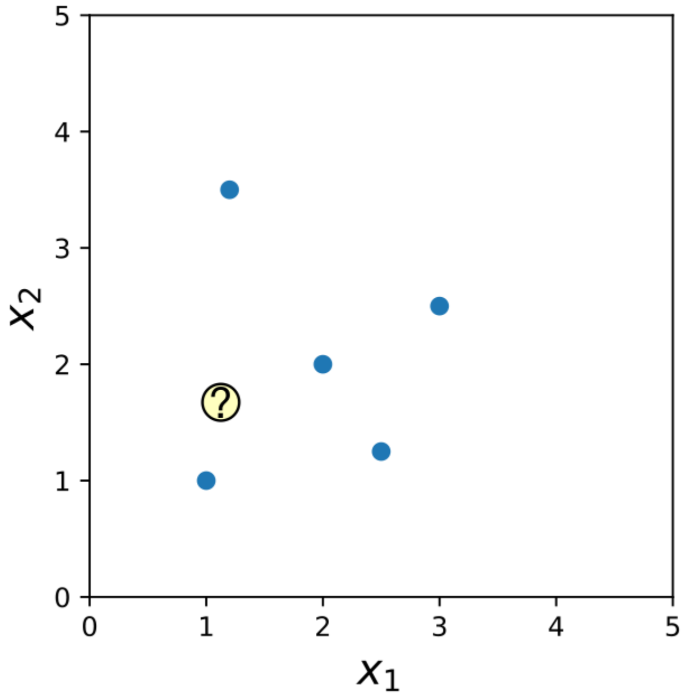
k-найближчих сусіди

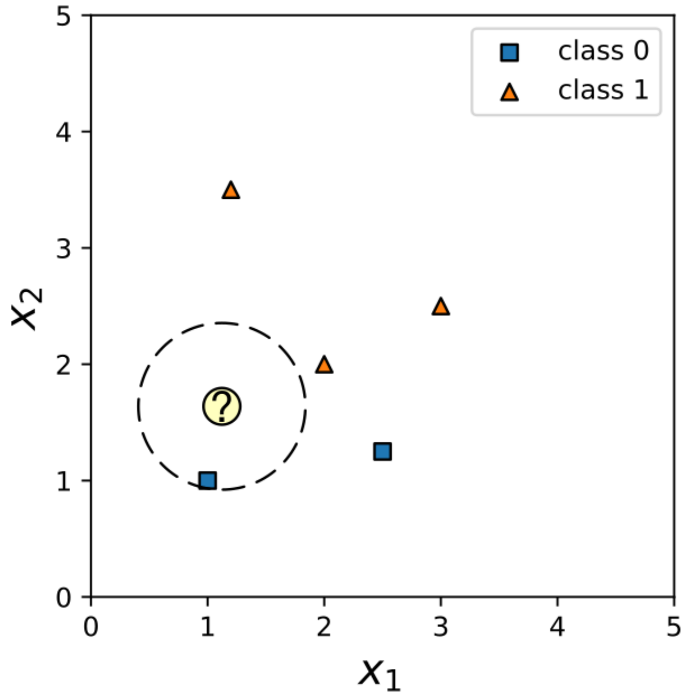
k-nearest neighbors (kNN)

1-найближчий сусід



1-найближчий сусід





Навчальна вибірка

$$\left(\mathbf{X}^{(i)}, y^{(i)}\right) \in \mathcal{D},$$

$$|\mathcal{D}| = n$$

Псевдокод: 1-найближчий сусід

closest_point := None

closest_distance := ∞

- for $i = 1, ..., n$:

  current_distance := $d(\mathbf{X}^{(i)}, \mathbf{X}^{(q)})$

  if current_distance $<$ closest_distance:

  - closest_distance := current_distance
  - closest_point := $\mathbf{X}^{(i)}$
  - return $f(\mathbf{X}^{(i)})$

Мітка найближчого сусіда (closest_point): $\left(\mathbf{X}^{(i)}, f(\mathbf{X}^{(i)})\right)$

Евклідова відстань **(L2)**

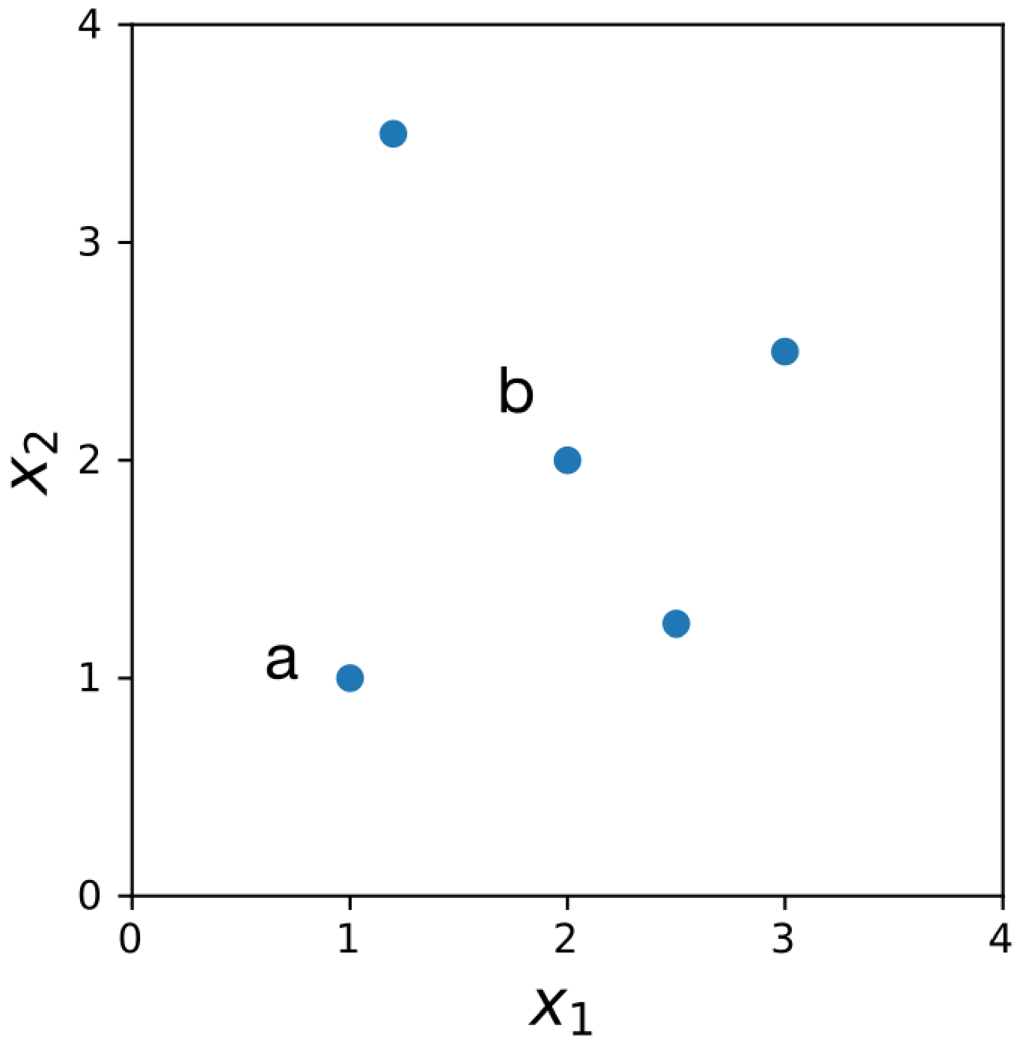
$$d\left(\mathbf{X}^{(a)}, \mathbf{X}^{(b)}\right) = \sqrt{\sum_{j=1}^{m} \left(\mathbf{X}_j^{(a)} - \mathbf{X}_j^{(b)}\right)^2} =$$

$$= \left\|\mathbf{X}^{(a)} - \mathbf{X}^{(b)}\right\|_2$$

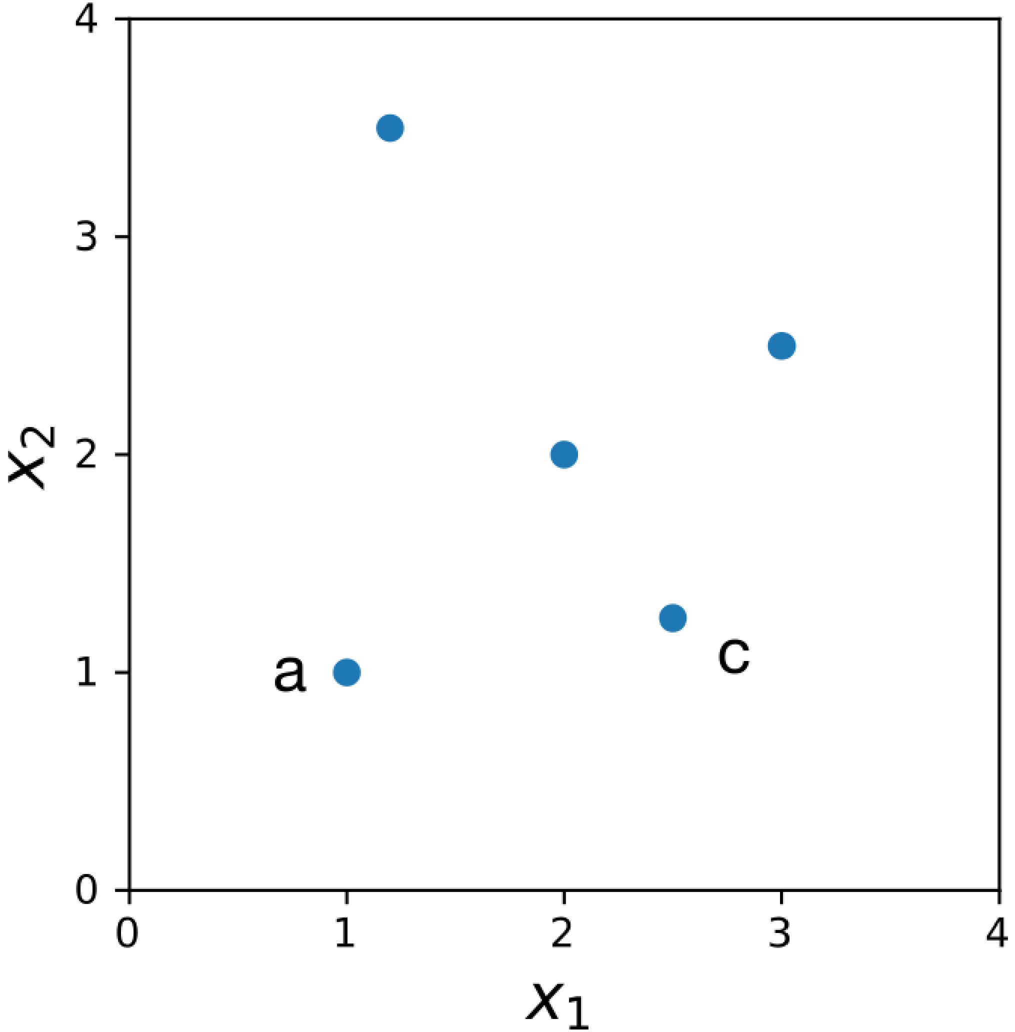
де $\mathbf{X}^{(i)} = (x_1^{(i)}, x_2^{(i)}, \dots, x_m^{(i)}) \in \mathbb{R}^m$ – m-вимірний вектор ознак.

Границa рішень методу найближчих сусідів

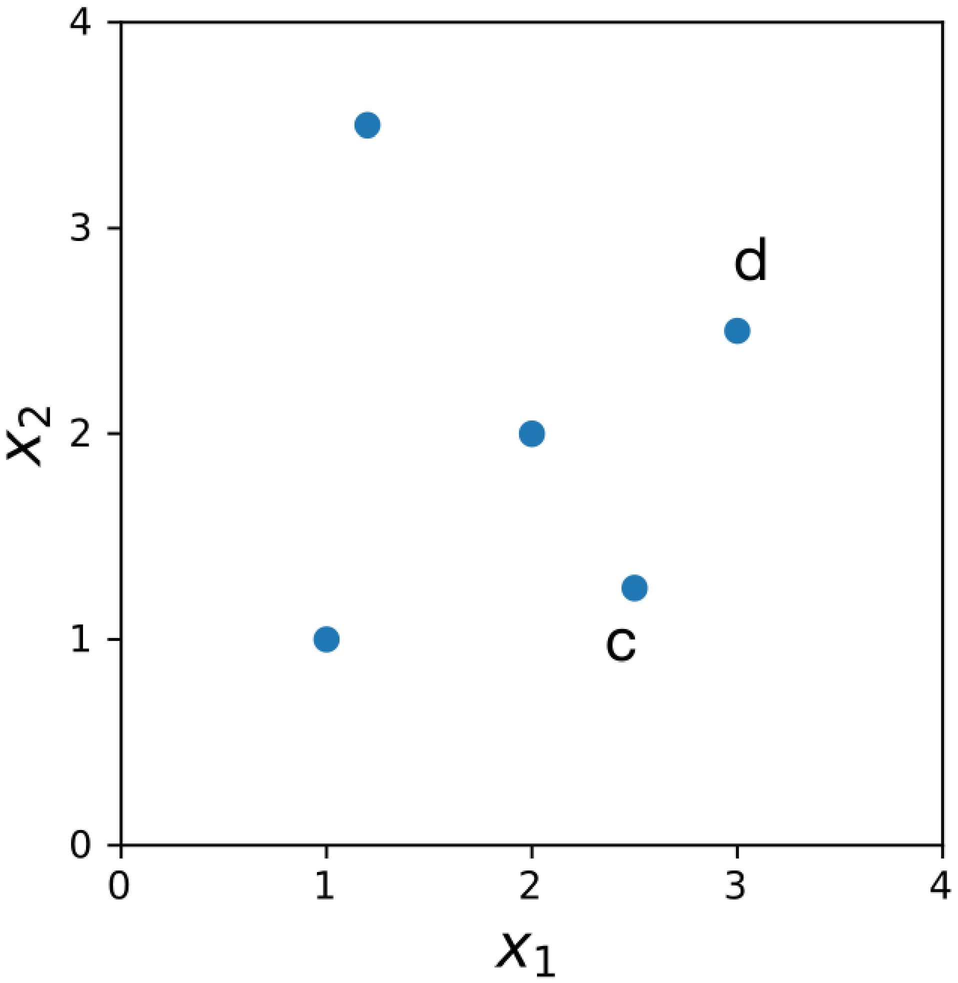
Границя рішень між **(a)** та **(b)**



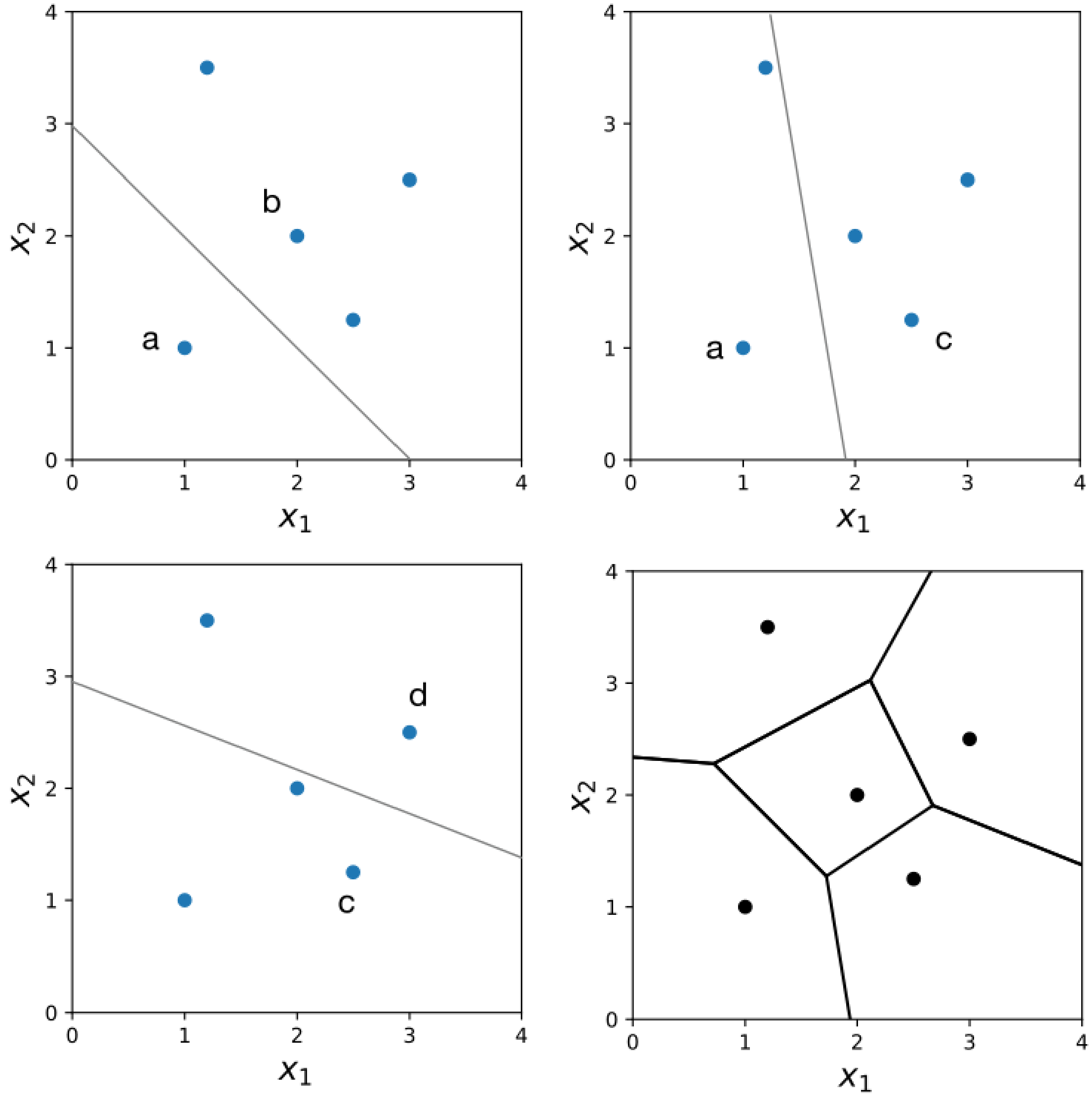
Границя рішень між **(a)** та **(c)**



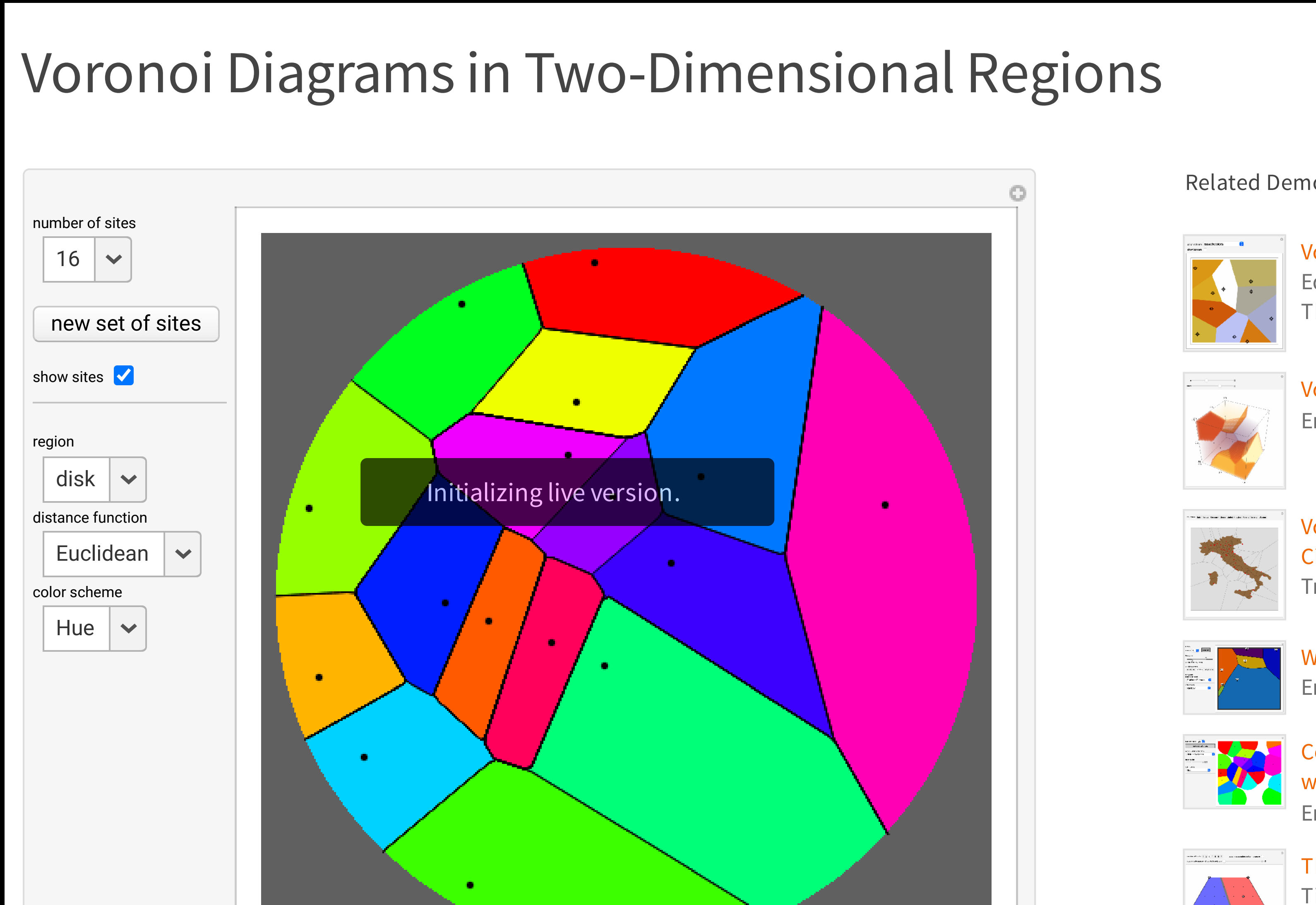
Границя рішень між **(c)** та **(d)**



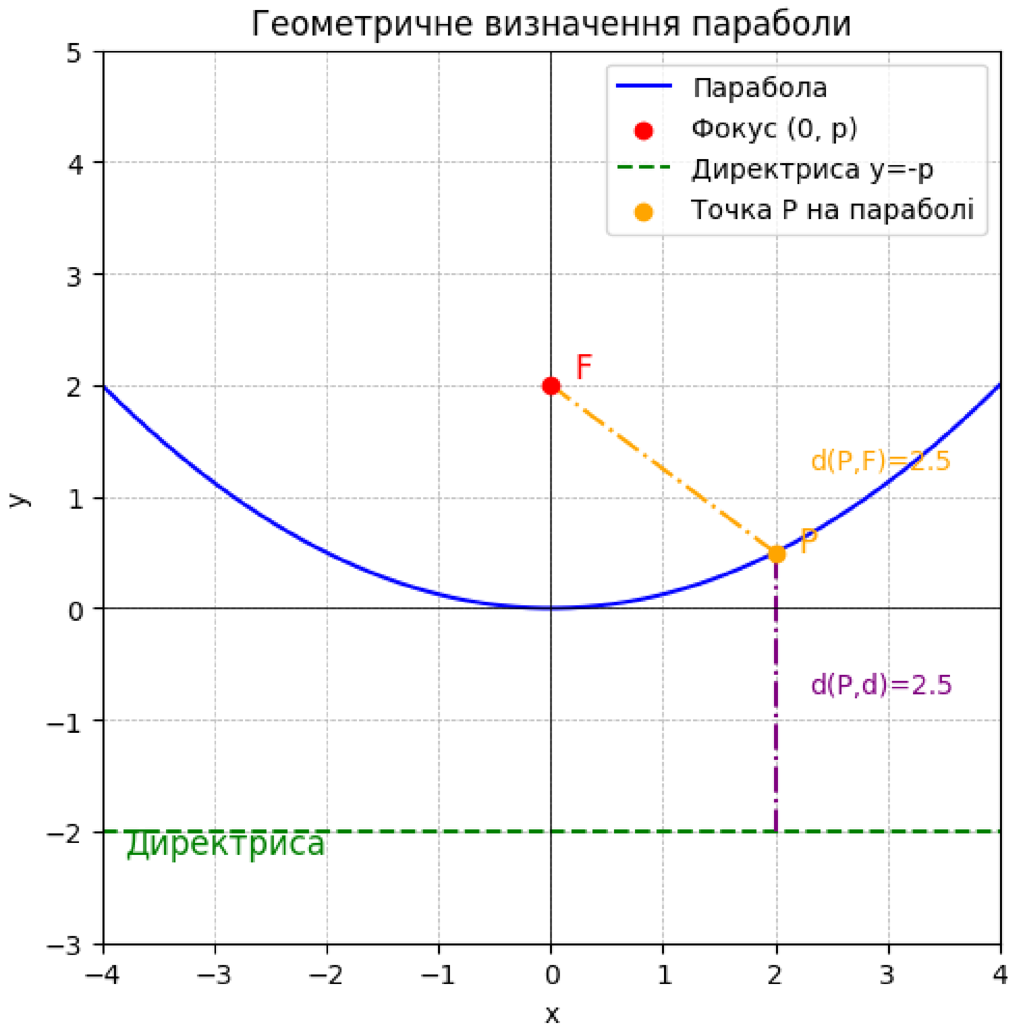
Границя рішень для **1-NN**



Георгій Феодосійович Вороний
(1868-1908)



Діаграми Вороного в двовимірних областях


Геометричне визначення параболи
Парабола
Фокус (0, p)
Директриса y=-p
Точка P на параболі
F
P
d(P,F)=2.5
d(P,d)=2.5
Директриса
x
y

$$y = \frac{x^2}{4p}$$

Fortune's Algorithm for Voronoi Diagrams

Open Notebook in Cloud | Copy Manipulate to Clipboard | Initializing live version | Source Code

Given a finite set of points (called sites) in a plane, a Voronoi diagram divides the plane into regions around
are closer to that site than to any of the others. This Demonstration shows Fortune's algorithm for drawin
diagrams[1].

Two auxiliary curves are involved in the procedure:

1. The sweep line (green) can move up or down; only sites above it are active.

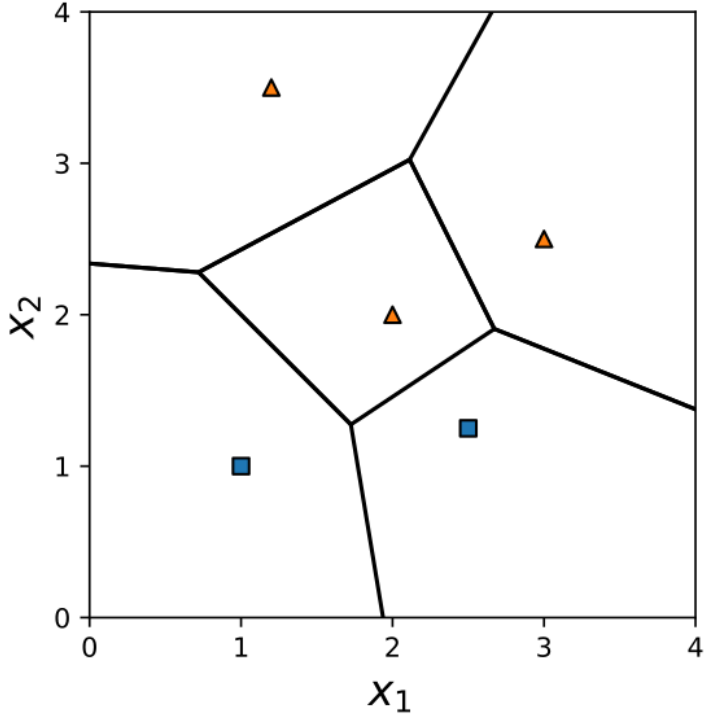
2. The beach line (red) is a sequence of parabolic arcs. Each parabola has an active site for its locus and the
directrix.

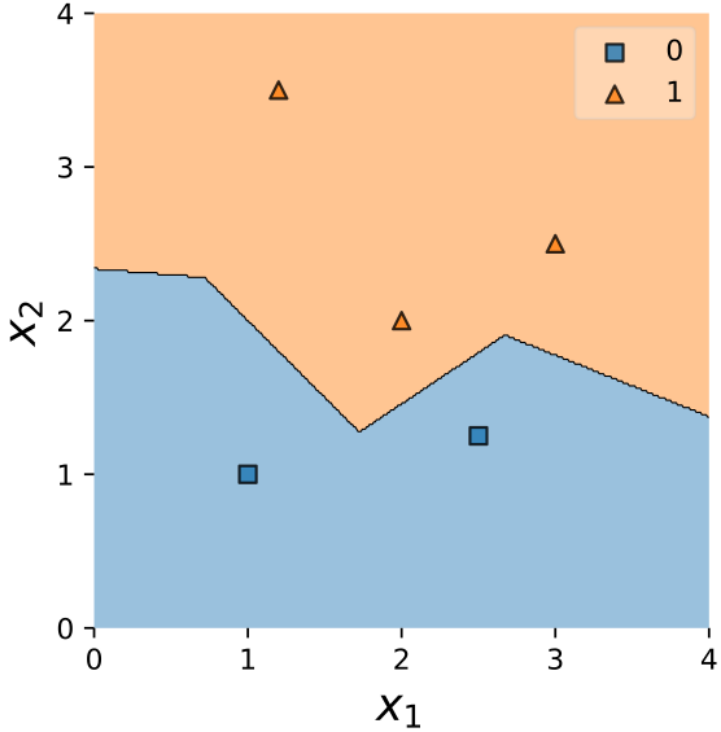
The intersections of two parabolic arcs (the breakpoints on the beach line) have equal distance to two site
the two parabolas) and to the sweep line. Thus, as the sweep line moves down, these intersection points d
rest of the diagram.

Contributed by: Erik Mahieu (2016)

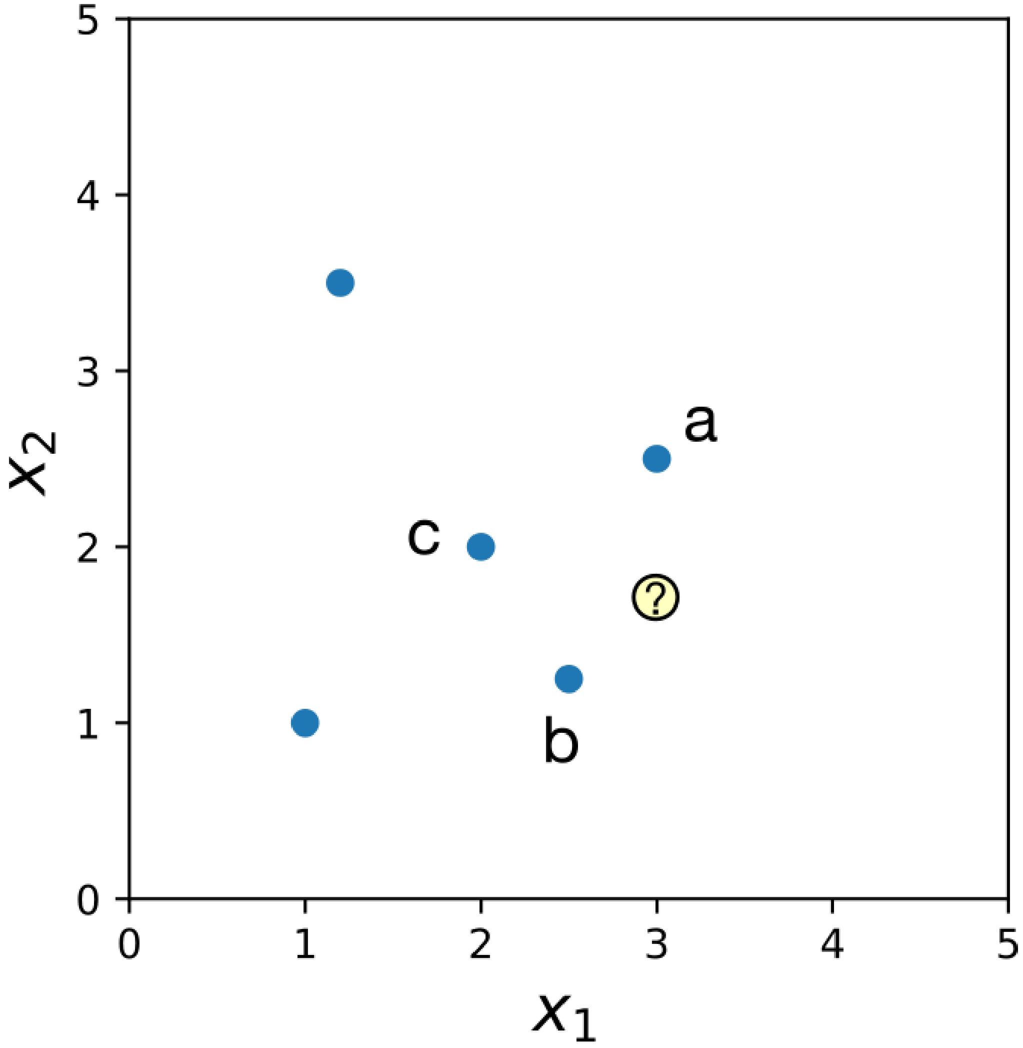
Алгоритм Форчуна для діаграм Вороного

Границя рішень для **1-NN**

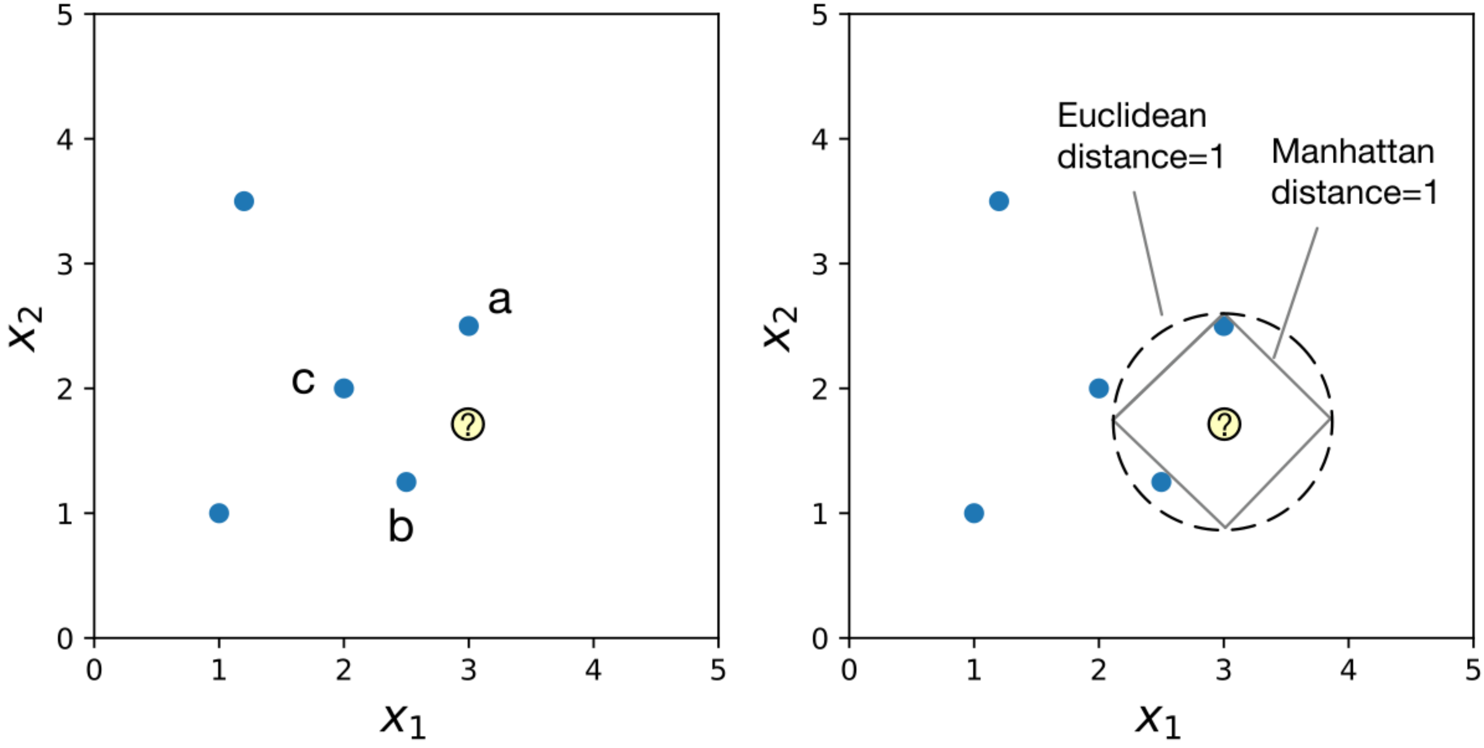




Який екземпляр найближчий?



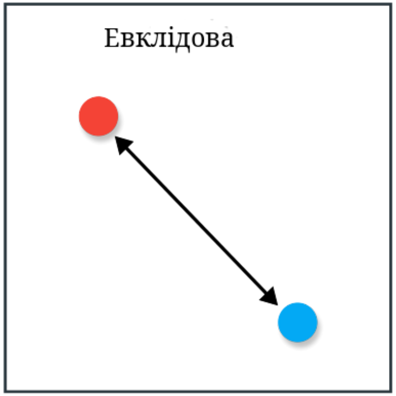
Залежить від міри відстані!



Неперервні міри відстані

Евклідова відстань:

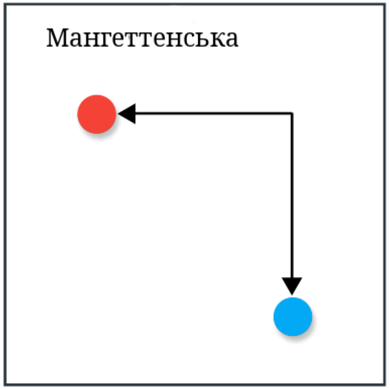
$$d\left(\mathbf{x}^{(a)}, \mathbf{x}^{(b)}\right) = \sqrt{\sum_{j=1}^{m} \left(\mathbf{x}_j^{(a)} - \mathbf{x}_j^{(b)}\right)^2}$$



Неперервні міри відстані

Мангеттенська відстань:

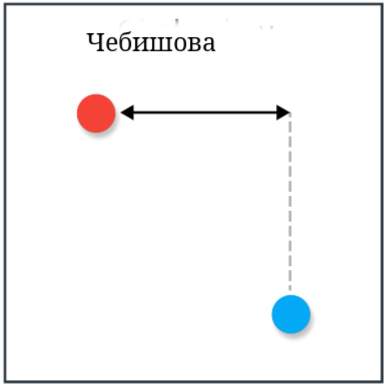
$$d(\mathbf{X}^{(a)}, \mathbf{X}^{(b)}) = \sum_{j=1}^{m} |\mathbf{X}_j^{(a)} - \mathbf{X}_j^{(b)}|$$



Неперервні міри відстані

Чебишова відстань:

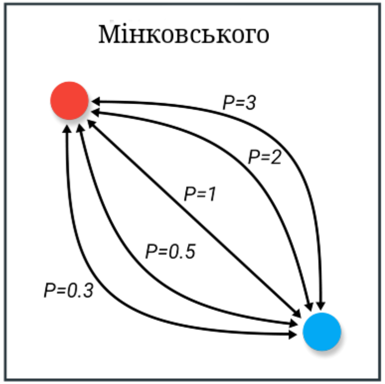
$$d(\mathbf{X}^{(a)}, \mathbf{X}^{(b)}) = \max_{j=1}^{m} |\mathbf{X}_j^{(a)} - \mathbf{X}_j^{(b)}|$$



Неперервні міри відстані

Мінковського відстань:

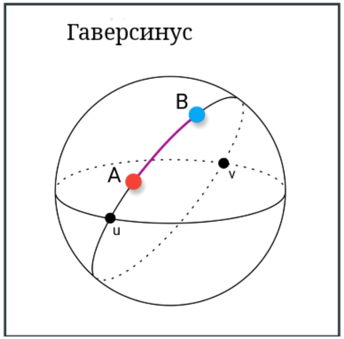
$$d(\mathbf{X}^{(a)}, \mathbf{X}^{(b)}) = \left[\sum_{j=1}^{m} \left(|\mathbf{X}_j^{(a)} - \mathbf{X}_j^{(b)}| \right)^p \right]^{\frac{1}{p}}$$



Неперервні міри відстані

Гаверсинус кута θ визначається як:

$$hav(\theta) = \sin^2\left(\frac{\theta}{2}\right)$$



Відстань гаверсинуса:

$$d = 2r \cdot arcsin\sqrt{hav(\phi_b - \phi_a) + \cos(\phi_a)\cos(\phi_b)hav(\lambda_b - \lambda_a)}$$

де r — радіус сфери, ϕ — широта, λ — довгота.

Неперервні міри відстані

Відстань Махаланобіса:

$$d\left(\mathbf{X}^{(a)}, \mathbf{X}^{(b)}\right) = \sqrt{\left(\mathbf{X}^{(a)} - \mathbf{X}^{(b)}\right)^T S^{-1} \left(\mathbf{X}^{(a)} - \mathbf{X}^{(b)}\right)},$$

де S^{-1} — коваріаційна матриця.

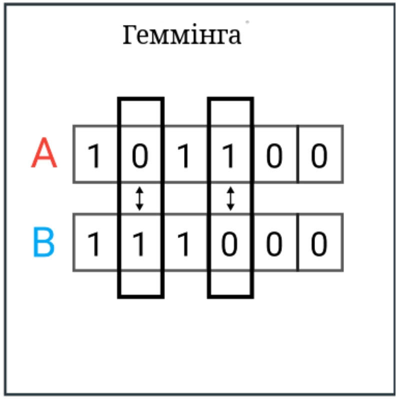
Дискретні міри відстані

Відстань Геммінга:

$$d(\mathbf{X}^{(a)}, \mathbf{X}^{(b)}) = \sum_{j=1}^{m} \delta(\mathbf{X}_j^{(a)}, \mathbf{X}_j^{(b)}),$$

де

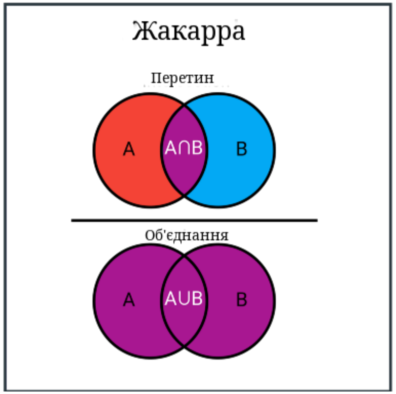
$$\delta(\mathbf{X}_j^{(a)}, \mathbf{X}_j^{(b)}) = \begin{cases} 1, & \text{якщо } \mathbf{X}_j^{(a)} \neq \mathbf{X}_j^{(b)} \\ 0, & \text{якщо } \mathbf{X}_j^{(a)} = \mathbf{X}_j^{(b)} \end{cases}$$



Дискретні міри відстані

Відстань Жаккара/Танімото:

$$d(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

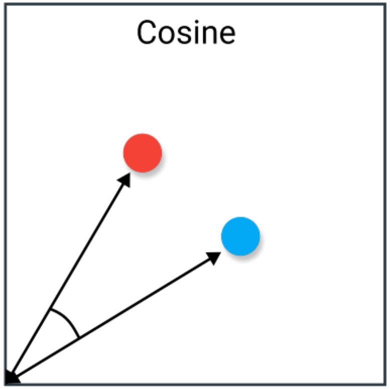


Дискретні міри відстані

Косинусна відстань:

$$d(A, B) = 1 - \cos(\theta),$$

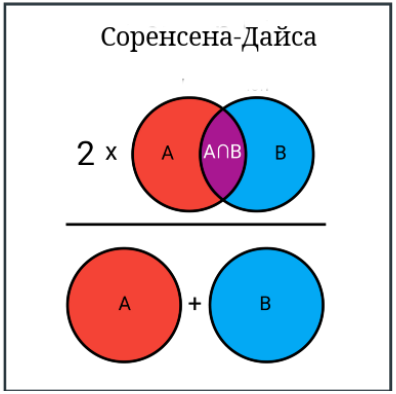
де $\cos(\theta) = \frac{A \cdot B}{||A|| \cdot ||B||}$



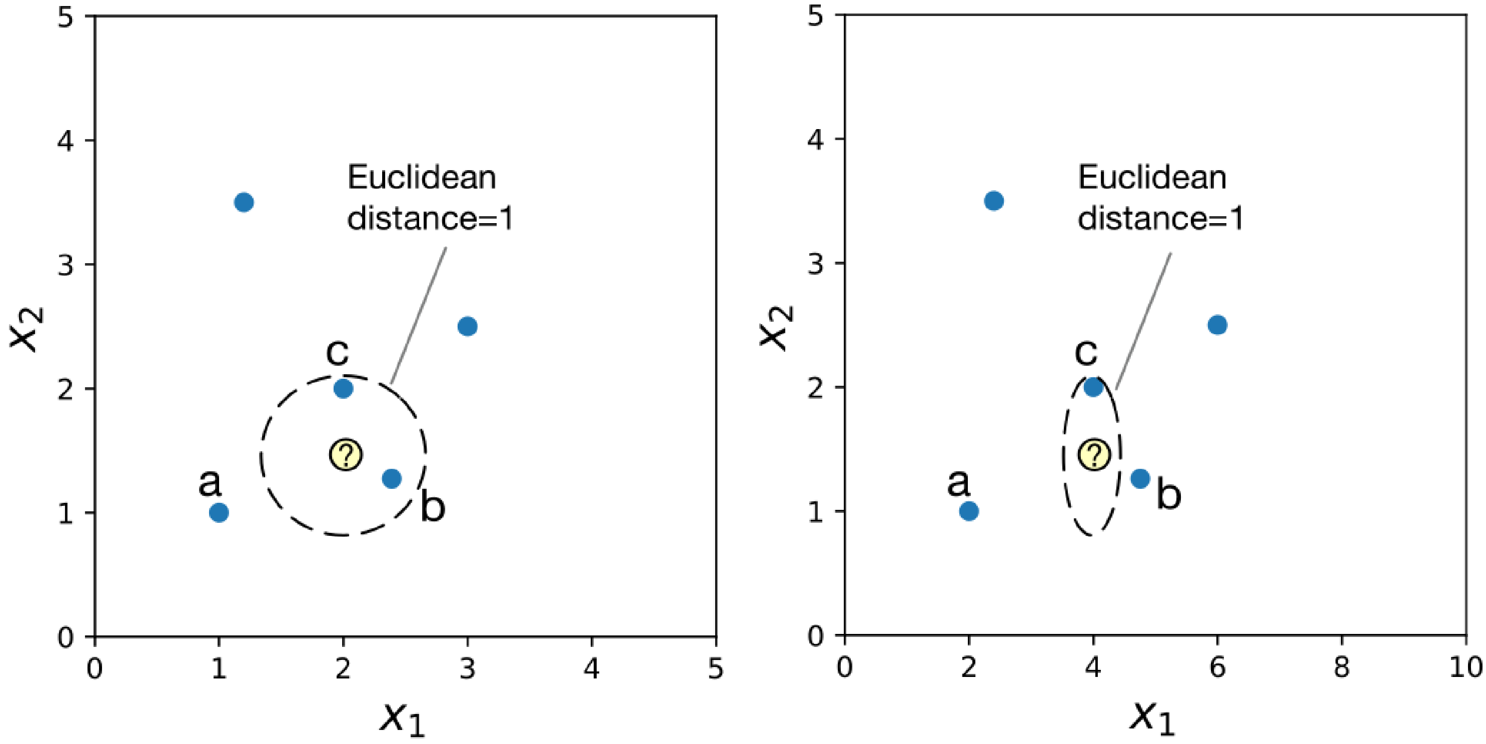
Дискретні міри відстані

Відстань Соренсена-Дайса:

$$d(A, B) = \frac{2|A \cap B|}{|A| + |B|}$$

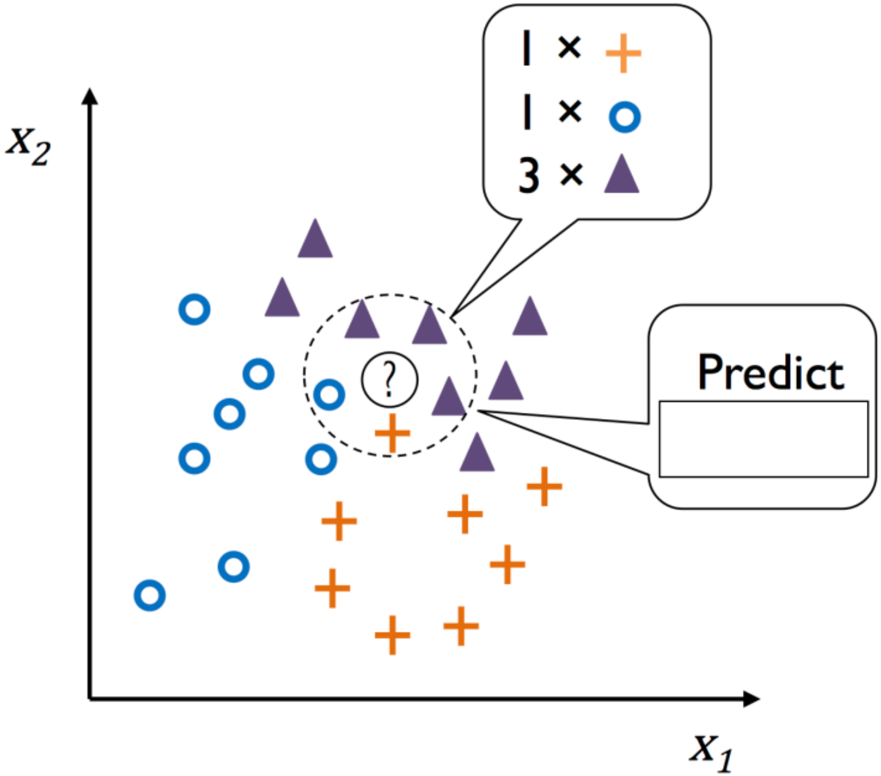


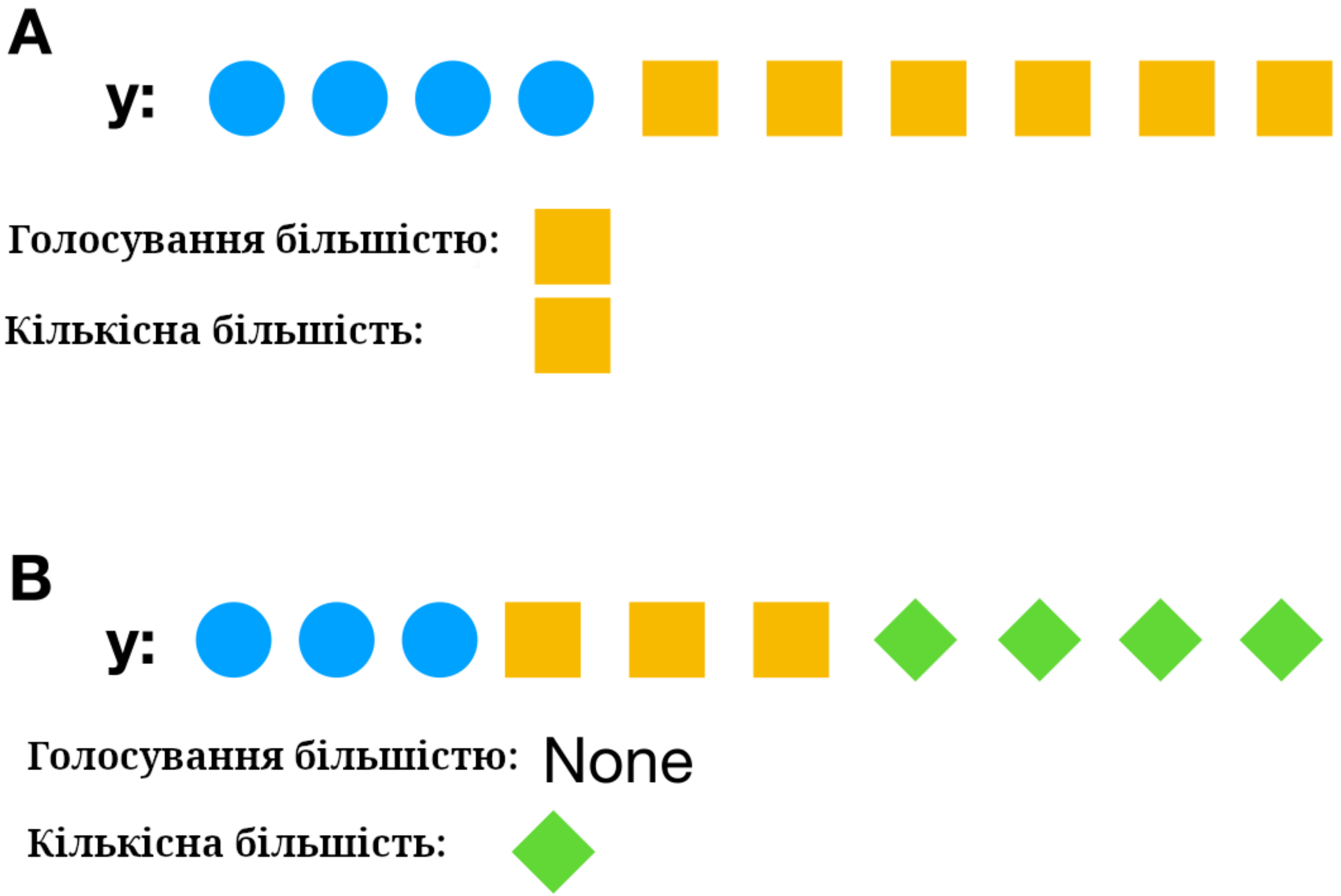
Масштабування ознак



k-найближчі сусіди

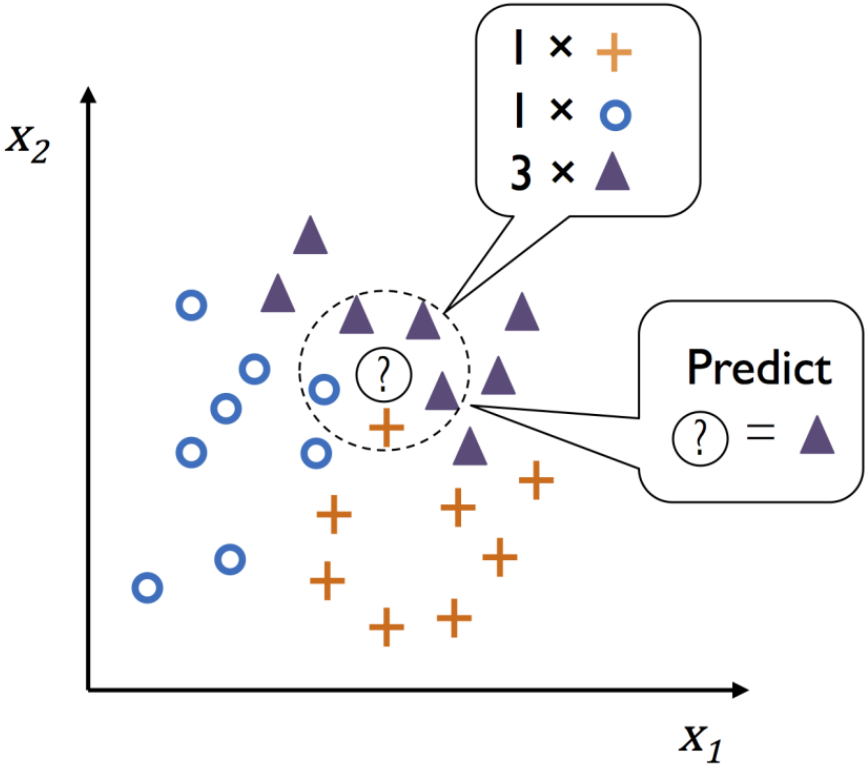
k = 5



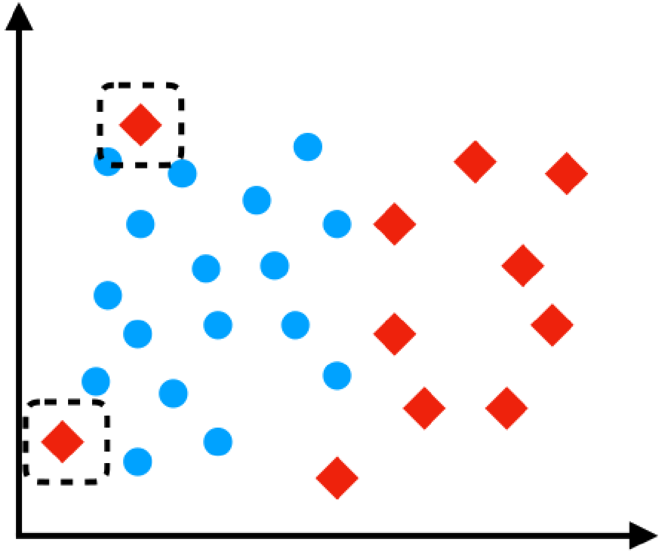

A
y:
Голосування більшістю:
Кількісна більшість:
B
y:
Голосування більшістю: None
Кількісна більшість:

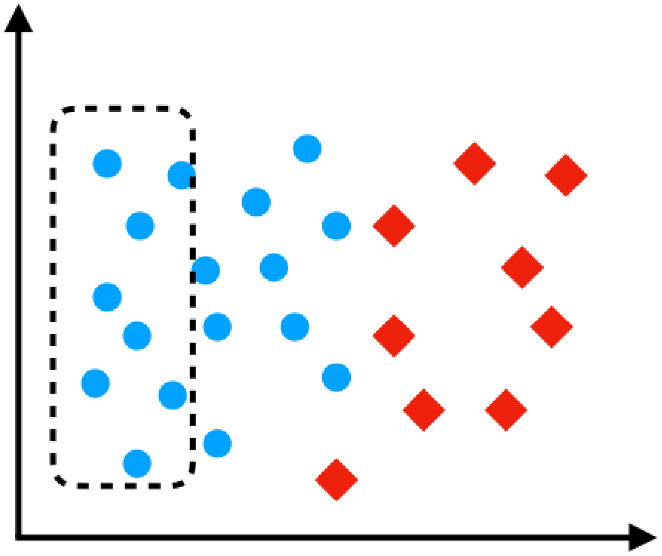
k-найближчі сусіди

k = 5



Редагування **kNN**





Покращення ефективності прогнозування

- Підібрати оптимальне значення k
- Масштабування осей ознак
- Вибір метрики для визначення відстані
- Зважування міри відстані

Кінець