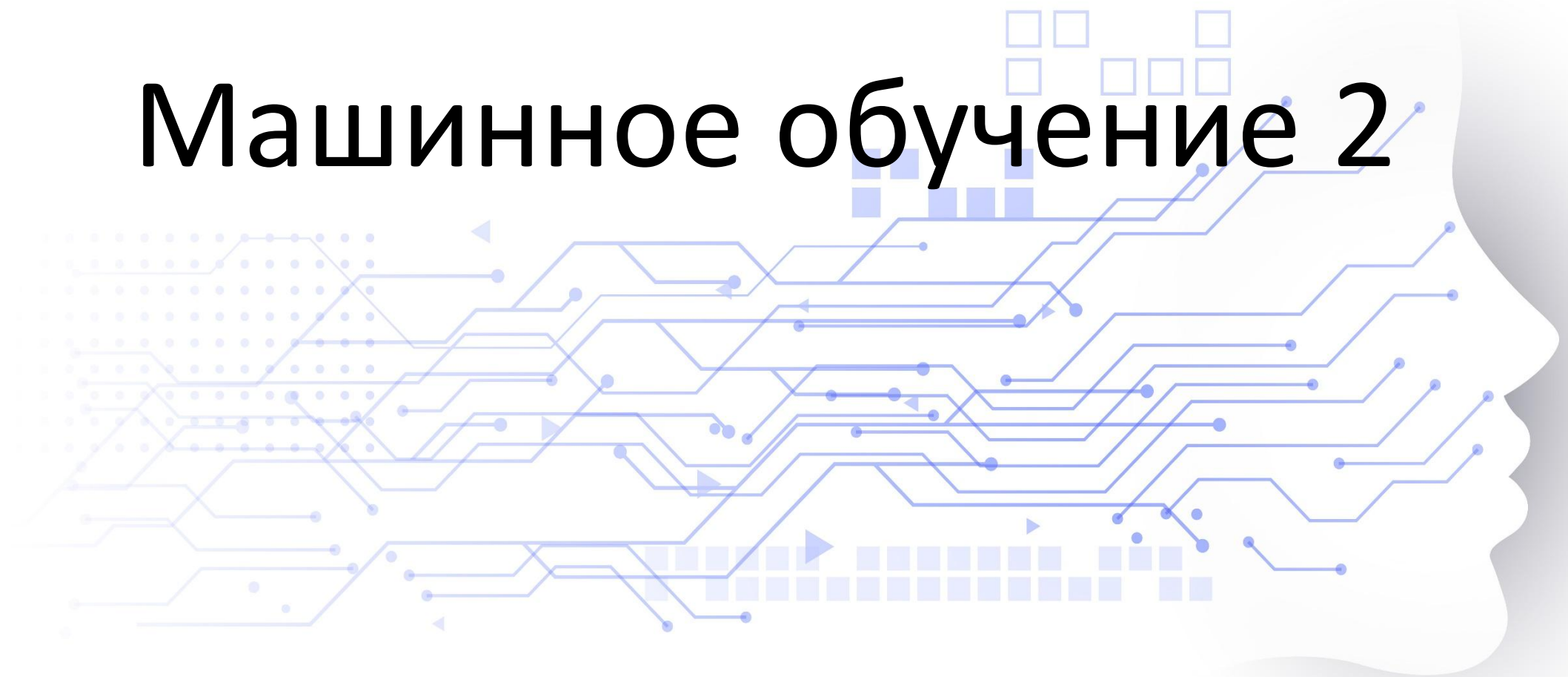


Машинное обучение 2





Резаиан Наим

E-mail: rezaian-n@rudn.ru

Telegram: [@NaeimRezaeian](https://www.t.me/NaeimRezaeian)

1. Заведующий лабораторией искусственного интеллекта
2. Руководитель направления разработок Центра развития цифровых технологий в образовательных процессах
3. Старший преподаватель факультета искусственного интеллекта

Метод k ближайших соседей

Как отличить ель от сосны?



Ель



Сосна

Как отличить ель от сосны?



Ель

Ветки

Смотрят вверх

Ствол

Не видно

Иголки

Густые

Цвет

Ближе к зеленому



Сосна

Параллельно земле

Видно

Более редкие

Ближе к желтому

Как отличить ель от сосны?



Ветки

Смотрят вверх

Ствол

Не видно

Иголки

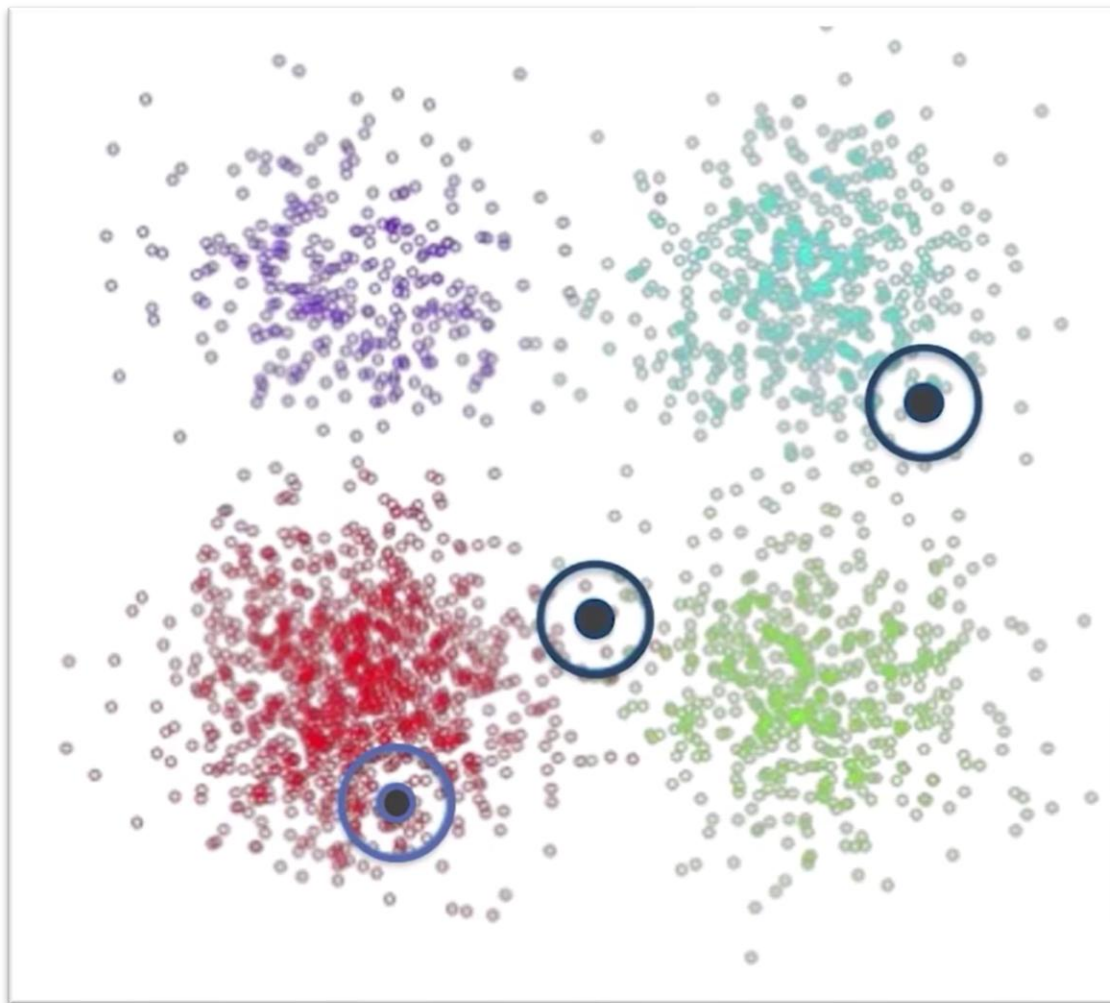
Густые

Цвет

Ближе к синему

Ель

Гипотеза компактности



Гипотеза компактности

**Если два объекта похожи друг на друга, то ответы
на них тоже похожи**

Метод k ближайших соседей (kNN): Обучение

Дано: обучающая выборка $X = (x_i, y_i)_{i=1}^{\ell}$

Задача классификации: ответы из множества $Y = \{1, \dots, K\}$

Обучение модели: запоминаем обучающую выборку X

Метод k ближайших соседей (kNN): Применение

Дано: новый объект x

Применение модели:

Сортируем объекты обучающей выборки по расстоянию до нового объекта:

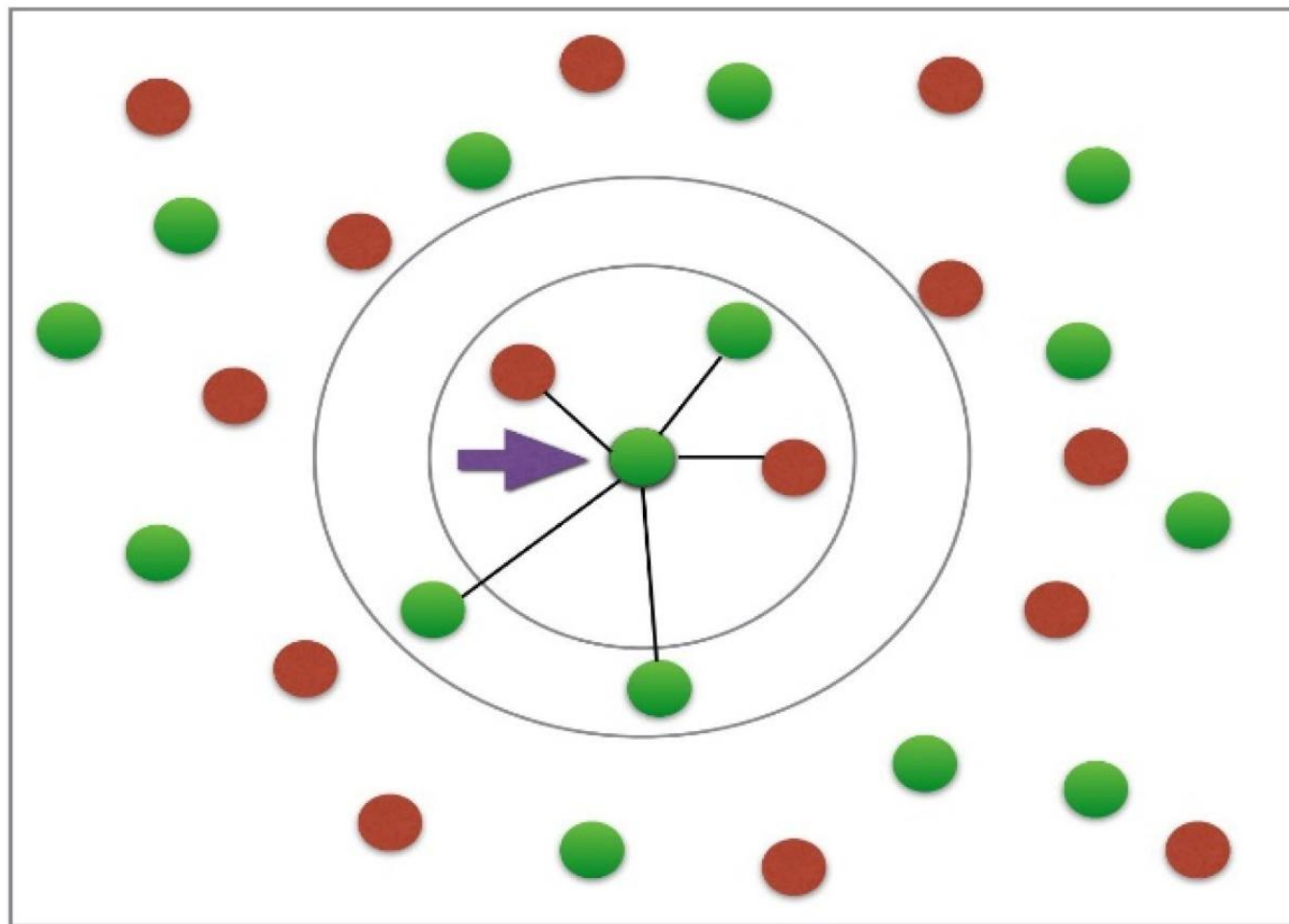
$$\rho(x, x_{(1)}) \leq \rho(x, x_{(2)}) \leq \dots \leq \rho(x, x_{(\ell)})$$

Выбираем k ближайших объектов: $x_{(1)}, \dots, x_{(k)}$

Выдаем наиболее популярный среди них класс

$$a(x) = \operatorname{argmax}_{y \in Y} \sum_{i=1}^k [y_{(i)} = y]$$

Метод k ближайших соседей (kNN): Применение



Метрика

Метрика – это функция ρ с двумя аргументами, удовлетворяющая трём требованиям:

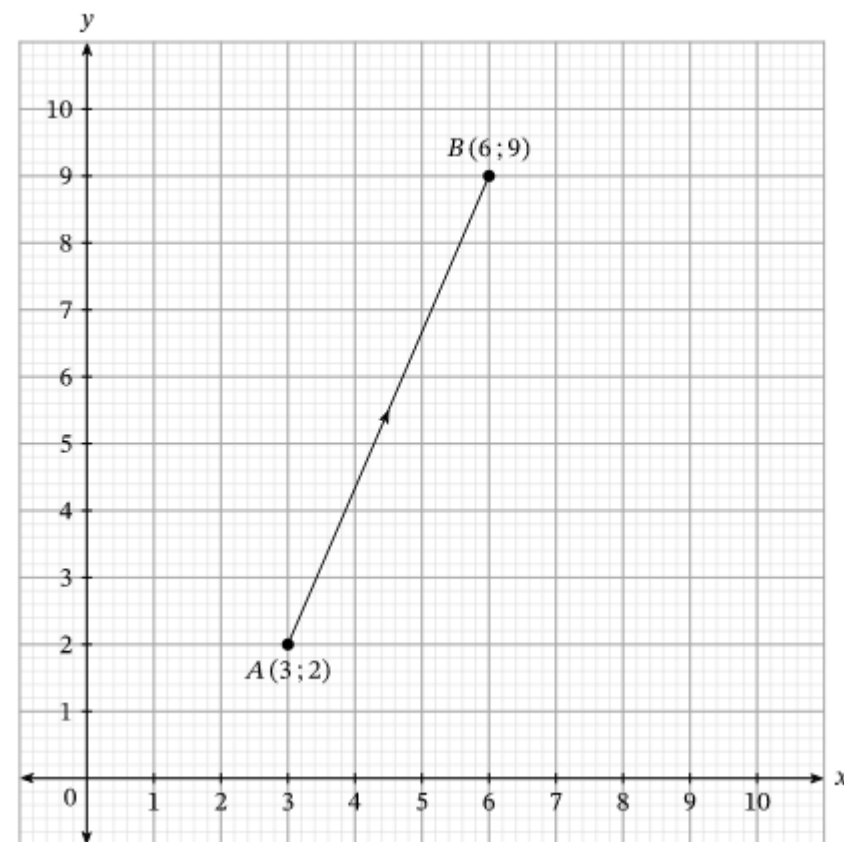
$\rho(x, z) = 0$ тогда и только тогда, когда $x=z$

$$\rho(x, z) = \rho(z, x)$$

$\rho(x, z) \leq \rho(x, v) + \rho(v, z)$ неравенство треугольника

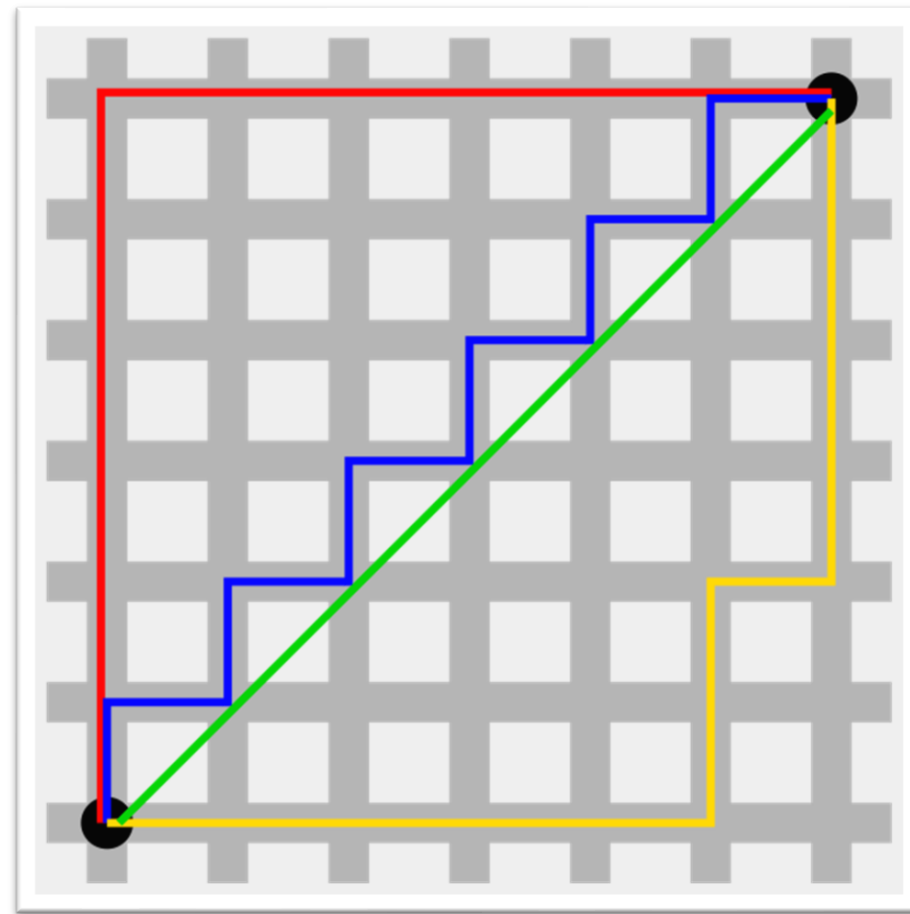
Евклидова метрика

$$\rho(x, z) = \sqrt{\sum_{j=1}^d (x_j - z_j)^2}$$



Манхэттенская метрика

$$\rho(x, z) = \sum_{j=1}^d |x_j - z_j|$$

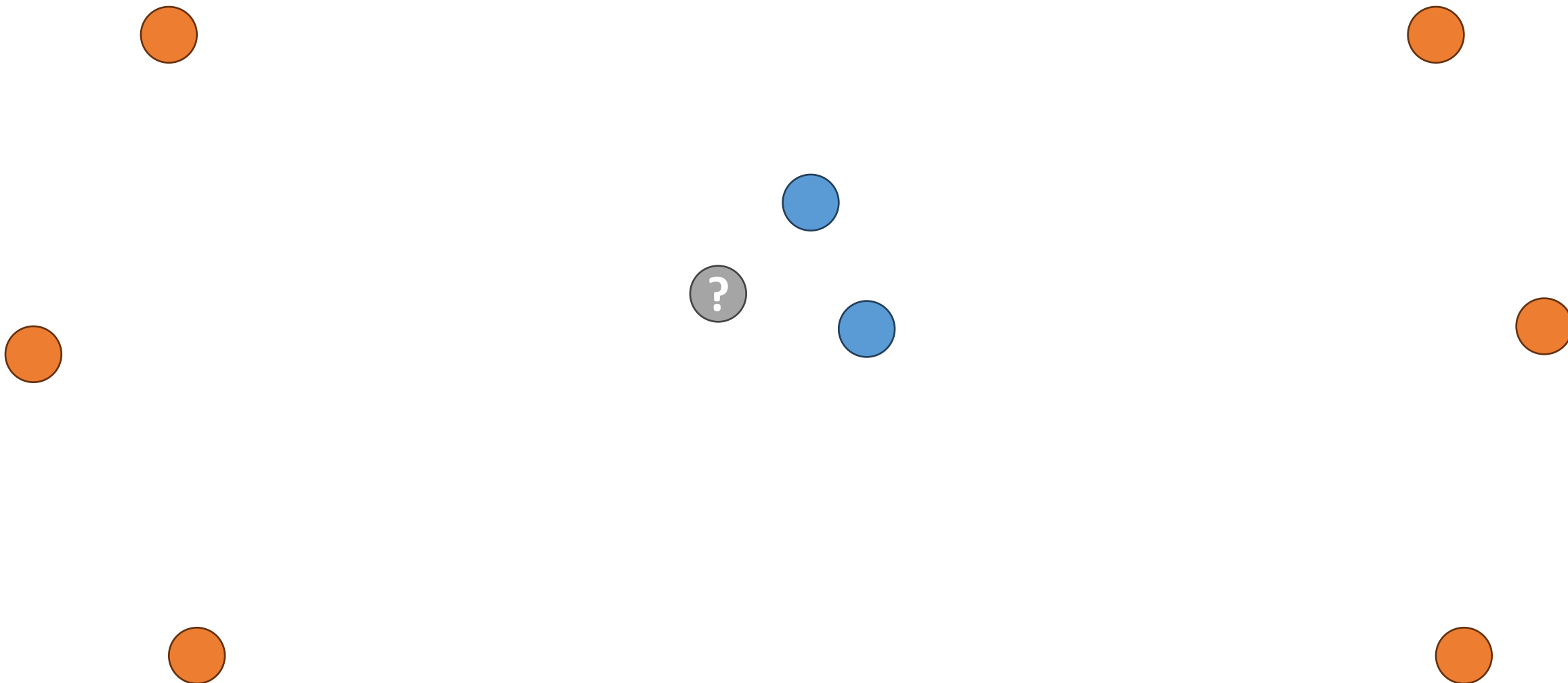


Функция потерь в классификации

Accuracy - это доля правильных ответов модели

$$accuracy = \frac{1}{\ell} \sum_{i=1}^{\ell} [a(x_i) = y_i]$$

Проблема kNN



Взвешенный knn

$$a(x) = \operatorname{argmax}_{y \in Y} \sum_{i=1}^k w_i [y_{(i)} = y]$$

$$w_i = \frac{1}{\rho(x, x_{(i)})}$$

Метод k ближайших соседей в sklearn

[illegible]

Метод k ближайших соседей для регрессии: Обучение

Дано: обучающая выборка $X = (x_i, y_i)_{i=1}^{\ell}$

Задача регрессии: ответы из множества $Y = \mathbb{R}$

Обучение модели: запоминаем обучающую выборку X

Метод k ближайших соседей для регрессии: **Применение**

Дано: новый объект x

Применение модели:

Сортируем объекты обучающей выборки по расстоянию до нового объекта:

$$\rho(x, x_{(1)}) \leq \rho(x, x_{(2)}) \leq \dots \leq \rho(x, x_{(\ell)})$$

Выбираем k ближайших объектов: $x_{(1)}, \dots, x_{(k)}$

Усредняем ответы

$$a(x) = \frac{1}{k} \sum_{i=1}^k y_{(i)}$$

Метод k ближайших соседей для регрессии в sklearn

[illegible]

Метод k ближайших соседей

[sklearn.neighbors.KNeighborsClassifier](#)

[sklearn.neighbors.KNeighborsRegressor](#)

[BallTree и KD-Tree](#)