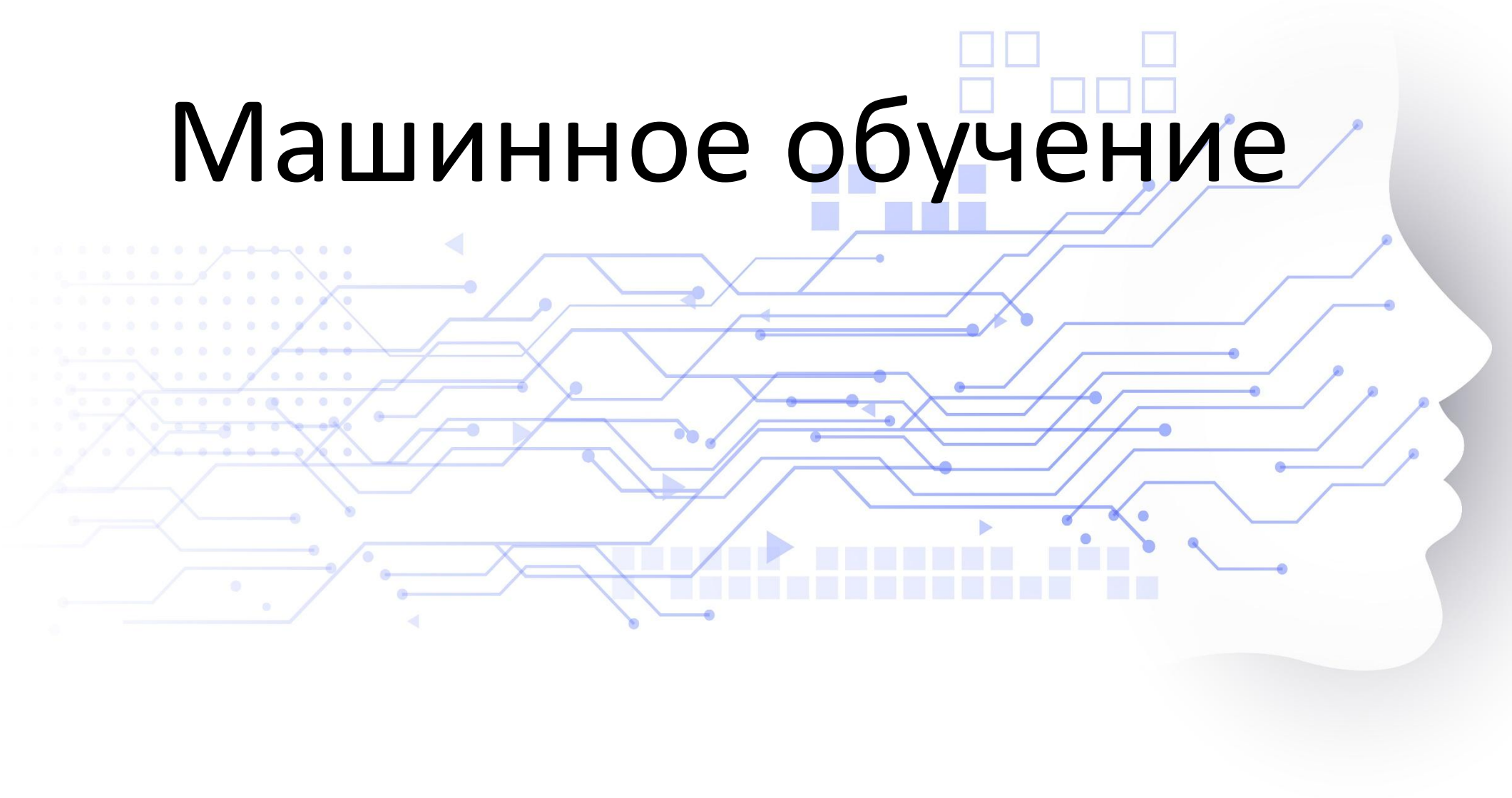
Машинное обучение



Метод k ближайших соседей (Knn)

Как отличить ель от сосны?

Ель

Сосна

Как отличить ель от сосны?

Ель

Сосна

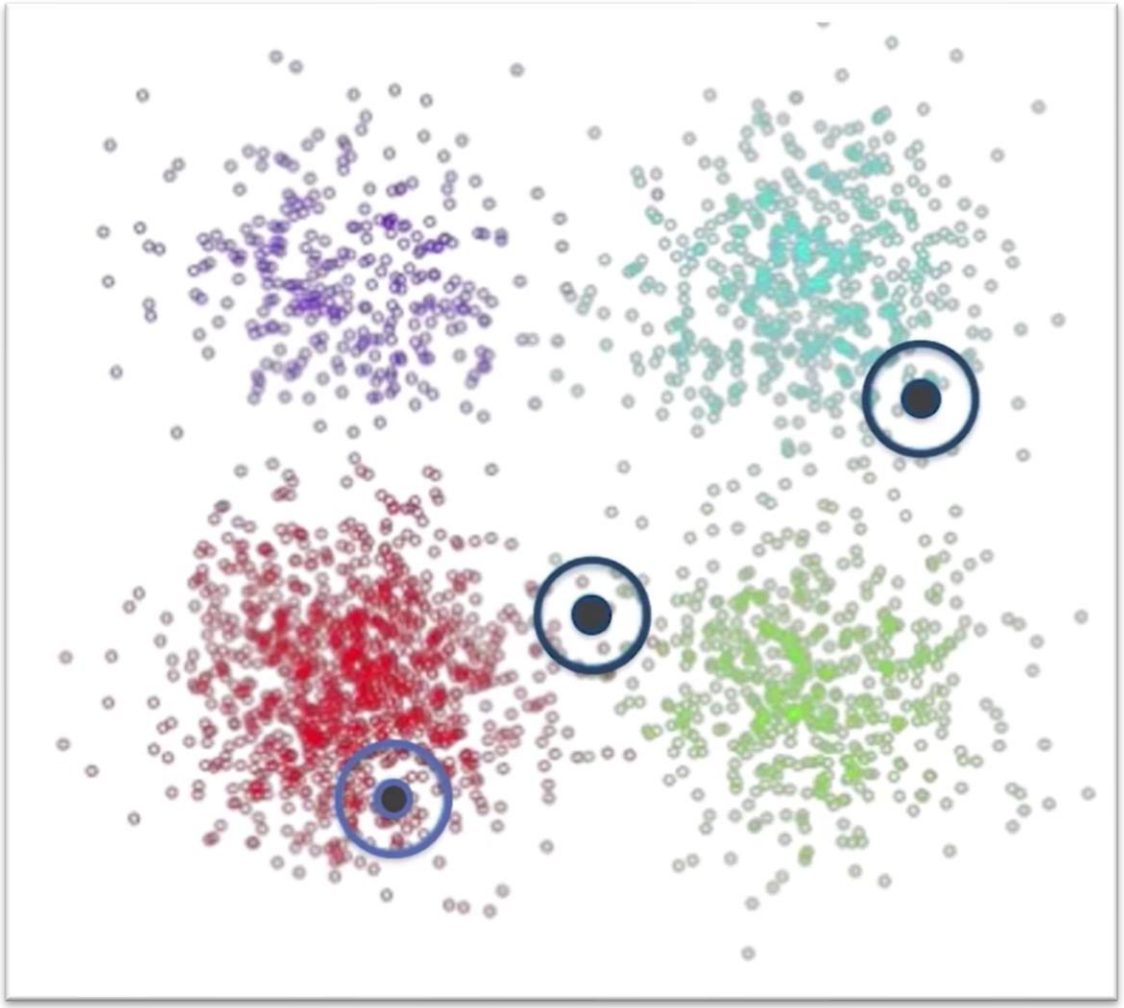
Ветки	Смотрят вверх	Параллельно земле
Ствол	Не видно	Видно
Иголки	Густые	Более редкие
Цвет	Ближе к зеленому	Ближе к желтому

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 обьект x

Применение модели:

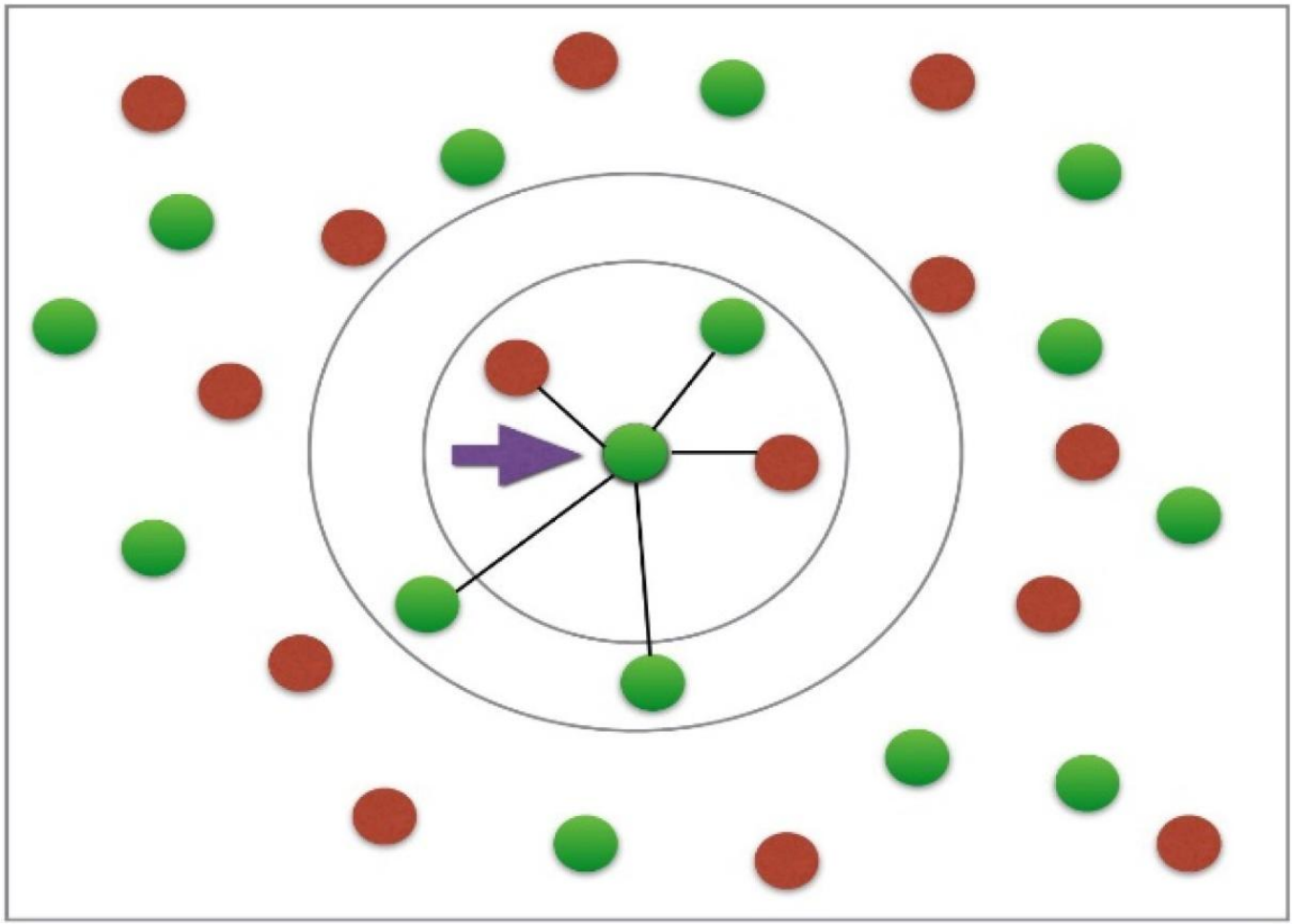
Сортируем объекты обучающей выборки по расстоянию до нового объекта:
$\rho(x, x_{(1)}) \leq \rho(x, x_{(2)}) \leq \ldots \leq \rho(x, x_{(\ell)})$

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

Выдаем наиболее популярный среди них класс

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

Метод k ближайших соседей (kNN): Применение



Метрика

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

1. Условие тождественности (Identity):

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

2. Симметрия (Symmetry)

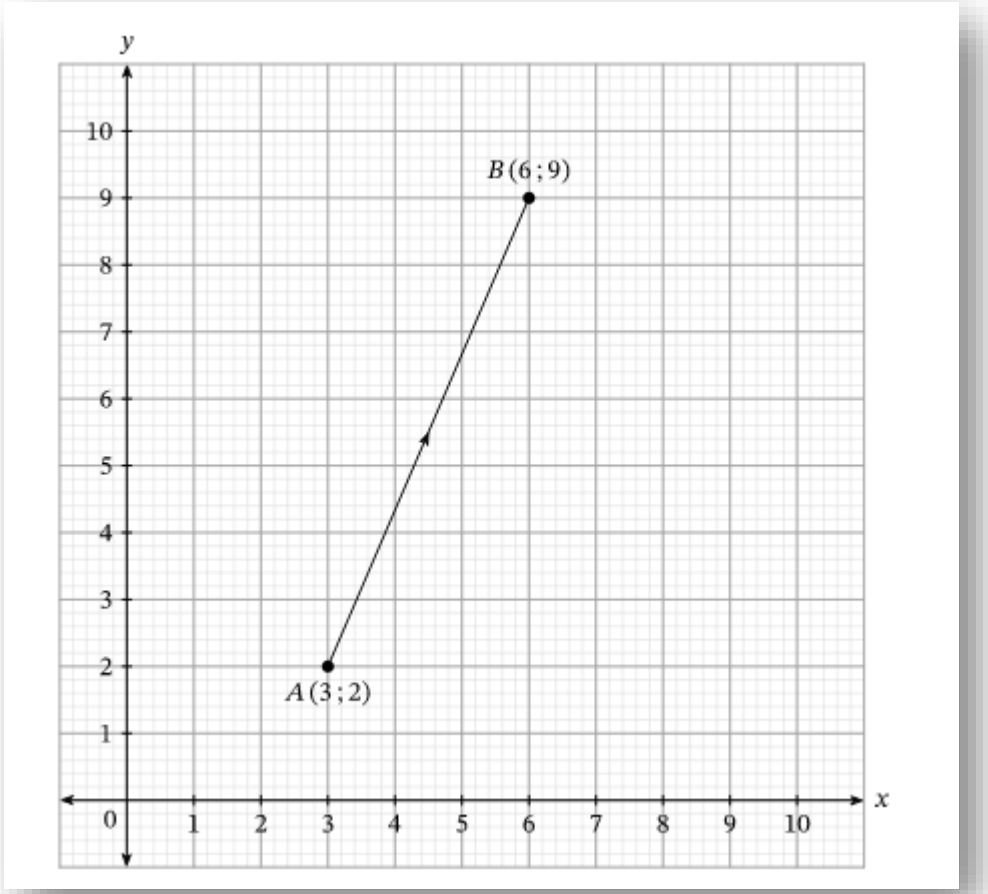
$\rho(x, z) = \rho(z, x)$

3. Неравенство треугольника (Triangle inequality):

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

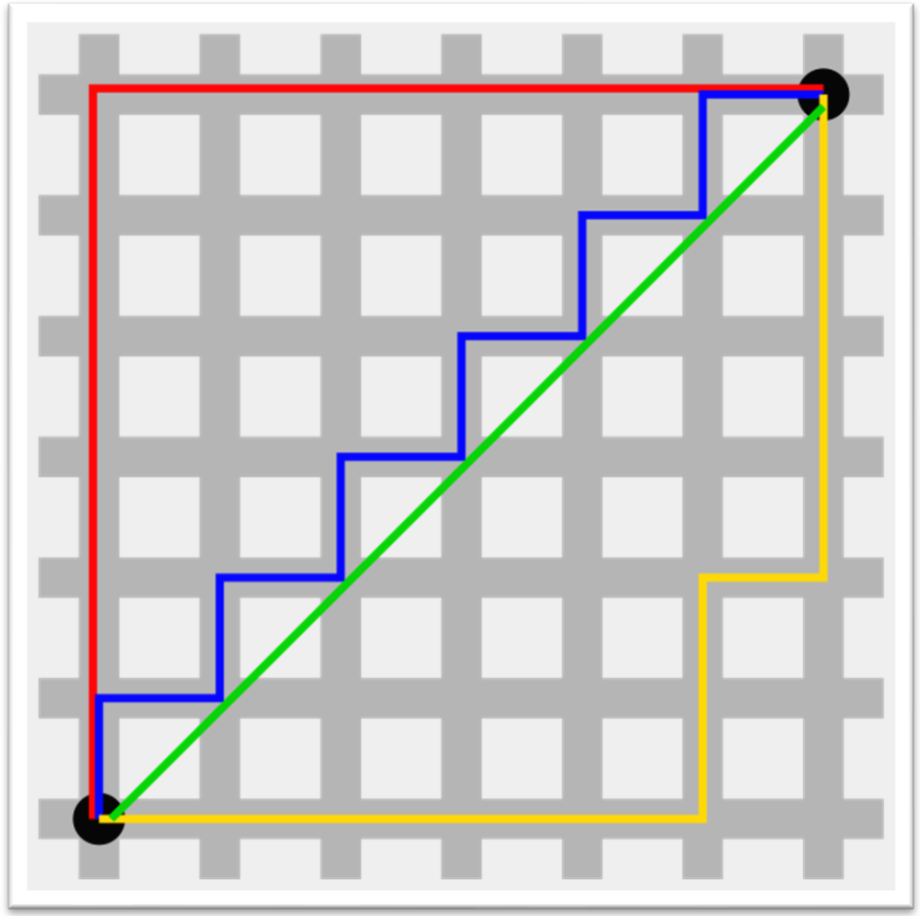
Евклидова метрика

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



Манхэттенская метрика

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

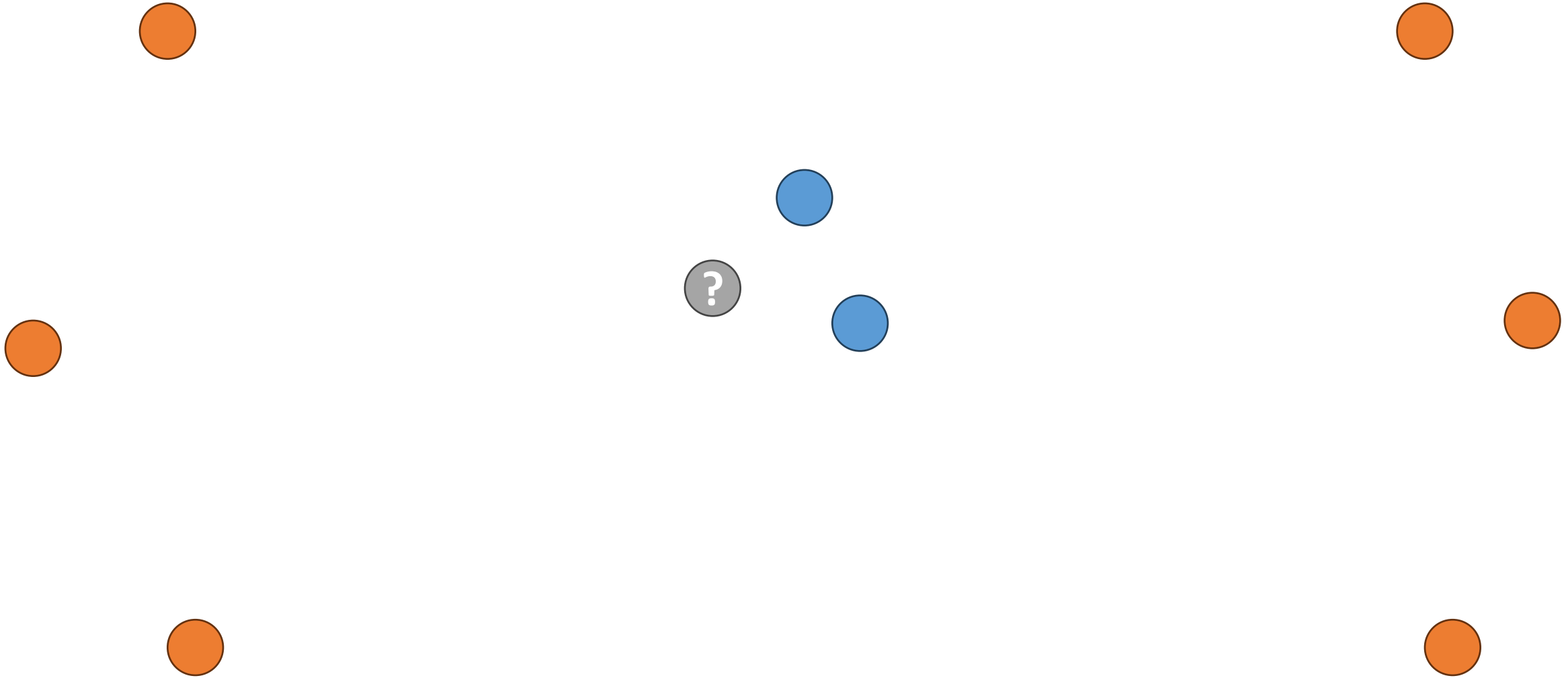


Функция потерь в классификации

Accuracy - это доля правильных ответов модели

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

Проблема kNN



Взвешенный knn

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

$w_i = \mathrm{K}(\frac{\rho(x, x_{(i)})}{h})$

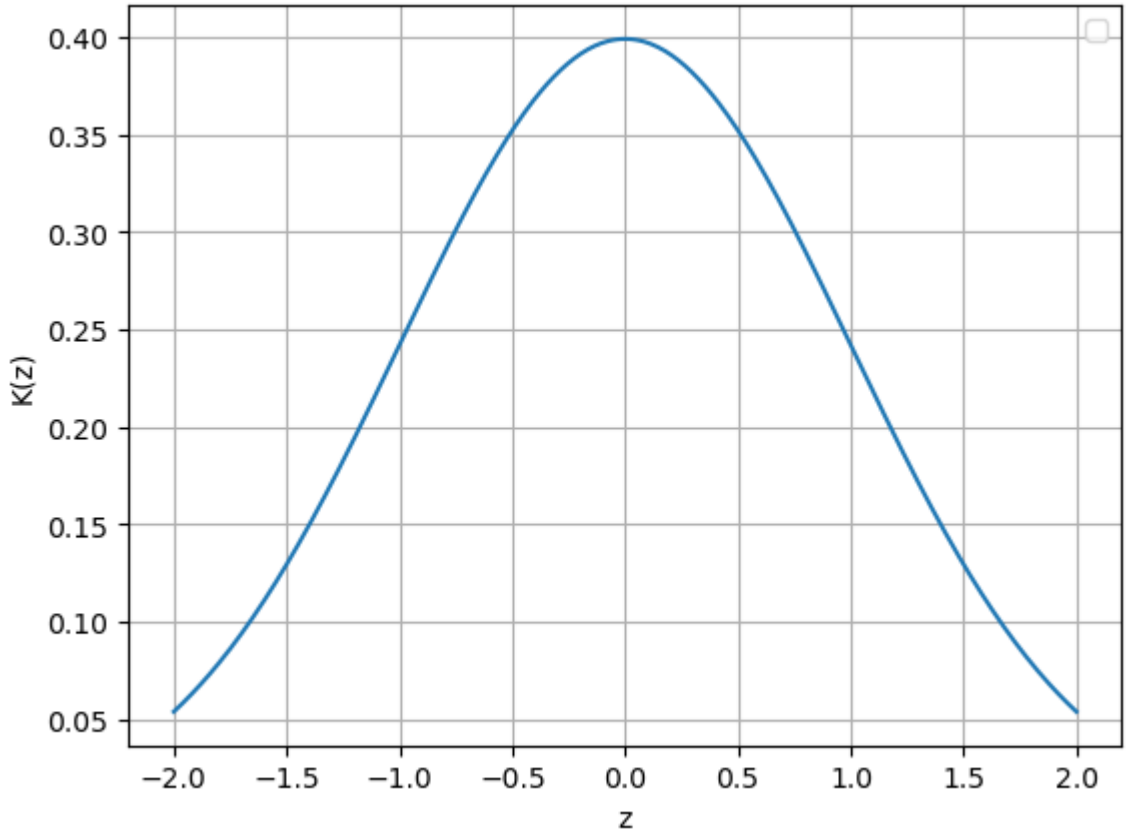
K – ядро

h – ширина окна

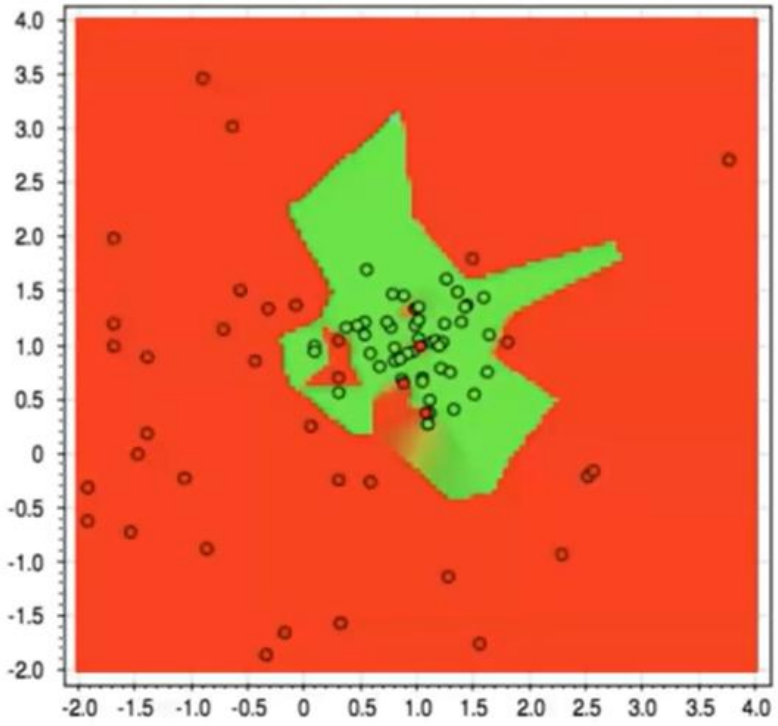
Ядра для весов

Гауссовское ядро

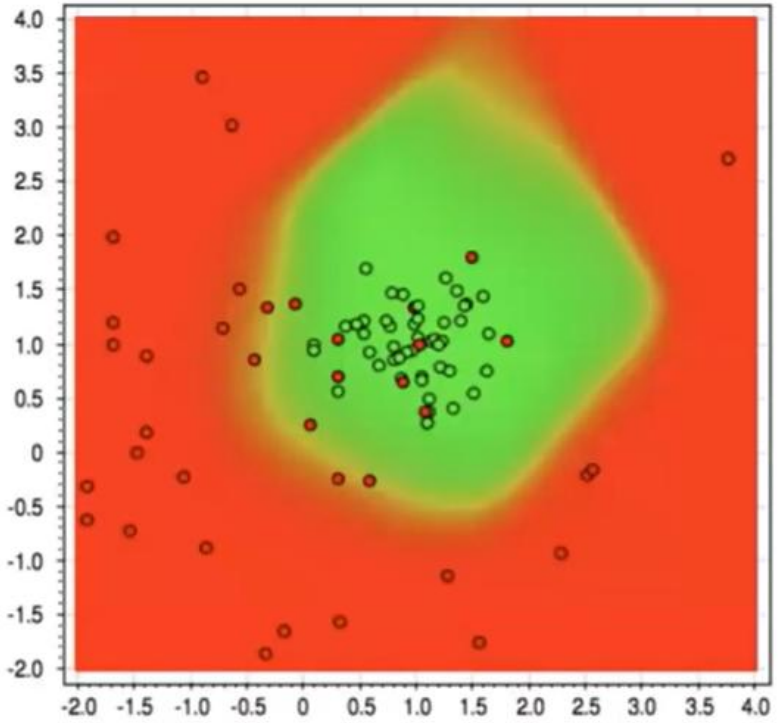
$$K(\mathrm{z}) = (2\pi)^{-0.5}\exp(\frac{1}{2}z^2)$$



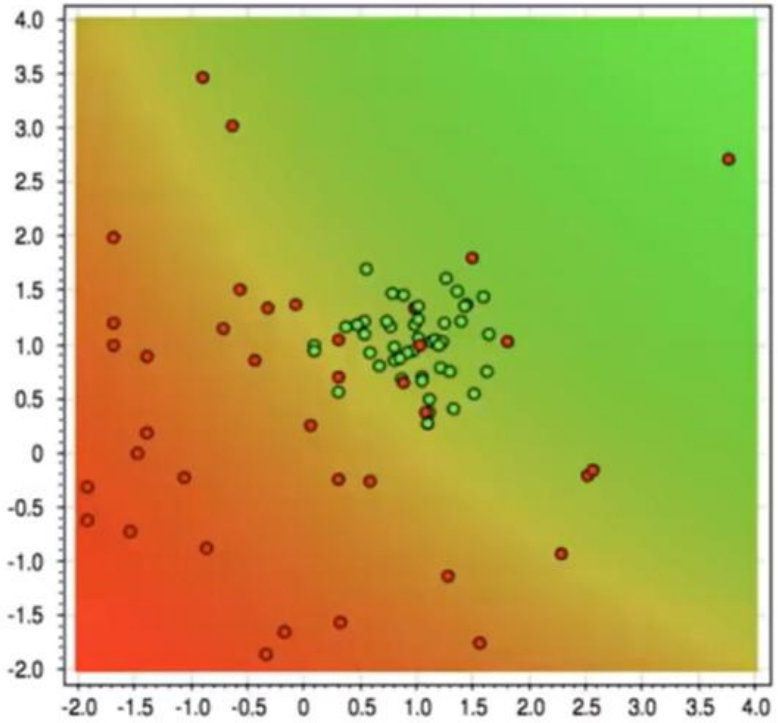
Ядра для весов



$h = 0.05$



$h = 0.5$



$h = 5$

Метод k ближайших соседей в sklearn

```
sklearn.neighbors.KneighborsClassifier(n_neighbors=5,
weights='uniform',p=2,metric='minkowski')
```

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 обьект x

Применение модели:

Сортируем объекты обучающей выборки по расстоянию до нового объекта:
$\rho(x, x_{(1)}) \leq \rho(x, x_{(2)}) \leq \ldots \leq \rho(x, x_{(\ell)})$

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

Усредняем ответы

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

Функция потерь для регрессии

Частый выбор – квадратичная функция потерь

$$L(y, a) = (a - y)^2$$

Функционал ошибки – спеднеквадратичная ошибка (mean squared error, MSE)

$$Q(\mathrm{a}, X) = \frac{1}{l}\sum_{i=1}^{l}(a(x^{(i)}) - y^{(i)})^2$$

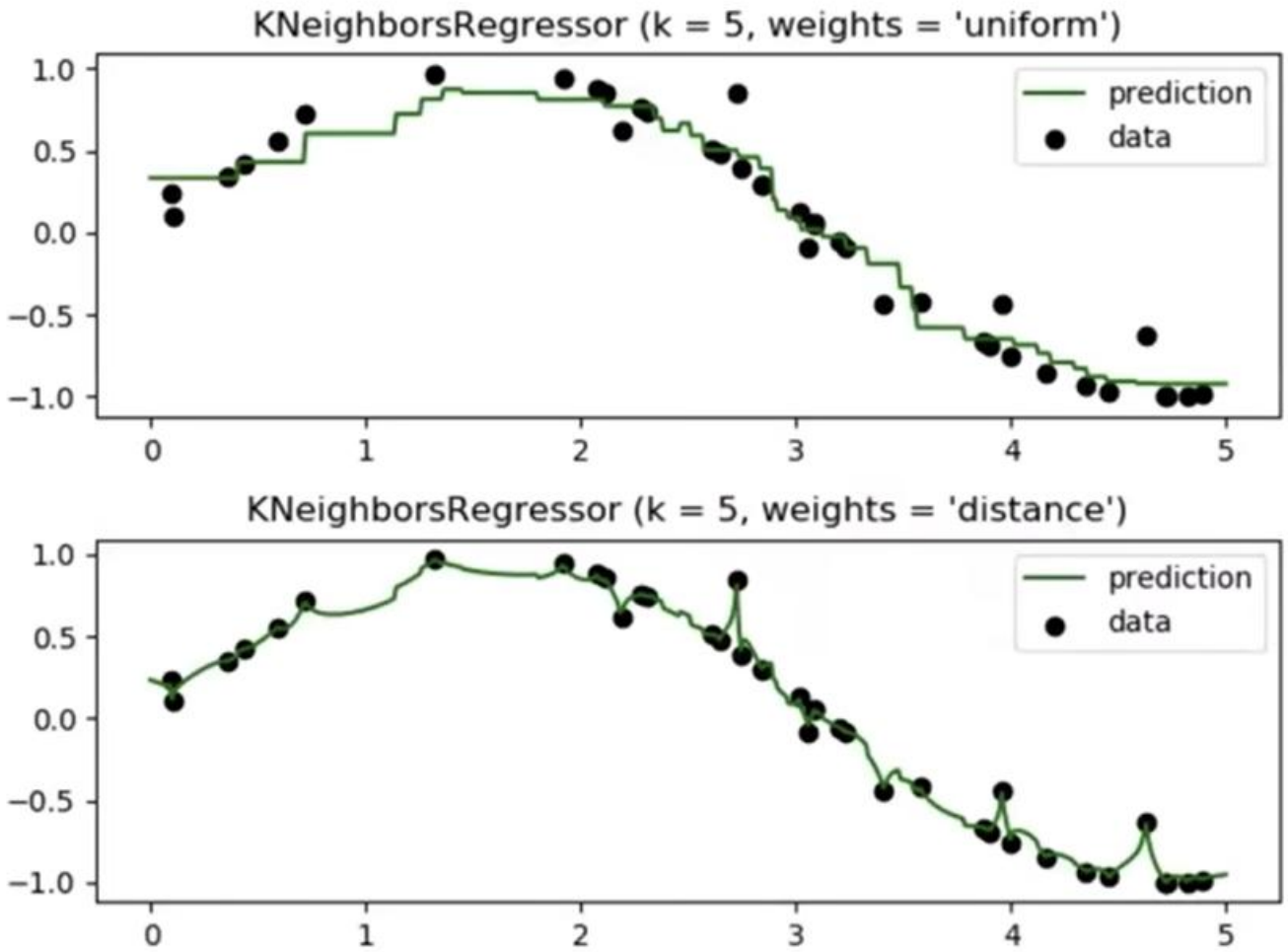
Взвешенный knn

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

$$w_i = \mathrm{K}\left(\frac{\rho\left(x, x_{(i)}\right)}{h}\right)$$

Формула Надарая-Ватсона

Применение


KNeighborsRegressor (k = 5, weights = 'uniform')
prediction
data
1.0
0.5
0.0
−0.5
−1.0
0
1
2
3
4
5
KNeighborsRegressor (k = 5, weights = 'distance')
prediction
data
1.0
0.5
0.0
−0.5
−1.0
0
1
2
3
4
5

Метод k ближайших соседей для регрессии в sklearn

```
sklearn.neighbors.KneighborsRegressor(n_neighbors=5, weights='uniform',
p=2,metric='minkowski')
```

Плюсы kNN

- Если данных много и для любого объекта найдётся похожий в обучающей выборке, то это лучшая модель
- Очень простое обучение
- Мало гиперпараметров

Минусы kNN

- Часто другие модели оказываются лучше
- Надо хранить в памяти всю обучающую выборку
- Искать k ближайших соседей довольно долго
- Мало способов настроить модель

Метод k ближайших соседей

sklearn.neighbors.KNeighborsClassifier

sklearn.neighbors.KNeighborsRegressor

BallTree и KD-Tree