



РАНХиГС

РОССИЙСКАЯ АКАДЕМИЯ НАРОДНОГО ХОЗЯЙСТВА
И ГОСУДАРСТВЕННОЙ СЛУЖБЫ
ПРИ ПРЕЗИДЕНТЕ РОССИЙСКОЙ ФЕДЕРАЦИИ



РАНХиГС
экономический
факультет



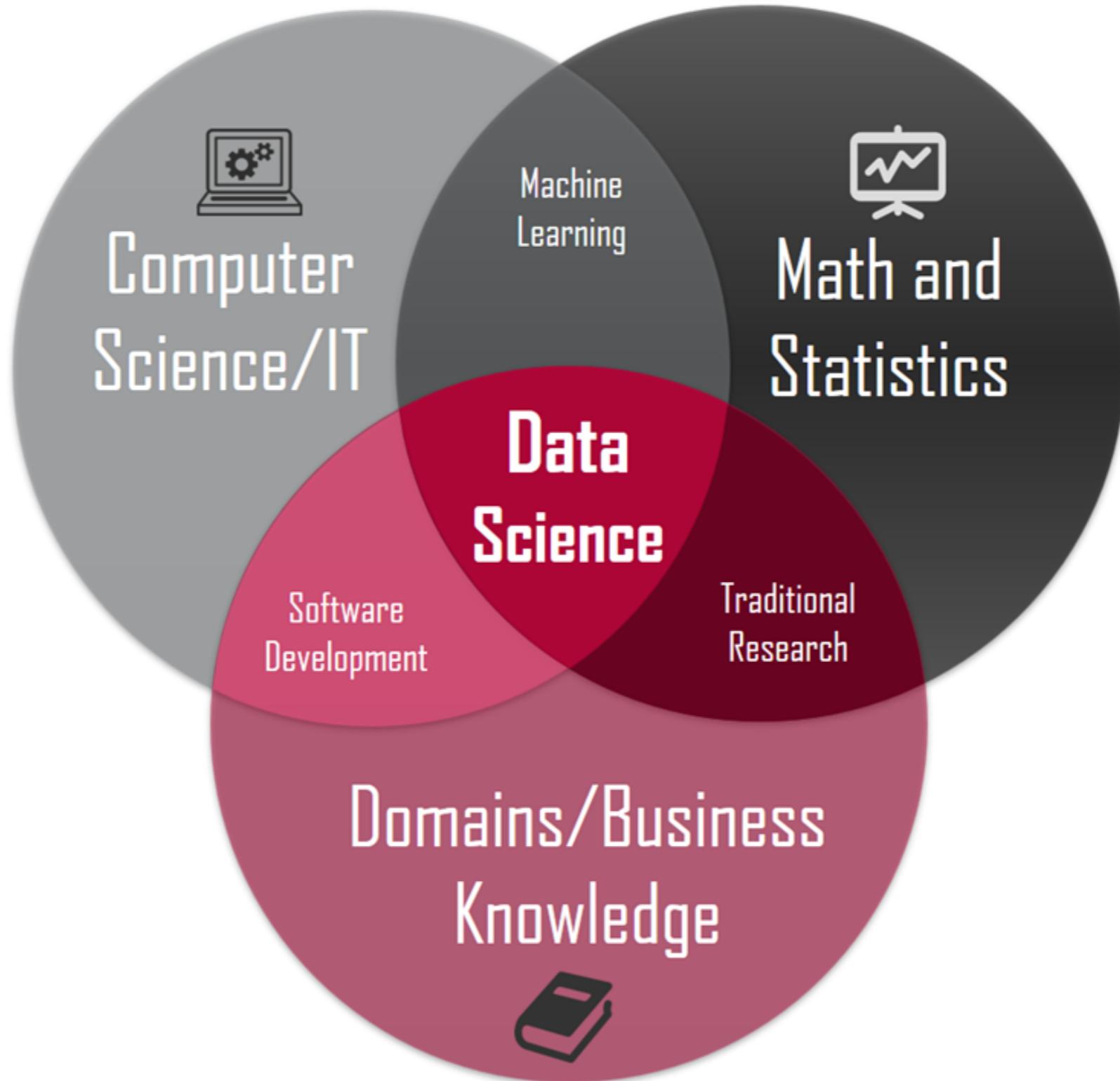
Введение в машинное обучение (Machine Learning)

проф. кафедры Эконометрики и математической экономики ЭФ
д.т.н. Шилин Кирилл Юрьевич

РАНХиГС каб. 419/3
email: kshilin@ranepa.ru



Что такое «наука о данных»?



Основной перечень задач

Машинное обучение предназначено для:

- 1. Классификация (поиск класса объекта)**
- 2. Регрессия (влияния независимых переменных на зависимую)**
- 3. Кластеризация (разделение объектов на классы)**
- 4. Понижение размерности данных (выявление значимых переменных)**
- 5. Одноклассовая классификация и выявление новизны**
- 6. Восстановление плотности распределения вероятности по набору данных**
- 7. Построение ранговых зависимостей**

Способы машинного обучения:

- 1. Обучение с учителем (1, 2)**
- 2. Обучение без учителя (3, 4, 5)**
- 3. Обучение с подкреплением**



Современное деление методов ML



Классические методы
(поиск признаков ручной)
scikit-learn, catboost, xgboost

Глубокое обучение
(поиск признаков в алгоритме)
pytorch, tensorflow

Вероятностное
программирование
pymc, pyro

Основные алгоритмы



Обучение с учителем

- 1. Классификация**
- 2. Регрессия (прогнозирование)**

основные алгоритмы

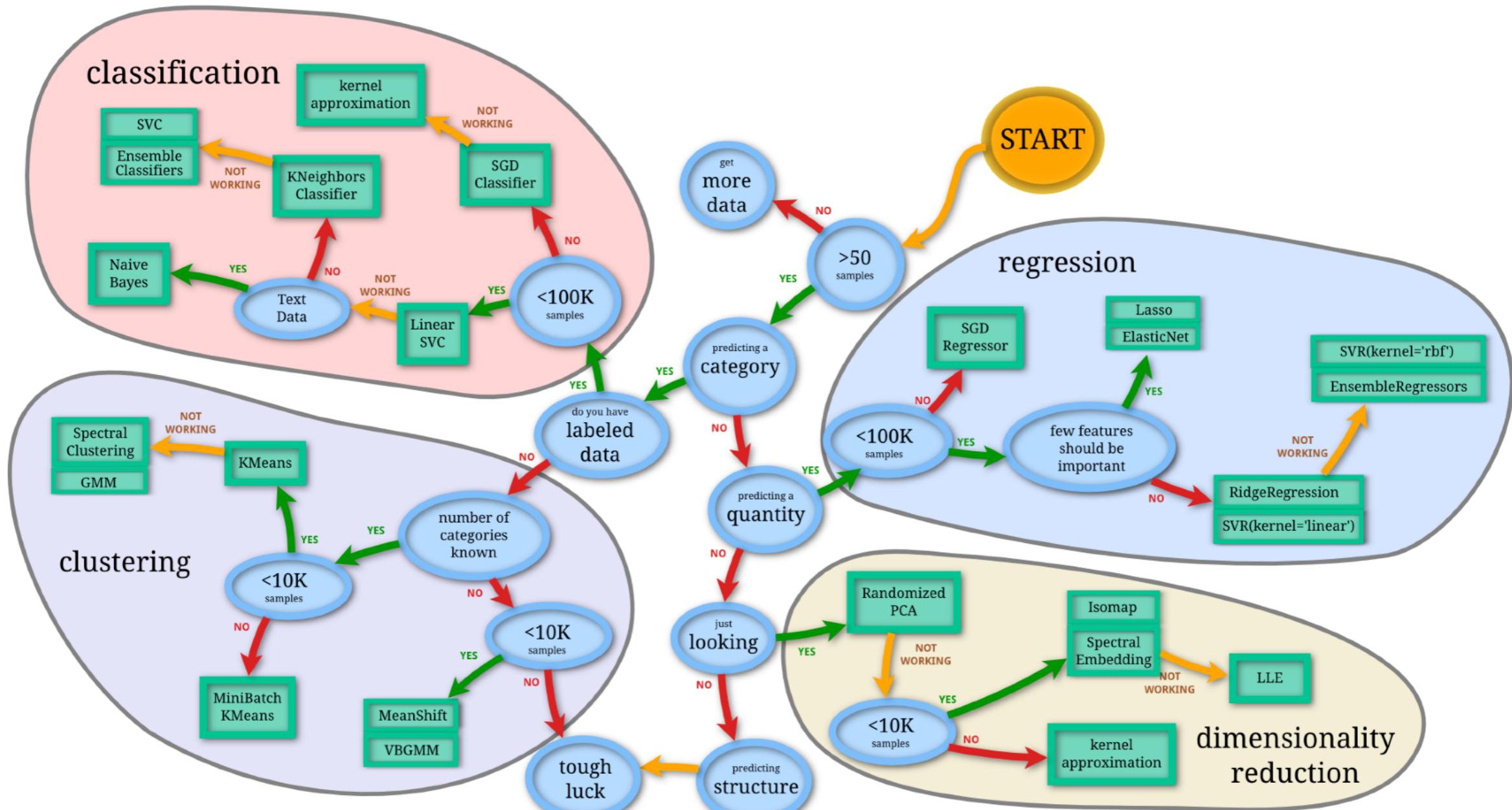
- Метод k ближайших соседей
- Линейные модели
- Наивные байесовские классификаторы
- Деревья решений
- Ансамбли деревьев решений
- Ядерный метод опорных векторов
- Нейронные сети (глубокое обучение)

Обучение без учителя

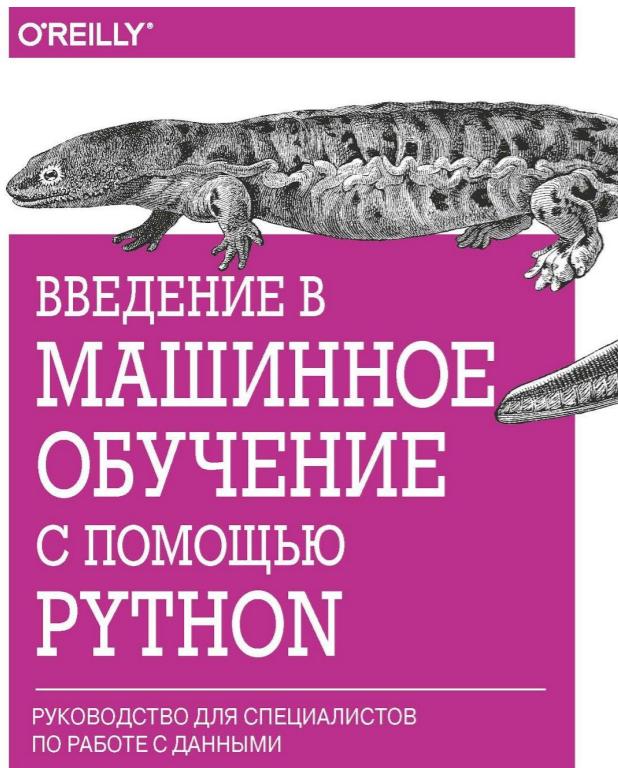
- 1. Кластеризация**
 - Кластеризация k-средних
 - Агломеративная кластеризация
 - DBSCAN
- 2. Понижение размерности**
 - Анализ главных компонент (PCA)
 - Факторизация неотрицательных матриц (NMF)
 - Множественное обучение с помощью алгоритма t-SNE

Машинное обучение в Python

Навигатор по работе с данными scikit-learn



Что читать и смотреть

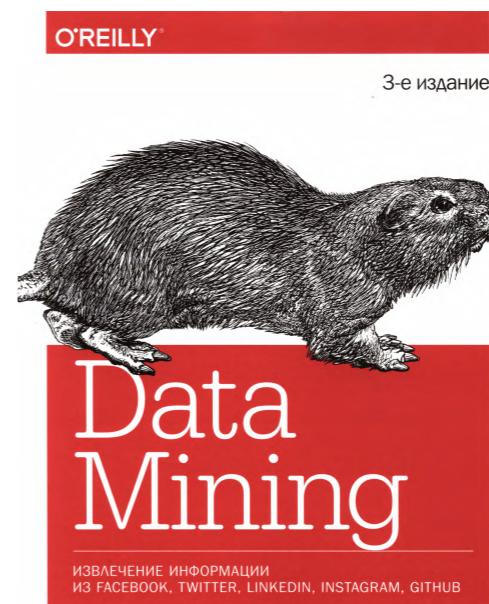


Андреас Мюллер, Сара Гвидо

Курс лекций Воронцова К.В.

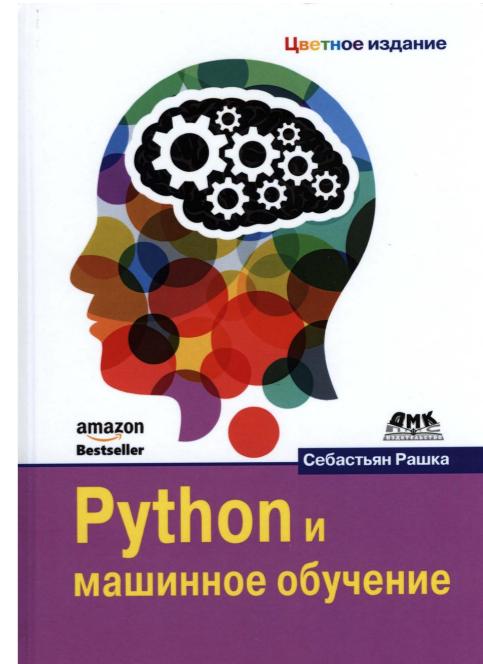


быв® Джоэл Грас

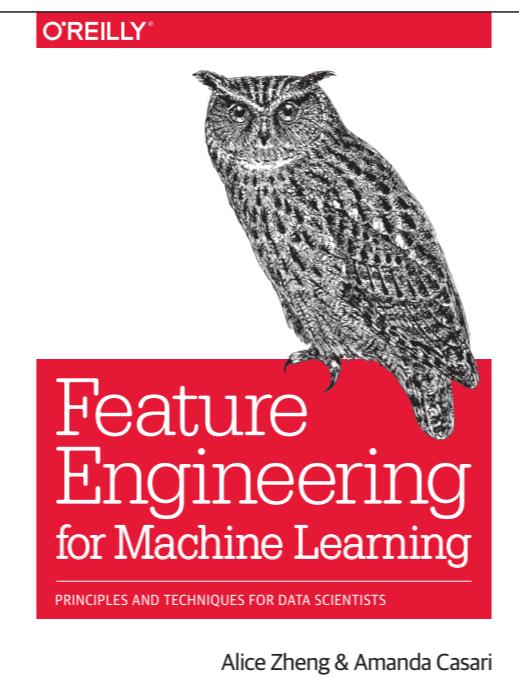


ИЗВЛЕЧЕНИЕ ИНФОРМАЦИИ
ИЗ FACEBOOK, TWITTER, LINKEDIN, INSTAGRAM, GITHUB

быв® Мэттью Рассел
Михаил Классен



Цветное издание
амазон
Bestseller
AMX
Себастьян Рашка



быв® Alice Zheng & Amanda Casari



быв® Практические решения от предобработки до глубокого обучения

Крис Элбон



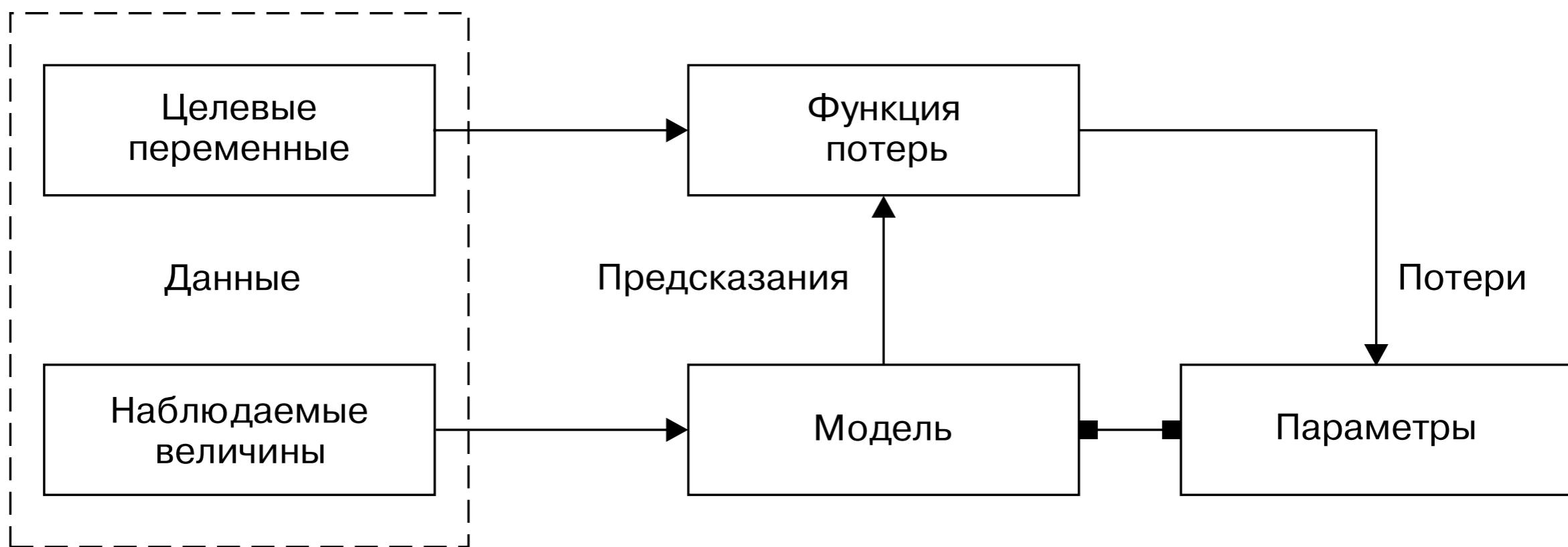
powered by
jupyter
быв® ПИТЕР®

Дж. Вандер Плас

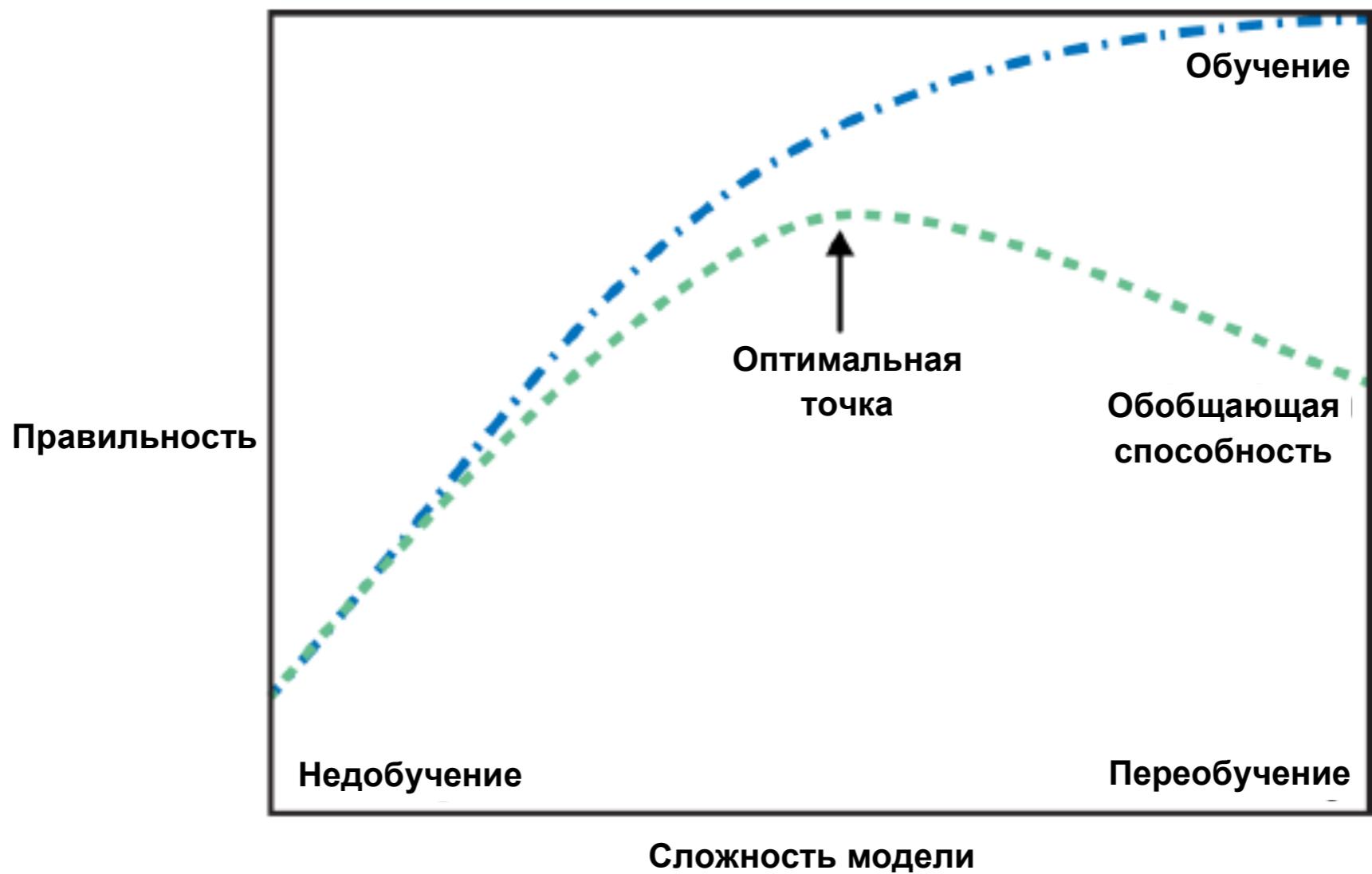
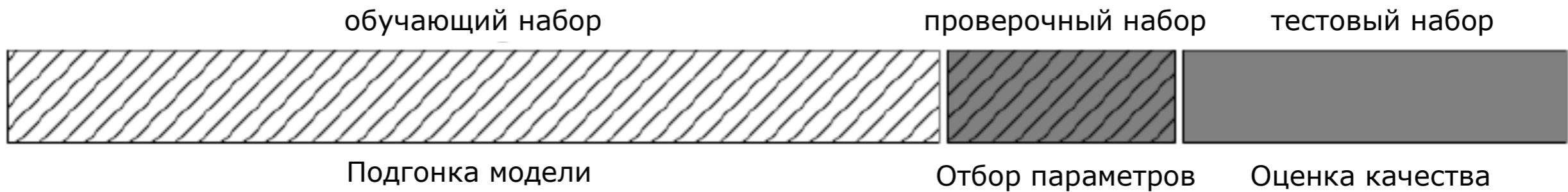
Машинное обучение в Python

Построение модели машинного обучения с учителем:

- Этап 1 Подготовка данных (препроцессинг)**
- Этап 2 Подбор алгоритма машинного обучения**
- Этап 3 Оценка качества модели и ее улучшение**



Обучение с учителем



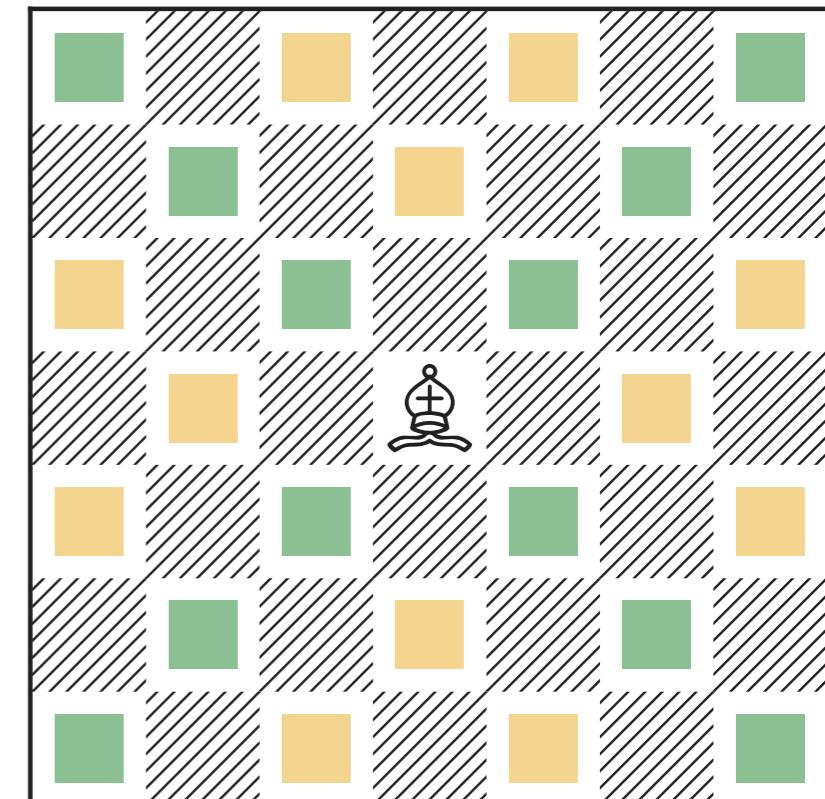
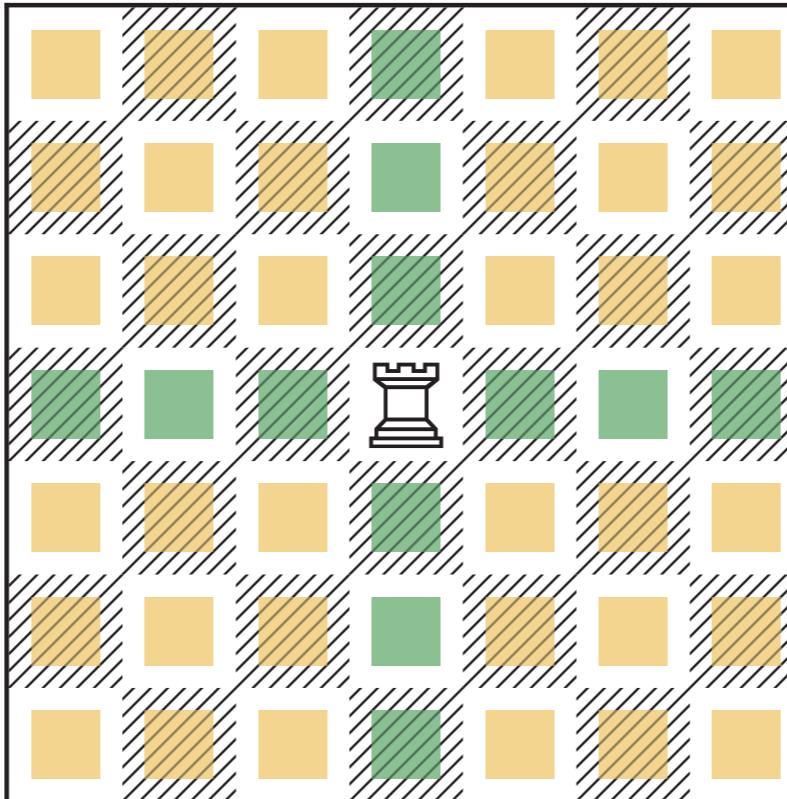
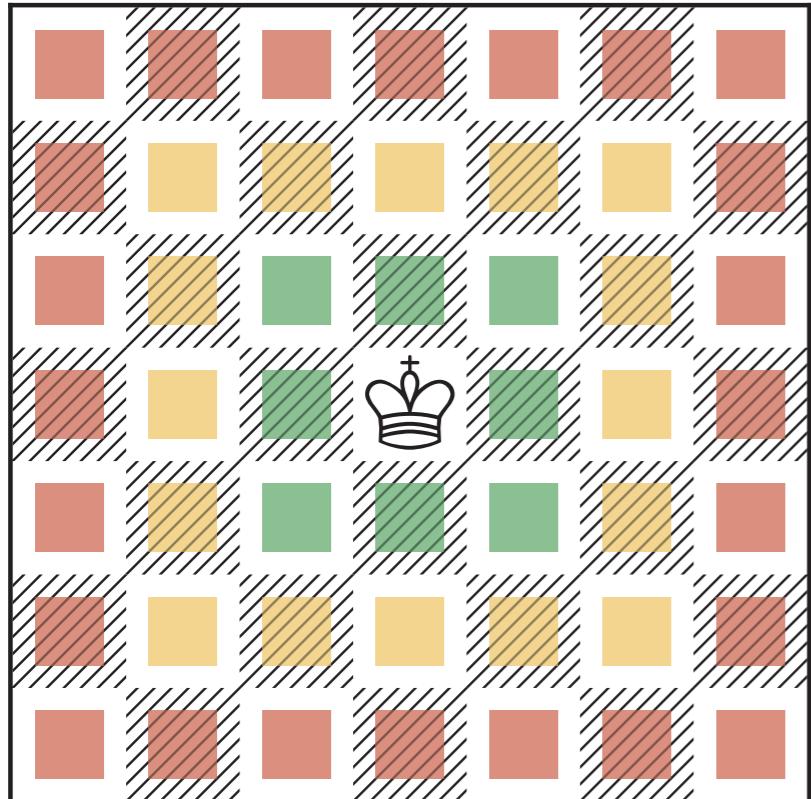


Пауза



Прервемся с презентацией и запишем формально задачу машинного обучения.

Шахматы





Метрическое пространство

Метрическое пространство есть пара (X, d) , где X – множество, а d – числовая функция, которая определена на декартовом произведении $X \times X$, принимает значения в множестве неотрицательных вещественных чисел, и такова, что

1. $d(x, y) = 0 \Leftrightarrow x = y$ (аксиома тождества).
2. $d(x, y) = d(y, x)$ (аксиома симметрии).
3. $d(x, z) \leq d(x, y) + d(y, z)$ (аксиома треугольника или неравенство треугольника).

При этом

- множество X называется подлежащим множеством метрического пространства.
- элементы множества X называются точками метрического пространства.
- функция d называется метрикой.

Расстояние Минковского порядка p между двумя точками определяется как:

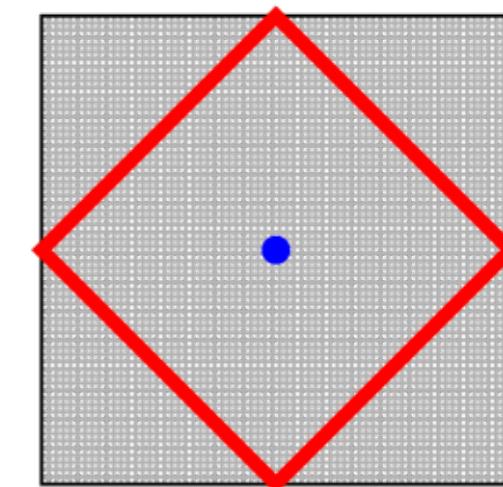
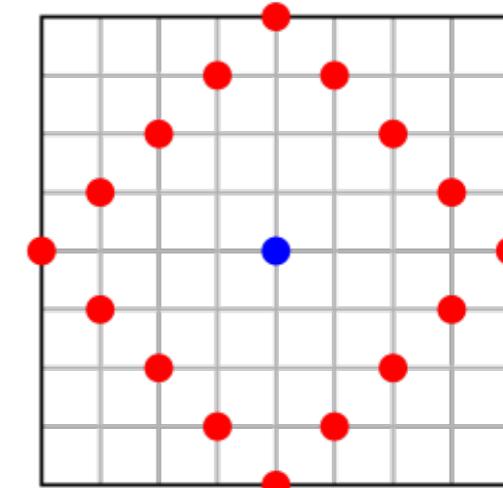
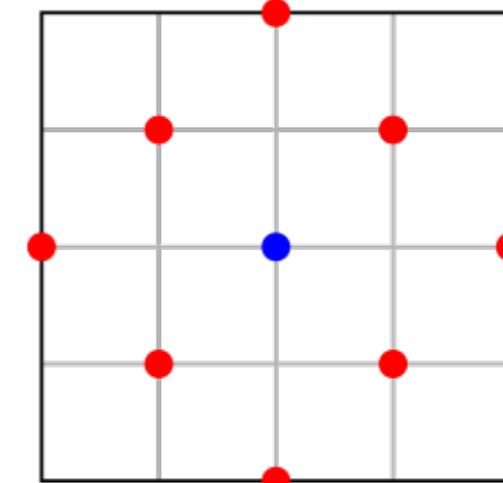
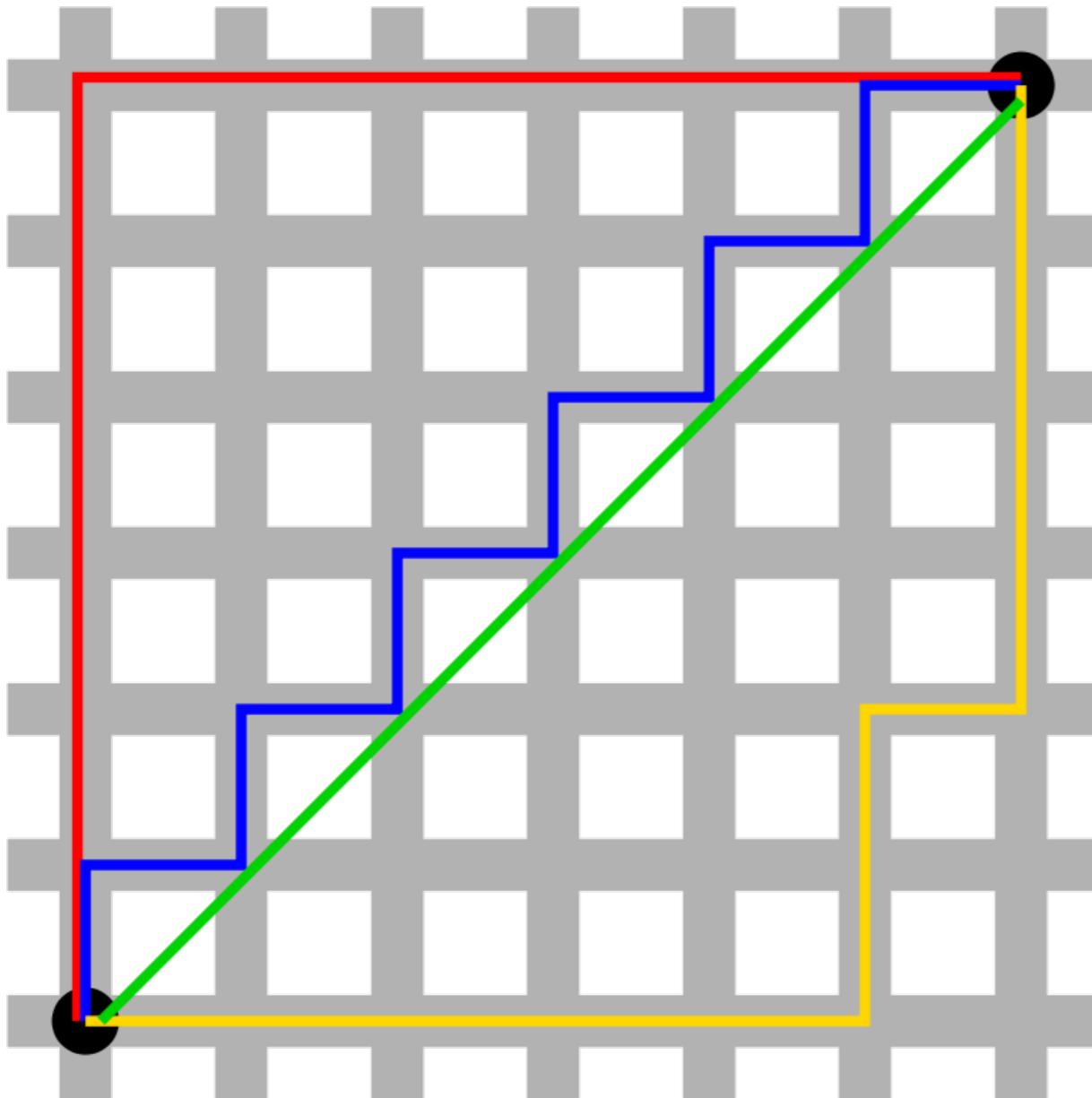
$$\rho(x, y) = \left(\sum_{i=1}^n |x_i - y_i|^p \right)^{1/p}.$$

Для $p \geq 1$ расстояние Минковского является метрикой вследствие неравенства Минковского.

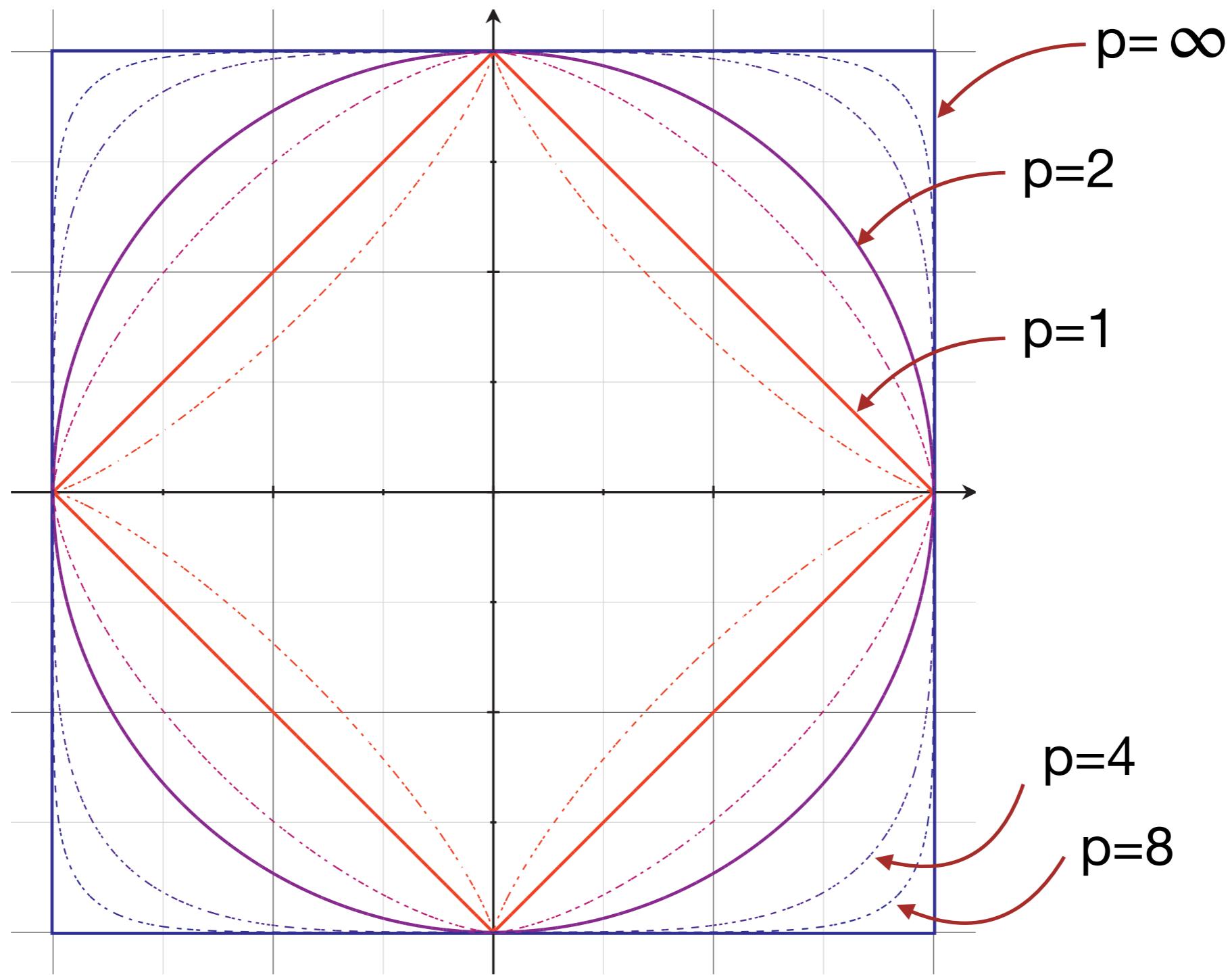
Для $p < 1$ расстояние не является метрикой, поскольку нарушается неравенство треугольника.

При $p = \infty$ метрика обращается в расстояние Чебышёва.

Манхэттенское расстояние

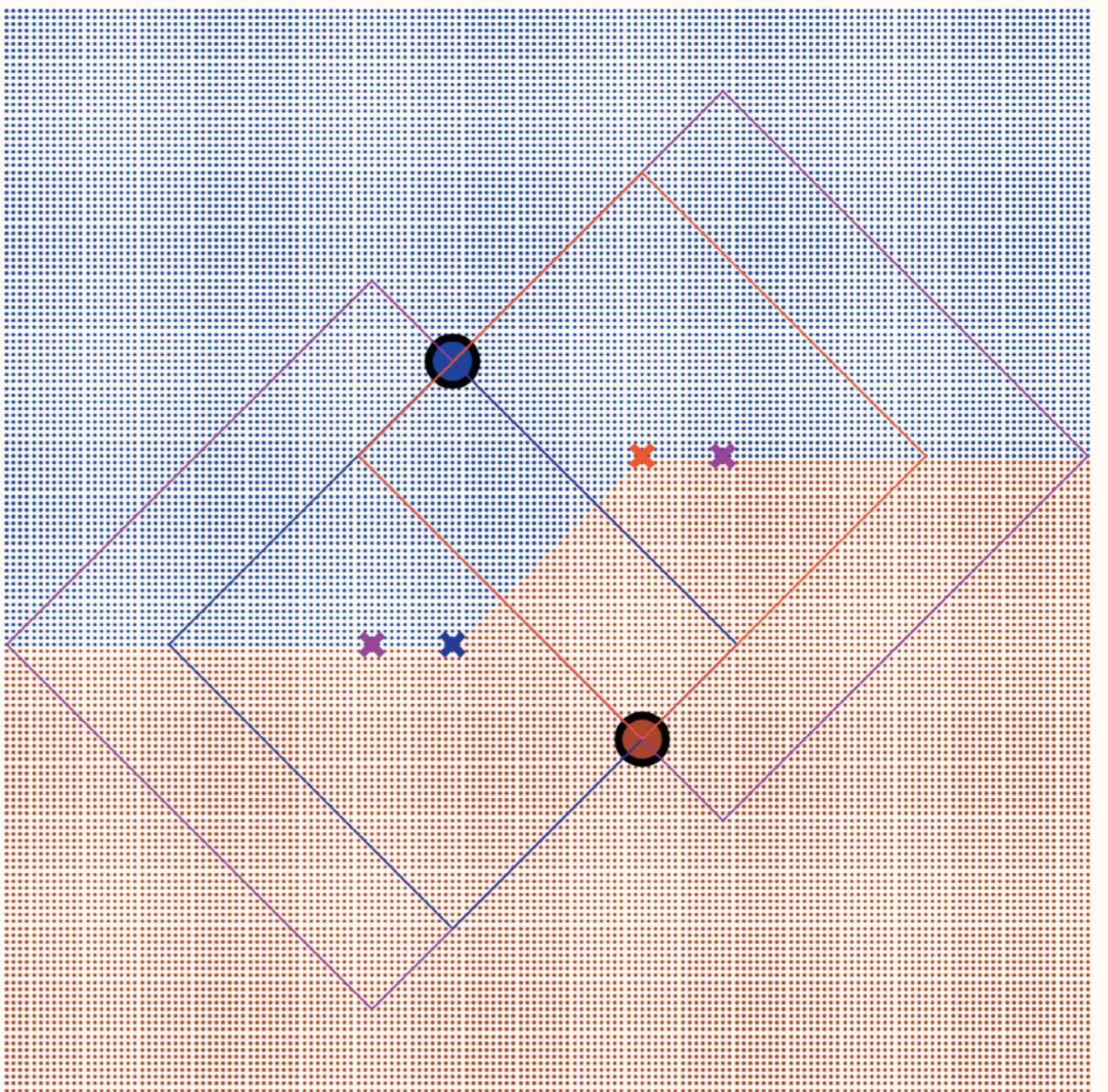
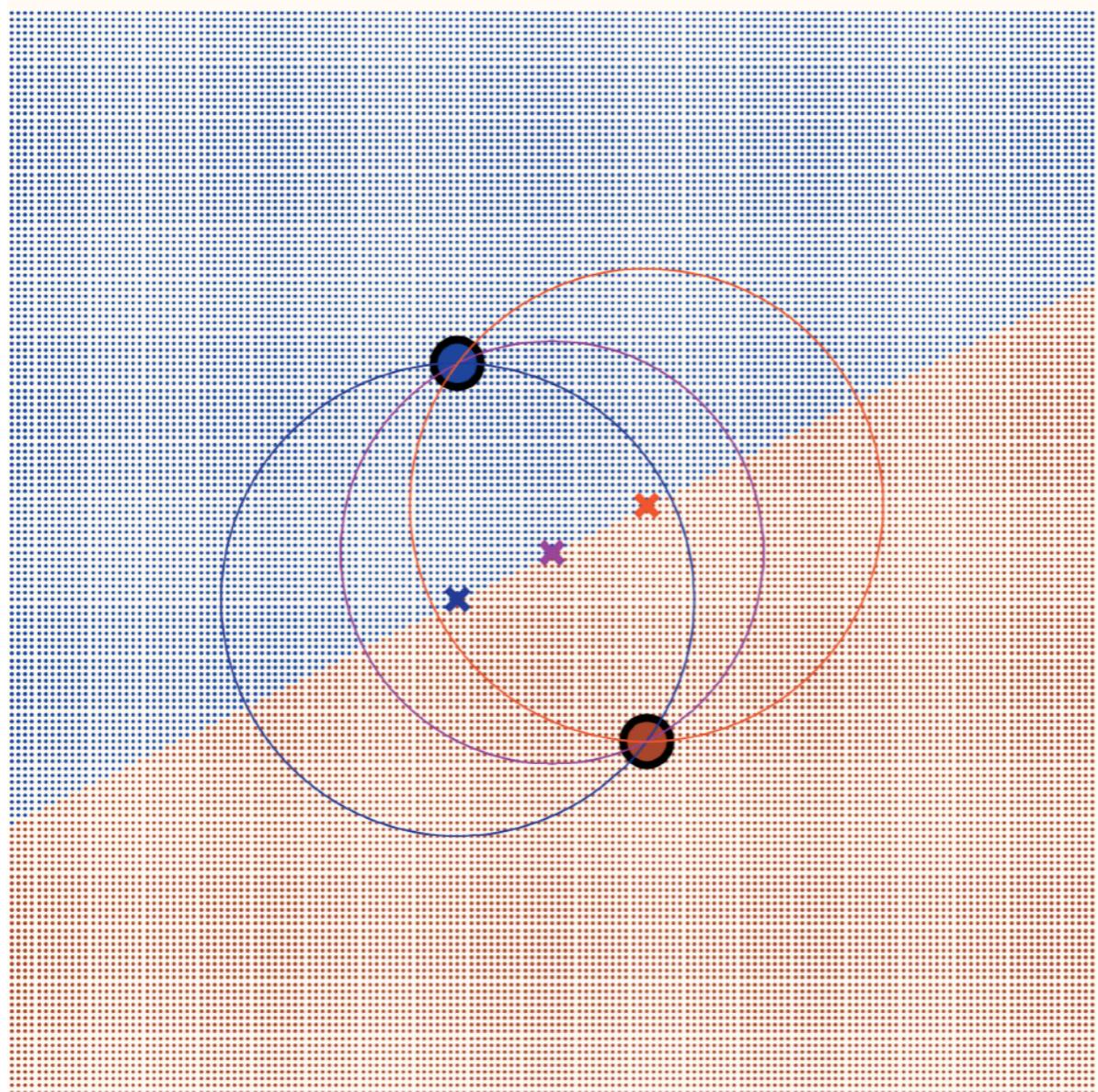


Расстояние Миньковского

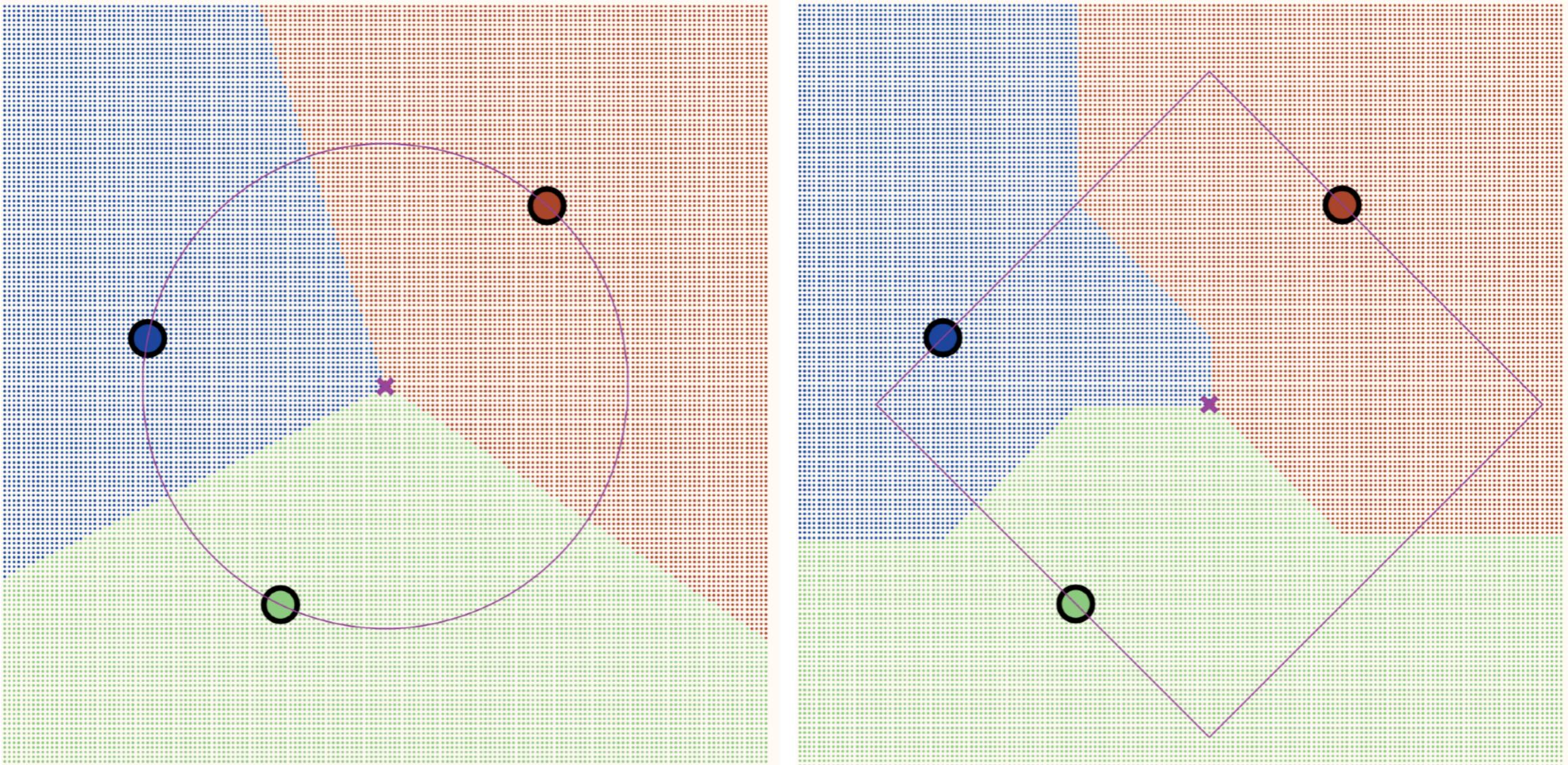




Решающая граница

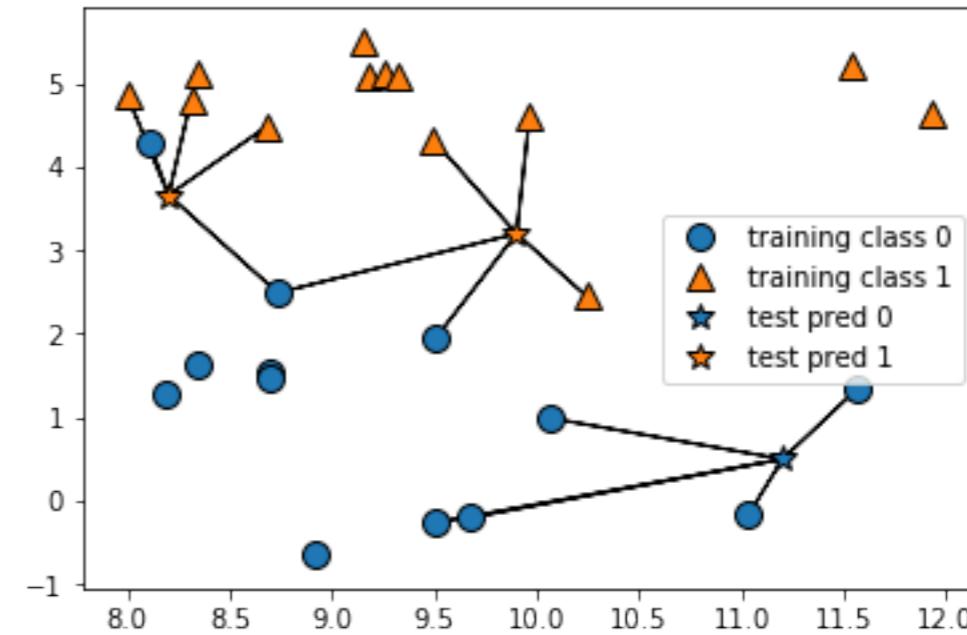
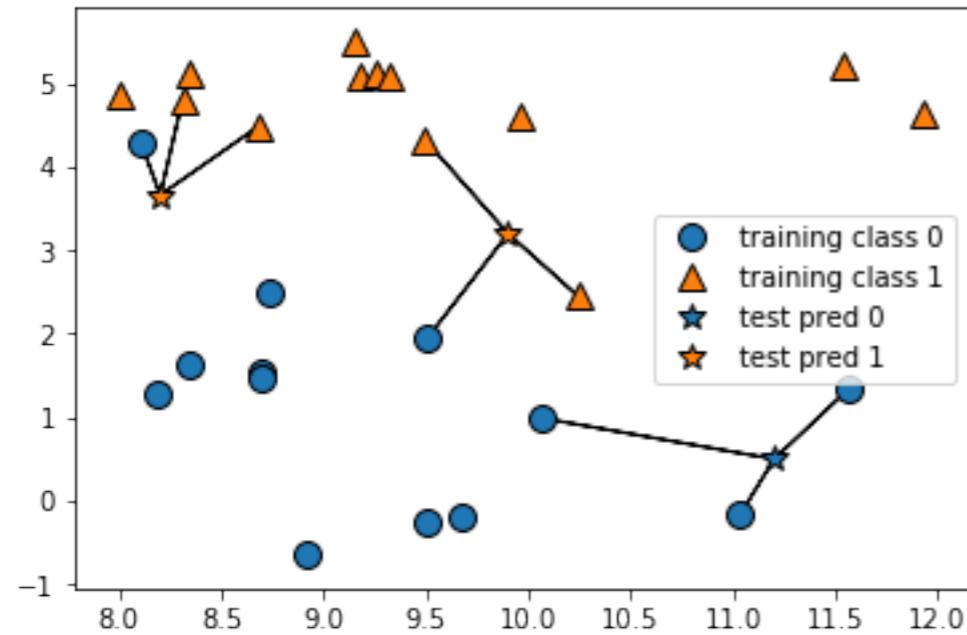
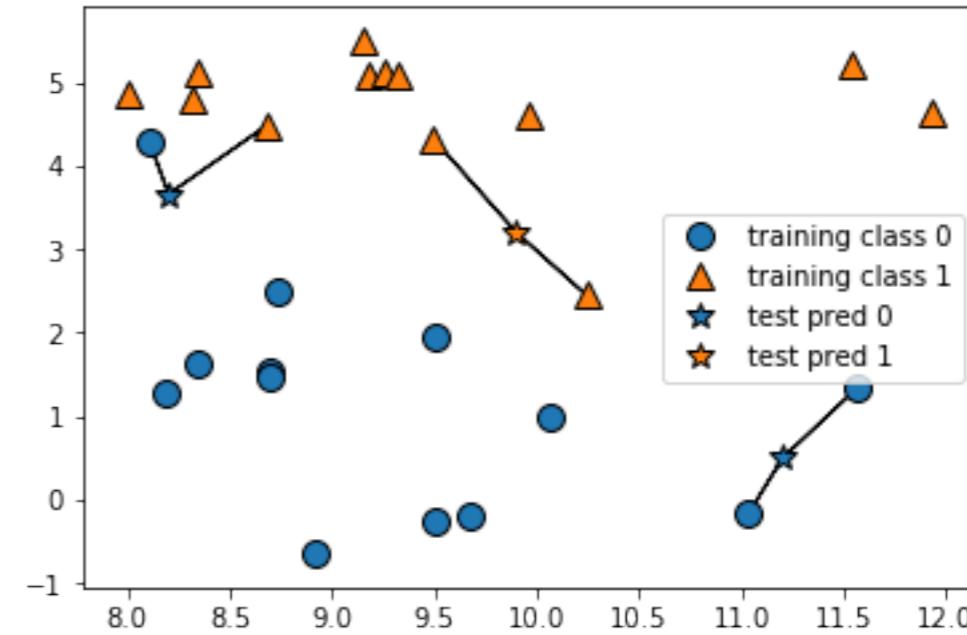
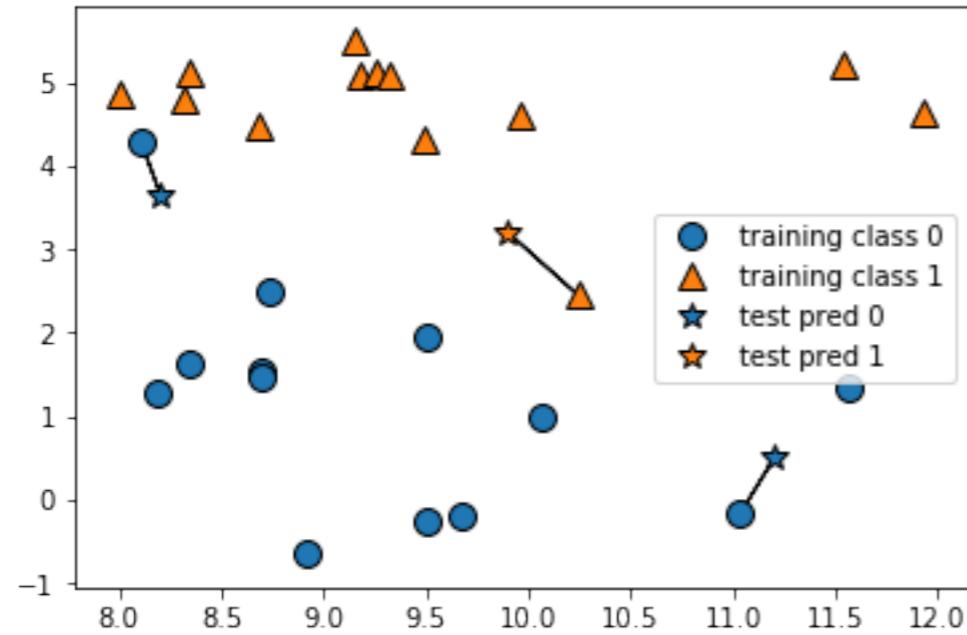


Решающая граница

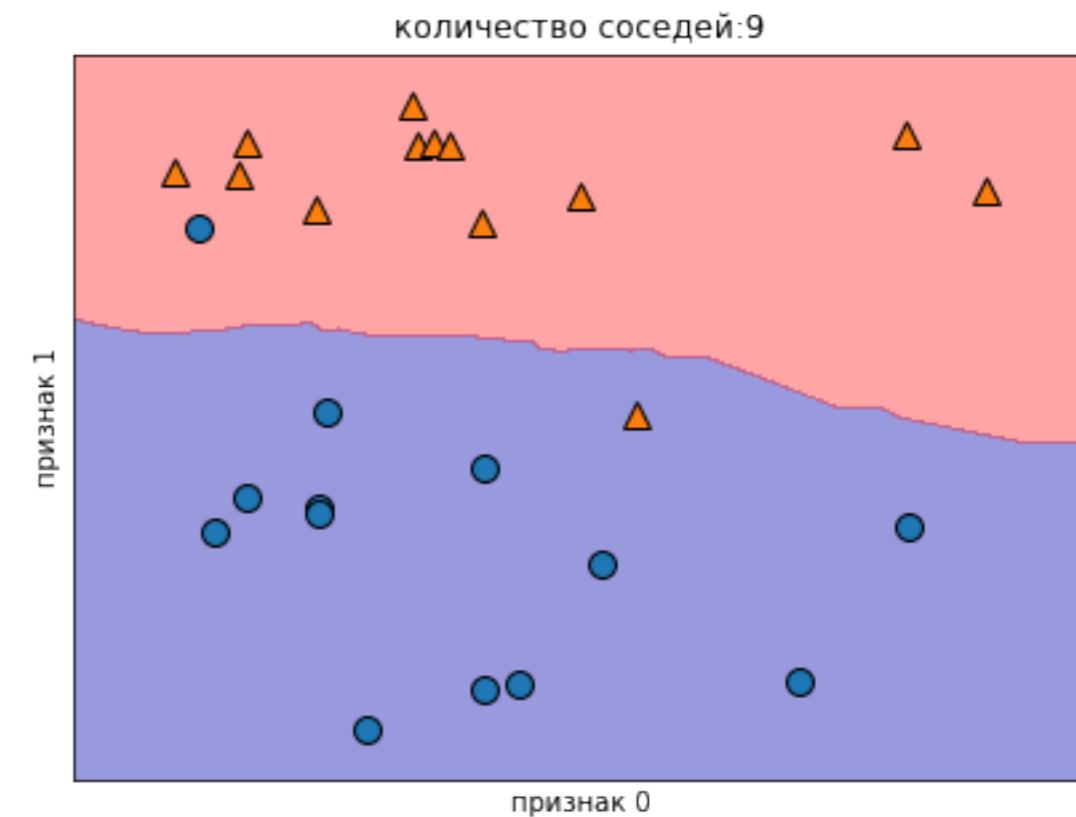
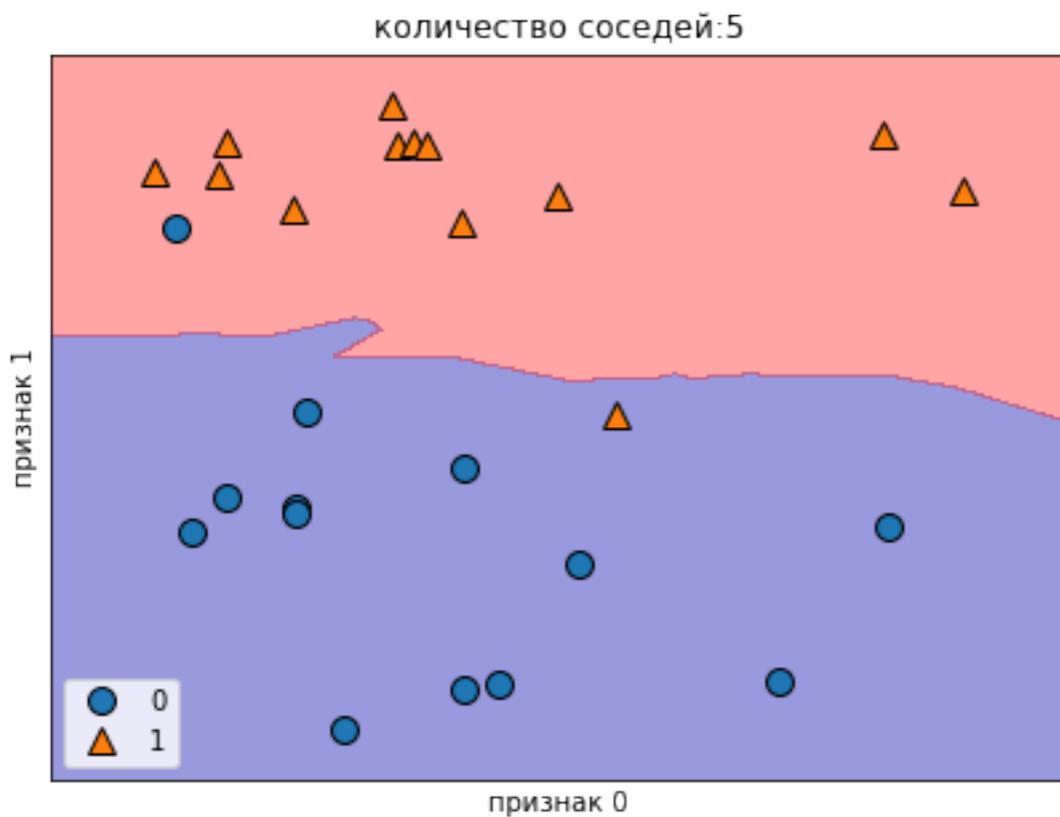
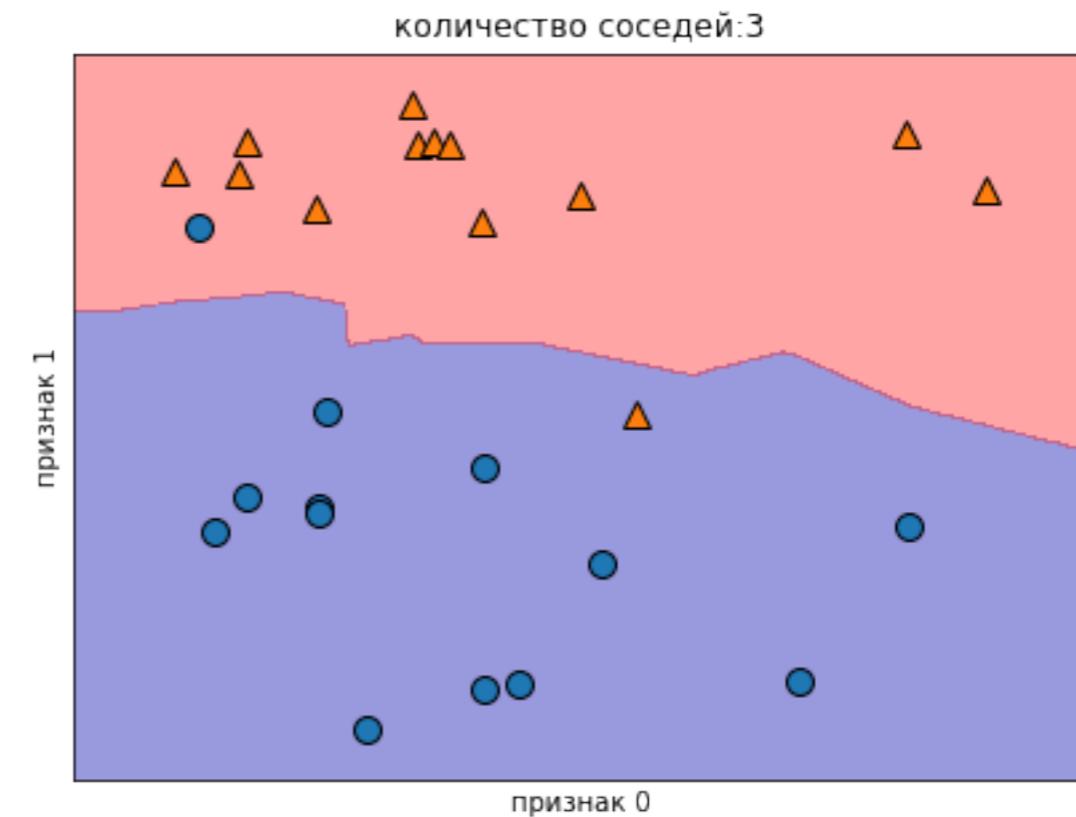
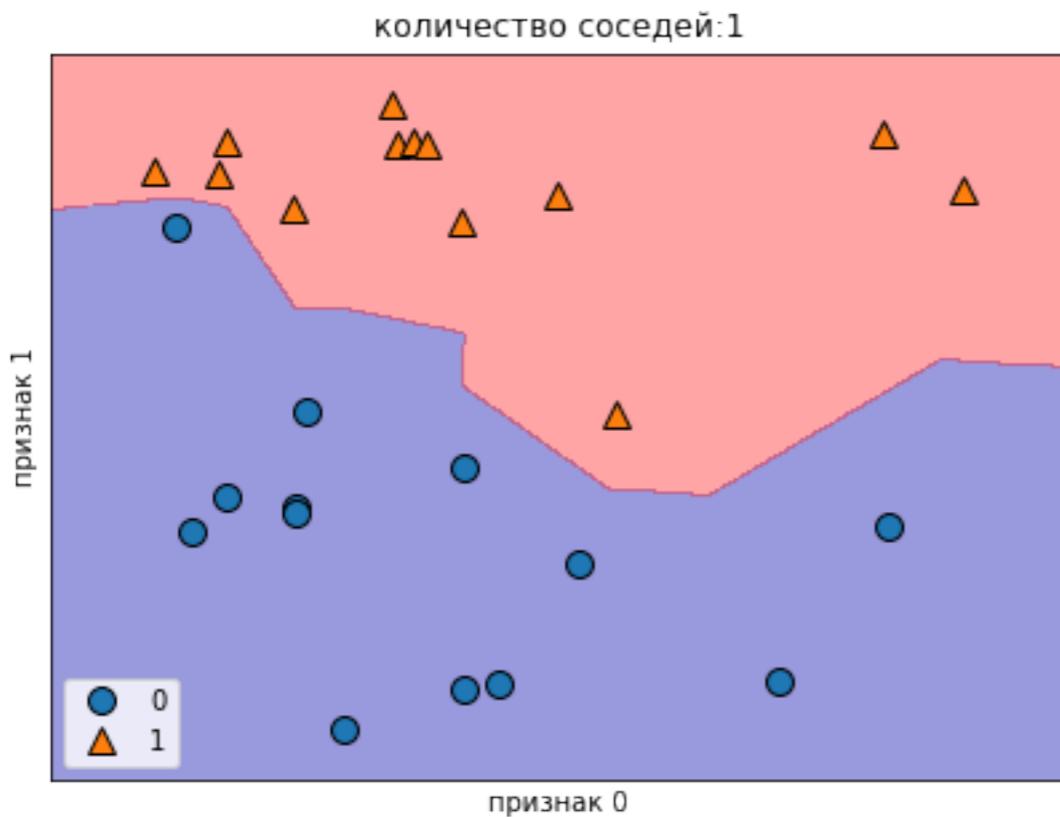


Алгоритм к ближайших соседей

Классификация

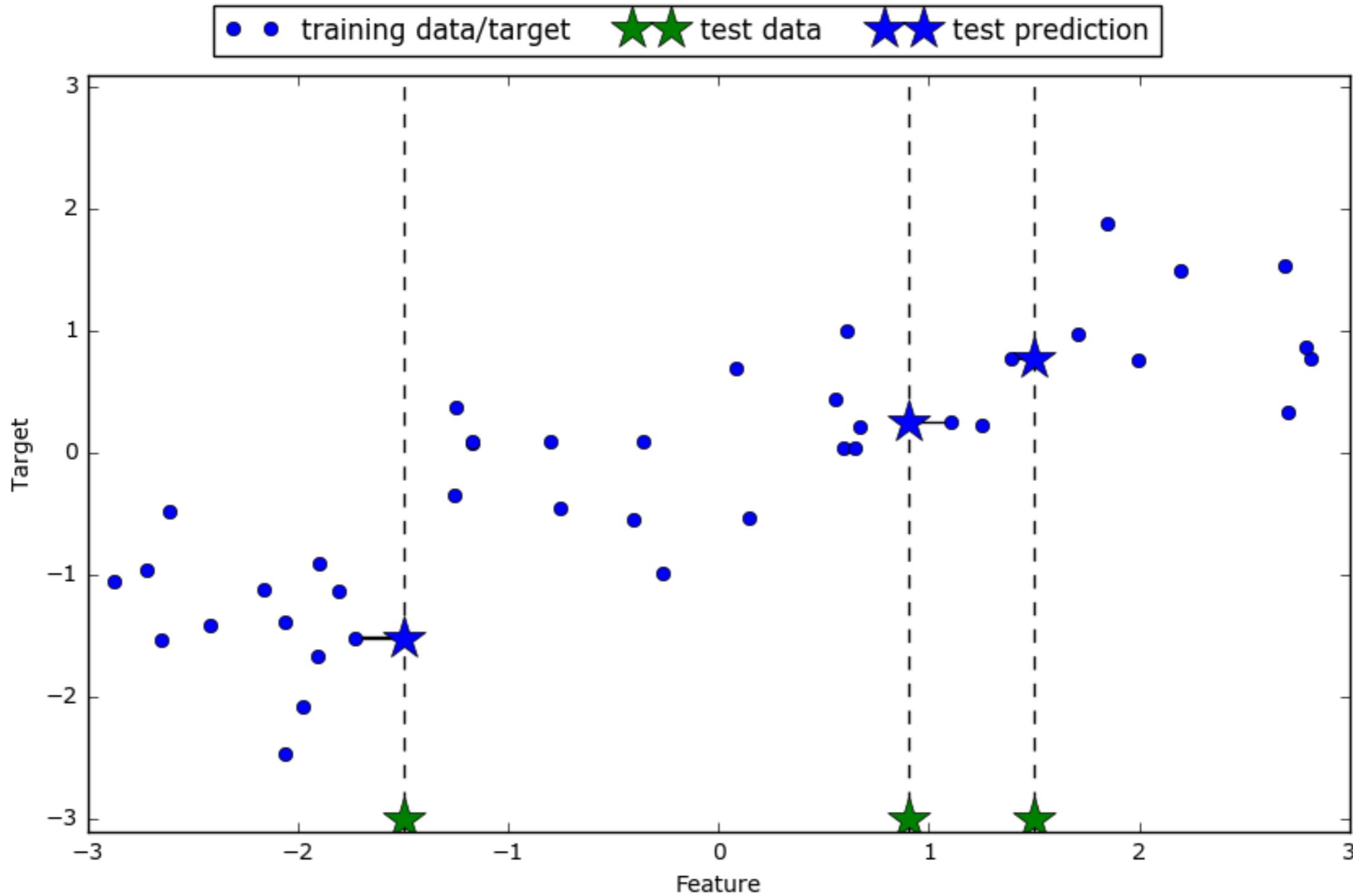


Классификация k-NN

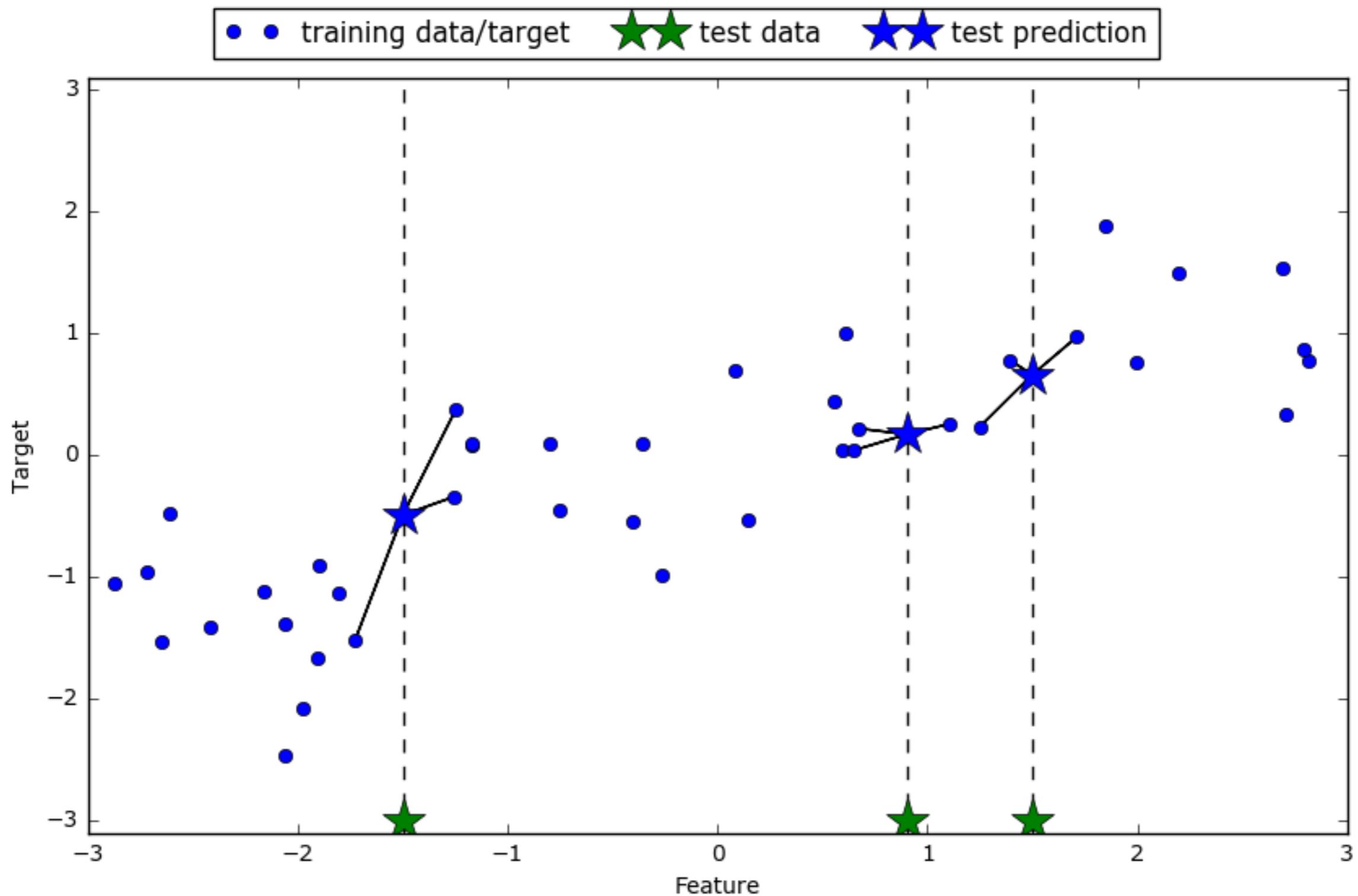


Алгоритм k ближайших соседей

Регрессия



Регрессия k-NN





Датасет (Dataset) ирисы Фишера

Набор данных «Ирисы Фишера» состоит из данных о 150 экземплярах ириса, по 50 экземпляров из трёх видов — Ирис щетинистый (*Iris setosa*), Ирис виргинский (*Iris virginica*) и Ирис разноцветный (*Iris versicolor*).

Для каждого экземпляра измерялись четыре характеристики (в сантиметрах):



- Длина лепестка (англ. sepal length);
- Ширина лепестка (англ. sepal width);
- Длина чашелистика (англ. petal length);
- Ширина чашелистика (англ. petal width).

На основании этого набора данных требуется построить правило классификации, определяющее вид растения по данным измерений. Это задача многоклассовой классификации, так как имеется три класса — три вида ириса.

Датасет Boston Housing

Задача, связанная с этим набором данных, заключается в том, чтобы спрогнозировать медианную стоимость домов в нескольких районах Бостона в 1970-е годы на основе такой информации, как уровень преступности, близость к Charles River, удаленность от радиальных магистралей и т.д. Набор данных содержит 506 точек данных и 13 признаков.

CRIM - уровень преступности на душу населения

ZN - доля жилой застройки для участков более 25 000 кв. футов

INDUS - доля не розничной торговли в районе

CHAS - вид на реку Charles River (1 - есть вид, 0 - нет вида)

NOX - концентрация оксидов азота (частей на 10 миллионов)

RM - среднее количество комнат в доме

AGE - доля построек на участке, построенных до 1940 г.

DIS - взвешенное расстояние до пяти бостонских центров занятости

RAD - индекс доступности радиальных магистралей

TAX - налоговая ставка на имущество на 10 000\$

PTRATIO - ученики/учителя

B - $1000(Bk-0.63)^2$, Bk - доля чернокожего населения в районе

LSAT - % населения с низким уровнем доходов

предсказываем MEDV - среднее стоимость домов в 1000\$



Алгоритм k ближайших соседей

Важные параметры:

1. количество соседей (n_neighbors)
2. мера расстояния между точками данных (metric) см. `sklearn.neighbors.DistanceMetric`

Достоинства:

1. Самый легко объясняемый алгоритм
2. Хорошо работает с небольшим количеством признаков
3. Необходимо предварительная обработка (проблема разнородных данных)

Недостатки:

1. Плохо работает при 100 и более признаков
2. Особенно плохо работает, если подавляющее большинство признаков равно нулю (проблема разряженных данных)

Вывод: несмотря на то что алгоритм ближайших соседей легко интерпретировать, на практике он не часто используется из-за скорости вычислений и его неспособности обрабатывать большое количество признаков.