



ПРЕЗИДЕНТСКАЯ
АКАДЕМИЯ

ВВЕДЕНИЕ В МАШИННОЕ ОБУЧЕНИЕ

О ПРЕДМЕТЕ И
БЛИЖАЙШИХ СОСЕДЯХ

Москва, 2024

Материалы занятия

<https://github.com/kshilin/machine-learning>

РАНХиГС

2024

Что такое машинное обучение?



Как люди видят
мою работу

$$\ell_0 = \|\theta\|_2^2 + \sum_{i=1}^n \log(1 + e^{-\theta^\top x_i}) + \sum_{j=1}^m \lambda_j$$

$$\theta_j \in \mathbb{R}, \forall j$$

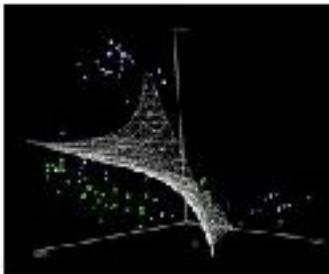
$$\theta = \sum_{i=1}^n x_i / n, \sum_{i=1}^n x_i = 0$$

$$V(\theta_k) = \frac{1}{n} \sum_{i=1}^n V(x_i, \theta_k; \theta_0) + V(x_i; \theta_0)$$

$$\nabla_{\theta_k} V(\theta_k) = \frac{1}{n} \nabla V(x_i, \theta_k; \theta_0) = (\theta - \nabla V(\theta_0))$$

$$E_{\theta_0}[V(x_i, \theta_k; \theta_0)] = \frac{1}{n} \sum_{i=1}^n E_{\theta_0}[x_i, \theta_k; \theta_0]$$


Как мои друзья
видят мою работу



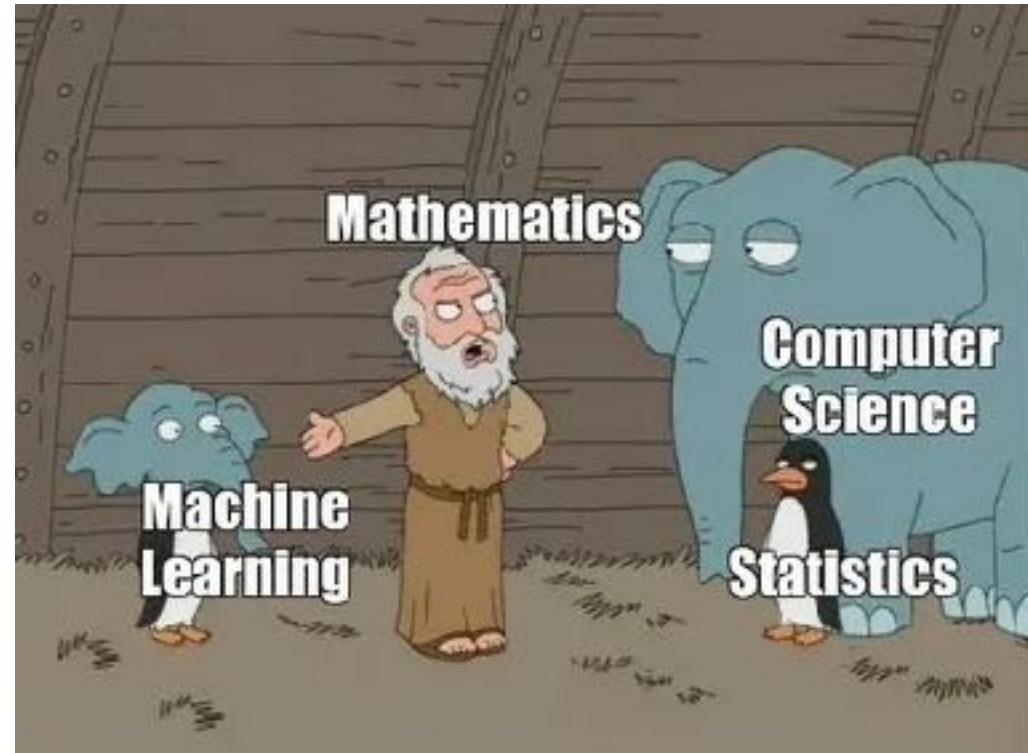
Как программисты
видят мою работу



Как мои родители
видят мою работу

```
>>> from sklearn import svm
```

Что я на самом
деле делаю



Что такое машинное обучение?



Какие пререквизиты?

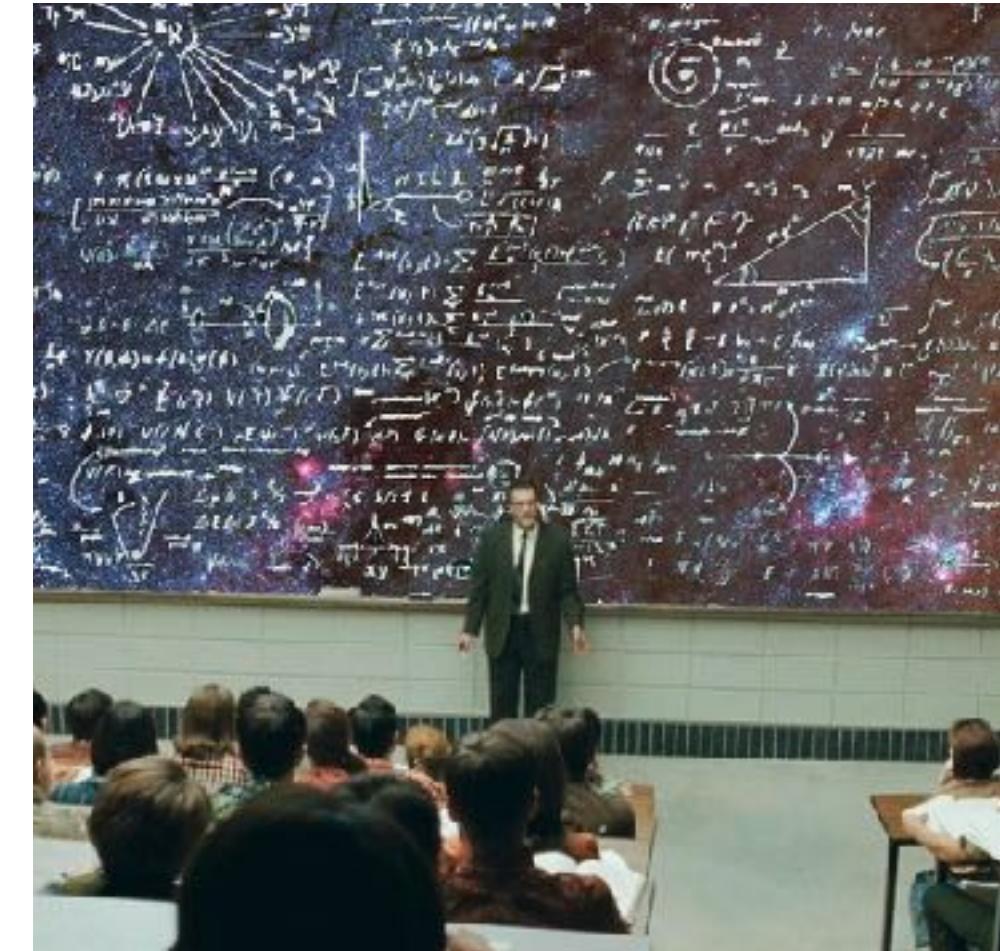
НЕОБХОДИМО

- Математический анализ
минимум 2 семестра
- Алгебра
минимум 1 семестра
- Теория вероятностей и
математическая статистика
минимум 1 семестр

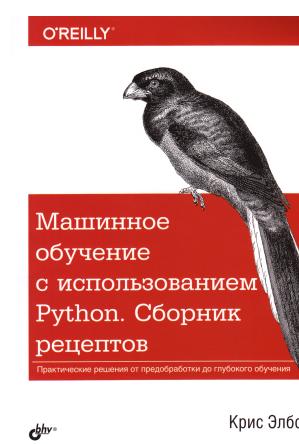
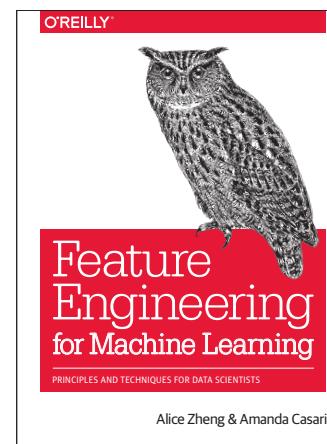
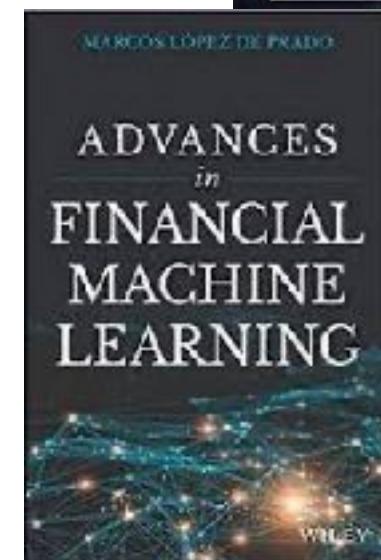
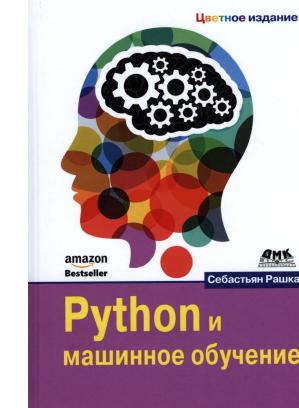
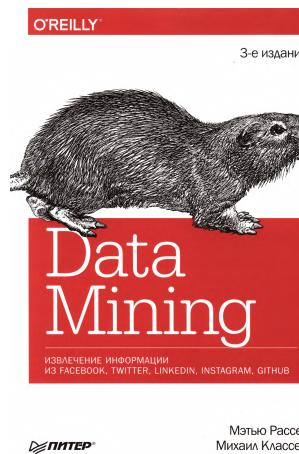
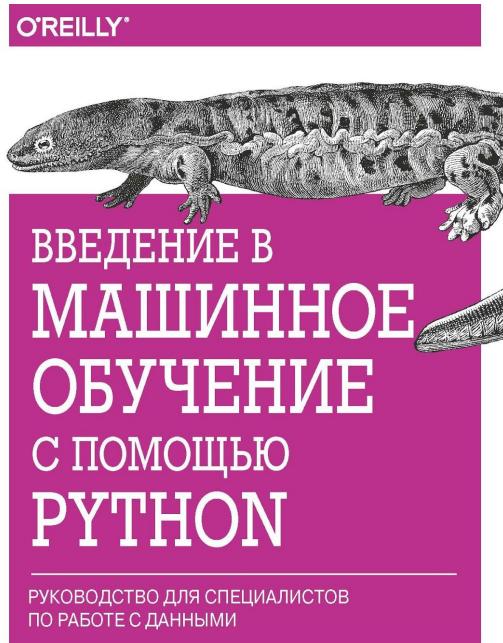
ЖЕЛАТЕЛЬНО

- Методы оптимизации
1 семестр
- Анализ данных (на Python)
1 семестр
- Программирование (на Python или R)
2 семестра

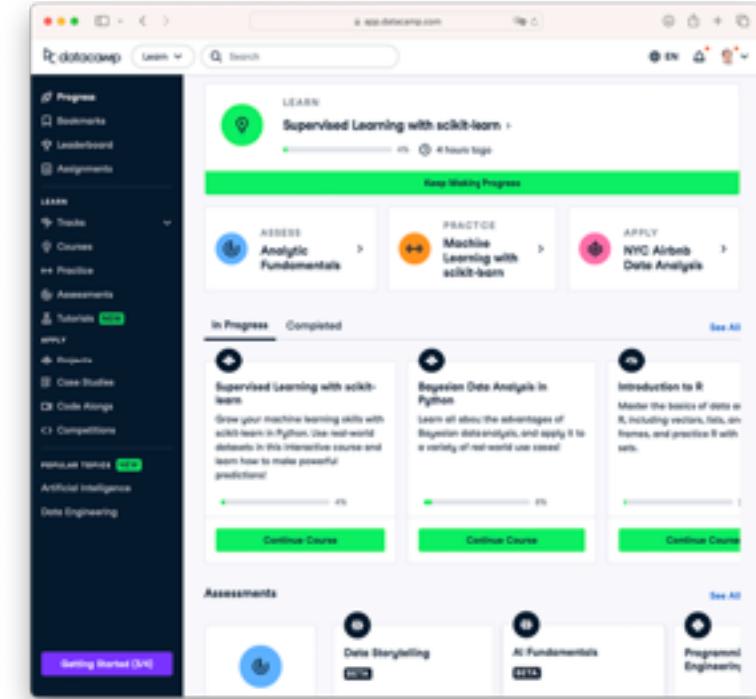
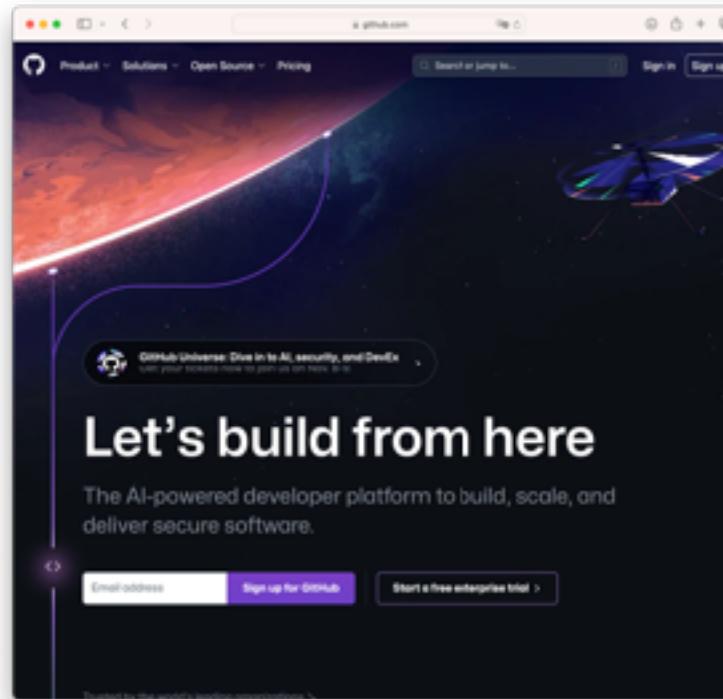
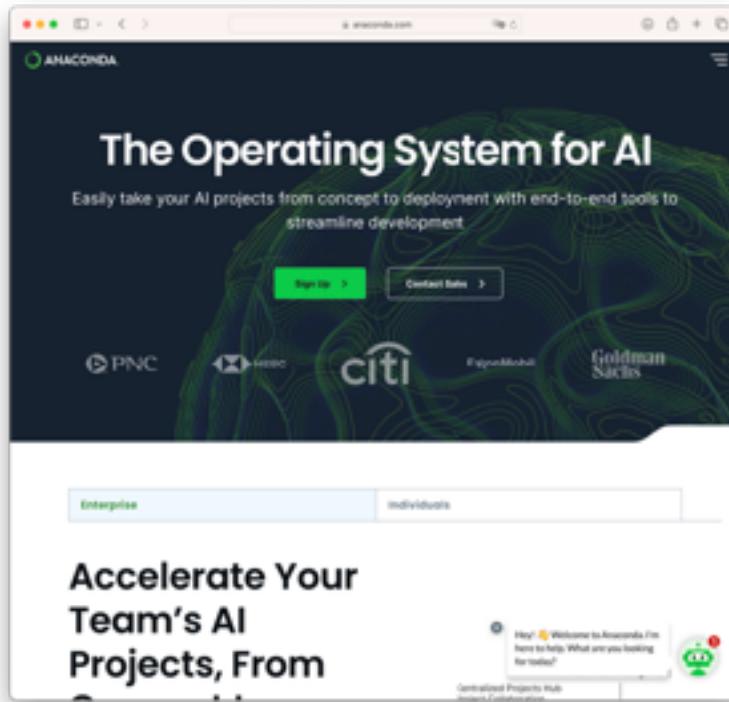
При минимуме входных знаний учится можно, но очень сложно



Что читать в процессе обучения?



В какой программной среде мы будем работать?



Какие алгоритмы изучим?

Обучение с учителем

Классификация и регрессия

- Ближайшие соседи
- Линейные модели
- Наивные байесовские модели
- Машины опорных векторов
- Деревья решений
- Случайные леса
- Градиентный бустинг на деревьях решений

Обучение без учителя

Кластеризация

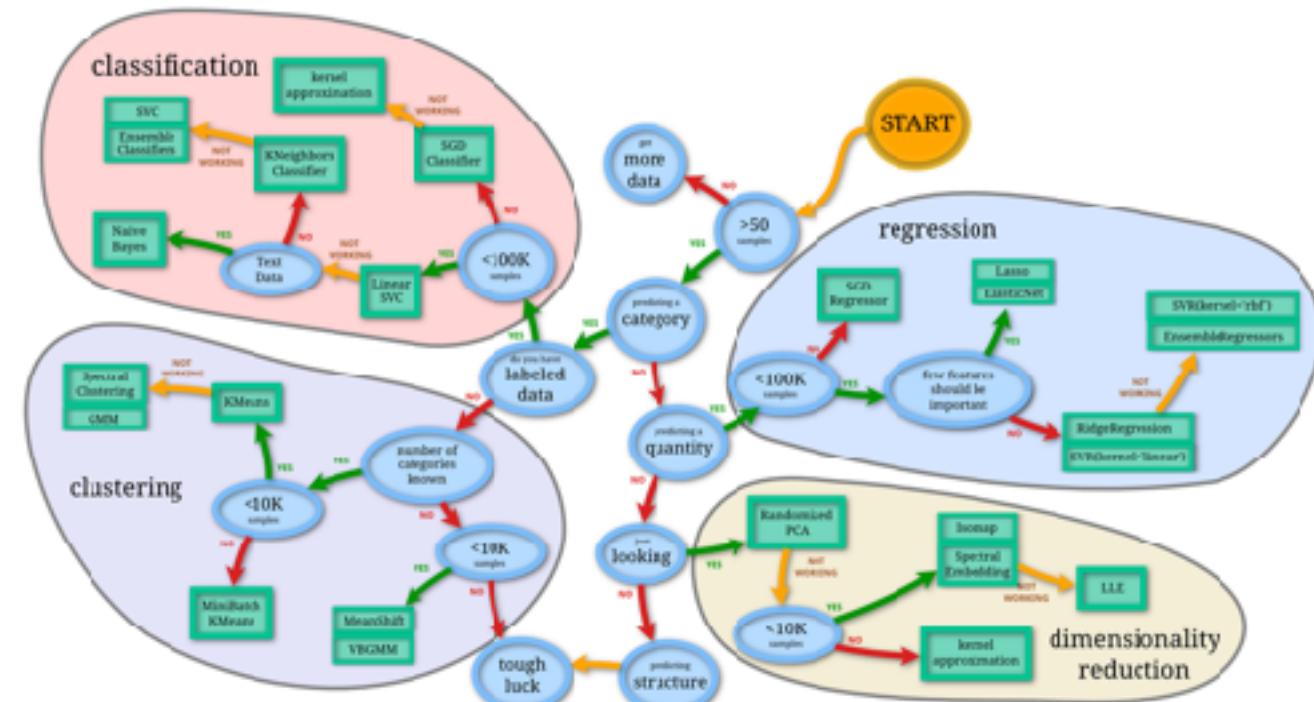
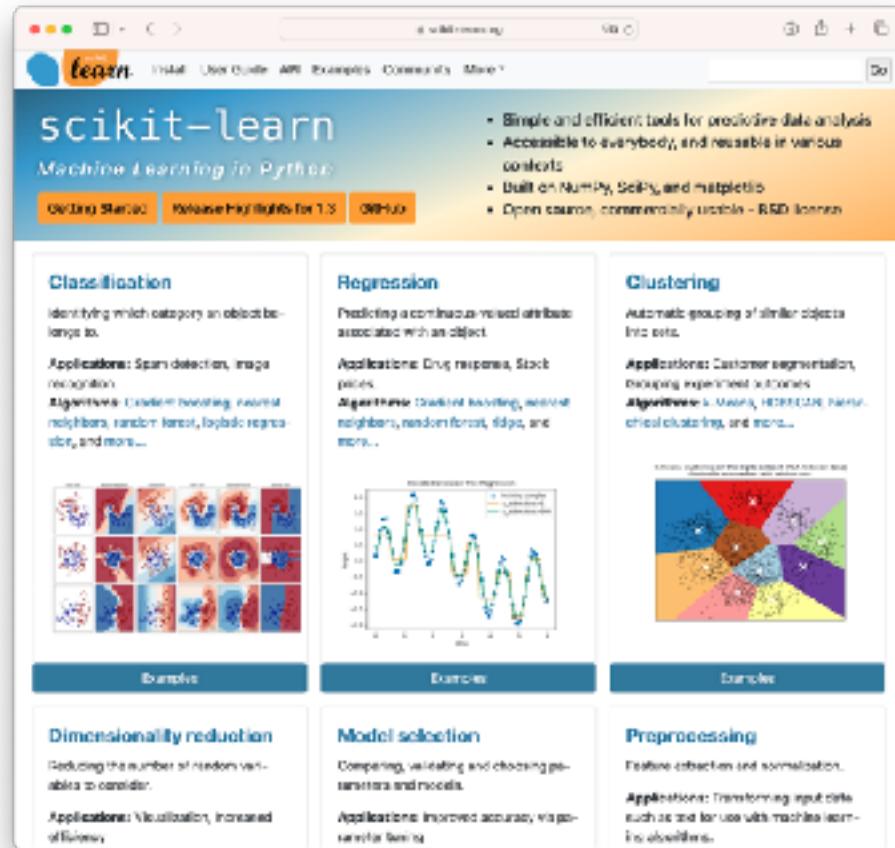
- Преобразование данных:
 - стандартизация
 - нормализация
 - обработка категориальных данных
 - анализ главных компонент
 - t-SNE
- Кластеризация:
 - k-средних
 - агломеративная
 - DBSCAN
- Поиск аномалий в данных

Дополнительно

Контейнеры

- Конструирование пайплайнов
- Отбор наиболее значимых признаков
- Трактовка значимости с использованием векторов Шепли
- Собственные метрики качества
- Регрессия и кластеризация временных рядов

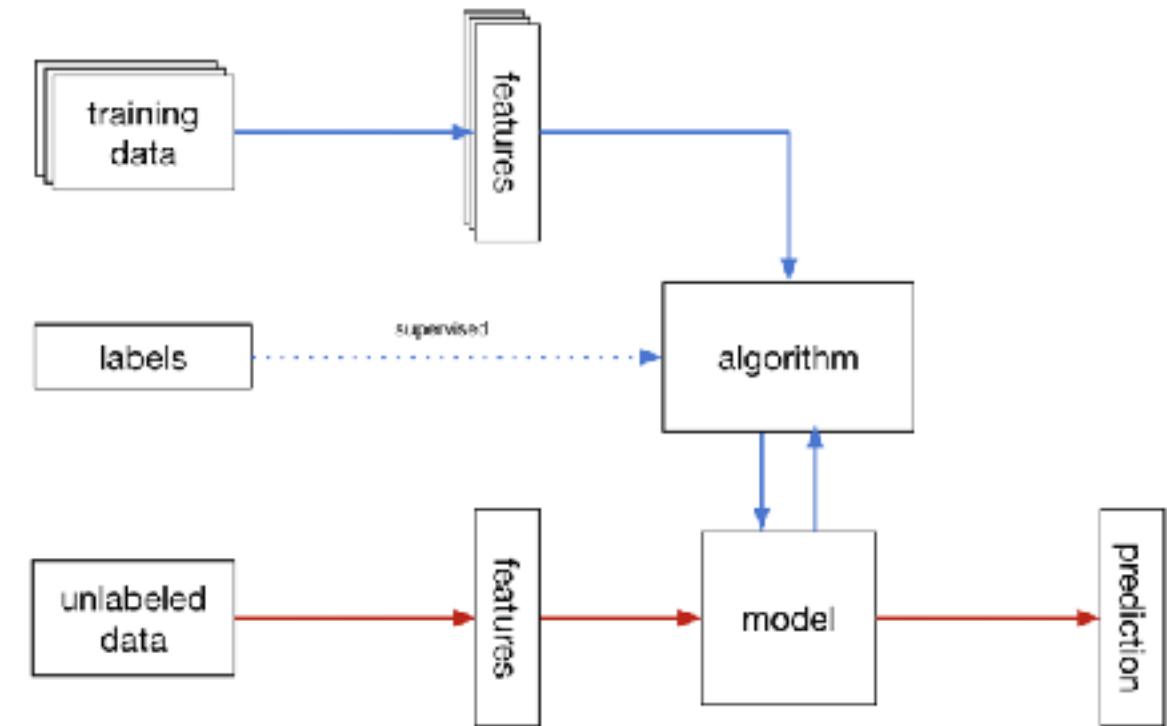
С каким фреймворком мы научимся работать?



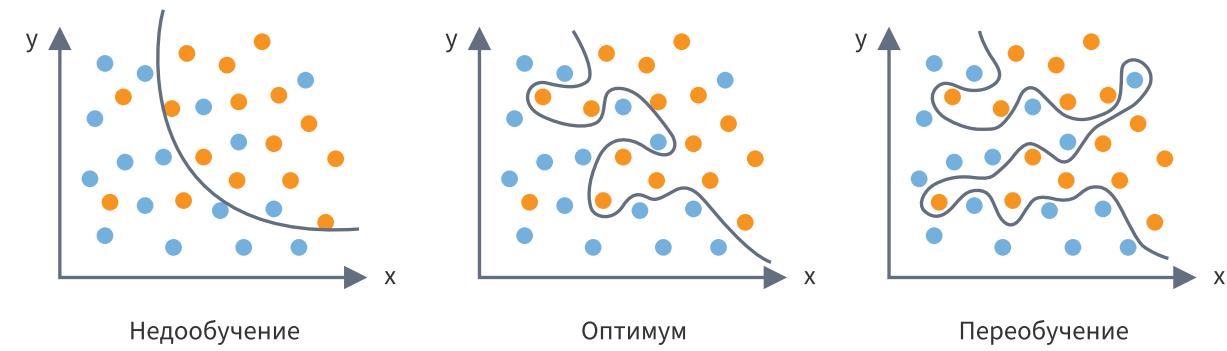
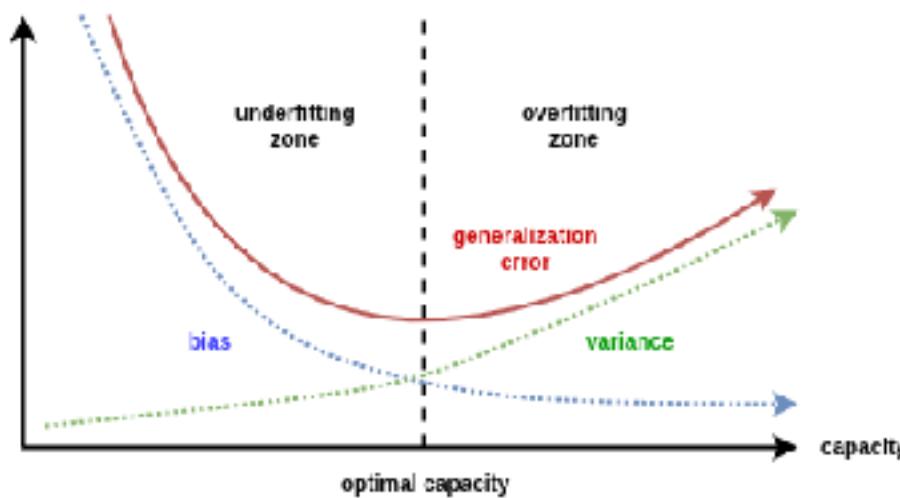
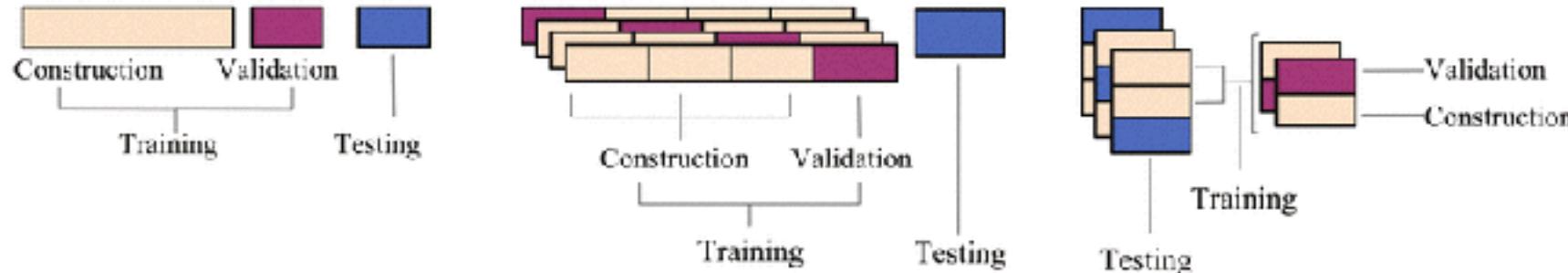
Как строится разработка модели ML

Обучение с учителем:

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



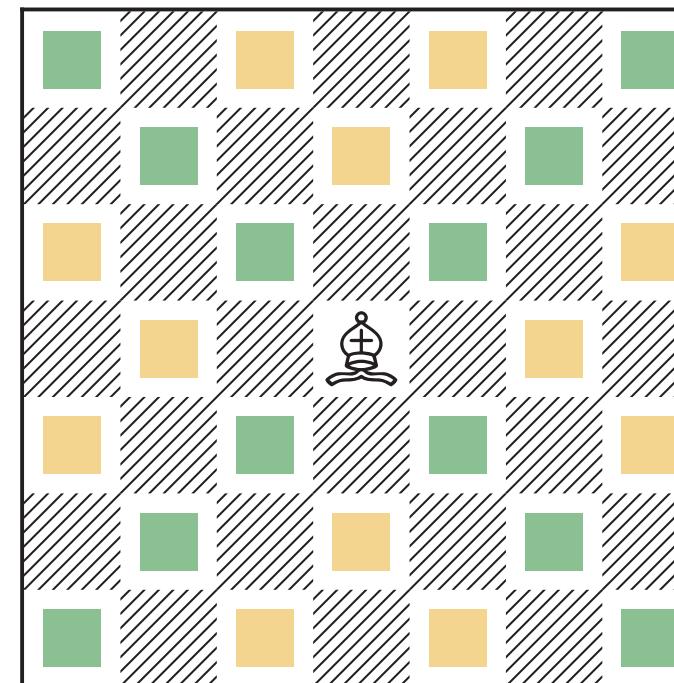
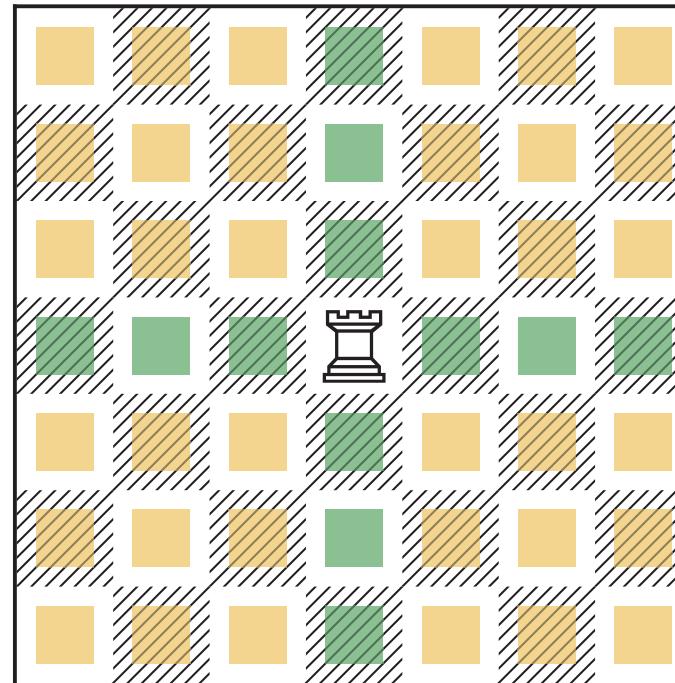
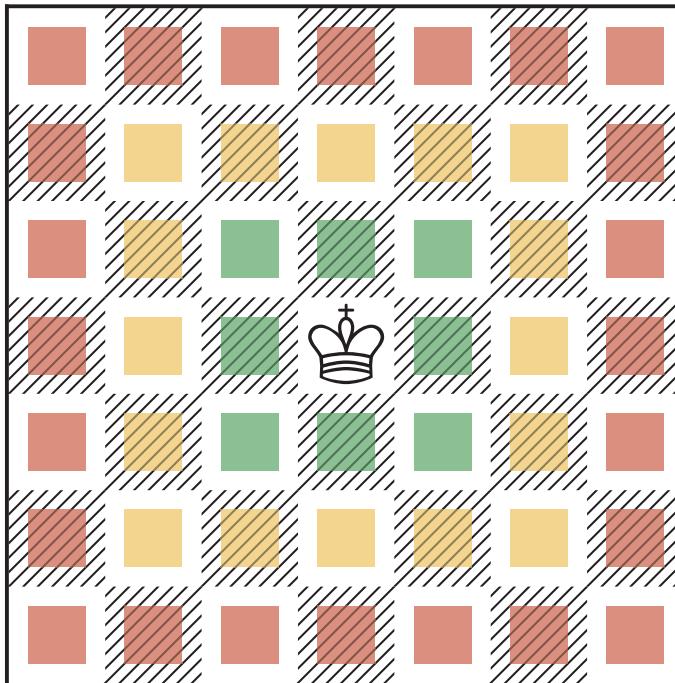
Как проверить качество алгоритма





Пятиминутка математики ...

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.

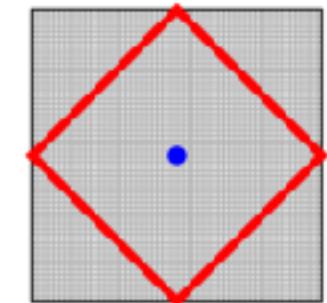
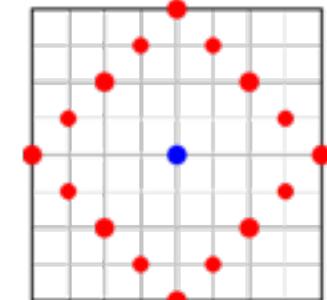
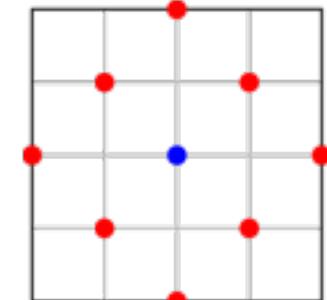
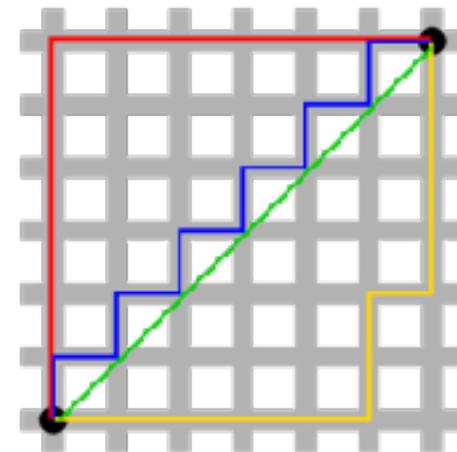
Расстояние Минковского порядка p между двумя точками определяется как:

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

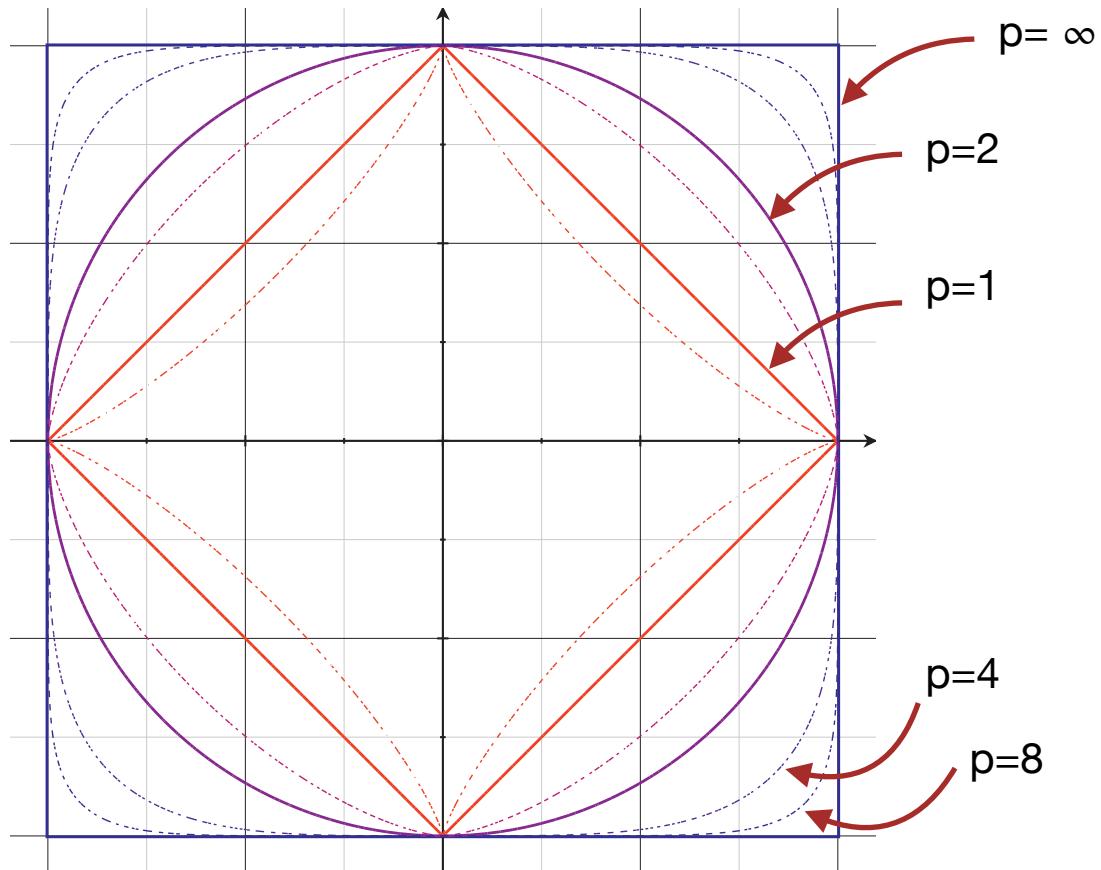
Для $p \geq 1$ расстояние Минковского является метрикой вследствие неравенства Минковского.

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

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

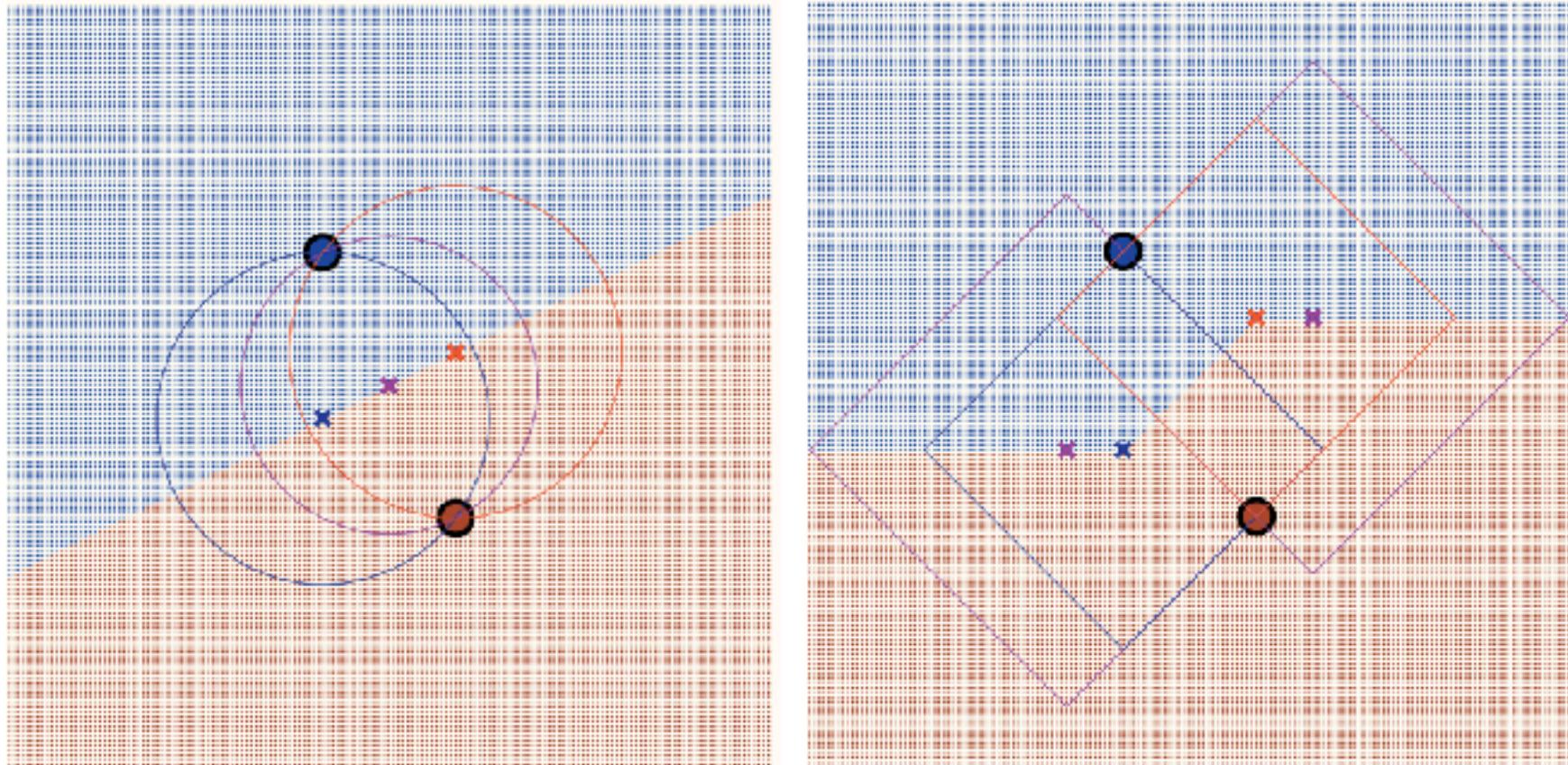


Расстояние Миньковского

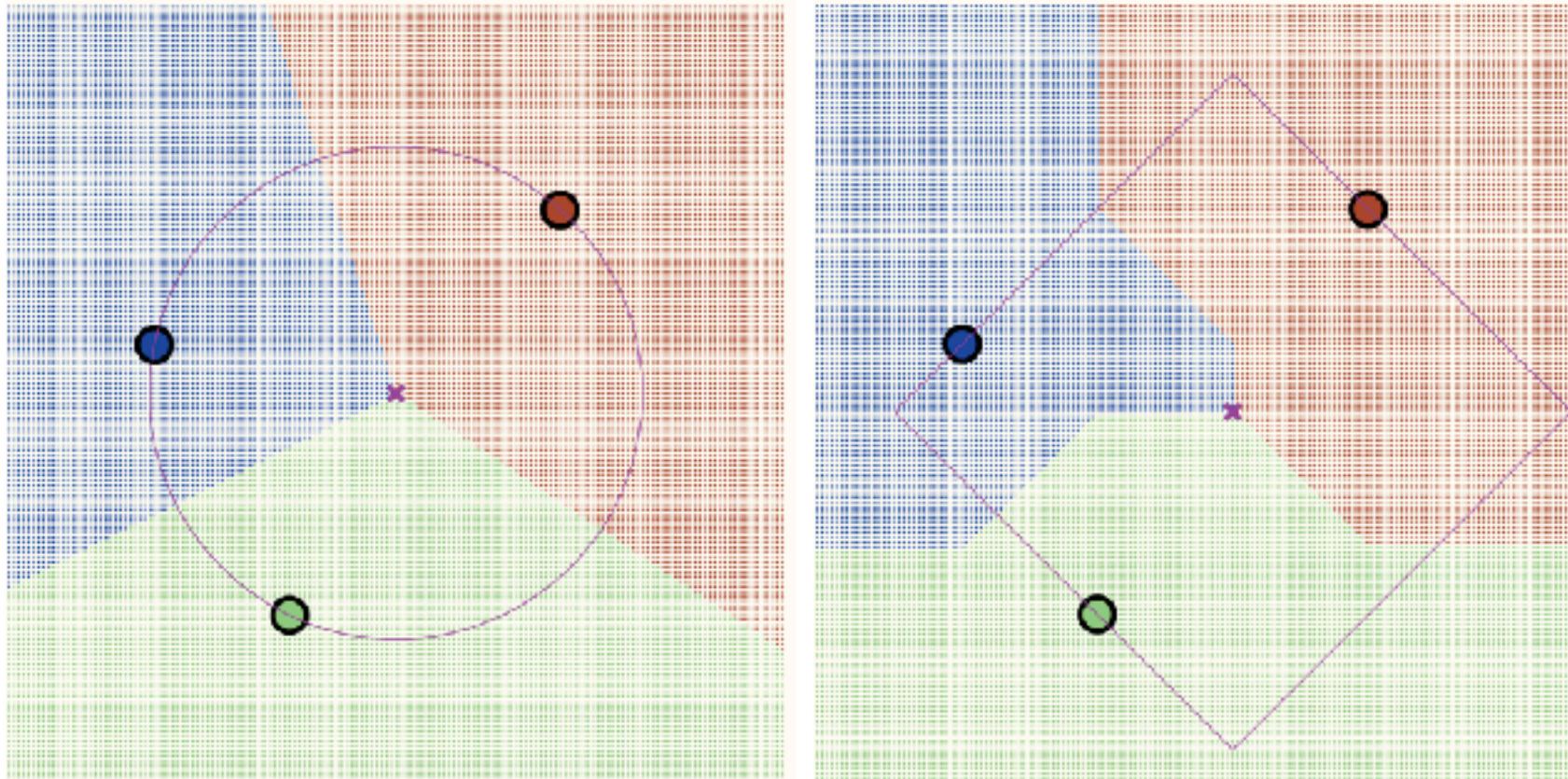


$$\rho(x, y) = \left(\sum_{i=1}^n |x_i - y_i|^p \right)^{1/p}.$$

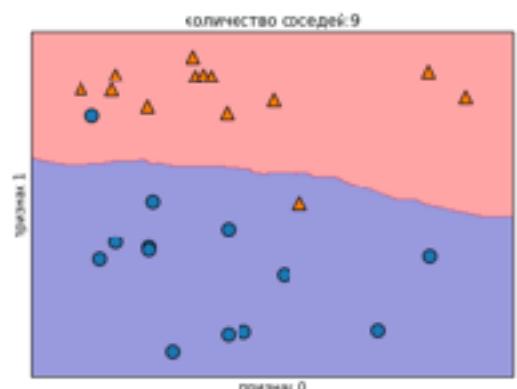
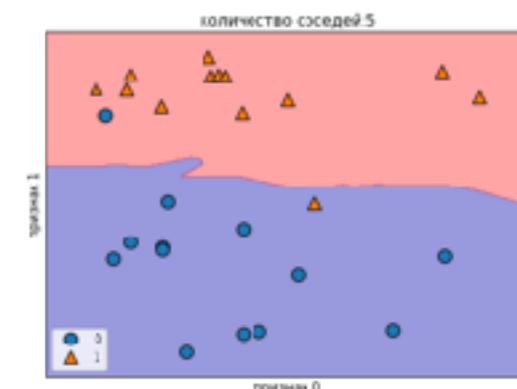
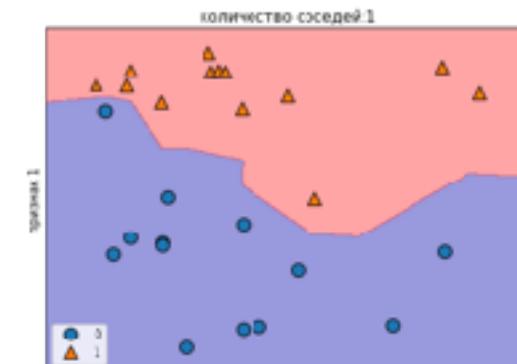
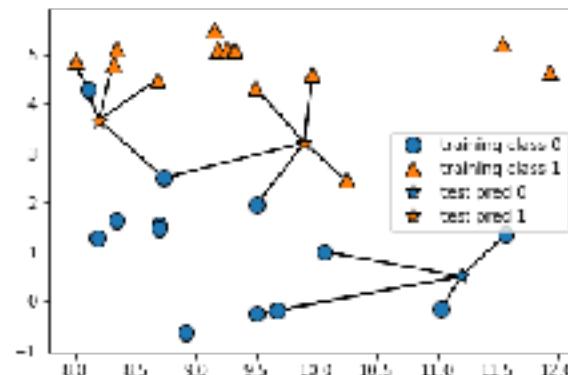
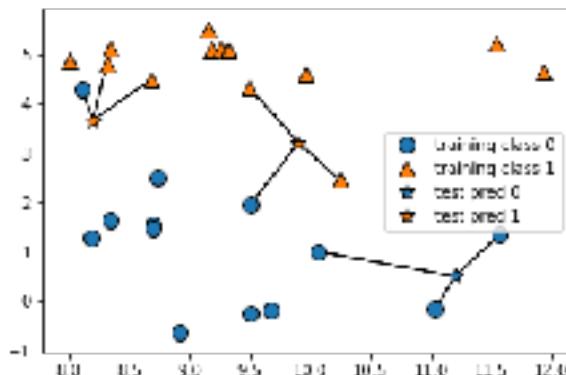
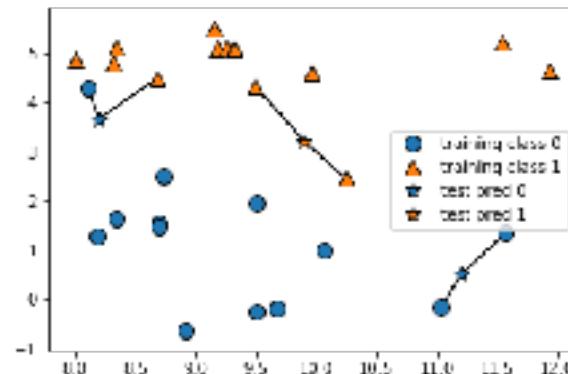
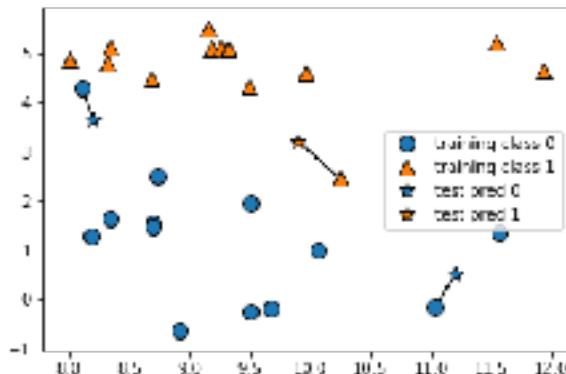
Решающая граница (2 класса)



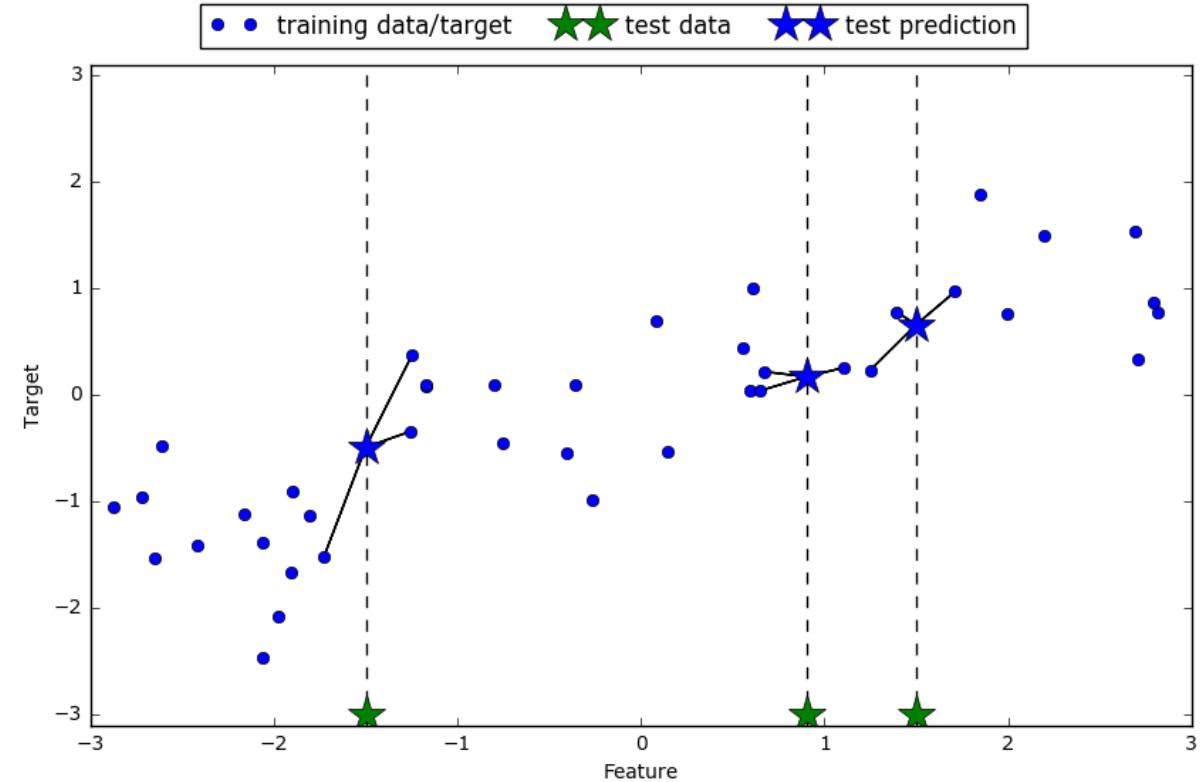
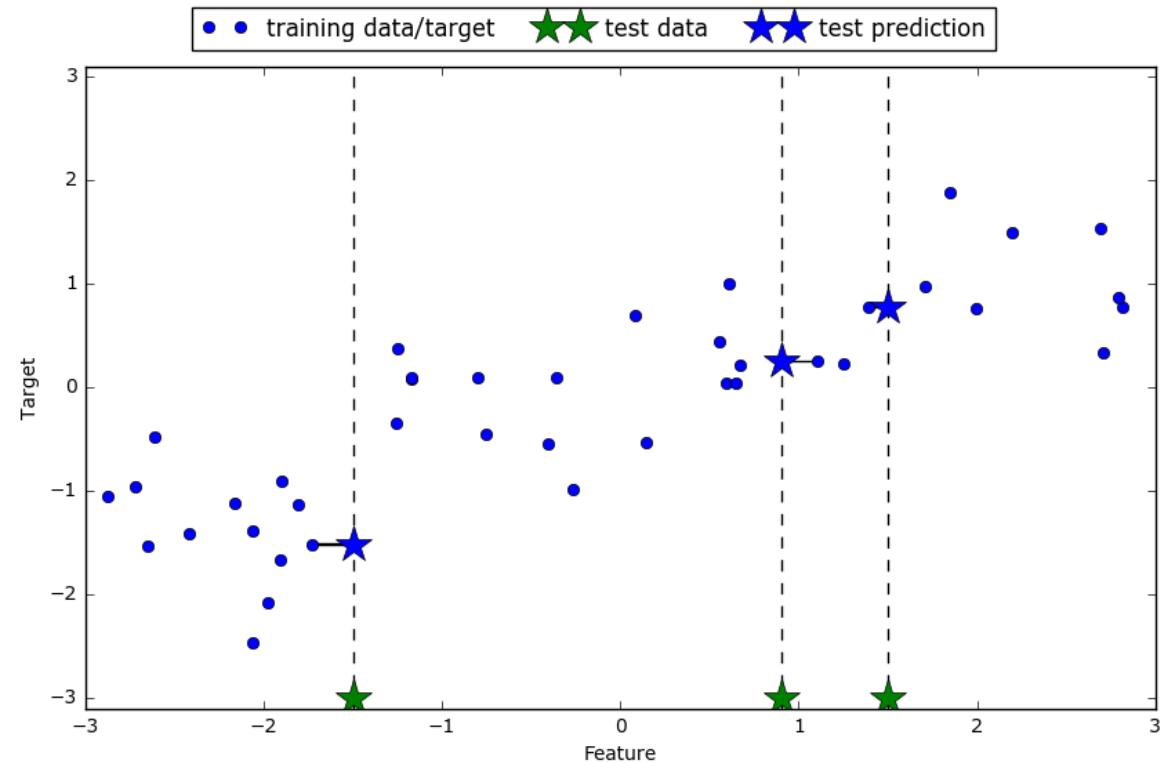
Решающая граница (3 класса)



Классификация к ближайших соседей



Регрессия к ближайших соседей



Ирисы Фишера

Набор данных «Ирисы Фишера» состоит из данных о 150 экземплярах ириса, по 50 экземпляров из трёх видов:

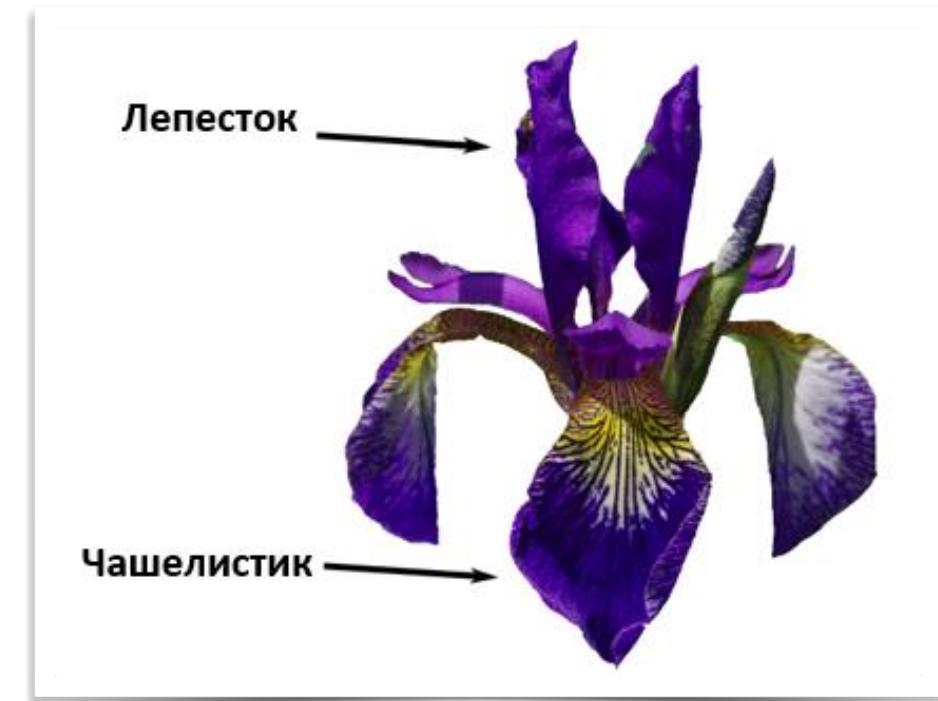
- Ирис щетинистый (*Iris setosa*),
- Ирис виргинский (*Iris virginica*)
- Ирис разноцветный (*Iris versicolor*).

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.





ПРЕЗИДЕНТСКАЯ
АКАДЕМИЯ

СПАСИБО ЗА ВНИМАНИЕ!

Москва

РАНХиГС

2024