

Лекция 10

Введение в ML

План занятия

1. Обзор ML
2. Данные
 - a. Как выглядят данные
 - b. Признаковое описание
 - c. Многомерные пространства и тензоры больших рангов
3. Модель
 - a. Обучение с учителем
 - b. Классификация и регрессия
 - c. Пример - KNN
4. Измерение качества
 - a. Примеры метрик
 - b. Разделение на тестовую и тренировочную выборки
 - c. Кросс-валидация

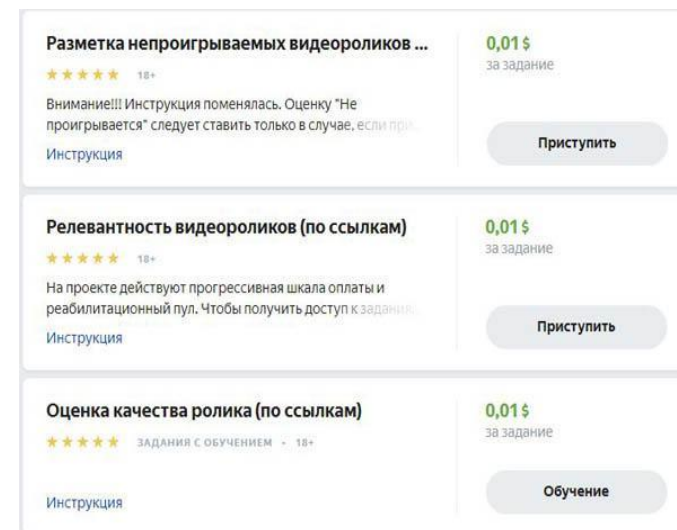
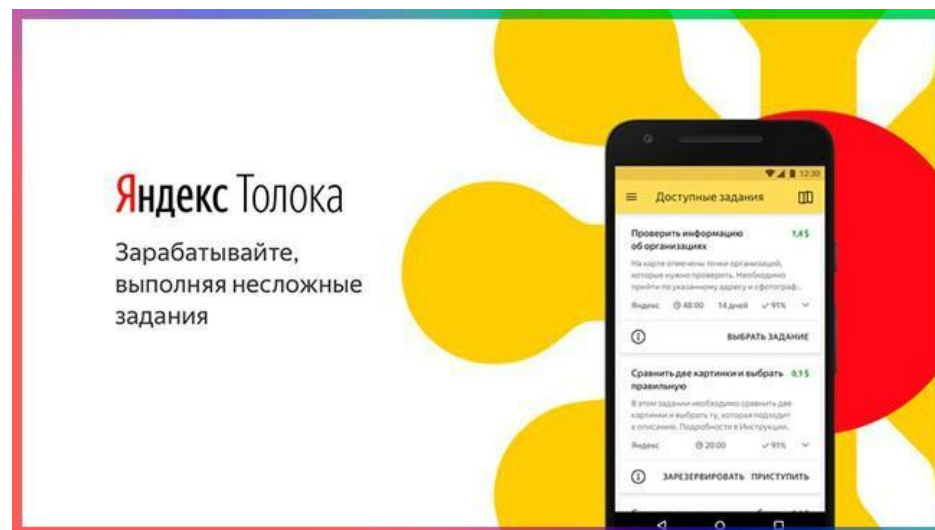
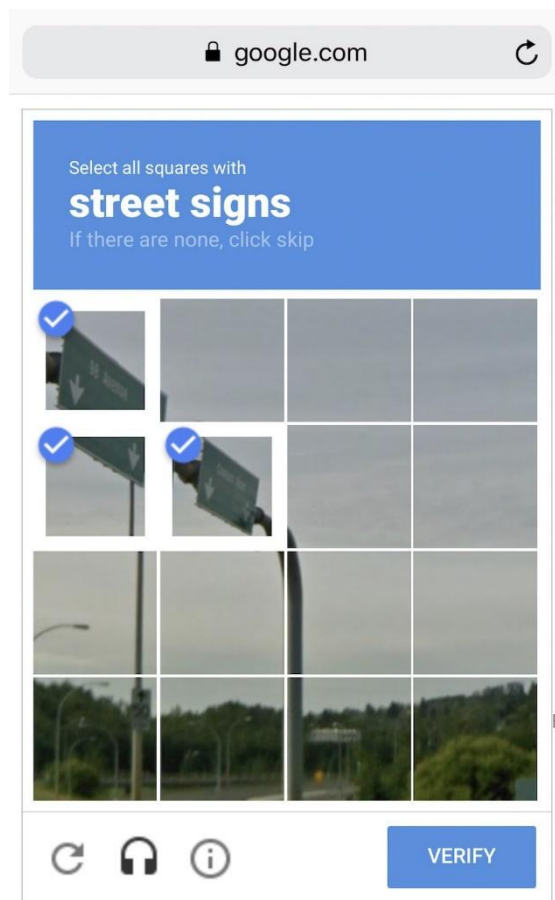
Полезные ссылки

- Введение в ML простыми словами https://vas3k.ru/blog/machine_learning/
- Метрики в задаче классификации
<https://medium.com/swlh/recall-precision-f1-roc-auc-and-everything-542aedef322b9>
- Метрики в задаче регрессии
<https://medium.com/usf-msds/choosing-the-right-metric-for-machine-learning-models-part-1-a99d7d7414e4>

Обзор ML

Как машинное обучение меняет мир

- “Nations with the strongest presence in AI R&D will establish leading positions in the automatization of the future.”
— отчет Белого дома США, октябрь 2016
- цифровая и распределенная экономика
- автоматизация и сокращение издержек
- автономный транспорт и роботизация
- автоматизация банковских услуг
- персональная медицина
- и многое другое



Google признал, что прослушивает голосовые команды пользователей



Марина Ибушева
12 Июля 2019, в 14:17



20



Шрифт:

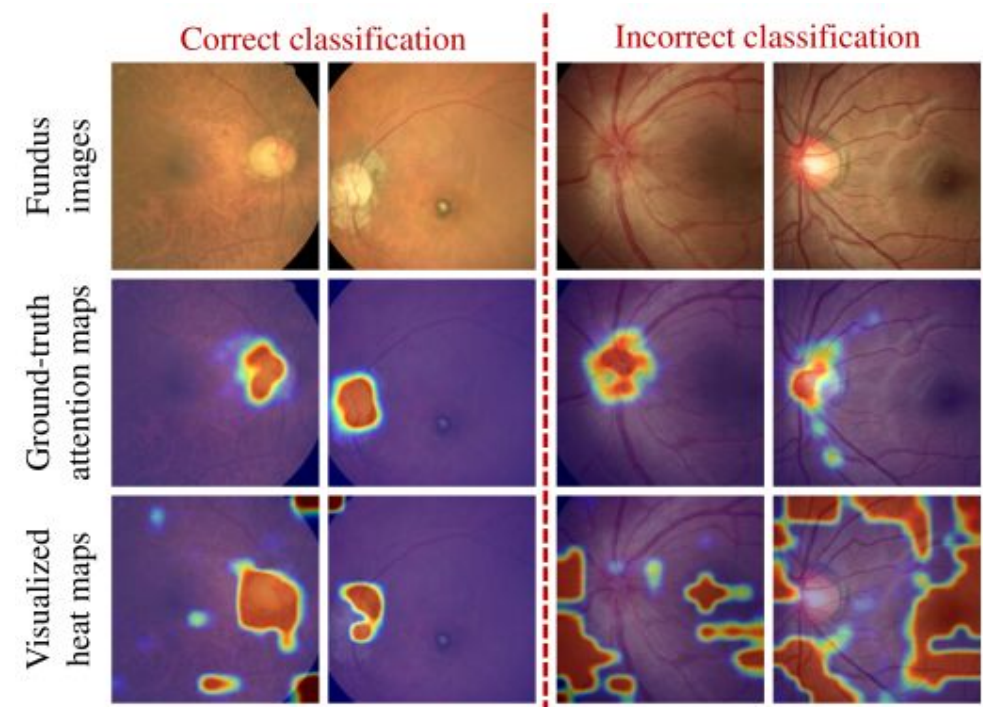


1



3934

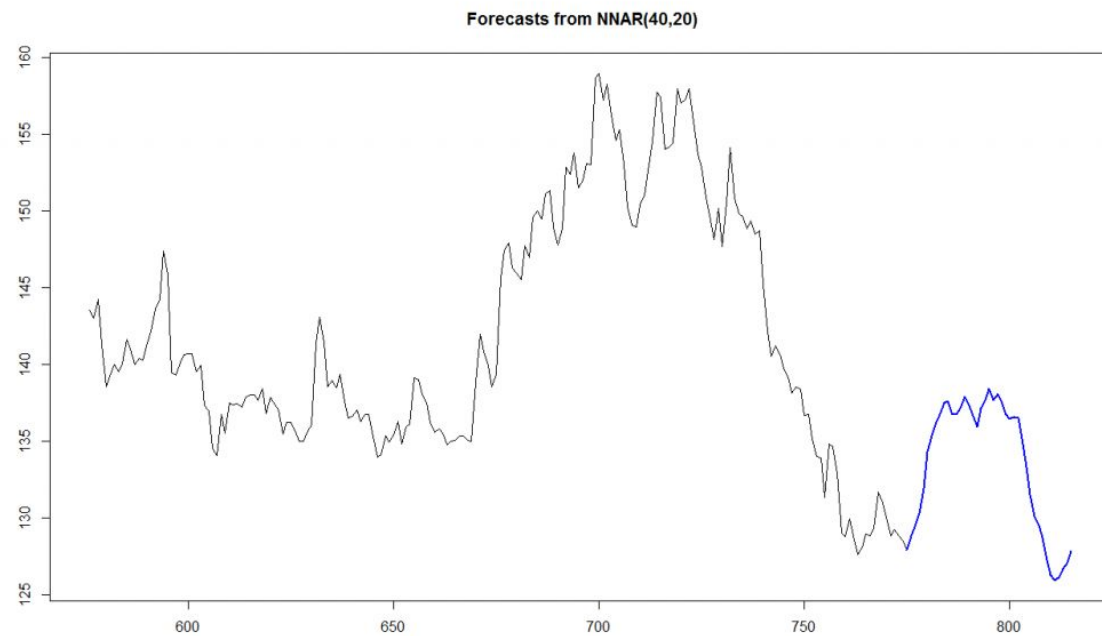
Google признался, что прослушивает команды, которые люди отдают голосовым помощникам Google Home и Google Assistant. Полученные данные собираются и обрабатываются экспертами-лингвистами, а потом используются для улучшения технологии распознавания. Доля анализируемых экспертами аудиоданных не превышает 0,2% от всех записей.




Давать ли кредит?



Дерево Решений



С ЭТИМ ТОВАРОМ ЧАСТО ПОКУПАЮТ



4.5


Браслет Xiaomi Mi Band 4

Умные часы и браслеты

от 1 689 Р

101 отзыв 252 предложения

Цвет корпуса: ●



4.5

Наушники Honor AM61


Наушники и Bluetooth-гарнитуры

от 1 848 Р

396 отзывов 59 предложений

Цвет: ● ● ● ●

Покупаем самокат



5.0

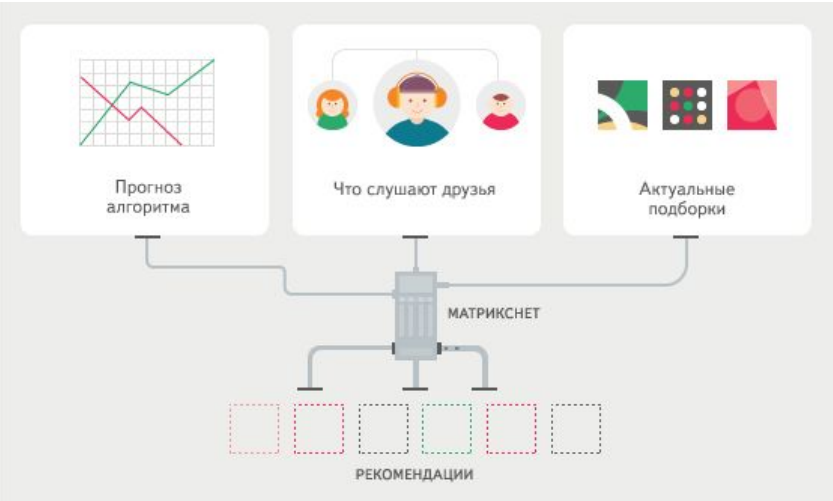
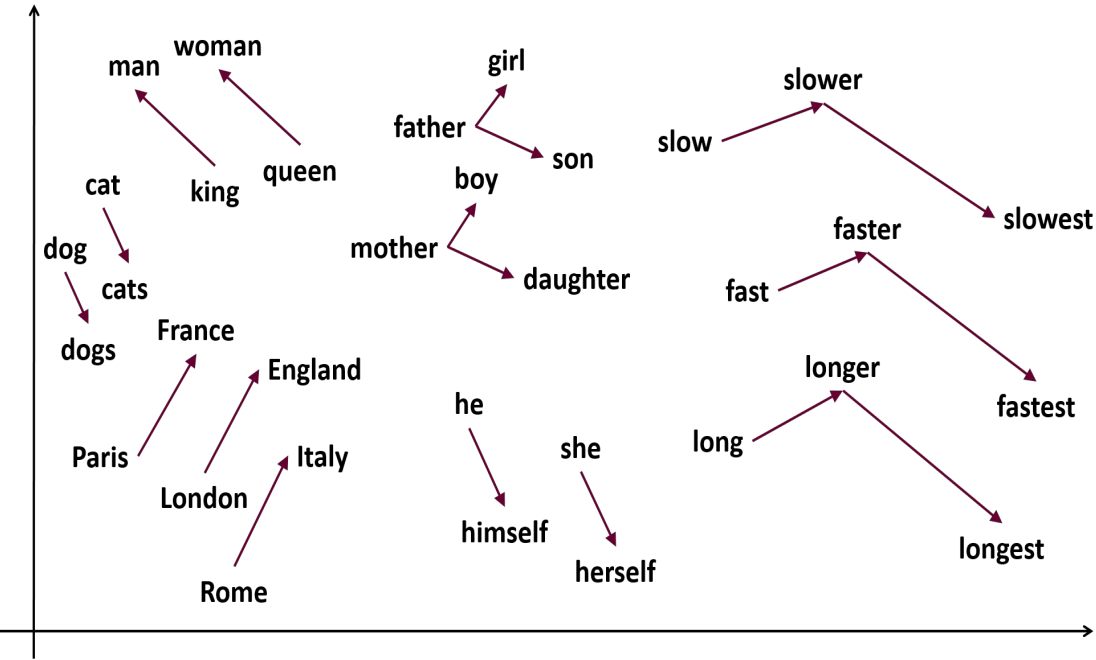
Рюкзак Xiaomi Mini 10

Рюкзаки

от 439 Р

16 отзывов 142 предложения

Цвет: ● ● ● Ещё



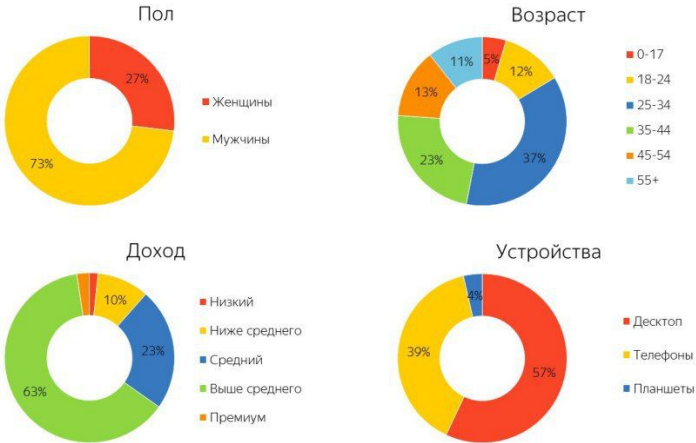
Портрет зрителей спортивных трансляций на Яндексе

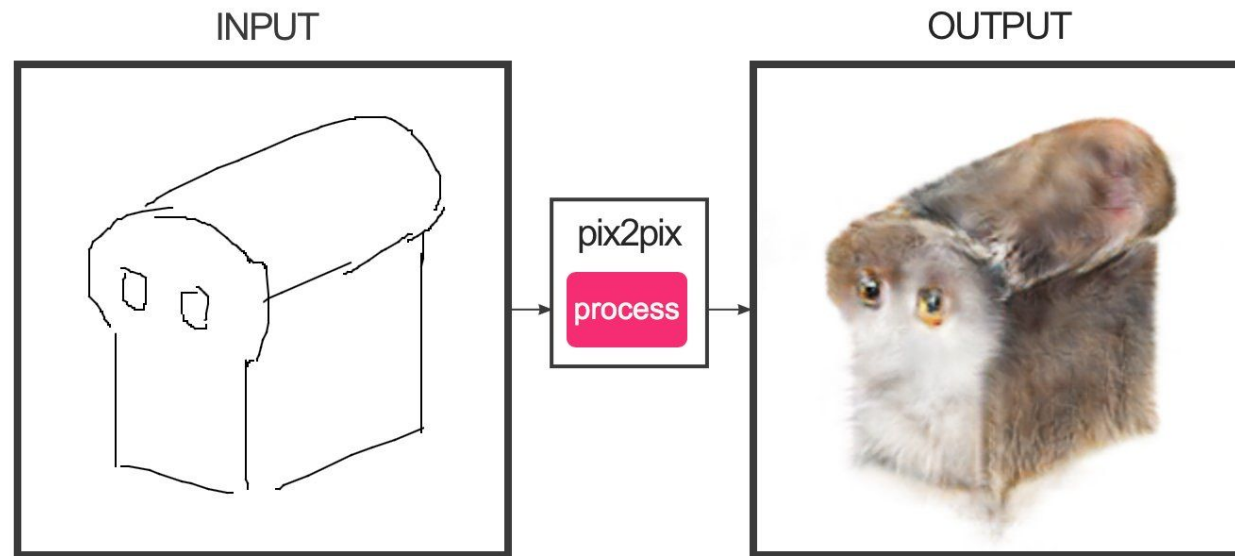
Посетители проектов в основном мужчины – 73%

Наибольшая часть пользователей относится к возрастной группе 25 - 34

Почти 63% пользователей имеют доход выше среднего

Большинство посещало проекты с десктопа (57%)

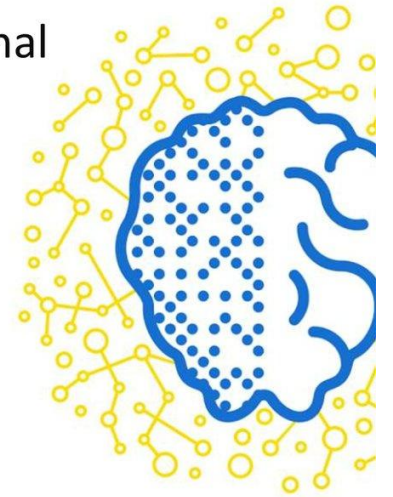




iPavlov: Conversational Intelligence Project

Mikhail Burtsev, PhD
Moscow Institute of Physics and Technology
(MIPT)

```
# Definition of iPavlov project
def iPavlov(talent, ideas):
    research = ideas * talent
    AI = development(research)
    return AI
```



" You need to hear me correctly Siri "

Thanks for that, Adam.






" I'm saying you're not doing a good
job "

It's nice to have one's work
appreciated.






?



Пример: оцените стоимость ноутбука

		Кол-во ядер	RAM (ГБ)	Объем жесткого диска (ГБ)	Диагональ/ разрешение	Работа от аккумулятора	Цена (руб.)
1		2	4	500 (HDD)	15"/1920x1080 пикс.	до 5 часов	31 490
2		4	8	256 (SSD)	14"/1920x1080 пикс.	до 12 часов	60 990
3		4	16	1000 (HDD)	17"/1920x1080 пикс.	до 3 часов	65 990
4		8	16	1000 (HDD) + 256 (SSD)	17"/1920x1080 пикс.	до 11 часов	109 990
5		4	16	1000 (HDD)+ 128 (SSD)	17"/1920x1080 пикс.	до 6 часов	?

Пример: оцените стоимость ноутбука

		Кол-во ядер	RAM (ГБ)	Объем жесткого диска (ГБ)	Диагональ/ разрешение	Работа от аккумулятора	Цена (руб.)
1		2	4	500 (HDD)	15"/1920x1080 пикс.	до 5 часов	31 490
2		4	8	256 (SSD)	14"/1920x1080 пикс.	до 12 часов	60 990
3		4	16	1000 (HDD)	17"/1920x1080 пикс.	до 3 часов	65 990
4		8	16	1000 (HDD) + 256 (SSD)	17"/1920x1080 пикс.	до 11 часов	109 990
5		4	16	1000 (HDD)+ 128 (SSD)	17"/1920x1080 пикс.	до 6 часов	86990

Постановка задачи машинного обучения

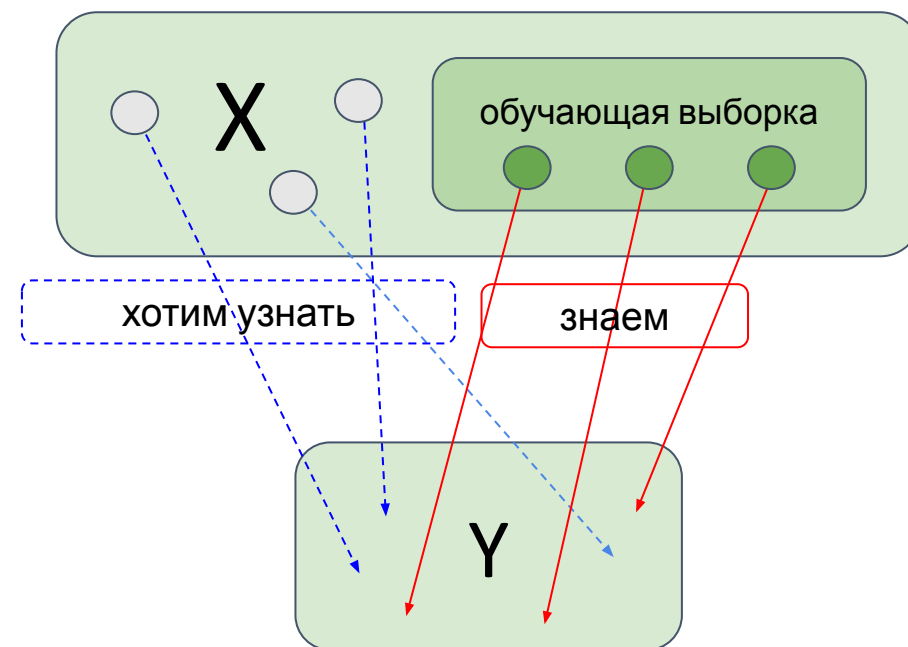
X — множество *объектов*

Y — множество *ответов* (например, два класса или произвольные числа)

$y: X \rightarrow Y$ — неизвестная закономерность

Дано: обучающая выборка, $\{x_1, x_2, \dots, x_n\}$ — подмножество множества X

Цель: подобрать *алгоритм*, приближающий функцию $y(x)$.



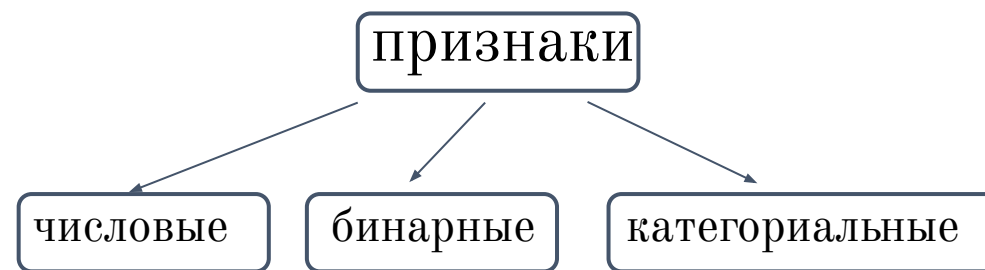
Как задаются объекты. Признаковое описание

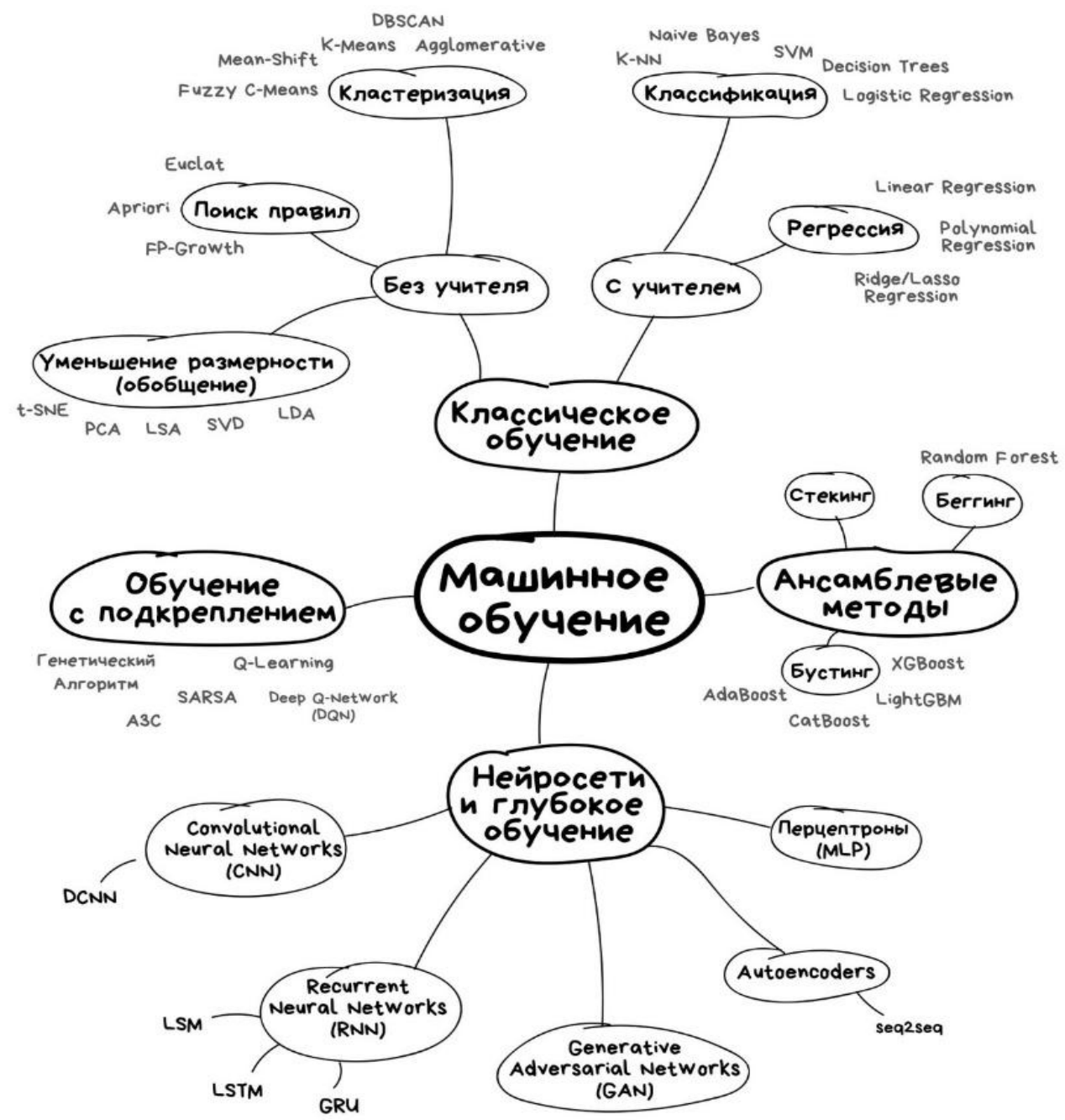
Объект x задаётся *признаковым описанием*

f_1, f_2, \dots, f_k — признаки (features) объекта x

$$\begin{bmatrix} f_1(x_1), & f_2(x_1), & \dots, & f_k(x_1) \\ f_1(x_2), & f_2(x_2), & \dots, & f_k(x_2) \\ \dots & & & \\ f_1(x_n), & f_2(x_n), & \dots, & f_k(x_n) \end{bmatrix}$$

— матрица “объекты-признаки”
объект, пригодный для применения
алгоритмов машинного обучения





Данные

Изображения

MNIST Dataset

- Изображения цифр, написанных от руки
- ~50к изображений
- Можно научить модель распознавать цифру



Табличные данные

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket
1	0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171
2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38	1	0	PC 17599
3	1	3	Heikkinen, Miss. Laina	female	26	0	0	STON/O2. 310128
4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35	1	0	113803
5	0	3	Allen, Mr. William Henry	male	35	0	0	373450
6	0	3	Moran, Mr. James	male		0	0	330877
7	0	1	McCarthy, Mr. Timothy J	male	54	0	0	17463
8	0	3	Palsson, Master. Gosta Leonard	male	2	3	1	349909

- Таблица как в Excel
- Как часто бывает: десятки столбцов, тысячи строк
- Один из столбцов - целевая переменная
- С данными такого вида мы будем работать на протяжении всего курса

Признаковое описание

Для того, чтобы работать с данными, нужно представить их в виде, пригодном для моделей ML

- Строка в таблице называется **объектом**
- Столбец в таблице называется **признаком**
- Признаки могут быть 3-х типов:
 - Числовые
 - Категориальные
 - Бинарные
- Столбец, который нужно предсказать, называется **целевой переменной**

X y* features

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket
1	0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171
2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38	1	0	PC 17599
3	1	3	Heikkinen, Miss. Laina	female	26	0	0	STON/O2. 310128
4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35	1	0	113803
5	0	3	Allen, Mr. William Henry	male	35	0	0	373450
6	0	3	Moran, Mr. James	male		0	0	330877
7	0	1	McCarthy, Mr. Timothy J	male	54	0	0	17463
8	0	3	Palsson, Master. Gosta Leonard	male	2	3	1	349909

Признаковое описание

Все признаки представляются в виде чисел:

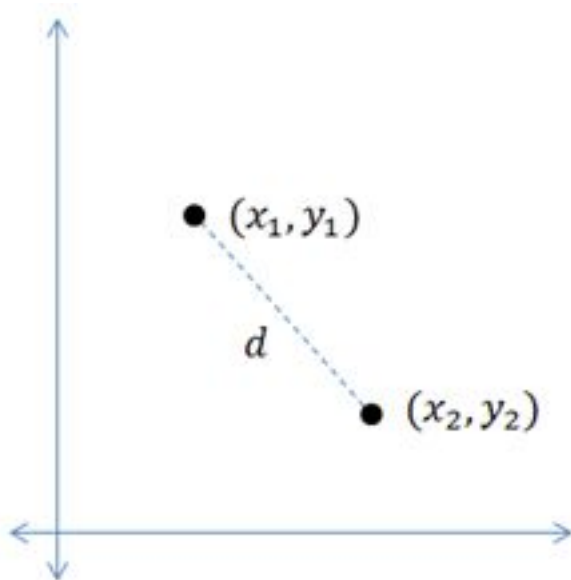
- Числовые признаки - это уже числа
- Бинарные признаки - как 0 и 1
- Категориальные признаки:
 - Как число от 0 до N , где N - число категорий
 - Как N -мерный вектор $\{0, 0, 1, 0, 0, 0\}$. Т.н. **one-hot vector**

Для каждого объекта набор его признаков собирается в один вектор

Вектор

- Вектор - это упорядоченный набор чисел
- Вектор - это координаты точки в пространстве

Для двух точек можно рассчитать расстояние между ними



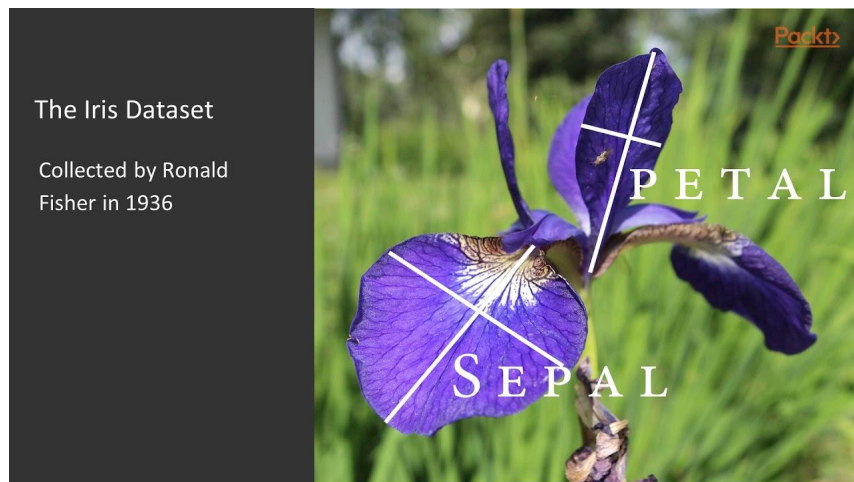
$$d = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

Матрица

- Матрица - это упорядоченный набор векторов одного размера
- Набор векторов - это набор точек в пространстве

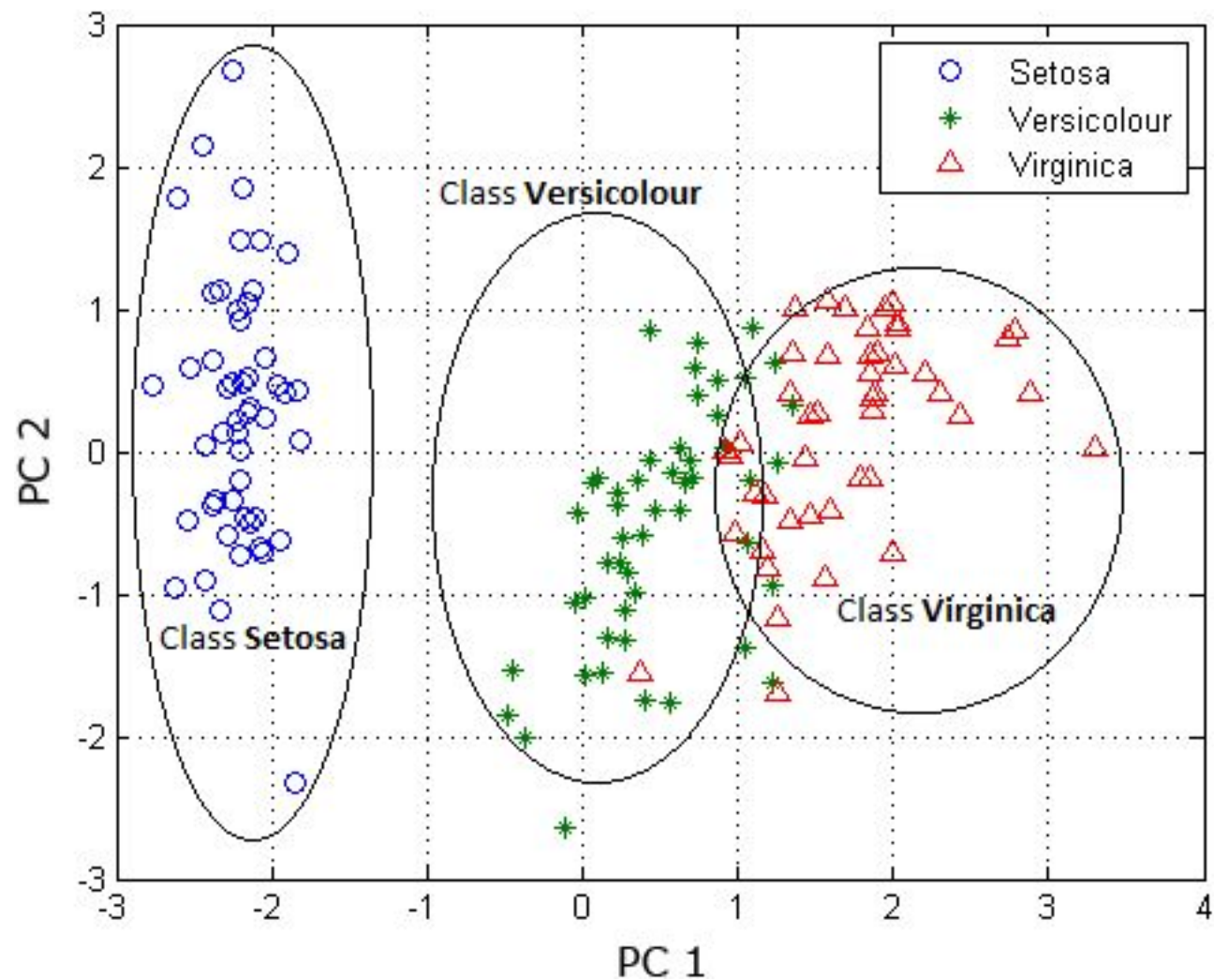
Датасет с подготовленными признаками - это матрица

Визуализация данных

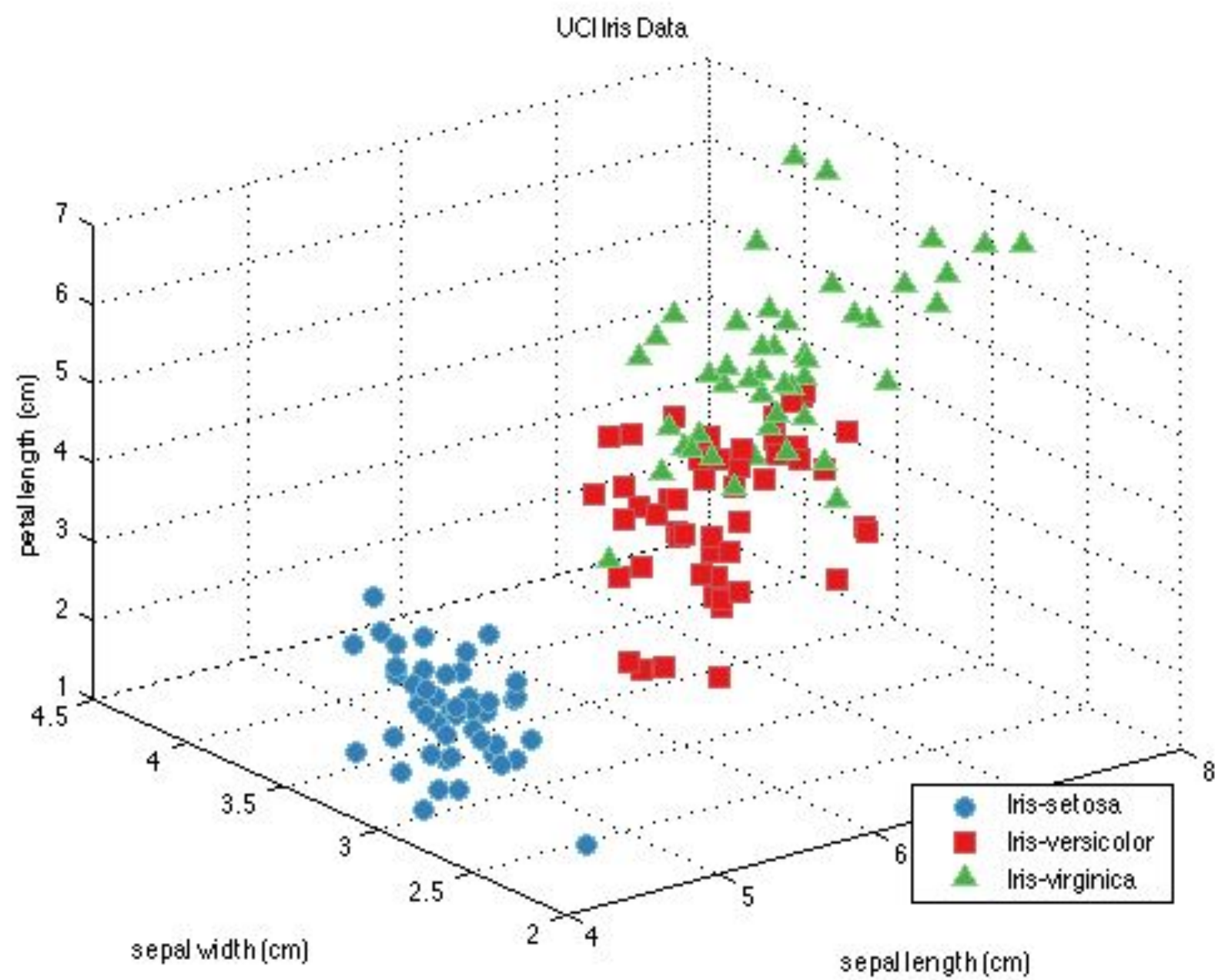


Датасет - это матрица

petal length	petal width	target
1.4	0.2	Iris-setosa
1.4	0.2	Iris-setosa
1.3	0.2	Iris-setosa
1.5	0.2	Iris-setosa
1.4	0.2	Iris-setosa

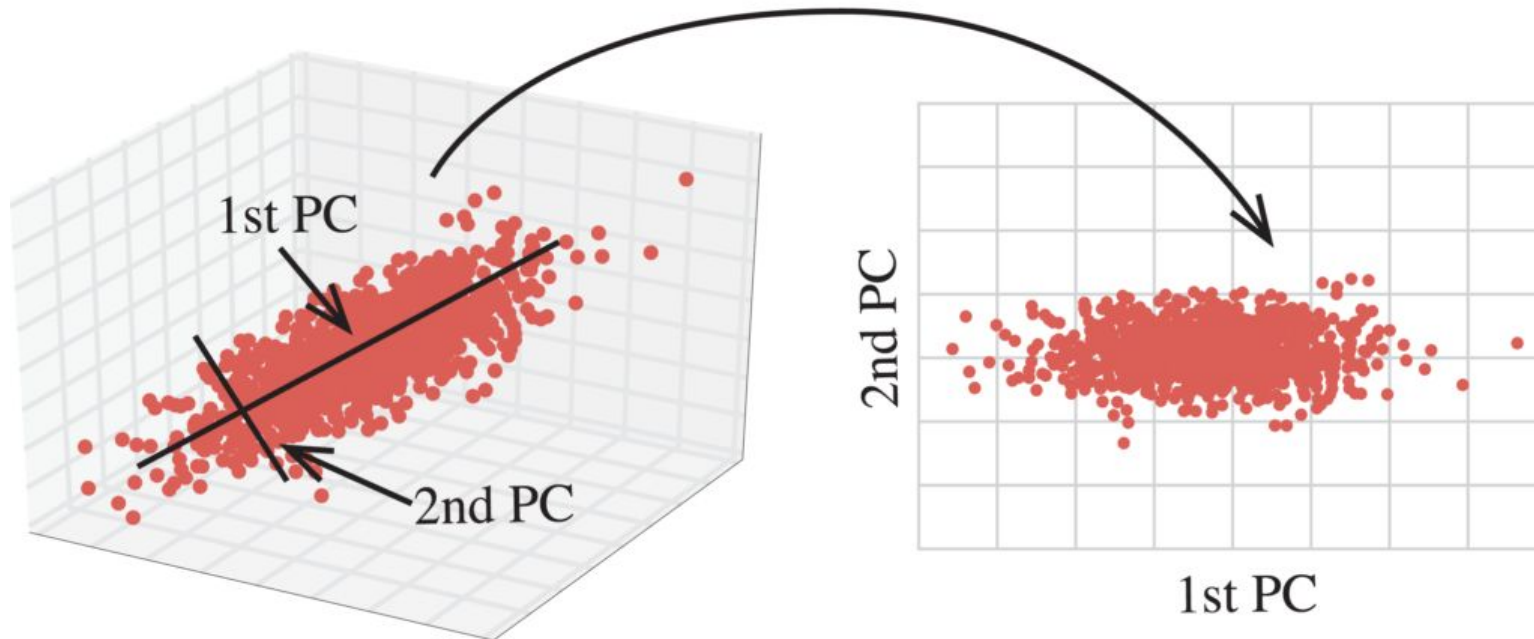


Вообще-то, в этом датасете признаков не 2, а 4. Как на них смотреть?

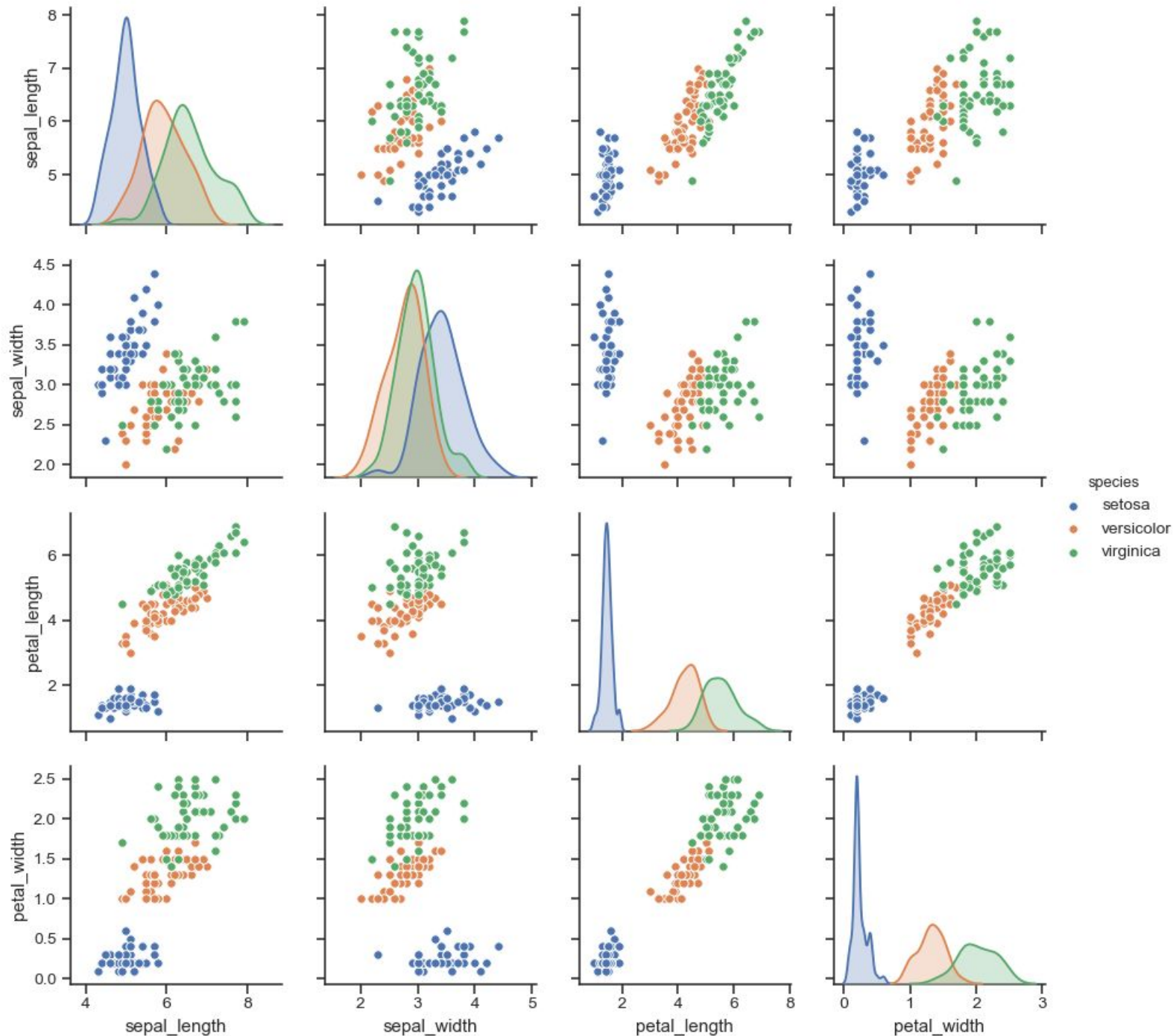


Большие размерности

- Двумерный набор точек можно нарисовать на плоскости
- Трёхмерный набор можно спроецировать на плоскость
- Размерности векторов могут быть порядков $100 \sim 100\,000$
- Их всё равно можно спроецировать!

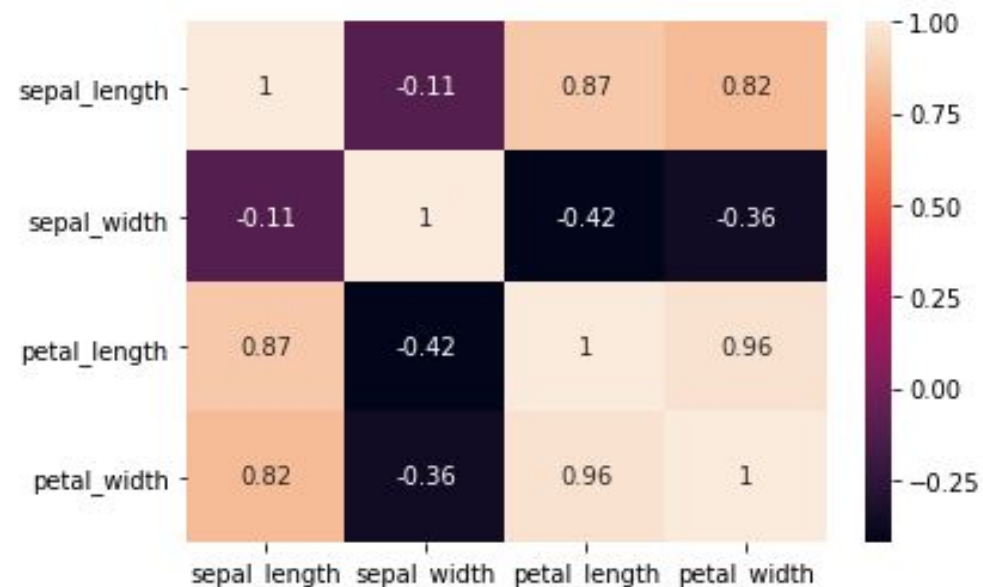


Визуализация реальных датасетов



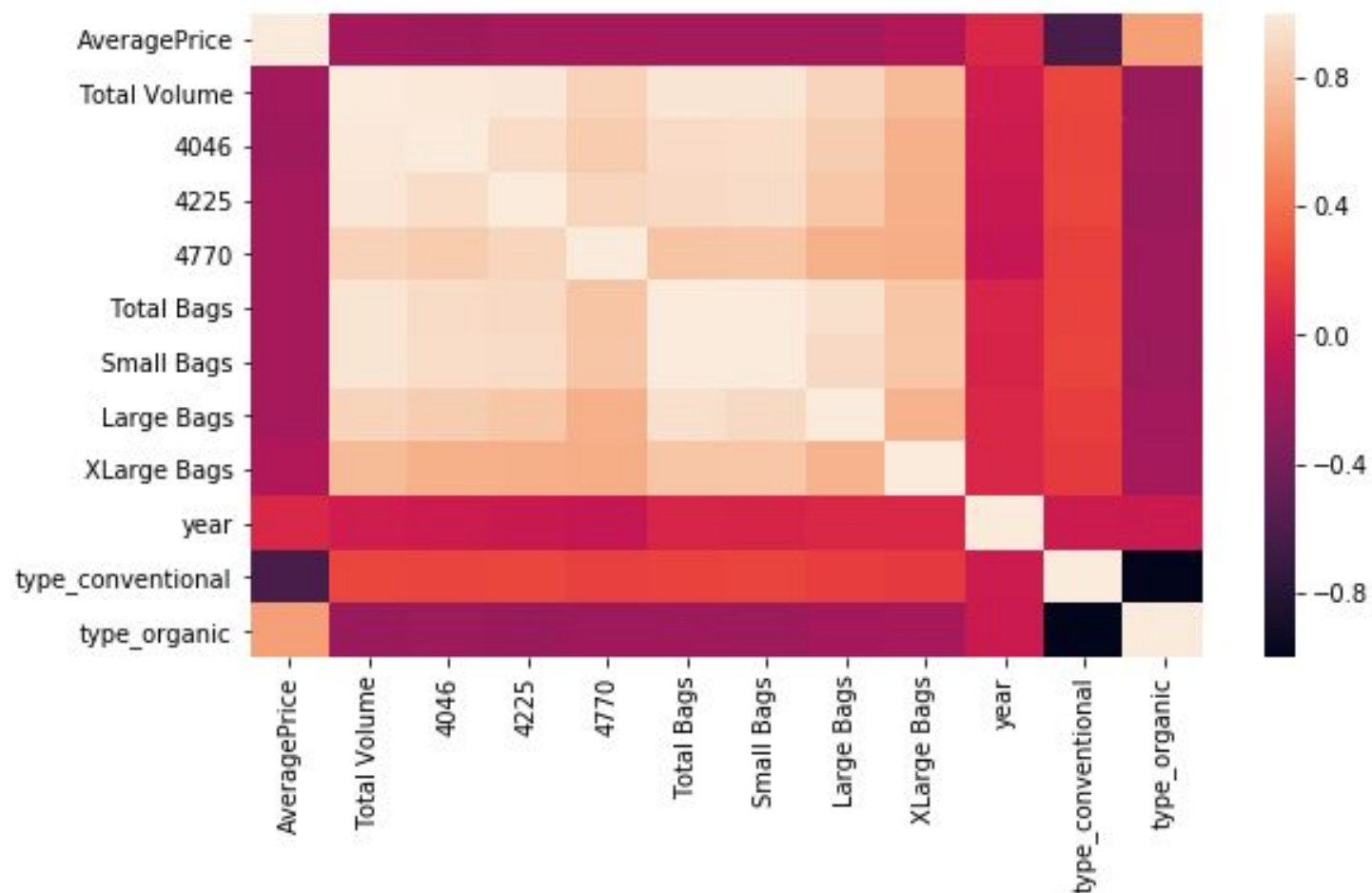
Матрица корреляции

- Если признаков много, то все парные графики не рассмотришь
- Визуально проанализировать признаки помогает матрица корреляции



```
plt.figure(figsize=(9,5))  
sns.heatmap(df.corr())
```

<matplotlib.axes._subplots.AxesSubplot at 0x1fa0e4d4a20>



Тензор

- Упорядоченный набор точек - это вектор
- Упорядоченный набор векторов одного размера - это матрица
- Упорядоченный набор матриц одного размера - это **тензор**

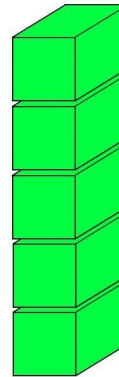
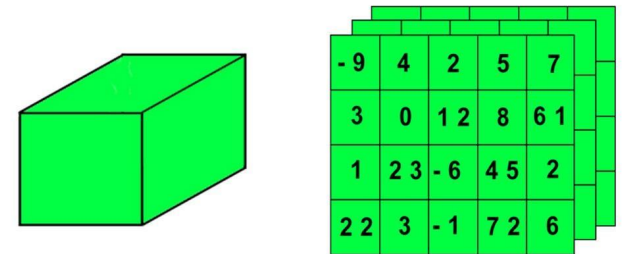
1D TENSOR /
VECTOR

5
7
4 5
1 2
-6
3
2 2
1
6
3
-9

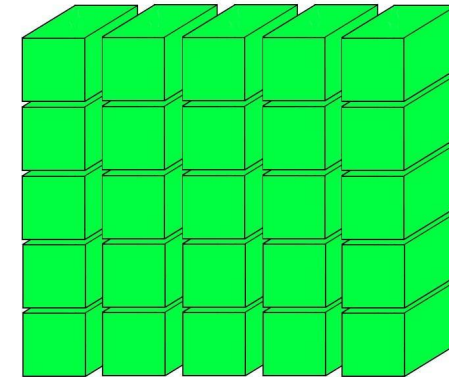
2D TENSOR /
MATRIX

-9	4	2	5	7
3	0	1 2	8	6 1
1	2 3	-6	4 5	2
2 2	3	-1	7 2	6

3D TENSOR /
CUBE



4D TENSOR
VECTOR OF CUBES



5D TENSOR
MATRIX OF CUBES

Tensor shape

У тензора k-го ранга есть k индексов

Форма тензора k-го ранга - это набор k чисел, каждое из которых означает: какое кол-во значений может пробегать данный индекс. Т.е. протяженность тензора в разных направлениях.

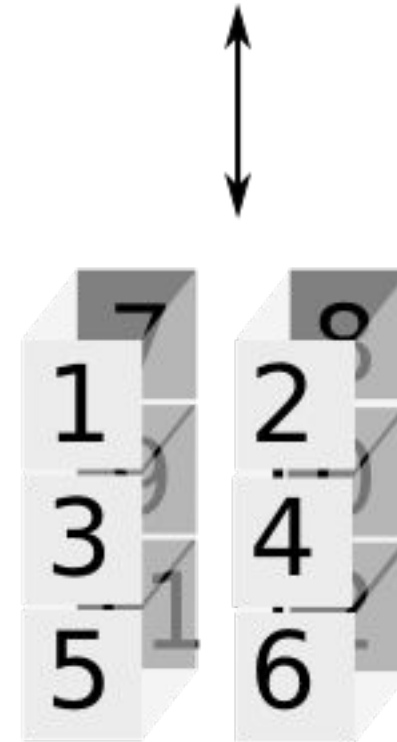
Форма тензора справа равна [2, 3, 2]

Форма цветного FullHD изображения: [1920, 1080, 3]

*Слово “форма” по-русски не применяется для тензоров. Это дословный перевод термина “shape”

This is how you represent a tensor in code

```
[ [[1,2], [3,4], [5,6]], [[7,8], [9,10], [11,12]] ]
```



And this is how we like to visualize tensors

MNIST

Вопрос: каким тензором является датасет?

Вопрос: что является объектом, признаком, и целевой переменной в задаче MNIST? Какого они типа?



Модель

Задача обучения с учителем

- Между объектом и целевой переменной существует **реальная зависимость**
- У нас есть только N сэмплов этой зависимости - **обучающая выборка**
- Задача - научиться **предсказывать** целевую переменную для новых точек
- Для этого строится **модель**

Модель - это функция, которой можно аппроксимировать реальную зависимость, имея конечное число примеров.

Классификация и регрессия

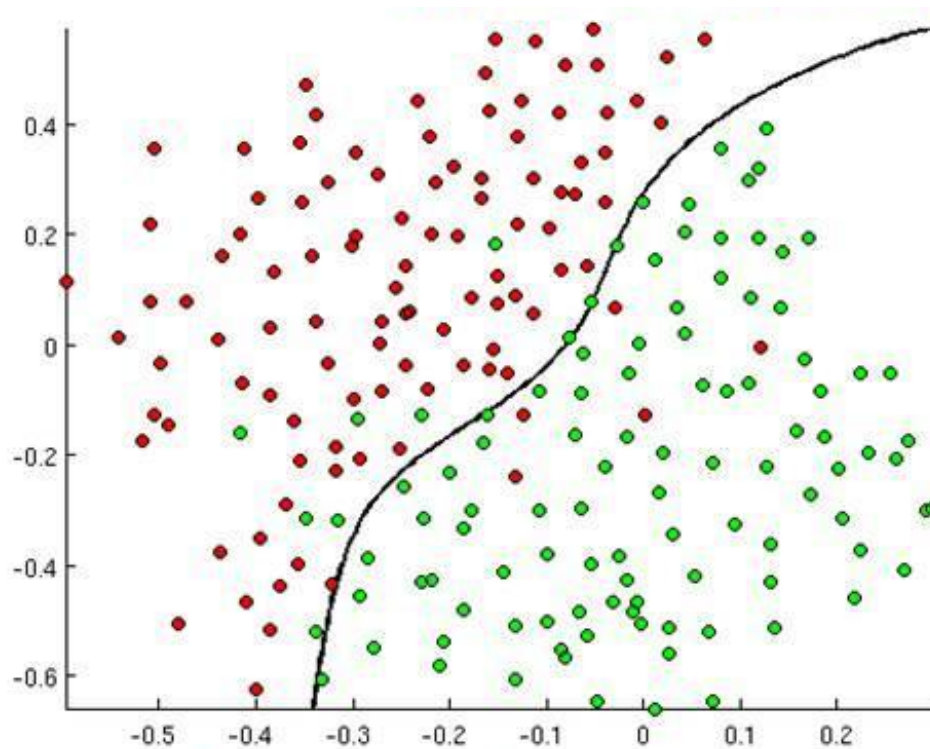
Целевая переменная, как и признаки, может быть трех типов:

- Числовая
- Бинарная
- Категориальная

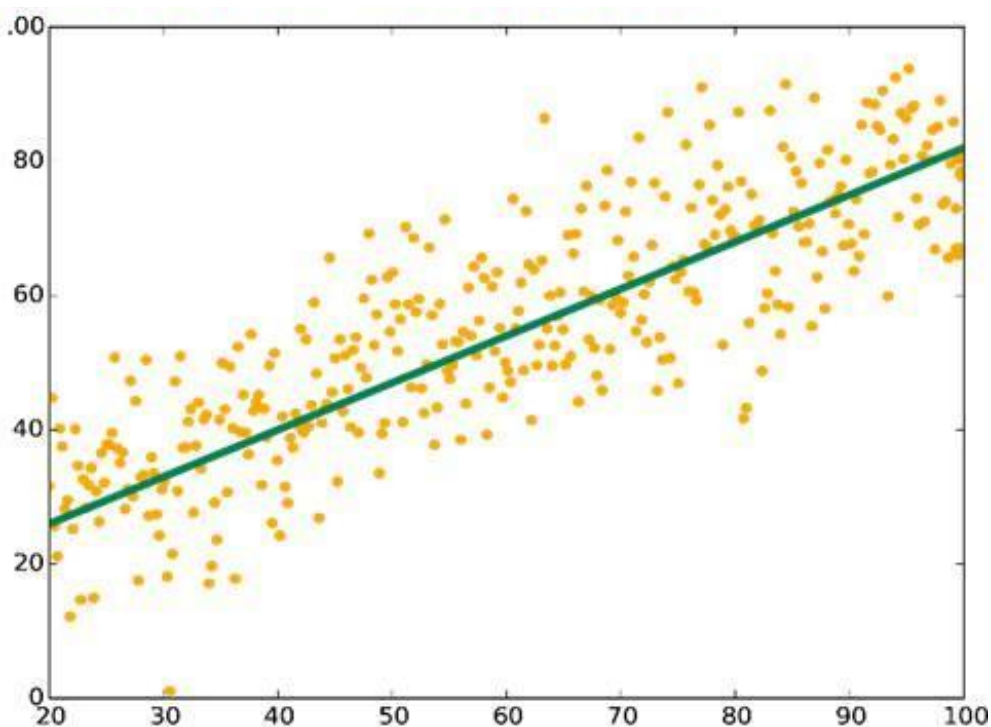
Предсказание числового значения называется **регрессия**

Предсказание одного из нескольких классов называется **классификация**

Классификация и регрессия

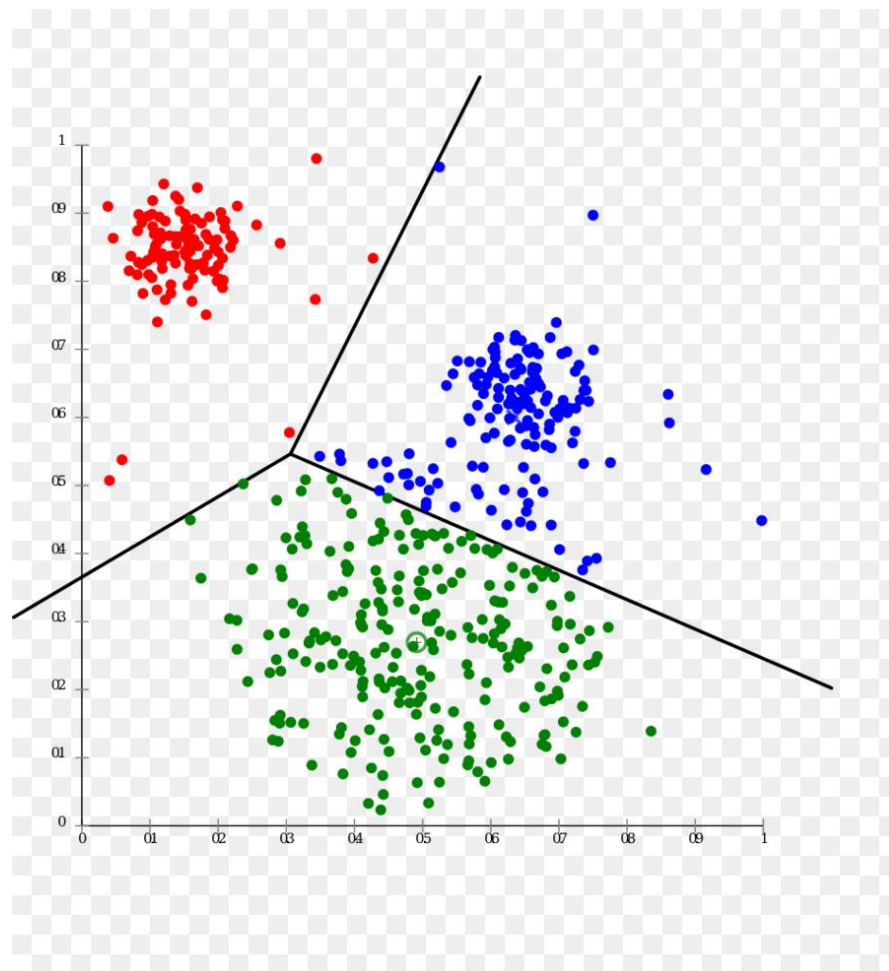


Classification



Regression

Кластеризация



KNN

K Nearest Neighbors

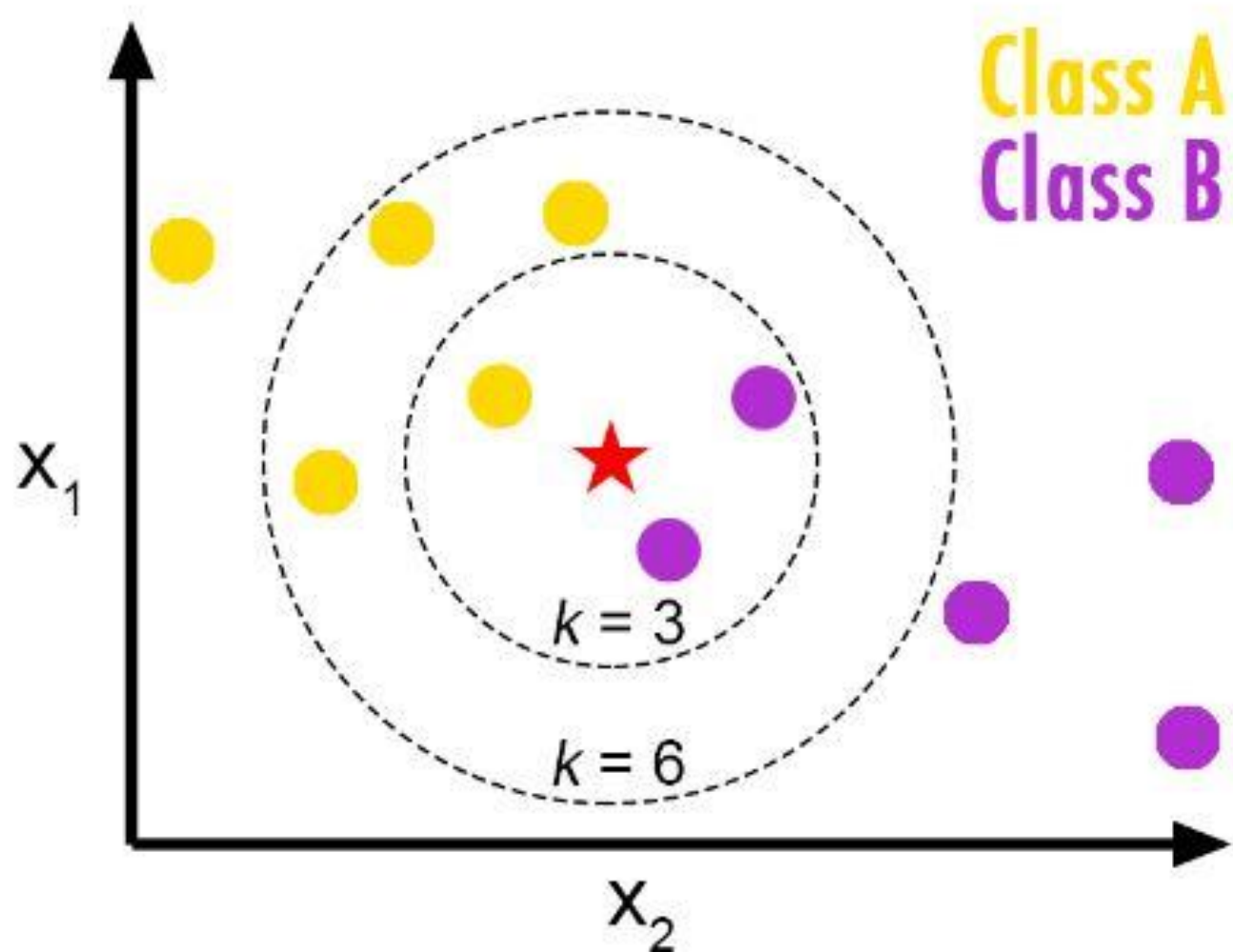
Метод K ближайших соседей

- На вход подается вектор - признаковое описание какого-то объекта
- Находится K ближайших к нему векторов, для которых ответ известен
- Ответ для новой точки выбирается с помощью
 - Усреднения в случае регрессии
 - Голосования в случае классификации
- Возможно также усреднение/голосование с весами

KNN классификация

K - внешний параметр. Он подбирается так, чтобы модель работала как можно лучше.

Результат предсказания для некоторых точек может зависеть от K



Метрики

Измерение качества модели

Чтобы понять, насколько адекватно ведет себя модель, нужно каким-то образом численно оценить ее качество.

Метрика - это функция вида:

$$metric(y, \hat{y})$$

где y - это правильное значение целевой переменной (**label**),

а \hat{y} - значение, предсказанное моделью (**prediction**).

Примеры метрик

Классификации:

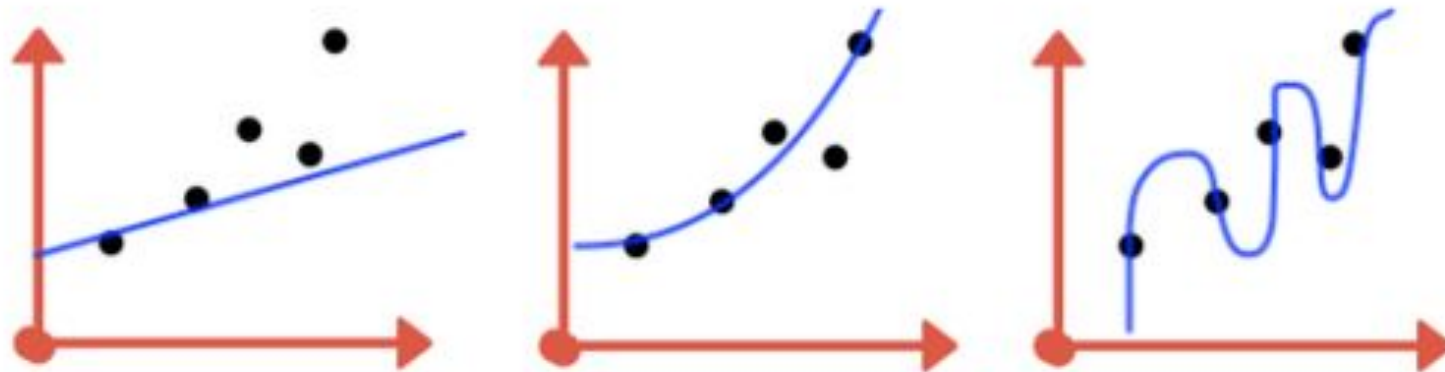
- **accuracy** - процент правильных предсказаний среди всех примеров
- **precision** - точность
- **recall** - полнота
- **f1** - объединяет полноту и точность
- **ROC-AUC** - вероятность правильного ранжирования двух случайных примеров

Регрессии:

- **MSE** - средний квадрат отклонения
- **RMSE** - стандартное отклонение
- **MAE** - средний модуль отклонения
- **R2** - коэффициент детерминации

Более подробно метрики будут рассмотрены после практического занятия

Несмещенная оценка



Вопрос: какое предсказание лучше по метрикам, а какое на самом деле?

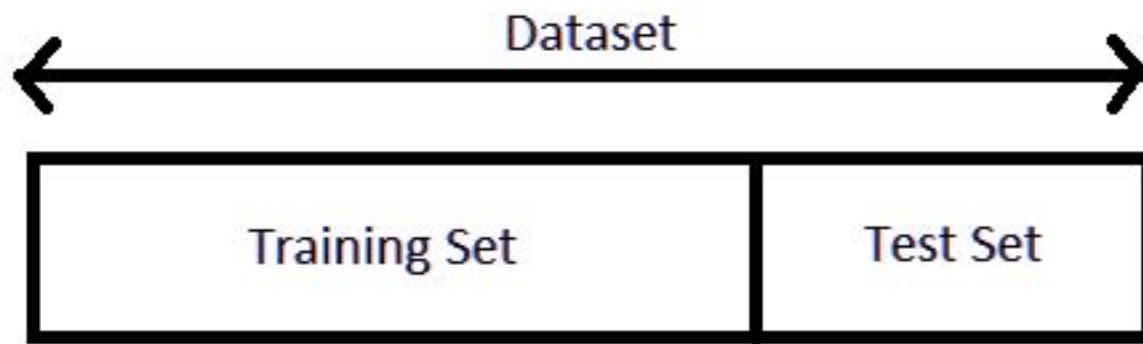
Если тестировать модель на той же выборке, на которой она обучалась, то оценка получится смещенной. В таком случае “самая лучшая” модель - это та, которая просто запомнила все данные.

Хорошая модель должна делать хорошие предсказания на **новых** для себя данных

Отложенная выборка

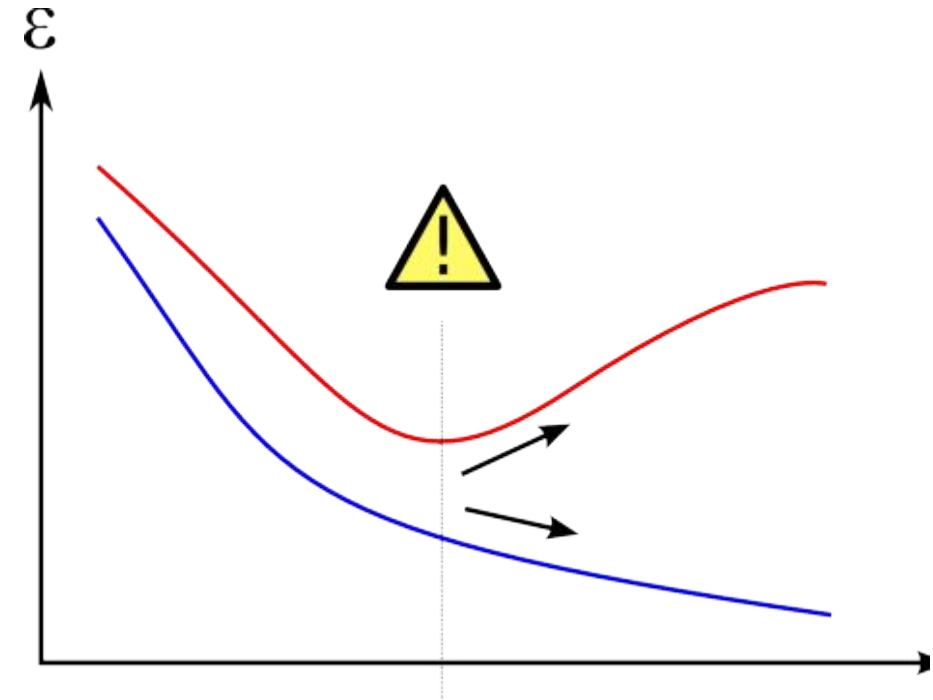
Можно “отложить”, скажем, 20% обучающей выборки для валидации модели. Использовать 80% выборки для обучения и 20% для тестирования.

- Оценка на тестовой выборке будет несмещенной
- Тестовая выборка маленькая - оценка будет иметь погрешности



Переобучение

- Как обнаружить? - Train/Test split
 - Разделить выборку на обучающую и контрольную
 - Следить за качеством на контрольной выборке
- Минусы?
 - Уменьшение размера обучающей выборки может негативно сказаться на качестве
 - Малый размер тестовой выборки может давать сильное смещение оценки.
 - Можно переобучиться под **тестовую выборку**



Кросс-валидация

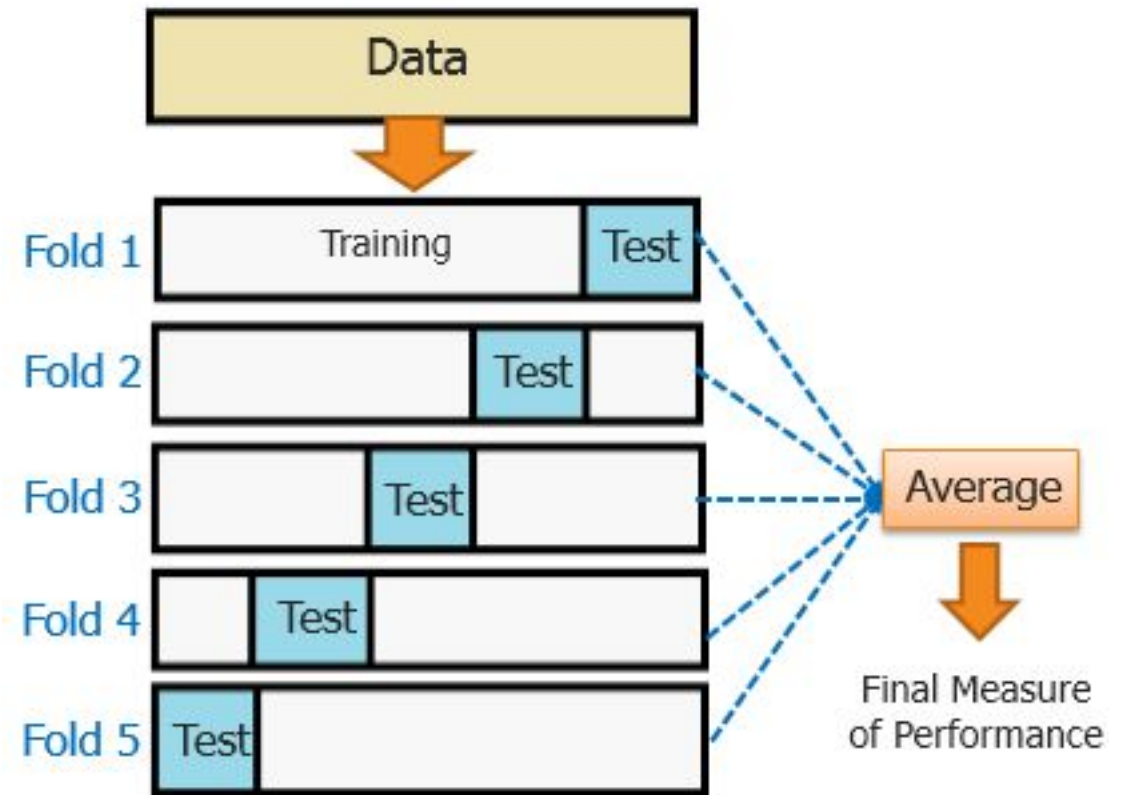
- Разбиваем выборку на k частей
- $k-1$ частей используются для обучения и одна - для тестирования
- Процесс повторяется k раз. Каждый раз для тестирования выбирается разная часть
- Результаты тестирования усредняются

Плюсы:

- Погрешность оценки уменьшается, т.к. используется весь набор

Минусы:

- Обучение производится k раз. Для некоторых моделей это может быть очень долго

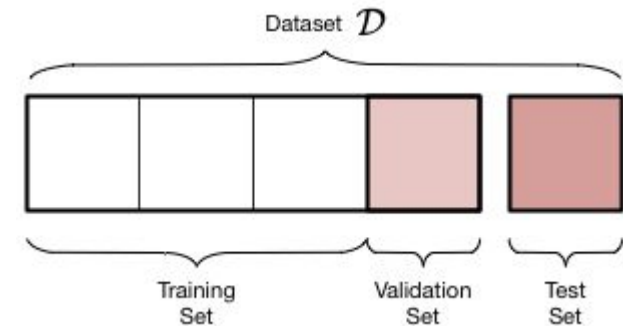


Кросс-Валидация

- Плюсы
 - Качество измеряется на всем наборе данных
 - Качество не зависит от выбора конкретного тестового набора
 - Сложнее переобучиться под тест
- Минусы
 - Скорость!
- Что выбрать
 - Мало обучающих данных => Кросс-Валидация
 - Много обучающих данных => Train/Validate split
- Не забыть
 - Отложить Test для замера итогового качества
 - Обучить итоговую модель на всех данных

Кросс-Валидация По Времени

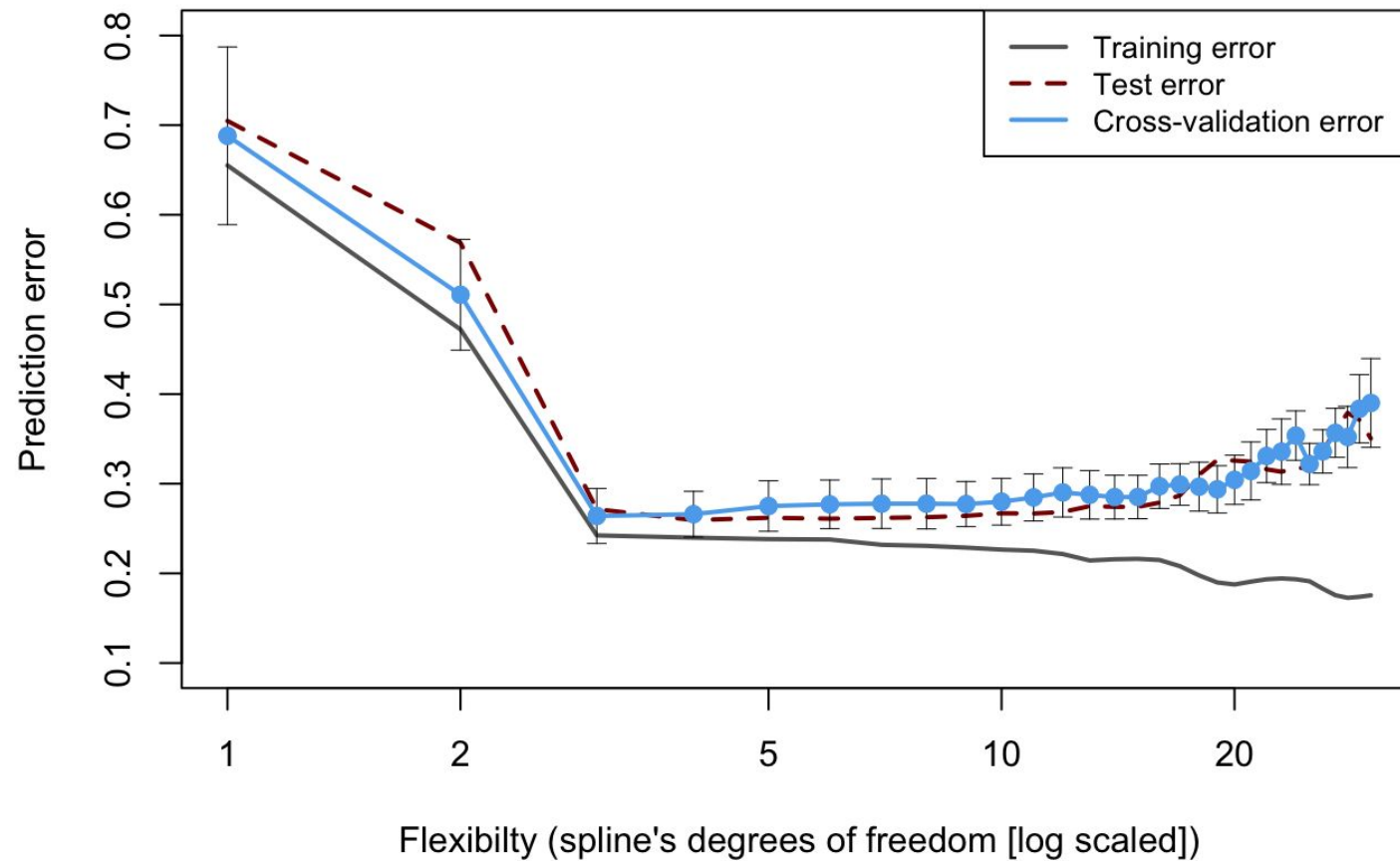
- Используется для анализа временных рядов
 - Тестовый набор выбирается из самых свежих данных. Обучение на более старых.
- Полезно в реальных задачах
 - Если в качестве признаков используется множество сигналов, которые могут меняться от времени.
 - Есть возможность определить дату наблюдения.



□ Время □

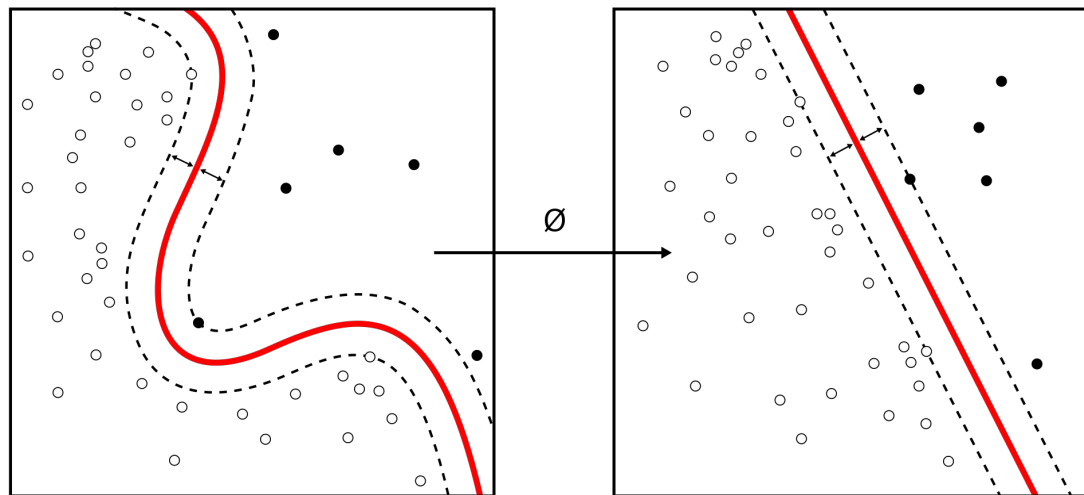
Кросс-Валидация, пример

10-fold Cross-Validation



Тезисы вводной лекции

- Данные нужно превращать в числа - признаковое описание
- В данных должна присутствовать целевая переменная
- Можно обучить модель предсказывать целевую переменную - это называется обучение с учителем
- Если предсказывается число - это регрессия, если класс - классификация
- Качество модели оценивается с помощью метрик



Colab