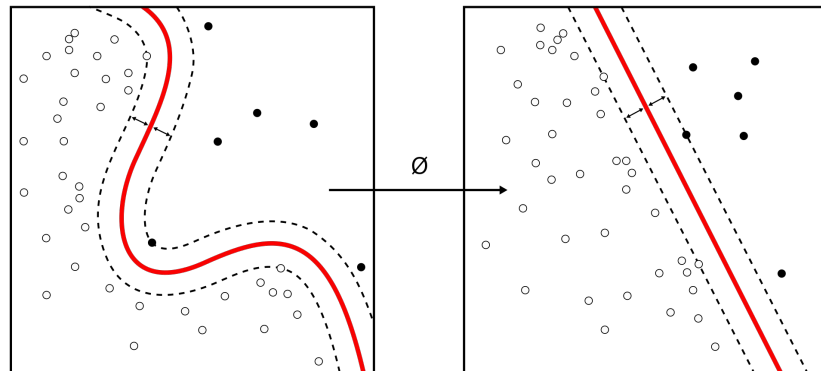


Лекция 1

Введение в ML



Аминов Тимур. email: aminov.tv@phystech.edu , telegram: @Sifonsoul

План

1. Данные

- a. Как выглядят данные
- b. Признаковое описание

2. Обзор ML

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
- d. Переобучение и недообучение

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>
- Демонстрация алгоритмов на простых примерах <https://ml-toy.herokuapp.com/>

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:

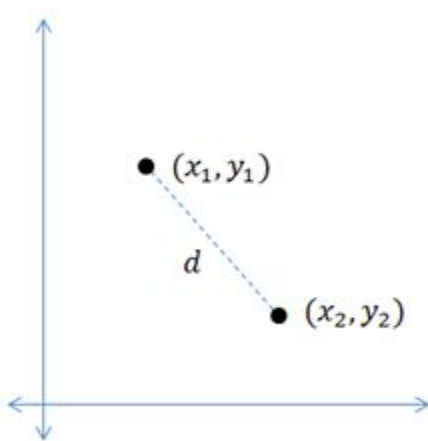
- Числовые признаки - это уже числа
- Бинарные признаки - как 0 и 1
- Категориальные признаки:
 - Как число от 0 до N, где N - число категорий
 - Как N-мерный вектор {0, 0, 1, 0, 0, 0}. Т.н. **one-hot vector**

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

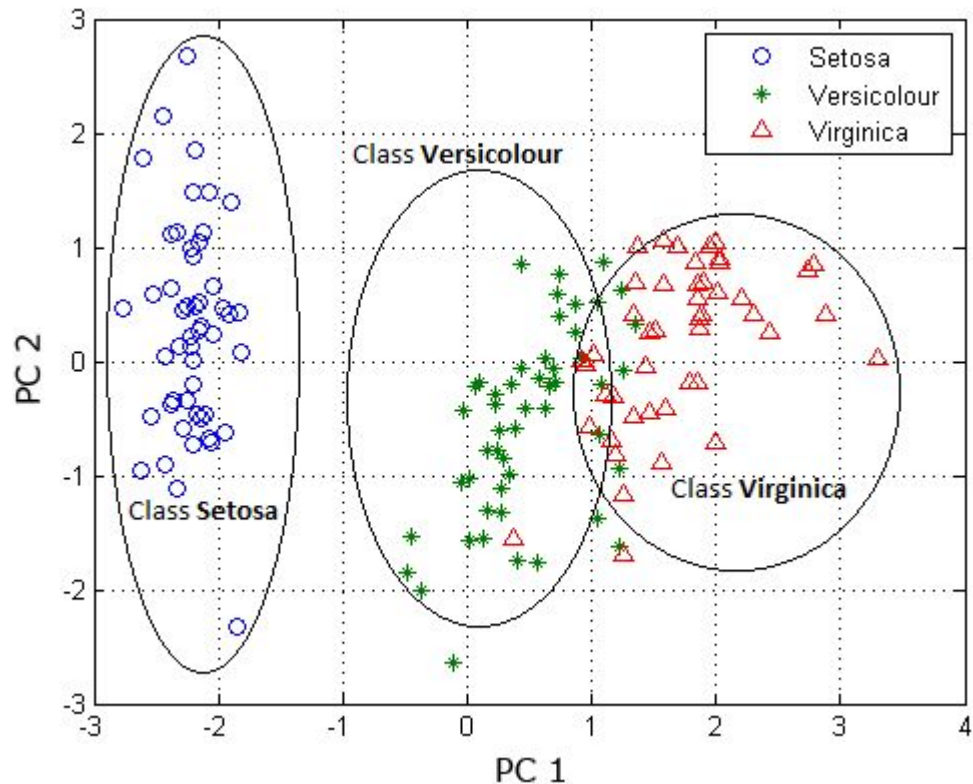
Датасет с подготовленными признаками - это матрица

Визуализация данных



Датасет - это матрица

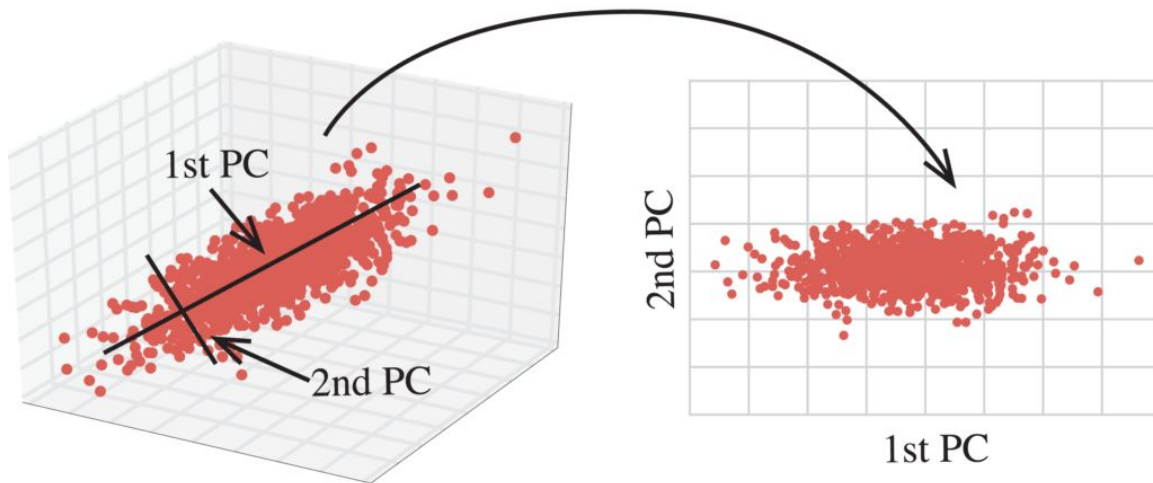
petal length	petal width	target
1.4	0.2	Iris-setosa
1.4	0.2	Iris-setosa
1.3	0.2	Iris-setosa
1.5	0.2	Iris-setosa
1.4	0.2	Iris-setosa



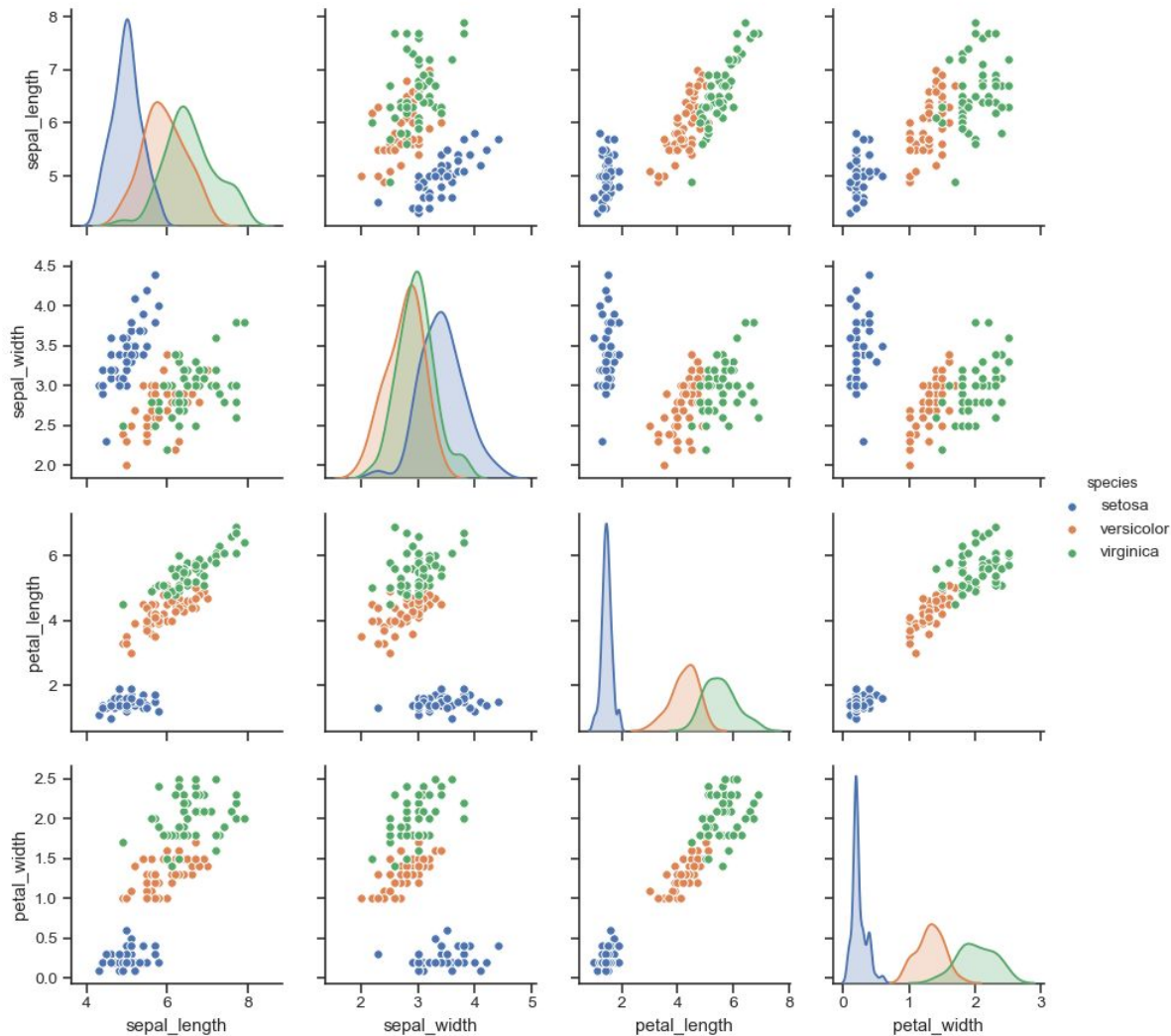
Вообще-то, в этом датасете признаков не 2, а 4. Как на них смотреть?

Большие размерности

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



Визуализация реальных датасетов

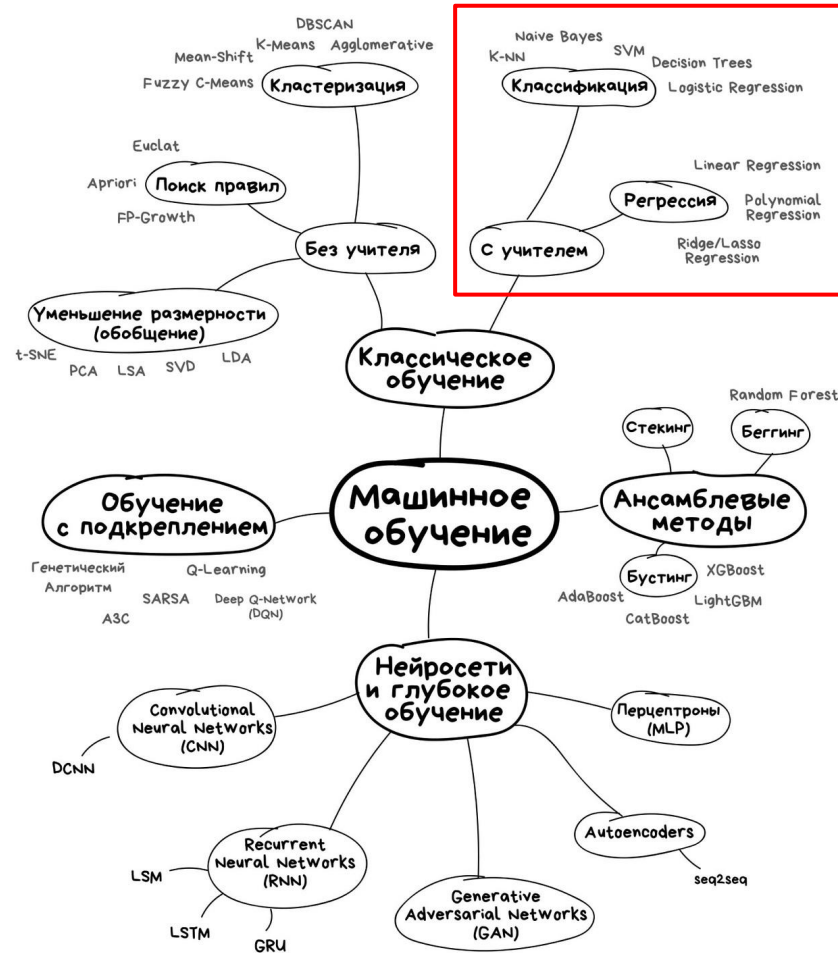


Обзор ML









Модель



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

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.

Формальная постановка задачи

X - Множество объектов

Y - Целевая переменная

$f: X \rightarrow Y$ - неизвестная зависимость

Дано

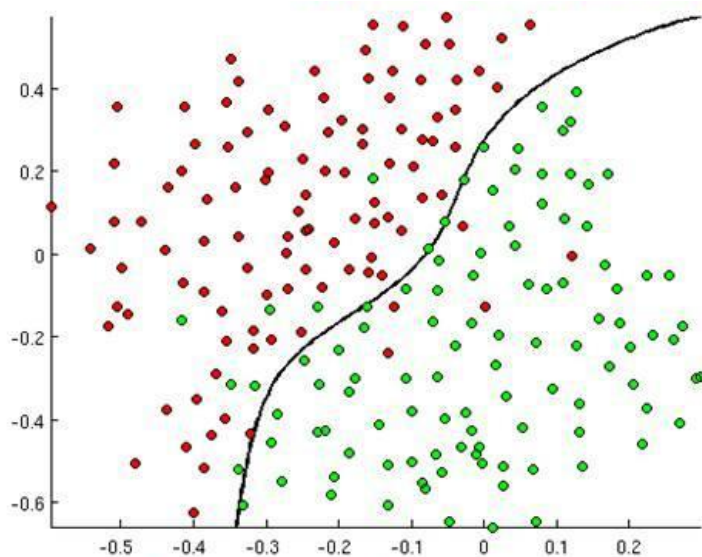
$\{x_1, \dots, x_l\}$ - обучающая выборка

$y_i = y(x_i)$, $i = 1, \dots, l$ - известные ответы

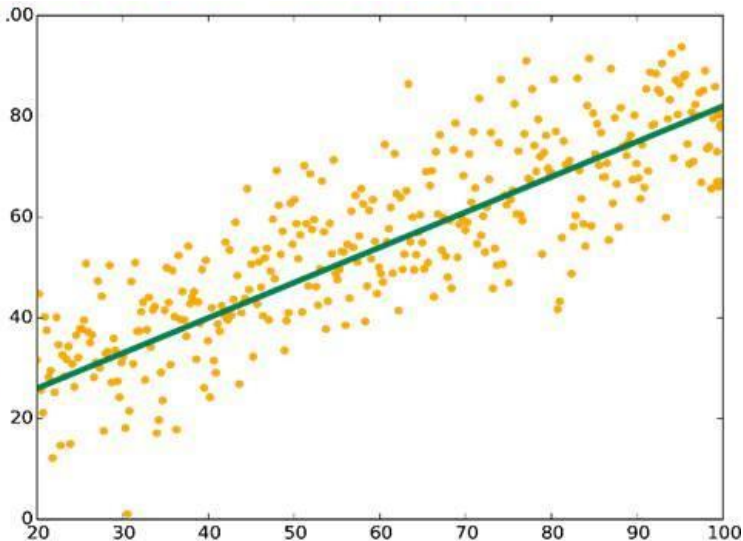
Найти

$a: X \rightarrow Y$ - алгоритм, решающую функцию, приближающую y на всем множестве X

Классификация и регрессия



Classification



Regression

Расстояние



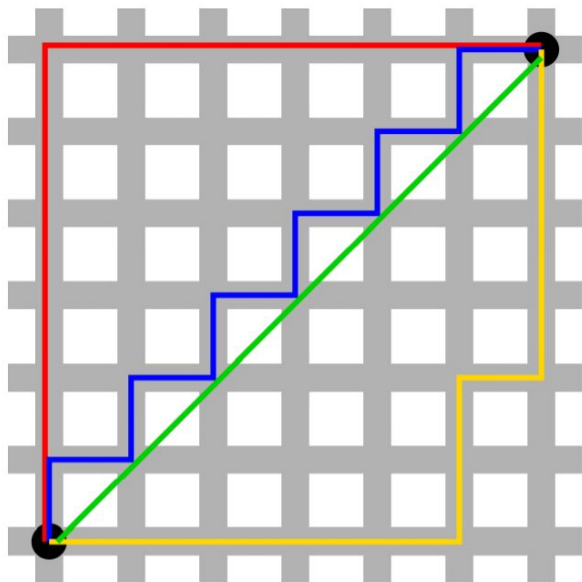
Что такое расстояние?

Метрическое пространство есть пара (X, d) , где X — множество, а d — числовая функция


функция d обладает следующими свойствами:

- $d(x, y) = 0$, значит $x = y$ (аксиома тождества)
- $d(x, y) = d(y, x)$ (аксиома симметрии)
- $d(x, y) \leq d(x, z) + d(z, y)$ (неравенство треугольника)

Примеры



$$D(x, y) = \sum_i |x_i - y_i|$$

	a	b	c	d	e	f	g	h	
8	5	4	3	2	2	2	2	2	8
7	5	4	3	2	1	1	1	2	7
6	5	4	3	2	1		1	2	6
5	5	4	3	2	1	1	1	2	5
4	5	4	3	2	2	2	2	2	4
3	5	4	3	3	3	3	3	3	3
2	5	4	4	4	4	4	4	4	2
1	5	5	5	5	5	5	5	5	1
	a	b	c	d	e	f	g	h	

$$D(x, y) = \max_i |x_i - y_i|$$

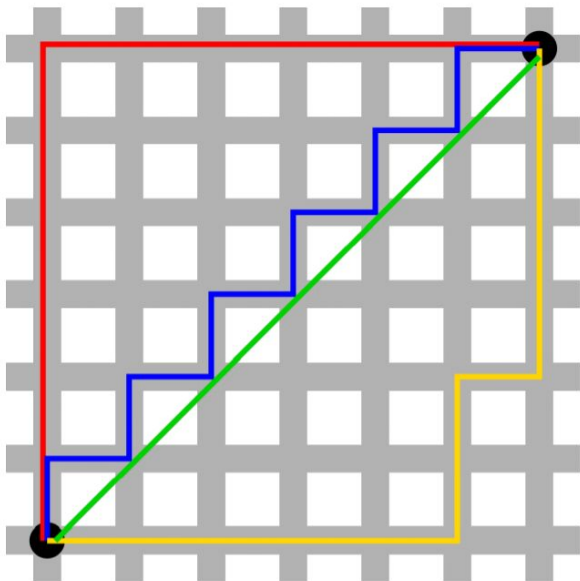
Семейство расстояний Менковского

$$D(x, y) = \sqrt[p]{|x_1 - y_1|^p + |x_2 - y_2|^p + \cdots + |x_n - y_n|^p}.$$

Можно показать, что для любого $p \geq 1$:


- ▶ $D(x, y) \geq 0$,
- ▶ $D(x, x) = 0$,
- ▶ $D(x, y) = D(y, x)$,
- ▶ $D(x, y) \leq D(x, z) + D(z, y)$.

Манхэттенское расстояние $p = 1$



$$D(x, y) = \sum_i |x_i - y_i|$$

Чебышевское расстояние $p = \infty$

	a	b	c	d	e	f	g	h	
8	5	4	3	2	2	2	2	2	8
7	5	4	3	2	1	1	1	2	7
6	5	4	3	2	1		1	2	6
5	5	4	3	2	1	1	1	2	5
4	5	4	3	2	2	2	2	2	4
3	5	4	3	3	3	3	3	3	3
2	5	4	4	4	4	4	4	4	2
1	5	5	5	5	5	5	5	5	1
	a	b	c	d	e	f	g	h	

$$D(x, y) = \max_i |x_i - y_i|$$

KNN



K Nearest Neighbors

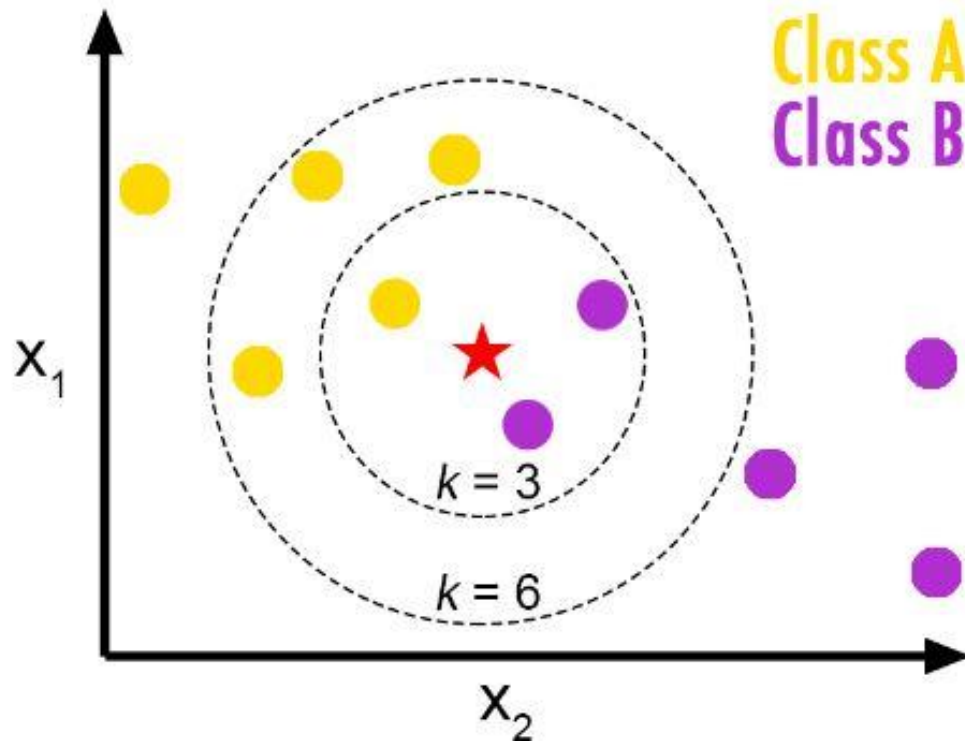
Метод K ближайших соседей

- 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

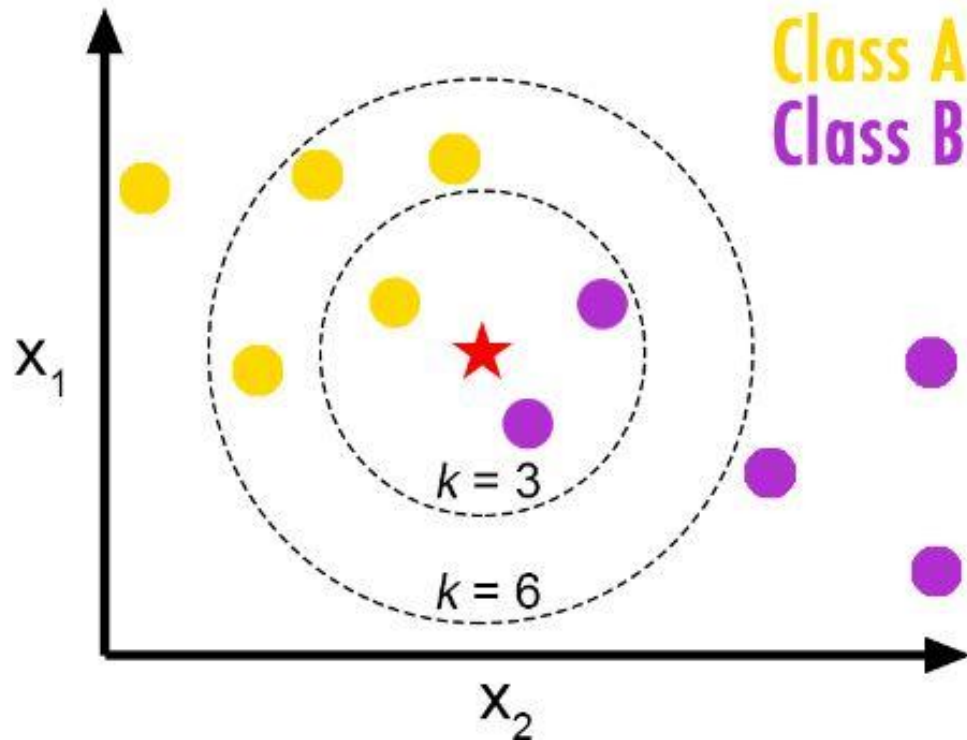


KNN регрессия

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

Результат предсказания для некоторых точек может зависеть от K

$$\hat{y} = \frac{\sum_{k=1}^K y_k}{K}$$



Метрики



Измерение качества модели

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

Метрика - это функция вида:

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

где \mathbf{y} - это правильное значение целевой переменной (**label**),

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

Примеры метрик

Классификации:

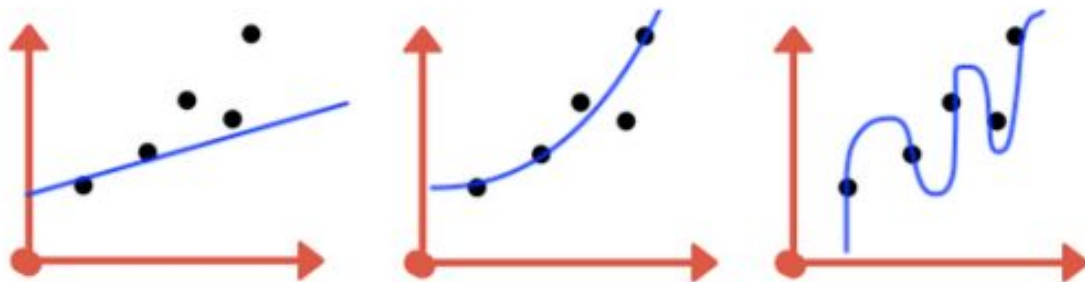
- **accuracy** - процент правильных предсказаний среди всех примеров
- precision - точность
- recall - полнота
- f1 - объединяет полноту и точность
- ROC-AUC - вероятность правильного ранжирования двух случайных примеров

Регрессии:

- MSE - средний квадрат отклонения
- RMSE - стандартное отклонение
- MAE - средний модуль отклонения
- MAPE - mean absolute percentage error
- R2 - коэффициент детерминации

Более подробно метрики будут рассмотрены в следующий раз

Смещенная оценка



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

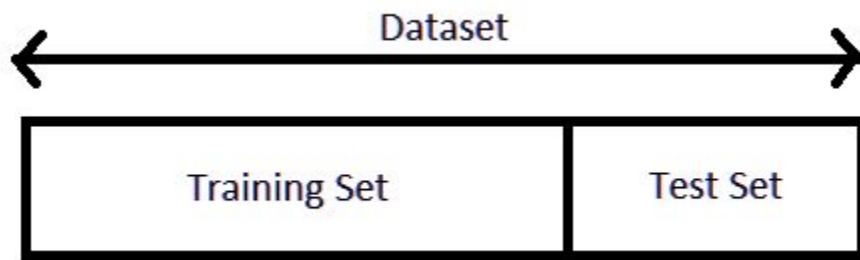
Если тестировать модель на той же выборке, на которой она обучалась, то оценка получится смещенной. В таком случае “самая лучшая” модель - это та, которая просто запомнила все данные.

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 погрешность



Кросс-валидация

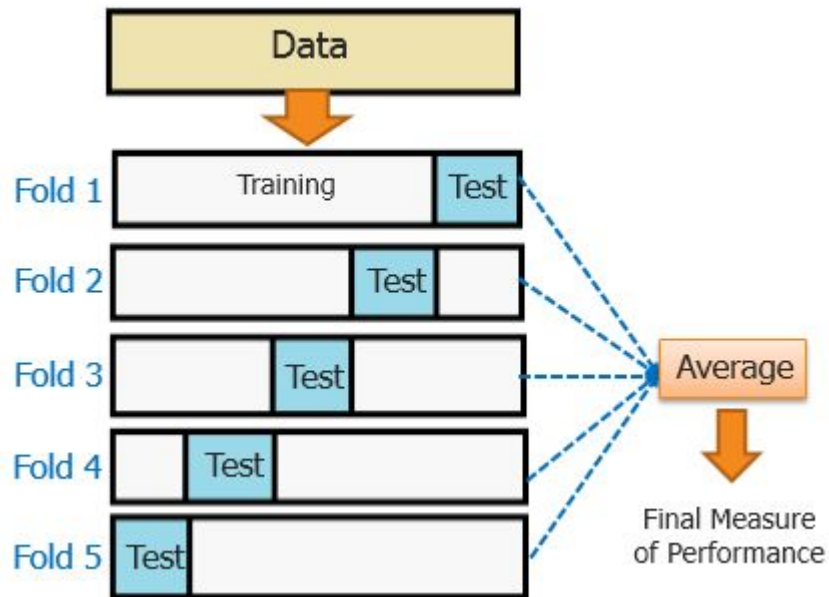
- Разбиваем выборку на k частей
- $k-1$ частей используются для обучения и одна - для тестирования
- Процесс повторяется k раз. Каждый раз для тестирования выбирается разная часть
- Результаты тестирования усредняются

Плюсы:

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

Минусы:

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



Переобучение и недообучение

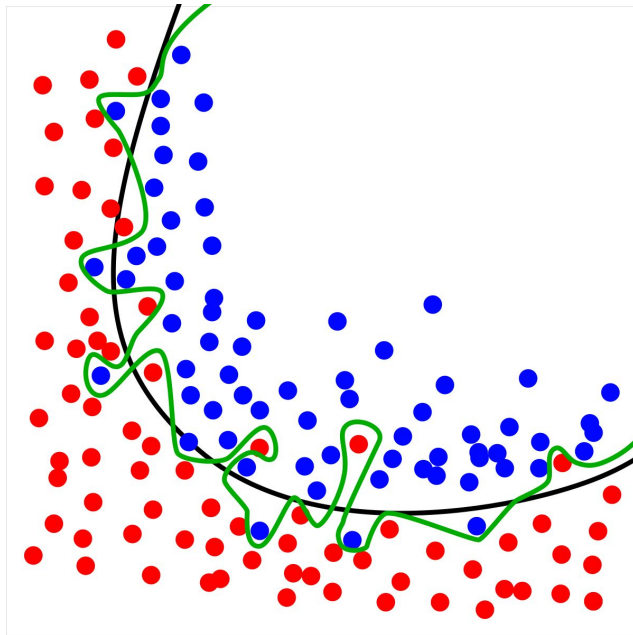
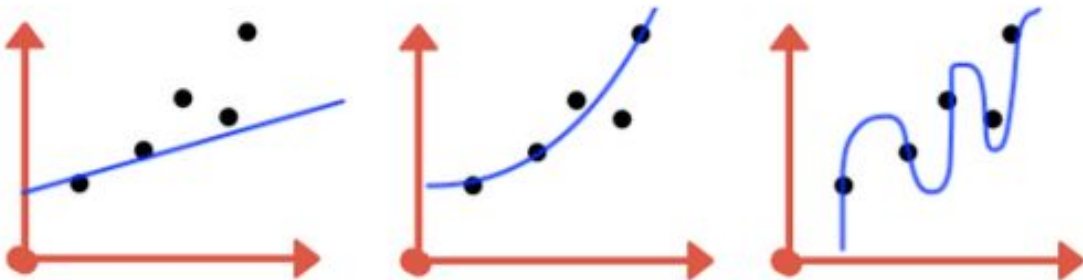


Переобучение и недообучение

Слишком простая “глупая” модель может уловить только общую закономерность, без деталей. Это называется **недообучение**.

Слишком “умная”, сложная модель может просто запомнить все точки обучающей выборки - это называется **переобучение**.

Для каждой задачи нужно найти свою **оптимальную** модель - не слишком простую и не слишком сложную.

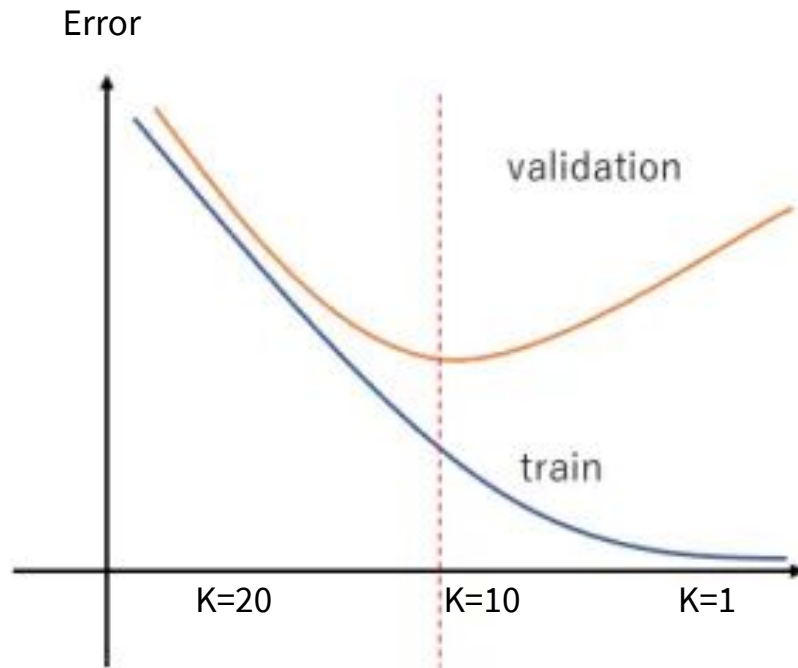


Сложность модели

Сложность модели регулируется внешними параметрами.

Например, в KNearestNeighbors сложность регулируется параметром K

Какой K самый лучший? Подскажет кросс-валидация



Summary



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

