Введение в ML

МАДМО Базовый, 2024

Что такое ML?

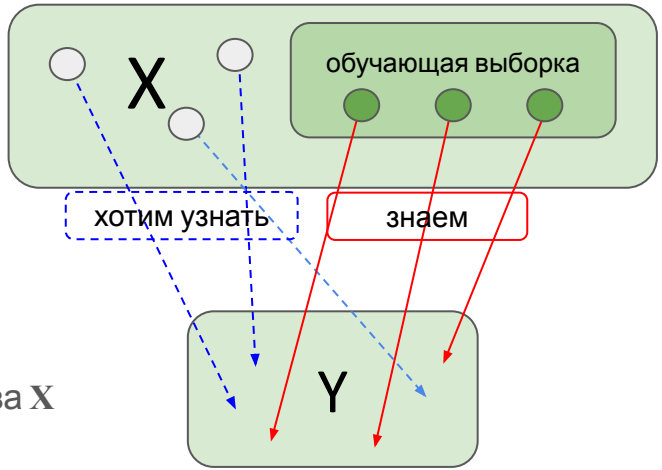
Машинное обучение изучает методы построения алгоритмов, способных обучаться - выявлять общие закономерности по частным эмпирическим данным.

`http://www.machinelearning.ru`

Более формально

- X — все множество объектов
  (данные, с которыми работаем)
- Y — все множество ответов
  (то, что хотим научиться предсказывать)
- $y: X \to Y$ — неизвестная закономерность

Дано:
обучающая выборка, $\{x1, x2, \ldots, xn\}$ — подмножество множества X
Цель: подобрать алгоритм a, приближающий функцию $y(x)$



Как задаются объекты?

Объект x_i задается признаковым описанием (как набор своих признаков) - $\{f_i^1,...,f_i^k\}$.

Получаем набор из $n \times k$ значений f_i^k - их обычно представляют в виде матрицы “объекты-признаки”:

$$\begin{pmatrix} f_1^1 & f_1^2 & \cdots & f_1^k \\ f_2^1 & f_2^2 & \cdots & f_2^k \\ \vdots & \vdots & \ddots & \vdots \\ f_n^1 & f_n^2 & \cdots & f_n^k \end{pmatrix}$$

Какие бывают признаки?

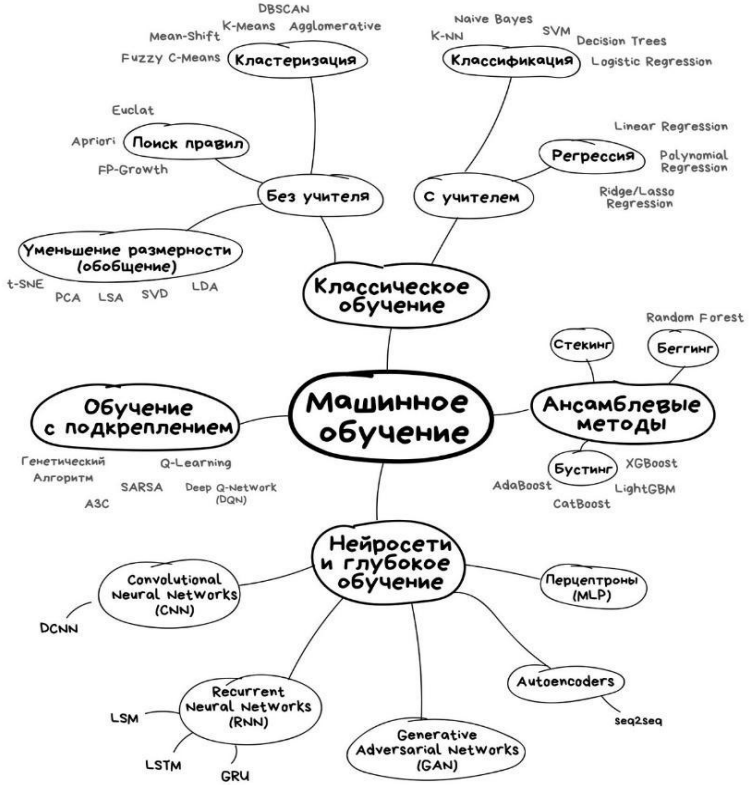
Выделяют три основных типа:

- **числовые** (действ. числа) - самые простые и понятные, компьютеры любят числа
- **бинарные** (0 и 1) - тоже отличный вариант, но есть некоторые нюансы
- **категориальные** (элементы множества) - обычно переводят в числовые или бинарные

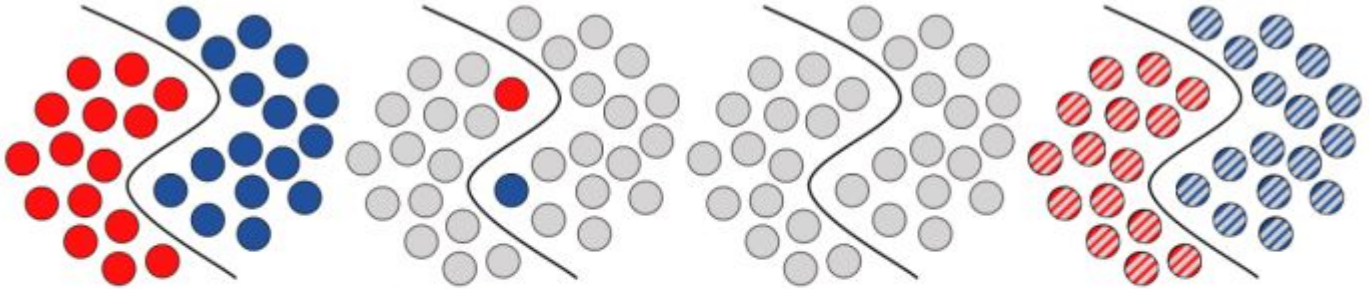
Задачи ML

На картинке показаны лишь основные задачи ML.

В каждой ветке есть и более узкие задачи.



Типы обучения



(a) Supervised (b) Semi-Supervised (c) Unsupervised (d) Self-Supervised

Figure 2: Illustrations of the four presented deep learning strategies - The red and dark blue circles represent labeled data points of different classes. The light grey circles represent unlabeled data points. The black lines define the underlying decision boundary between the classes. The striped circles represent datapoints which ignore and use the label information at different stages of the training process.

Supervised Learning

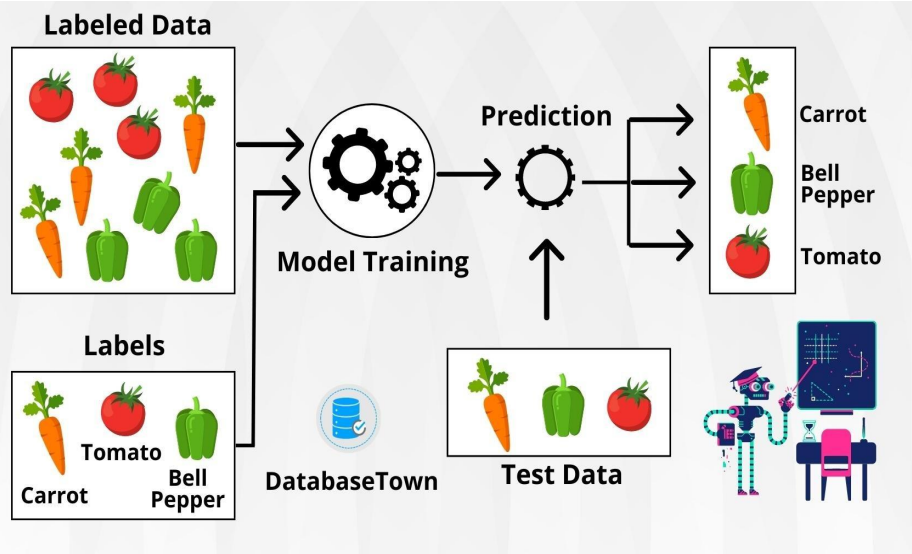
Обучающая выборка - пары (x, y), где:

- x – описание объекта
- y – его целевое значение (число или метка)

Учим модель $y = f(x)$ - целевое значение по описания

Типичные задачи:

- классификация
- регрессия
- ранжирование



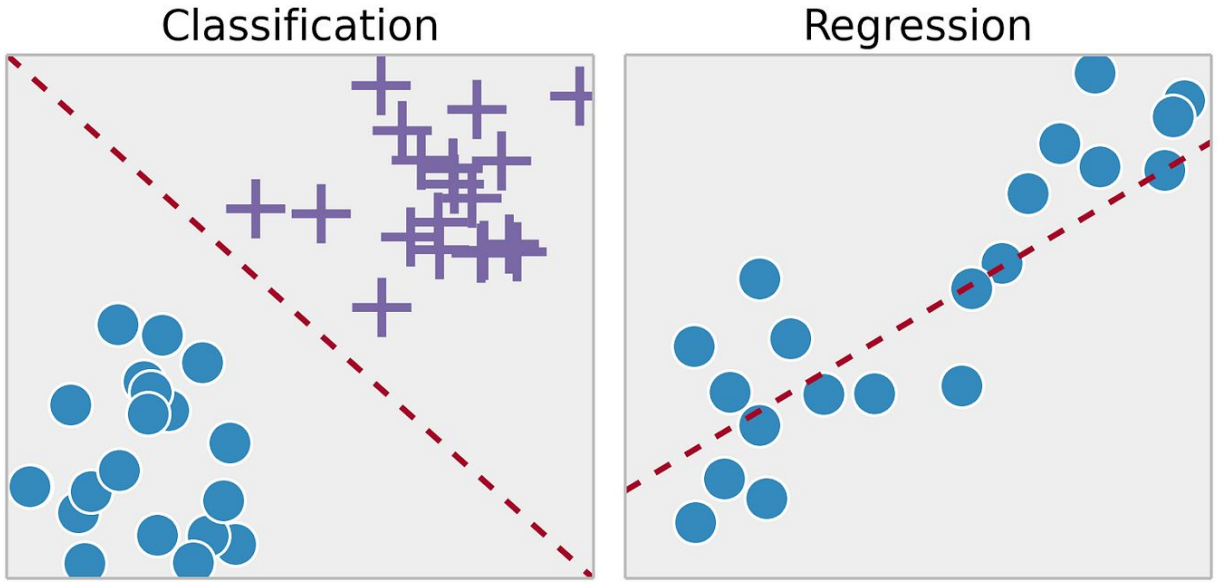
Классификация и регрессия

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

- Числовая
- Бинарная
- Категориальная

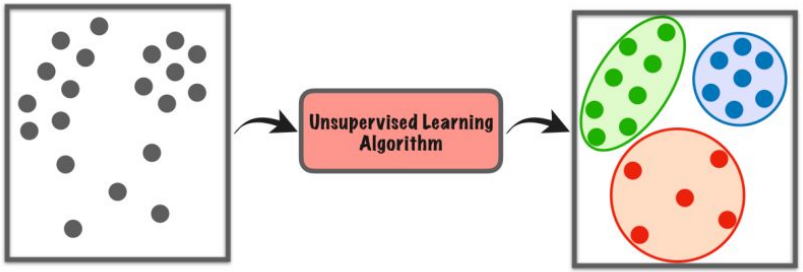
Предсказание числового значения называется **регрессия**.
Предсказание одного из нескольких классов называется **классификация**.

Классификация и регрессия



Unsupervised Learning

Обучающая выборка - x (только объекты, без меток),
Хотим эффективно описать объекты в пространстве описа



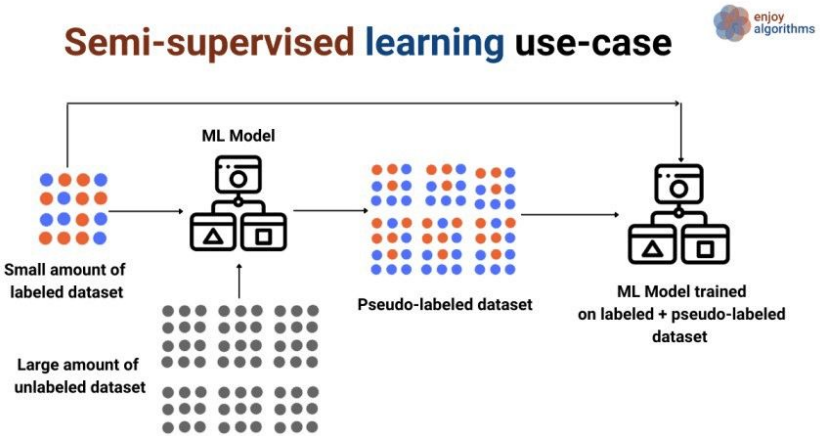
Типичные задачи:

- кластеризация
- понижение размерности
- детектирование аномалий
- оценка плотности

Semi-Supervised Learning

Обучающая выборка - данных с метками и без меток (последних, как правило, существенно больше).

Хотим модель $y=f(x)$, но с учетом информации о том, как объекты располагаются в пространстве описаний



Self-Supervised Learning

Обучающая выборка - неразмеченные данные + псевдо-метка
(ее создали мы)

- **Зачем?**
  Решить полученную SL-задачу и получить представление (representation) объектов
- **Зачем?**
  Это представление используют в SL-задачах с похожими данными (downstream task)
- **Зачем?**
  Размеченных данных мало, хотим использовать и неразмеченные

Данные

Какие бывают данные?

Простые, например, табличные данные

Датасет Titanic

- Таблица с пассажирами Титаника
- Десятки столбцов, множество строк
- Проще всего работать

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

Какие бывают данные?

Сложнее! Например, картинки.

Датасет MNIST

- Ч/б изображения рукописных цифр
- ~50к изображений 28 на 28 пикселей
- Для каждой картинки известна цифра, изображенная на ней

Общие шаги

Нужно выделить признаки объектов!

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

Нужно привести все признаки к числам!

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

Для каждого объекта набор его признаков собирается в один вектор

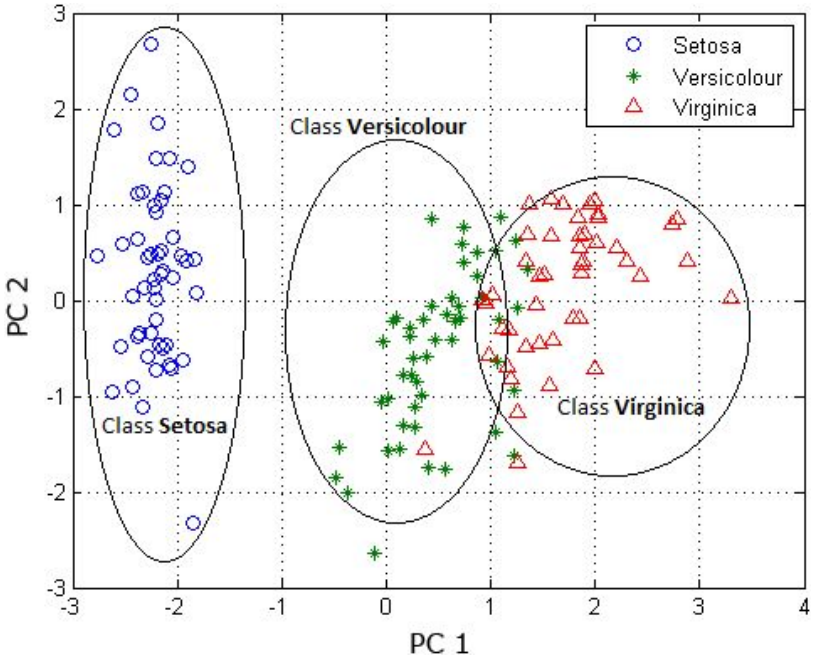
Визуализация данных

Нужно познакомиться с данными,
проще всего с помощью визуализаций!

Пример - датасет Iris:

- график зависимости 2х признаков,
  разные метки - разные классы (сорты ирисов)
- видим явные кластеры!
- что можно сказать о полезности признаков?

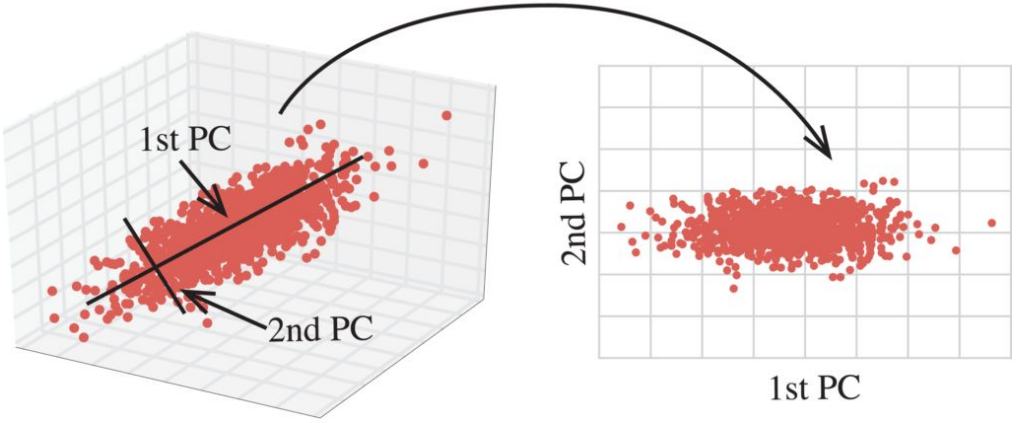
А как смотреть на большее количество признаков?



Понижение размерности

Данные в многомерном пространстве расположены неслучайно,
есть зависимости, значит можно представить меньшим количеством признаков!

Пример - проекция данных из 3D в 2D:

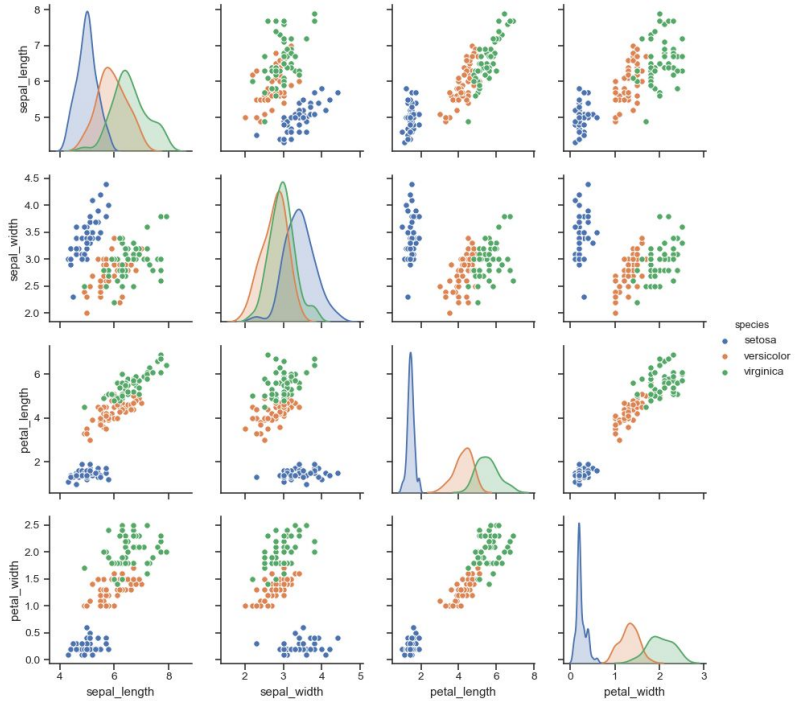


Попарные графики

Уменьшили кол-во признаков?
Можем смотреть попарно!

Для каждого объекта оставляем признаки i и j,
затем рисуем точку с их значениями на плоскости.

Разные классы = разные цвета

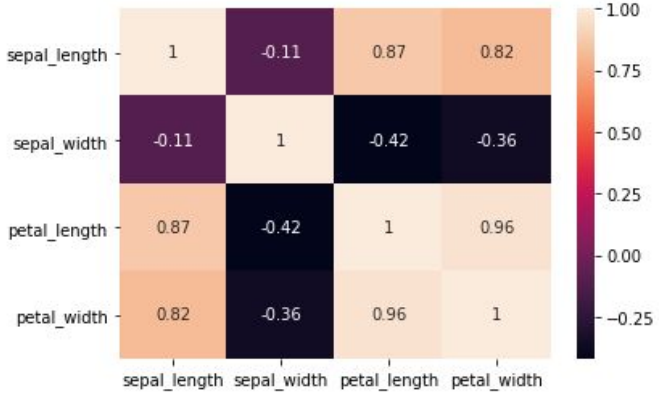


Матрица корреляции

Есть ли у нас (линейно) зависимые признаки?
(Линейно) зависимые признаки - плохо,
т.к. дают степень свободы, но не дают доп. информацию.

Для поиска поможет подсчет попарной корреляции признаков!

Визуализируем с помощью матрицы корреляции!



Модель

Модели в SL

Recap:

- Между объектом и целевой переменной **существует реальная зависимость**
- У нас есть **только N примеров** 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.

Модели и данные

Выбор модели обусловлен входными данными:

- **Табличные** -> классические (линейные методы, SVM, наивный Байес, деревья, леса и ансамблевые методы), градиентный бустинг, простейшие полносвязные нейросети
- **Временные ряды** (+звук) -> классические (Хольта–Уинтерса, ARIMA/SARIMA, линейные методы, деревья), рекуррентные нейросети (LSTM, GRU)
- **Изображения** -> сверточные сети, трансформеры
- **Текст** -> рекуррентные нейросети (LSTM), но обычно архитектуры на основе трансформеров

Метрики

Измерение качества

Имеем 2 модели. Как понять какая лучше?

Нужно оценить качество!

Метрика - это функция вида: $metric(y, \hat{y})$

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

а $\hat{y} = a(x)$ - 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

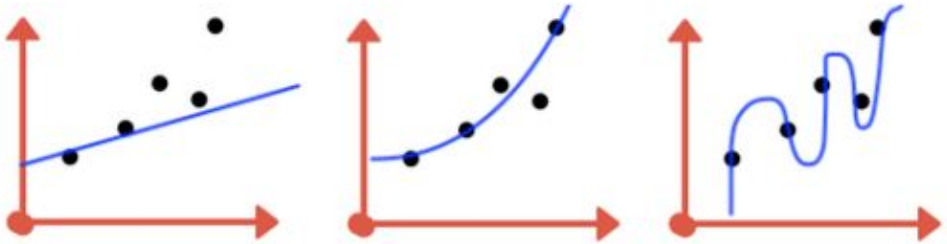
Регрессии:

- MSE - средний квадрат отклонения
- RMSE - стандартное отклонение
- MAE - средний модуль отклонения
- R^2 - коэффициент детерминации

Более подробно метрики будут рассмотрены на след. занятиях.

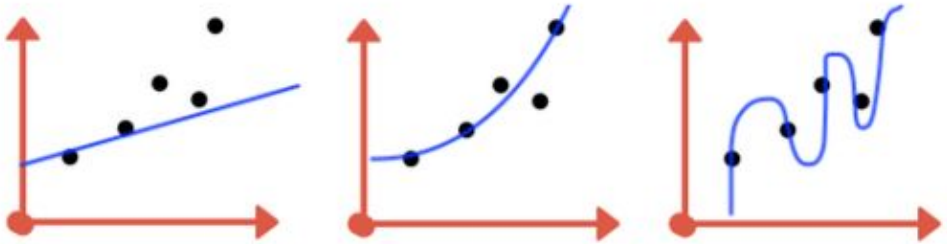
Валидация

Несмещенная оценка



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

Несмещенная оценка

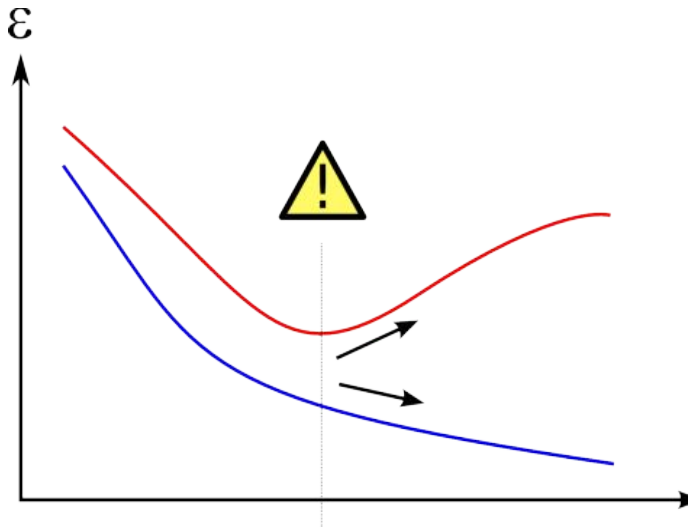


Тестируем на обучающей выборке -> получаем смещенную оценку.
Тогда “самая лучшая” модель = это та, которая просто запомнила все данные.

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

Переобучение

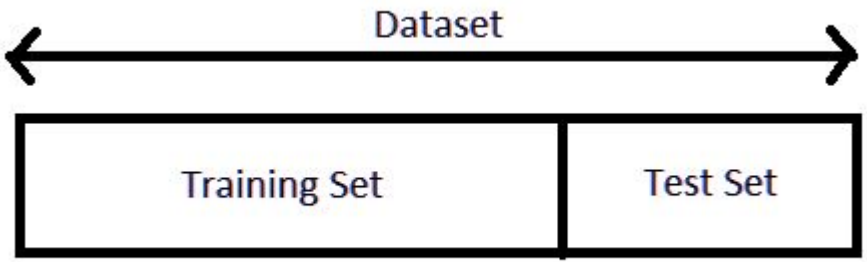
Переобучение - метрики на обучающей выборке хуже, чем на тестовой.



Отложенная выборка

Выход: можно “отложить” часть обучающей выборки для валидации модели. Например, использовать 80% выборки для обучения и 20% для тестирования.

Оценка на тестовой выборке будет несмещенной!



Отложенная выборка

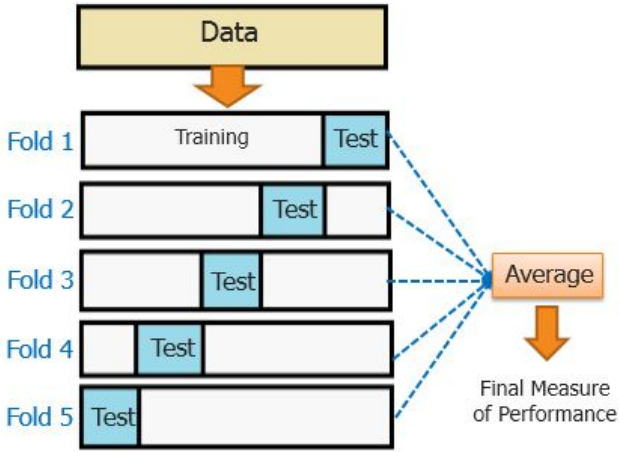
Недостатки:

- Уменьшение размера обучающей выборки может негативно сказаться на качестве
- Малый размер тестовой выборки может давать сильное смещение оценки
- Можно переобучиться под тестовую выборку
- Тестовая выборка маленькая - оценка будет иметь погрешность

Кросс-валидация

Можем подогнаться под одну отложенную выборку?
Сделаем больше!

- Разбиваем выборку на k частей
- k-1 частей - для обучения и одна - для теста
- Процесс повторяется k раз - тестируем на разных частях
- Результаты тестирования усредняются



Кросс-валидация

Плюсы:

- Погрешность оценки уменьшается - используем весь набор
- Качество измеряется на всем наборе данных
- Качество не зависит от выбора конкретного тестового набора
- Сложнее переобучиться под тест

Минусы:

- Обучение производится k раз - это может быть очень долго