Введение в машинное обучение

План

1. Введение

2. Основные понятия

3. Типы задач и примеры

Введение

Организационные моменты

Чат в телеграмме

Преподаватель - Ильдар Сафило @Ildar_Saf

Ассистент - Петр Григорьев @petragrig

Ссылка на гитхаб https://github.com/irsafilo/HSE/tree/main/lect_sem1

За каждое дз можно получить 10 баллов

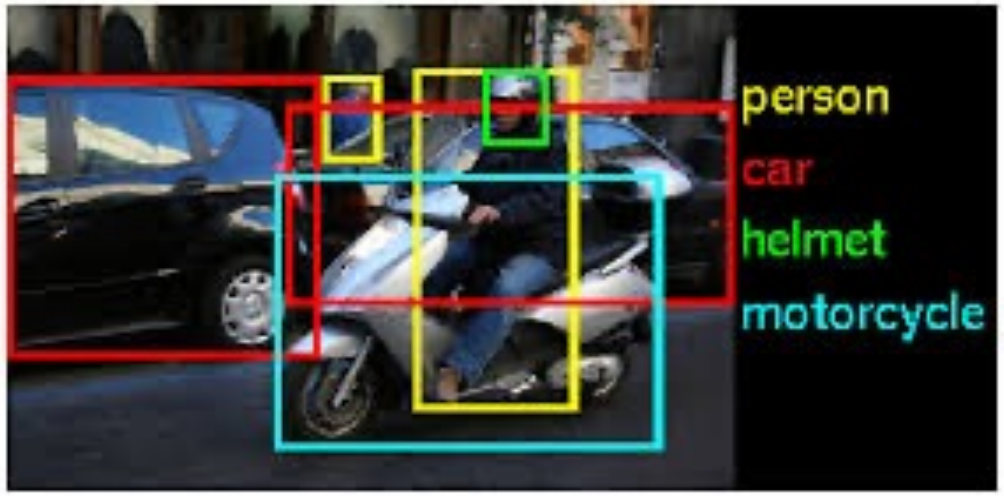
Оценка за курс = 0.7 * ДЗ + 0.3 * Контрольная работа

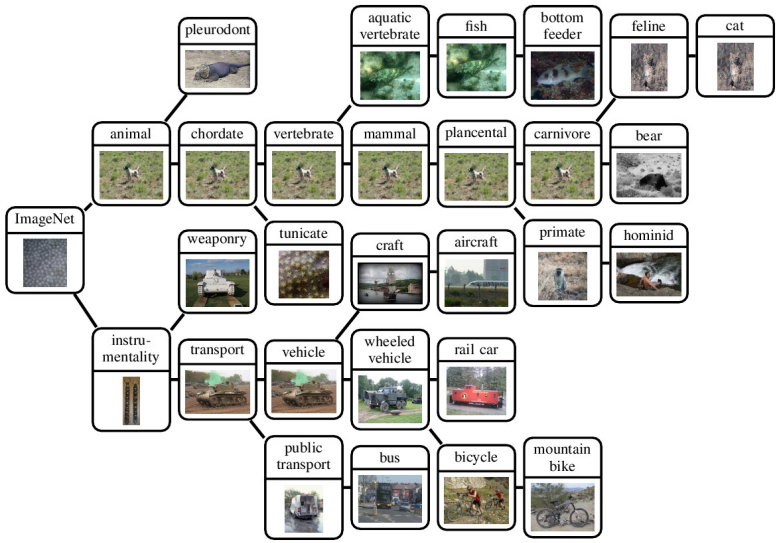
Зачем мл дата-аналитику?

DeepFake

Что такое машинное обучение?

- Распознавание обьектов на изображении
- Нейронные сети могут решать эту задачи точнее, чем люди
- Популярная модель - ImageNet



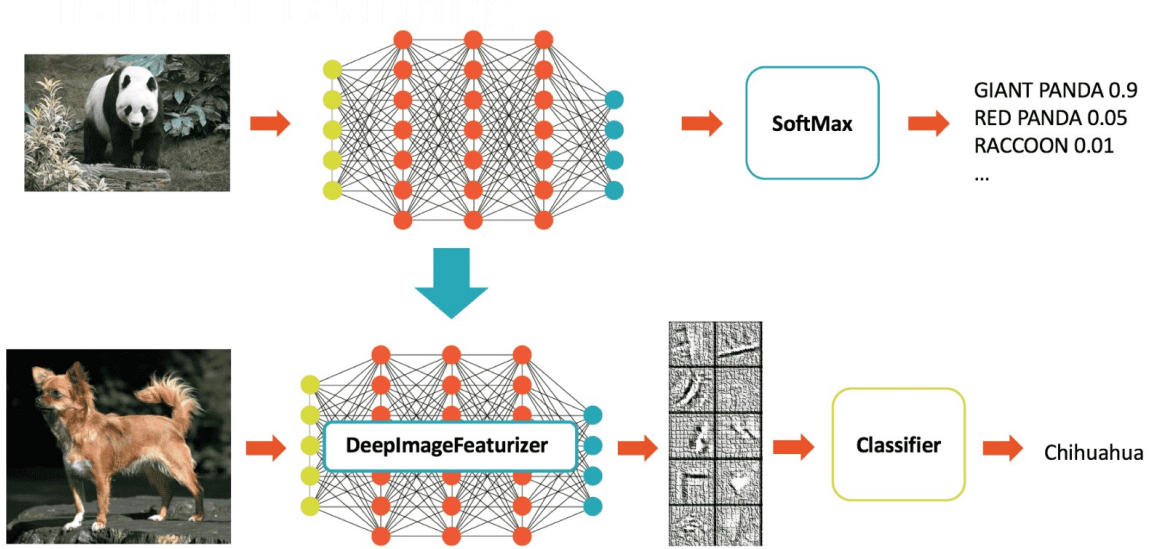


Что такое машинное обучение?

- Аннотирование изображений

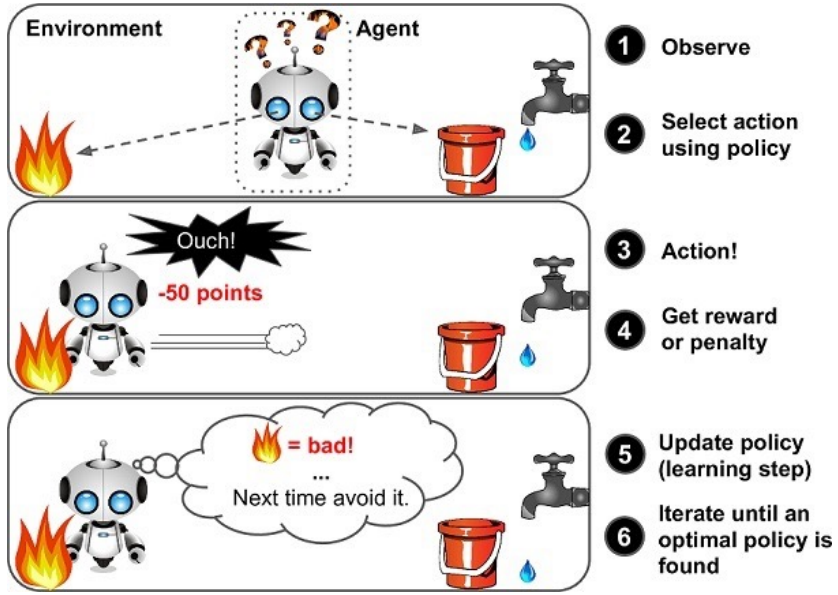
Sia Kate Isobelle Furler (/ˈsiə/ SEE-ə; born 18 December 1975) is an Australian singer, songwriter and music video director.[1] She started her career as a singer in the acid jazz band Crisp in the mid-1990s in Adelaide. In 1997, when Crisp disbanded, she released her debut studio album titled OnlySee in Australia. She moved to London, England, and provided lead vocals for the British duo Zero 7. In 2000, Sia released her second studio album, Healing Is Difficult, on the Columbia label the following year, and her third studio album, Colour the Small One, in 2004, but all of these struggled to connect with a mainstream audience.

Sia relocated to New York City in 2005 and toured in the United States. Her fourth and fifth studio



Что такое машинное обучение?

- Reinforcment learning – нейронные сети могут играть в игры и почти во все обыгрывать человека

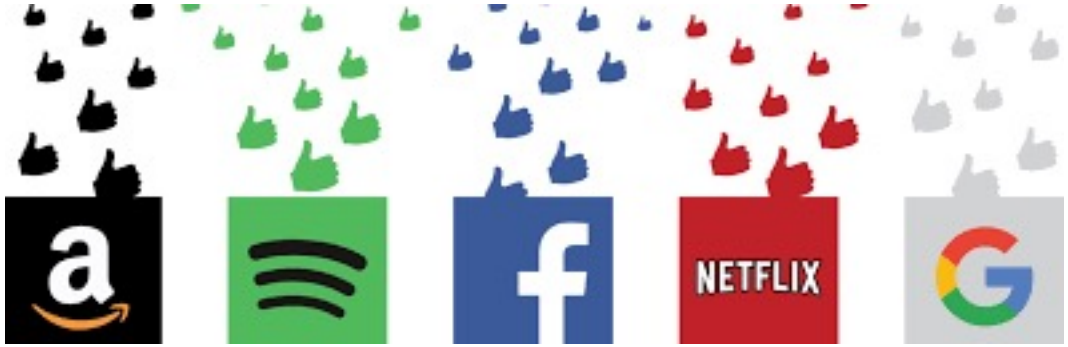


Примеры

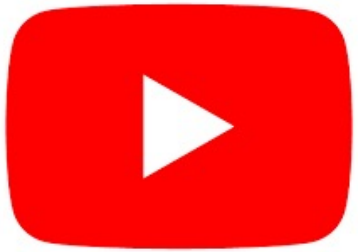
- Любой голосовой помощник использует ai
- Задачи NLP – одни из самых популярных сегодня

Примеры

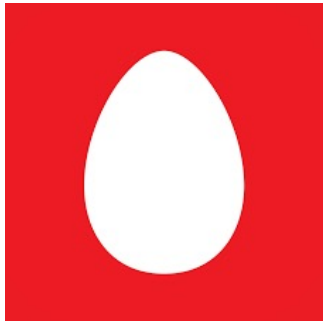
- Персонализация и рекомендательные системы – новый тренд в мл



Кто использует машинное обучение?



amazon



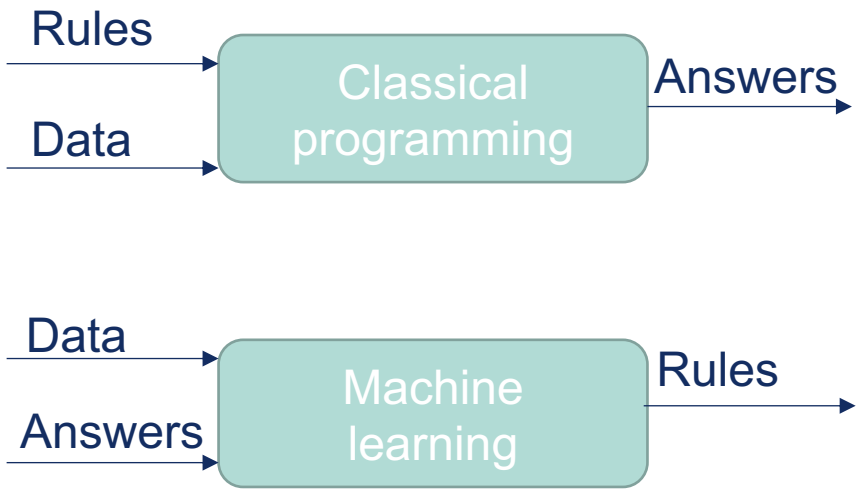
Какие бизнес-задачи решает мл?

- Какой фильм будет наиболее релевантным для пользователя?
- Вернет ли клиент кредит?
- Кому какой баннер с рекламой показать?
- Кто из клиентов уйдет в отток?
- Сколько потратит пользователь в следующие 3 месяца?
- Сколько молока купят клиенты завтра в магазине?
- Кому какой пуш отправить?

Что такое машинное обучение?

Машинное обучение – набор способов воспроизведения связей между событиями
И результатом

Машинное обучение – обширный подраздел искусственного интеллекта, изучающий методов
Построения алгоритмов, способных обучаться



Основные понятия

Пример: задача о ресторанах

- Сеть ресторанов
- Хотим открыть еще один
- Есть несколько вариантов размещения
- Какой из вариантов принесет максимальную прибыль?

https://www.kaggle.com/c/restaurant-revenue-prediction

Обозначения

- X– множество обьектов (характеристики ресторанов)
- Y – множество ответов, целевая переменная, target (прибыль от каждого ресторана)
- $a : X \to Y$ – неизвестная зависимость
- Какой из вариантов принесет максимальную прибыль?

Обучающая выборка

Дано:

- $\{x_1, \dots, x_n\} \subset X$ – обучающая выборка
- $\{y_1, \dots, y_n\}, y_i = y(x_i)$ – известные ответы

Найти:

- $a : X \to Y$ – алгоритм (решающая функция), приближающая

y на всем множестве X

Признаки

- Признаки обьекта x можно записать в виде вектора $(f_1(x), \dots, f_d(x)) = x_i$
- d – количество признаков
- Признаки/features/факторы в модель передаются числами

Feature / Case No.	Accidents/experience	Weight/Height	Optical status	Hearing ability	General health	Adherence to safety	Education	Overtime work
S1	1/12	71/160	6/18	5	5	4	5	2
S2	5/12	77/170	6/60	5	4	3	5	3
S3	1/21	90/175	6/6	4	5	5	5	5
S4	0/8	54/165	6/60	4	5	2	5	20
S5	6/3.5	68/187	6/36	3	4	3	5	0
S6	10/11	85/177	6/6	4	4	3	5	10
S7	2/19	76/173	6/60	4	5	5	5	14
S8	18/25	72/170	6/18	3	4	4	4	4
S9	45/14	80/176	6/6	4	4	3	4	8
S10	3/20	81/174	6/6	4	4	5	4	1

Признаки пример

- Демографические:
  - Средний возраст жителей ближайших кварталов
  - Количество жителей

- Про недвижимость:
  - Средняя стоимость квартир
  - Кол-во магазинов/школ/банков
  - Кол-во конкурентов

- Про дороги:
  - Среднее кол-во машин, проезжающих мимо за день
  - Кол-во пешеходов, проходящих мимо точки

Функционал/мера качества

- Как понять, какой алгоритм полезен ?
- $a(x) = 0$ – полезен ли такой алгоритм
- Примеры
  - Среднеквадратичная ошибка (MSE):
    - $Q(a, X) = \frac{1}{n}\sum_{i=1}^{n}(a(x_i) - y_i)^2$
  - Среднеабсолютная ошибка (MAE):
    - $Q(a, X) = \frac{1}{n}\sum_{i=1}^{n}|a(x_i) - y_i|$
  - Доля правильных ответов – для классификации
    - $accuracy(a, X) = \frac{1}{l}\sum_{i=1}^{l}[a(x_i) = y_i]$

Функция потерь

Функция потерь – функция, измеряющая ошибкуна одном Обьекте

- y – истинный ответ на обьекте x
- $a(x)$ - предсказание алгоритма на обьекте x

Как измерить ошибку предсказания?

Пример (квадратичная функция потерь):

$$L\big(y, a(x)\big) = (a(x) - y)^2$$

Алгоритм

- $a(x)$ — алгоритм, модель — функция, предсказывающая ответ для любого объекта

- Отображает X в Y

- Например, линейная модель: $a(x) = w_0 + w_1 * x_1 + \ ... + w_d * x_d$
  $a(x) = 1.000.000 + 100.000 \ *$ расстояние до конкурента – (расстояние до метро) * 100.000

Как получить предсказания?

В задачах обучения с учителем (обучение по прецедентам) всегда два этапа:

- Этап обучения(training):
  по выборке $X = \{(x_i, y_i)\}$ строим алгоритм $a(x)$

- Этап применения(testing):
  алгоритм a для новых обьектов x_i выдает ответы $a(x_i)$

Определения

- **Признаки, факторы (features)** - количественные характеристики обьекта

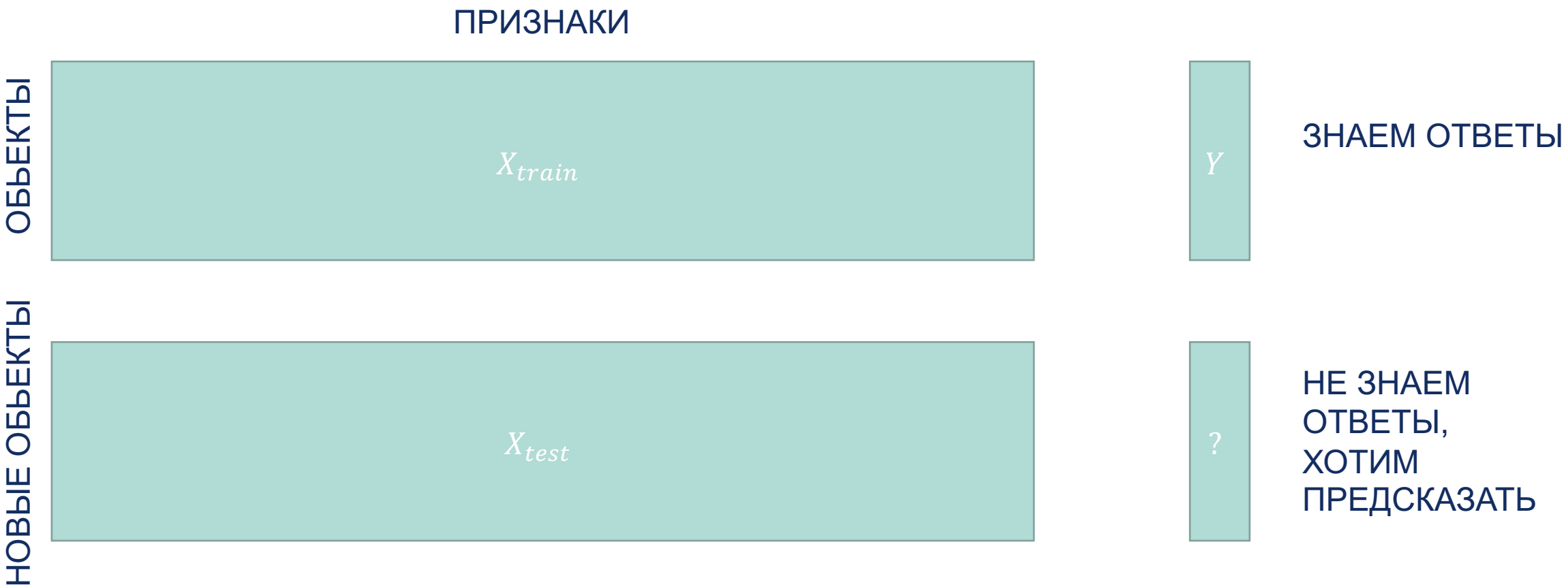
- **Обучающая выборка (training set)** – конечный набор обьектов, для которых известны значения целевой переменной

Пример: набор ресторанов, открытых более года назад, для которых известна их прибыль За первый год

Признаки описывают обьекты с помощью чиселз

Специалист по анализу данных не является экспертом в предметной области – вся необходимая информация содержится в обучающей выборке. Эксперты нужны при формировании признаков.

Стандартная постановка задачи



Обучение алгоритма

- Есть обучающая выборка и функционал качества
- Семейство алгоритмов $\mathcal{A}$
  - Из чего выбираем алгоритм
  - Пример: все линейные модели
  - $\mathcal{A} = w0 + w1x1 + \cdots + wdxd\ w0, w1, \ldots, wd \in \mathbb{R}\}$
- Обучение: поиск оптимального алгоритма с точки зрения функционала качества

Обучение алгоритма

Параметры w_0, w_1, w_2 подбираются так, чтобы на них достигался Минимум функции потерь:

$$Q(a, X) = \frac{1}{l}\sum_{i=1}^{l}(w_0 + w_1 x_1 + w_2 x_2 - y_i)^2 \rightarrow \min(w_0, w_1, w_2)$$

Процесс поиска оптимального алгоритма называется обучением

Обучение алгоритма

Предположим, что мы хотим предсказать стоимость квартиры (y)
по его площади (x_1) и количеству комнат (x_2)

Как правило, алгоритм $a(x)$ выбирают из некоторого семейства алгоритмов A

Используем линейную модель для предсказания стоимости.
Она будет выглядеть так:

$$a(x) = w_0 + w_1 x_1 + w_2 x_2,$$

где w_0, w_1, w_2 - параметры модели(веса)
Общий вид линейных моделей:

$$A = \{a(x) = w_0 + w_1 x_1 + \ \ldots + w_d x_d | w_0, w_1, \ldots, w_d \in R$$

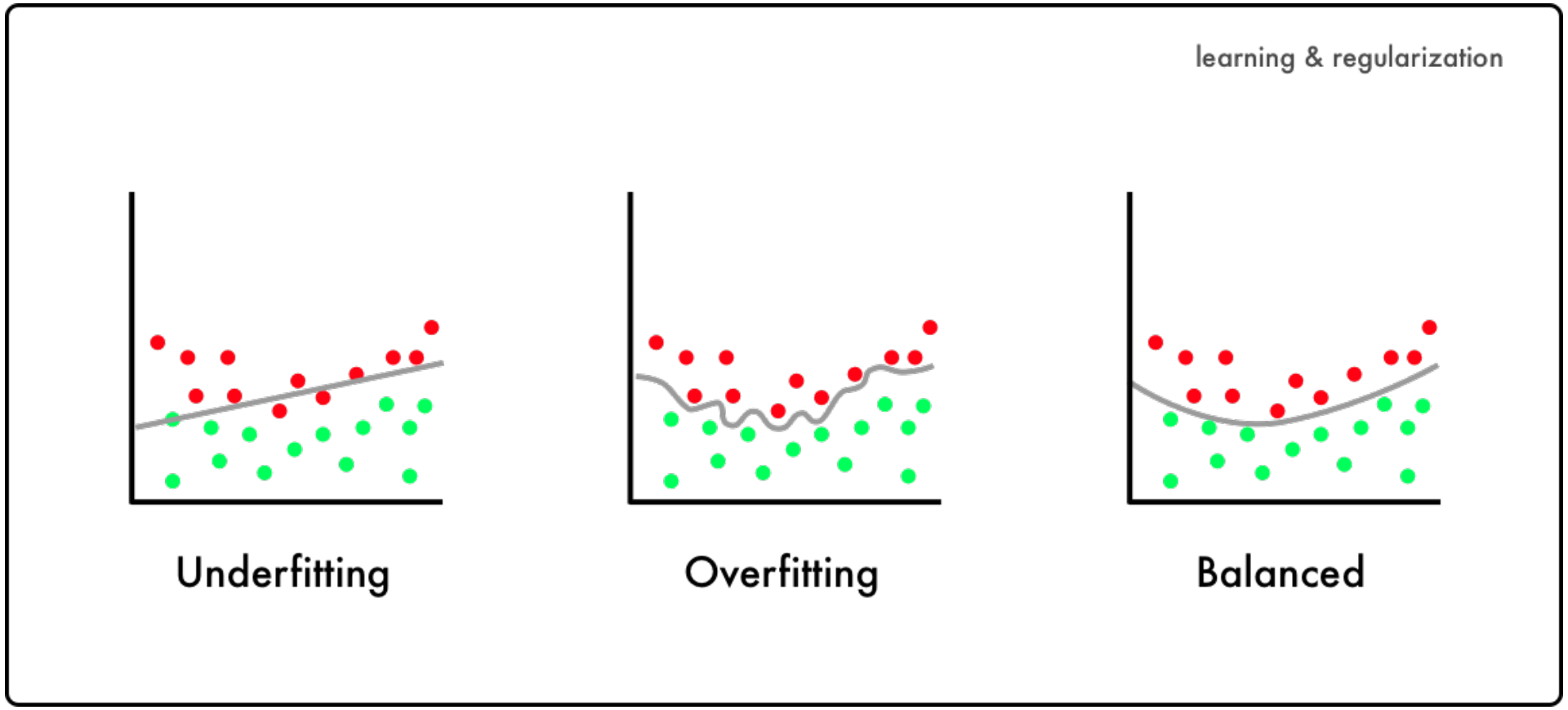
Функционал ошибки:

$$Q(a, X) = \frac{1}{l} \sum_{i=1}^{l} (a(x_i) - y_i)^2$$

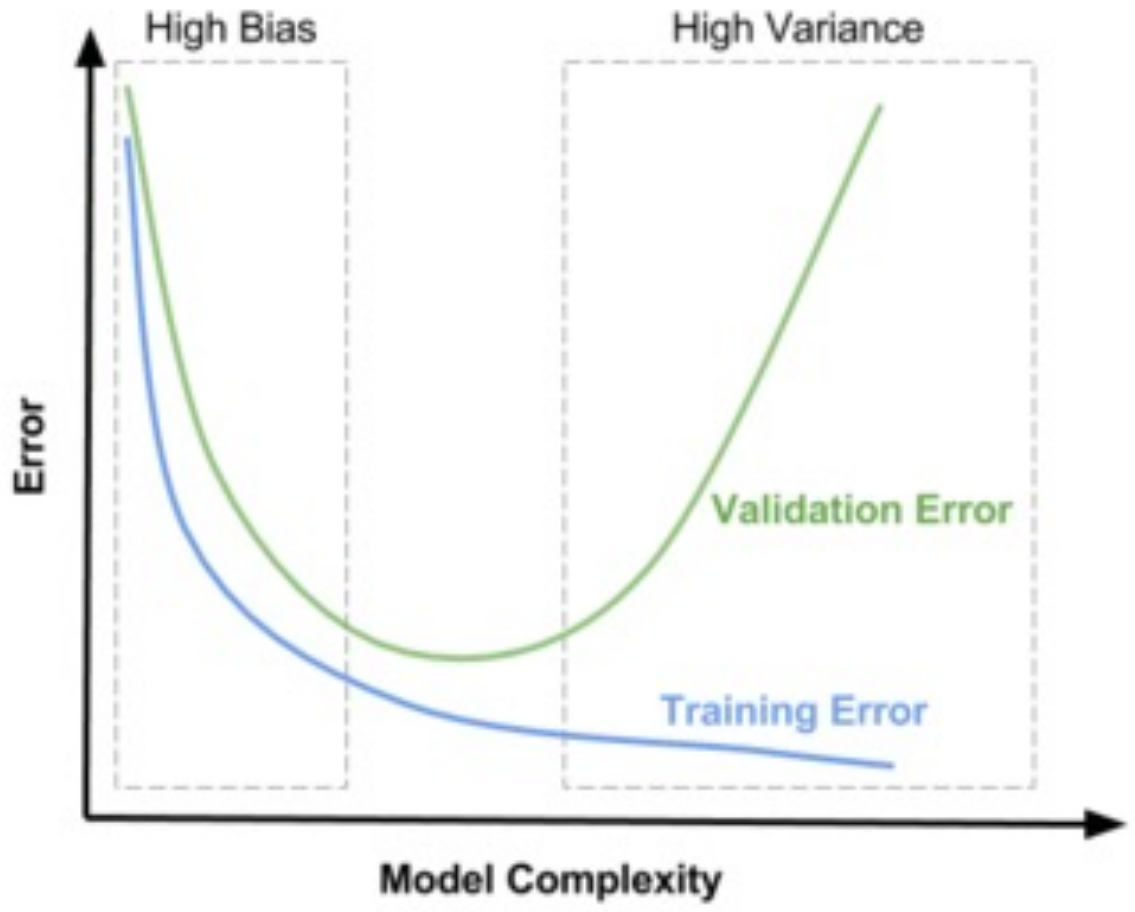
В нашем случае:

$$Q(a, X) = \frac{1}{l} \sum_{i=1}^{l} (w_0 + w_1 x_1 + w_2 x_2 - y_i)^2$$

Переобучение алгоритма



Переобучение алгоритма



Переобучение алгоритма

- Избыточная сложность пространства параметров, лишние степени свободы в модели $a(x, w)$ идут на подгонку под обучающую выборку
- Переобучение есть всегда, когда есть оптимизация параметров по конечной(заведомо неполной) выборке
- Если качество на отложенной выборке сильно ниже качества на обучающих данных, то происходит переобучение

Типы задач и примеры

Виды признаков

- Числовые
- Бинарные
- Категориальные (название города, марка машины)
- Признаки со сложной внутренней структурой (изображение)

Виды данных

- Таблицы
- Текстовые данные
- Изображения
- Звук
- Логи

Большинство алгоритмов машинного обучения работает с числовыми данными,
Поэтому все виды данных необходимо переводить в числа

Типы задач в зависимости от целевой переменной

Классификация

- $Y = \{0,\ 1\}$ – классификация на 2 класса
- $Y = \{1, \dots, M\}$– классификация на M непересекающихся классов
- $Y = \{0, 1\}^M$ - классификация на M классов, которые могут пересекаться

Примеры задач классификации

- Задача медицинской диагностики
- Задача кредитного скоринга
- Задача предсказания оттока клиентов
- Предсказание бинарного поведения пользователя
- Классификация изображений

Примеры задач классификации

- Мультиклассовая классификация

- Определение наиболее подходящей профессии для данного кандидата

Типы задач в зависимости от целевой переменной

Классификация

- $Y = \{0,\ 1\}$ – классификация на 2 класса
- $Y = \{1, \dots, M\}$– классификация на M непересекающихся классов
- $Y = \{0, 1\}^M$ - классификация на M классов, которые могут пересекаться

Регрессия

- $Y = R$
- $Y = R^n$

Примеры задач регрессии

- Предсказание стоимости недвижимости (стоимость квартиры квартиры в Москве)
- Предсказание прибыли рестораны
- Предсказание поведения временного ряда в будущем (стоимость акций)
- Предсказание зарплаты выпускника вуза по его оценкам

Типы задач в зависимости от целевой переменной

Классификация

- $Y = \{0,\ 1\}$ – классификация на 2 класса
- $Y = \{1, \dots, M\}$– классификация на M непересекающихся классов
- $Y = \{0, 1\}^M$ - классификация на M классов, которые могут пересекаться

Регрессия

- $Y = R$
- $Y = R^n$

Ранжирование

- $\boldsymbol{Y}$ – конечное упорядоченное множество

Примеры задач ранжирования

- Вывести подходящие запросу документов в порядке уменьшения релевантности
- Вывести кандидатов на должность в порядке уменьшения релевантности

Задачи без целевой переменной

- Кластеризация – задача разделения обьектов на группы,
  при этом целевые переменные для обьектов неизвестны
(или не существуют). Разделение происходит только на основе
 признаковых описаний обьектов

Примеры задач кластеризации

- Разбить пользователей на группы, внутри каждой из которых будут похожие пользователи
- Разбить тесктовые документы на группы по похожести документов

Задачи без целевой переменной

- Кластеризация – задача разделения обьектов на группы, при этом где целевые переменные для обьектов неизвестны (или не существуют). Разделение происходит только на основе признаковых описаний обьектов

- Понижение размерности – задачи генерации новых признаков (по числу меньше, чем старых), так что с помощью их задача решается не хуже чем с исходными

Задачи без целевой переменной

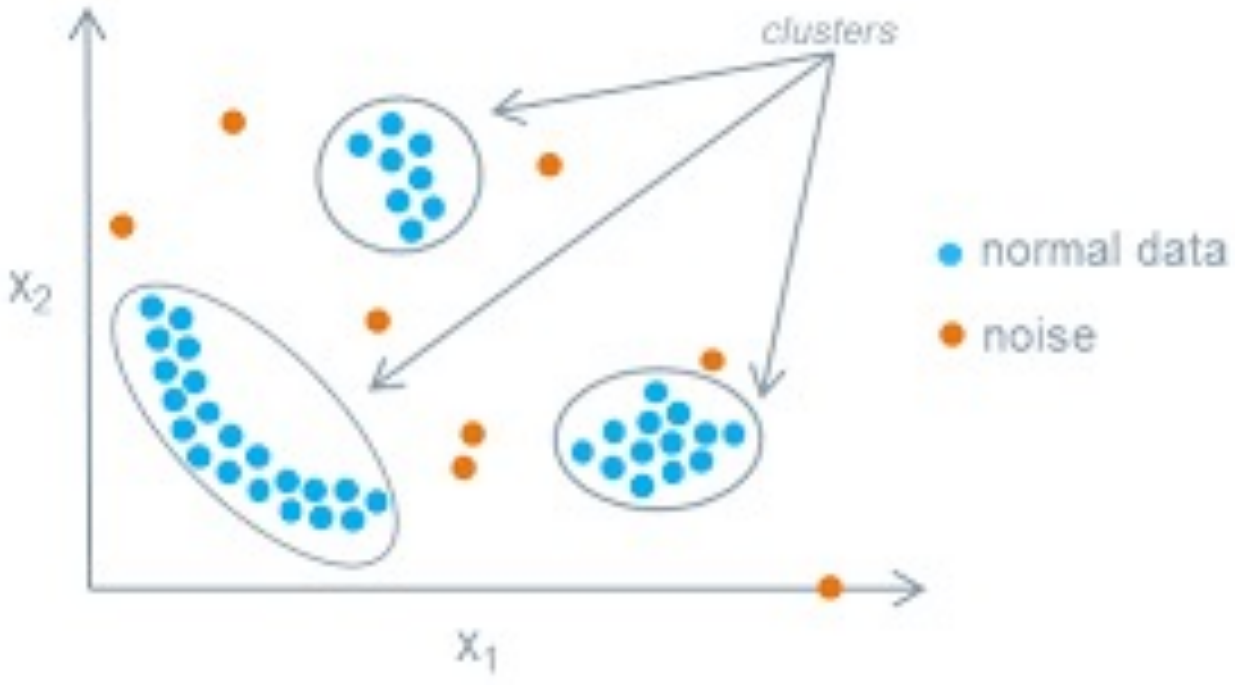
- Кластеризация – задача разделения обьектов на группы,
  при этом где целевые переменные для обьектов неизвестны
(или не существуют). Разделение происходит только на основе
 признаковых описаний обьектов

- Понижение размерности – задачи генерации новых признаков
  (по числу меньше, чем старых), так что с помощью их задача
  решается не хуже чем с исходными

- Оценивание плотности – задача приближения распределения обьектов

Пример оценивая плотности

- Поиск аномалий с помощью оценивания плотностей



Задачи без целевой переменной

- Кластеризация – задача разделения обьектов на группы,
  при этом где целевые переменные для обьектов неизвестны
(или не существуют). Разделение происходит только на основе
 признаковых описаний обьектов

- Понижение размерности – задачи генерации новых признаков
  (по числу меньше, чем старых), так что с помощью их задача
  решается не хуже чем с исходными

- Оценивание плотности – задача приближения распределения обьектов

- Визуализация – задача изображения многомерных обьектов
  в 2х или 3х мерном пространстве с сохранением зависимостей
  между ними

Обучение с учителем

- Если нам известны значения целевой переменной, то есть алгоритм обучается так, чтобы правильно предсказывать целевую переменную – это обучение с учителем :
  - Классификация
  - Регрессия
  - Ранжирования

Обучение без учителя

- Если нам неизвестны значения целевой переменной или целевая переменная вообще отсутствует, то есть алгоритм обучается только по признакам обьектов, то это обучение без учителя. Примерами обучения с учителем являются кластеризация, понижение размерности и др.

Алгоритм решения задачи

1. Постановка задачи
2. Выделение признаков
3. Формирование выборки
4. Выбор функции потерь и метрики качества
5. Предобработка данных
6. Построение модели
7. Оценивание качества модели

Спасибо за внимание!

Ildar Safilo

@Ildar_Saf
irsafilo@gmail.com
https://www.linkedin.com/in/isafilo/