Лекция 1
Введение в машинное обучение

Кантонистова Елена Олеговна

elena.kantonistova@yandex.ru

ekantonistova@hse.ru

ВШЭ, 2022

ПЛАН РАССКАЗА

- Про курс
- Введение в машинное обучение
- Основные понятия
- Типы задач
- Обучение и оценка качества модели
- Структура реального проекта по ML
- Математика в машинном обучении

ПРО КУРС

Команда курса:

- Лекции: Елена Кантонистова @murr4a
- Семинары: Максим Некрашевич @MaxNekrashevich
- Ассистенты: Даша Саламашенкова @salamashenkovadasha

  Слава Овчинников @Mega_Serega

ПРО КУРС

- **16 лекций и 16 семинаров:**

- Линейные модели классификации и регрессии
- Работа с данными
- Деревья и композиции моделей
- Кластеризация
- Снижение размерности
- Введение в глубинное обучение
- Анализ временных рядов

ПРО КУРС

- 16 лекций и 16 семинаров
- **6 домашних заданий**
- **бонусный проект (весит как 3 домашних задания)**

ПРО КУРС

- 16 лекций и 16 семинаров
- 6 домашних заданий
- бонусный проект (весит как 3 домашних задания)
- **12 пятиминуток (на семинарах)**
- **коллоквиум (письменный)**
- **экзамен (письменный, на отл – ещё устная часть)**

ПРО КУРС

Информация о курсе:

- Канал курса: https://t.me/ML_math
- Github с материалами: https://github.com/Murcha1990/ML_math_2022
- Вики-страничка (скоро будет)

ИСТОРИЯ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

1956 ГОД — ПЕРВЫЙ СЕМИНАР ПО ПРОБЛЕМАМ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

Задача: моделирование интеллекта человека математическими методами

ИСТОРИЯ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

1956 ГОД — ПЕРВЫЙ СЕМИНАР ПО ПРОБЛЕМАМ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

Задача: моделирование интеллекта человека математическими методами

50-Е – 70-Е ГОДЫ — ПРОСТЕЙШИЕ СИСТЕМЫ ИИ

Системы дедукции для доказательства теорем

Робот-психотерапевт ELIZA

ИСТОРИЯ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

1956 ГОД — ПЕРВЫЙ СЕМИНАР ПО ПРОБЛЕМАМ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

Задача: моделирование интеллекта человека математическими методами

50-Е – 70-Е ГОДЫ — ПРОСТЕЙШИЕ СИСТЕМЫ ИИ

Системы дедукции для доказательства теорем

Робот-психотерапевт ELIZA

80-Е ГОДЫ — РАЗВИТИЕ ЭКСПЕРТНЫХ СИСТЕМ

Моделирование работы эксперта

с помощью правил

ИСТОРИЯ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

1956 ГОД — ПЕРВЫЙ СЕМИНАР ПО ПРОБЛЕМАМ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

Задача: моделирование интеллекта человека математическими методами

50-Е – 70-Е ГОДЫ — ПРОСТЕЙШИЕ СИСТЕМЫ ИИ

Системы дедукции для доказательства теорем

Робот-психотерапевт ELIZA

80-Е ГОДЫ — РАЗВИТИЕ ЭКСПЕРТНЫХ СИСТЕМ

Моделирование работы эксперта

с помощью правил

Слишком много ручного труда

для создания системы

ИСТОРИЯ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

1956 ГОД — ПЕРВЫЙ СЕМИНАР ПО ПРОБЛЕМАМ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

Задача: моделирование интеллекта человека математическими методами

50-Е – 70-Е ГОДЫ — ПРОСТЕЙШИЕ СИСТЕМЫ ИИ

Системы дедукции для доказательства теорем

Робот-психотерапевт ELIZA

80-Е ГОДЫ — РАЗВИТИЕ ЭКСПЕРТНЫХ СИСТЕМ

Моделирование работы эксперта с помощью правил

Слишком много ручного труда для создания системы

90-Е ГОДЫ — РАЗВИТИЕ МАШИННОГО ОБУЧЕНИЯ КАК ОБЛАСТИ ИИ

Нейронные сети

Генетические алгоритмы

Автоматический поиск сложных закономерностей

ИСТОРИЯ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

1956 ГОД — ПЕРВЫЙ СЕМИНАР ПО ПРОБЛЕМАМ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

Задача: моделирование интеллекта человека математическими методами

50-Е – 70-Е ГОДЫ — ПРОСТЕЙШИЕ СИСТЕМЫ ИИ

Системы дедукции для доказательства теорем

Робот-психотерапевт ELIZA

80-Е ГОДЫ — РАЗВИТИЕ ЭКСПЕРТНЫХ СИСТЕМ

Моделирование работы эксперта с помощью правил

Слишком много ручного труда для создания системы

90-Е ГОДЫ — РАЗВИТИЕ МАШИННОГО ОБУЧЕНИЯ КАК ОБЛАСТИ ИИ

Нейронные сети

Генетические алгоритмы

Автоматический поиск сложных закономерностей

НАЧАЛО 21 ВЕКА — ГЛУБИННОЕ ОБУЧЕНИЕ (DEEP LEARNING)

Решение сложных задач распознавания с точностью, близкой к человеку

ЧТО ТАКОЕ МАШИННОЕ ОБУЧЕНИЕ

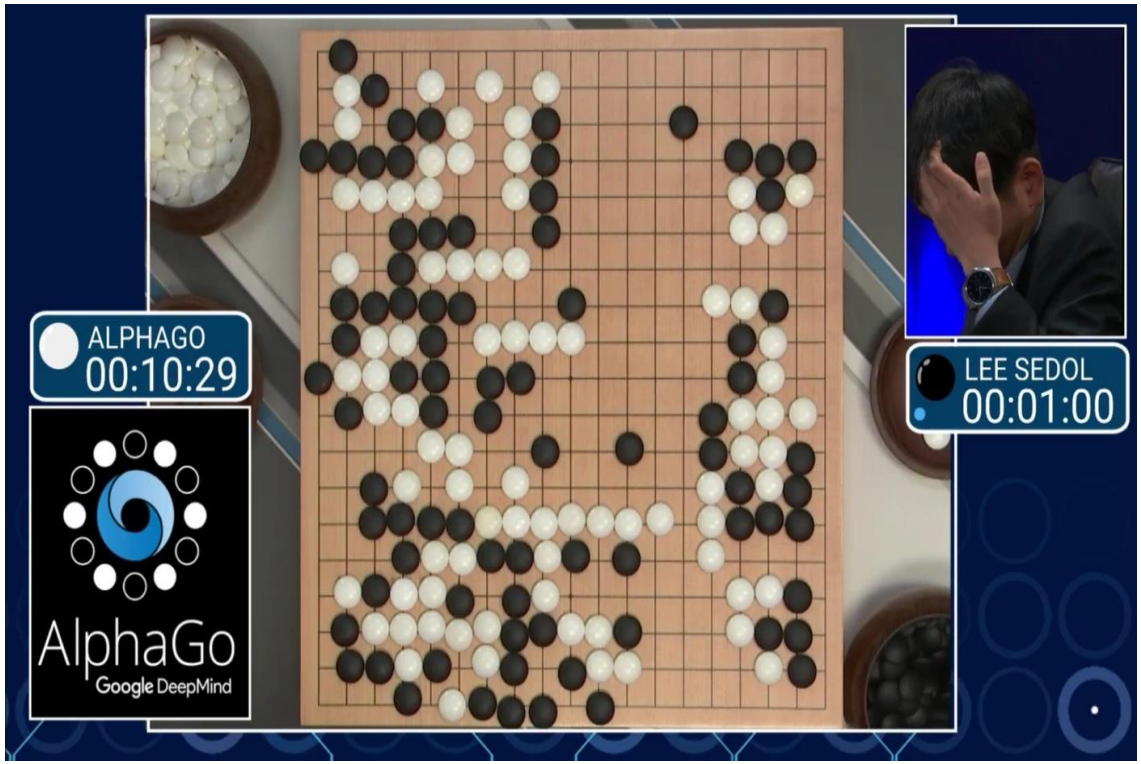
Машинное обучение – набор способов воспроизведения связей между событиями и результатом.

Машинное обучение – обширный подраздел искусственного интеллекта, изучающий методы построения алгоритмов, способных обучаться.

Machine learning – the field of study that gives computers the ability to learn without being explicitly programmed.

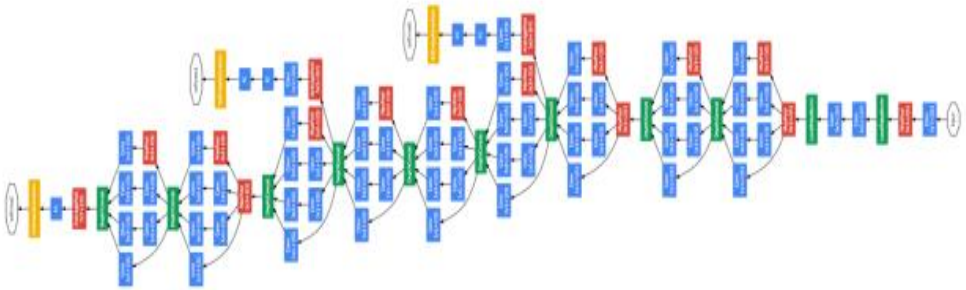
ПРИМЕРЫ

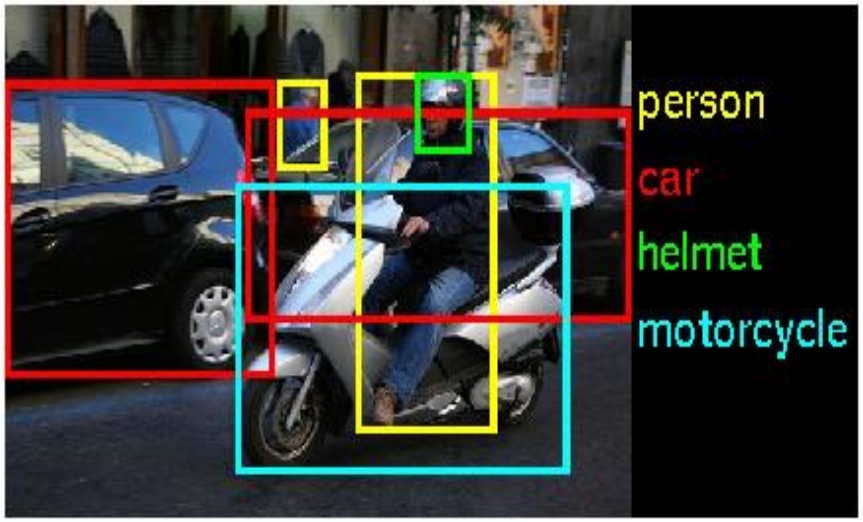
- **Нейронная сеть, играющая в Го**
- **Март 2016 –** победа над мировым чемпионом
- Нейронная сеть обучалась, играя сама с собой для увеличения объёмов входных данных (принцип обучения с подкреплением, reinforcement learning)



ПРИМЕРЫ

- **ImageNet** — задача распознавания объектов на изображении
- Решается с помощью нейронных сетей с точностью, превышающей точность работы человека





ПРИМЕРЫ

- Аннотирование изображений

"man in black shirt is playing guitar."

"construction worker in orange safety vest is working on road."

"two young girls are playing with legos toy."

"boy is doing backflip on wakeboard."

"girl in pink dress is jumping in air."

"black and white dog jumps over bar."

"young girl in pink shirt is swinging on swing."

"man in blue wetsuit is surfing on wave."

ПРИМЕРЫ

- Чтение по губам

*Google Deepmind в **2017** году создали модель, обученную на телевизионном датасете, которая смогла превзойти профессионального lips reader'а с канала BBC.*

ТРАНСФОРМЕРЫ В ЗАДАЧАХ АНАЛИЗА ТЕКСТОВ

✓ В **2017** году в Google

Выпустили статью *“Attention is all you need”*, описывающую механизм внимания – механизм, используемый в нейронных сетях для извлечения информации из текста.

✓ Трансформер – архитектура нейронной сети, основанная на механизме внимания. Она даёт state-of-the-art (SOTA) результаты во многих задачах машинного обучения, связанных с обработкой естественного языка:

- Определение тональности текста
- Перевод с одного языка на другой
- Определение связности предложений в тексте и др.

ПРИМЕР: ОТВЕТЫ НА ВОПРОСЫ ПО ТЕКСТУ

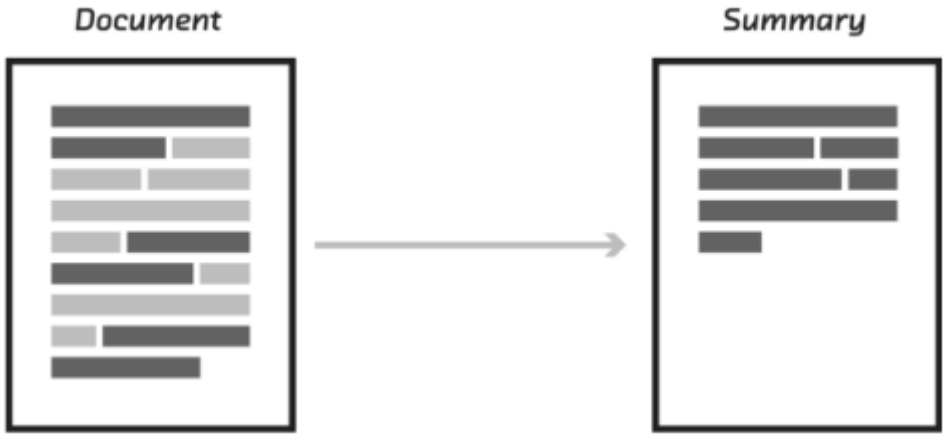
- Context: *Architecturally, the school has a Catholic character. Atop the Main Building's gold dome is a golden statue of the Virgin Mary. Immediately in front of the Main Building and facing it, is a copper statue of Christ with arms upraised with the legend "Venite Ad Me Omnes". Next to the Main Building is the Basilica of the Sacred Heart. Immediately behind the basilica is the Grotto, a Marian place of prayer and reflection. It is a replica of the grotto at Lourdes, France where the Virgin Mary reputedly appeared to Saint Bernadette Soubirous in 1858. At the end of the main drive (and in a direct line that connects through 3 statues and the Gold Dome), is a simple, modern stone statue of Mary.*
- Question: *The Basilica of the Sacred heart at Notre Dame is beside to which structure?*

ПРИМЕР: ОТВЕТЫ НА ВОПРОСЫ ПО ТЕКСТУ

- Context: *Architecturally, the school has a Catholic character. Atop the Main Building's gold dome is a golden statue of the Virgin Mary. Immediately in front of the Main Building and facing it, is a copper statue of Christ with arms upraised with the legend "Venite Ad Me Omnes". Next to the Main Building is the Basilica of the Sacred Heart. Immediately behind the basilica is the Grotto, a Marian place of prayer and reflection. It is a replica of the grotto at Lourdes, France where the Virgin Mary reputedly appeared to Saint Bernadette Soubirous in 1858. At the end of the main drive (and in a direct line that connects through 3 statues and the Gold Dome), is a simple, modern stone statue of Mary.*
- Question: *The Basilica of the Sacred heart at Notre Dame is beside to which structure?*
- Answer: *start_position: 49, end_position: 51*

ПРИМЕР: СУММАРИЗАЦИЯ ТЕКСТОВ

Суммаризация – получение смысловой выжимки из текста. С 2019 года сильно улучшилось качество суммаризации, благодаря внедрению Deep Learning подхода, основанного на механизме внимания.



сервис для сокращения текста

ПРИМЕР: РЕКОМЕНДАЦИИ

Рекомендации Netflix:

Profile Type	Score Image A	Score Image B
Comedy	5.7	6.3
Romance	7.2	6.5

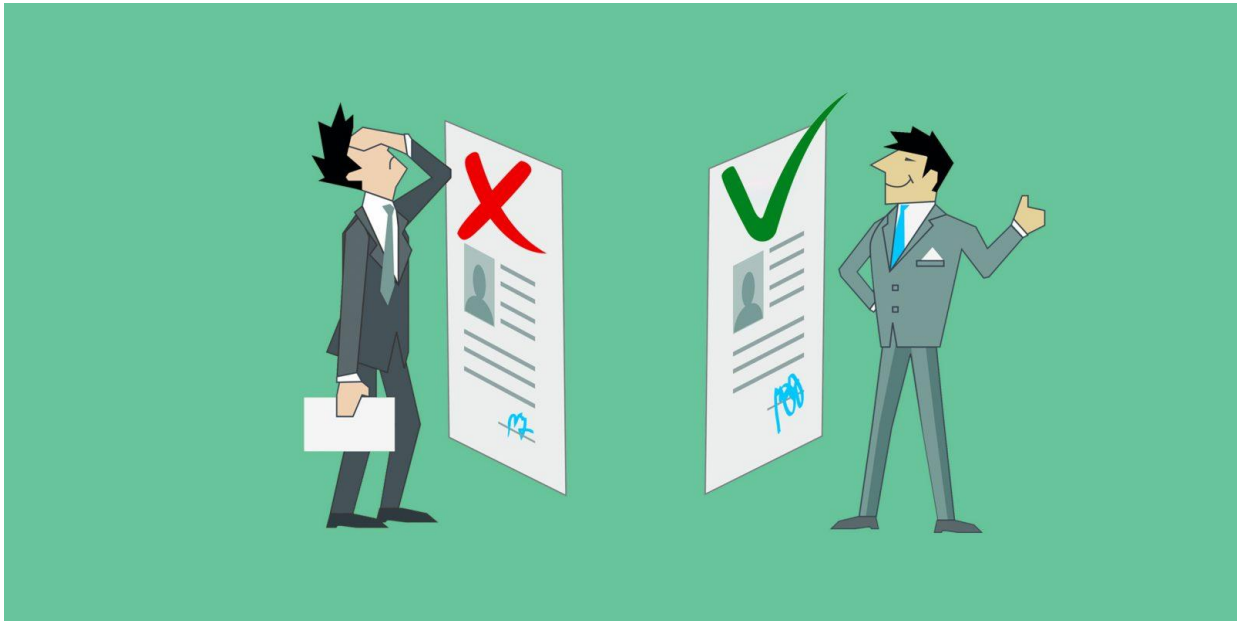
Image A

Image B

ОСНОВНЫЕ ПОНЯТИЯ МАШИННОГО ОБУЧЕНИЯ

ПРИМЕР: ЗАДАЧА СКОРИНГА

- Пусть по характеристикам клиента (пол, возраст, средний доход, рейтинг кредитной истории и так далее) мы хотим предсказать, **вернёт клиент кредит или не вернёт**.



ПРИМЕР: ЗАДАЧА СКОРИНГА

- **Целевая переменная (target)**, то есть величина, которую хотим предсказать - это число (например, 1 - если человек вернет кредит, и 0 иначе).
- Характеристики клиента, а именно, его пол, возраст, доход и так далее, называются **признаками (features)**.
- Сами же клиенты - сущности, с которыми мы работаем в этой задаче - называются **объектами (objects)**.

ОБУЧЕНИЕ АЛГОРИТМА

- На этапе обучения происходит анализ большого количества данных, для которых у нас имеются правильные ответы (например, клиенты, про которых мы знаем - вернули они кредит или нет; пациенты и их анализы, где про каждого пациента мы знаем, болен он или здоров и так далее).

- Модель машинного обучения изучает эти данные и старается научиться делать предсказания таким образом, чтобы для каждого объекта предсказывать как можно более точный ответ. Все данные с известными ответами называются **обучающей выборкой**.

ПРИМЕНЕНИЕ АЛГОРИТМА

- На этом этапе готовая (уже обученная) модель применяется для того, чтобы получить ответ на новых данных. Например, у нас есть подробная информация о клиентах, и мы применяем модель, чтобы она предсказала, кто из них вернет кредит, а кто нет.

ФОРМАЛИЗАЦИЯ

X – множество объектов

Y – множество ответов

$a: X \to Y$ – неизвестная зависимость

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 (решающую функцию), приближающую y на всем множестве X

ПРИЗНАКОВОЕ ОПИСАНИЕ ОБЪЕКТОВ

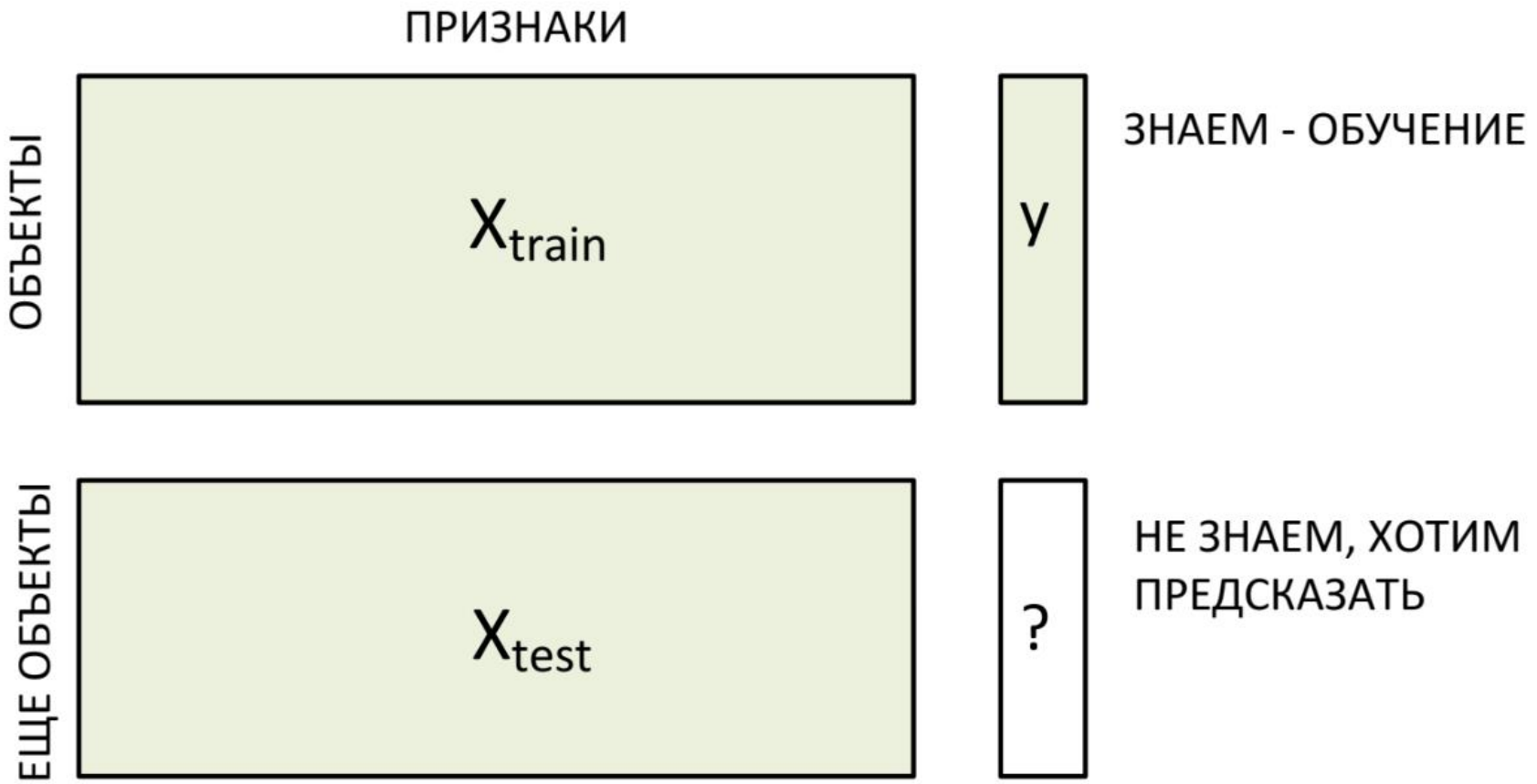
Признаки объекта x можно записать в виде вектора

$$(f_1(x), \dots, f_n(x))$$

Матрица “объекты-признаки”:

$$\begin{pmatrix} f_1(x_1) & \dots & f_n(x_1) \\ \dots & \dots & \dots \\ f_1(x_\ell) & \dots & f_n(x_\ell) \end{pmatrix}$$

СТАНДАРТНАЯ ПОСТАНОВКА ЗАДАЧ МАШИННОГО ОБУЧЕНИЯ



ВИДЫ ПРИЗНАКОВ

- Числовые
- Бинарные (0/1)
- Категориальные (название города, марка машины)
- Признаки со сложной внутренней структурой (изображение)

ВИДЫ ДАННЫХ

- Таблицы (-xls, -csv и другие форматы, содержащие данные)
- Текстовые данные
- Изображения
- Звук
- Логи

Большинство алгоритмов машинного обучения работает с числовыми данными, поэтому все виды данных необходимо переводить в числовые.

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

- Задачи медицинской диагностики (пациент здоров или болен)
- Задачи кредитного скоринга (выдаст банк кредит данному клиенту или нет)
- Задача предсказания оттока клиентов (уйдет клиент в следующем месяце или нет)
- Предсказание поведения пользователя (кликнет пользователь по данному баннеру или нет)
- Классификация изображений (на изображении кошка или собака)

ПРИМЕРЫ ЗАДАЧ КЛАССИФИКАЦИИ

Мультиклассовая классификация

- Определение типа объекта на изображении

Pedestrian

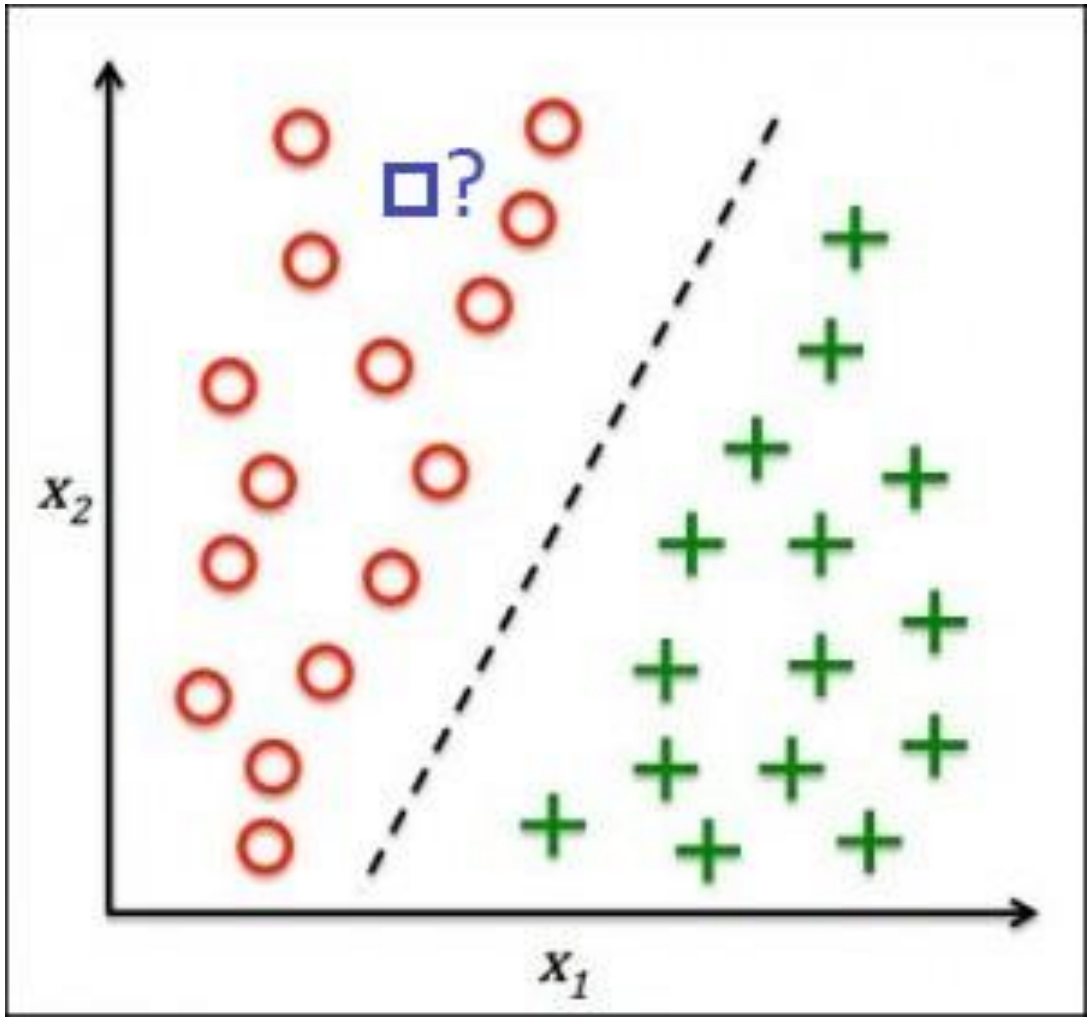
Car

Motorcycle

Truck

- Определение наиболее подходящей профессии для данного кандидата

ЗАДАЧА КЛАССИФИКАЦИИ



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

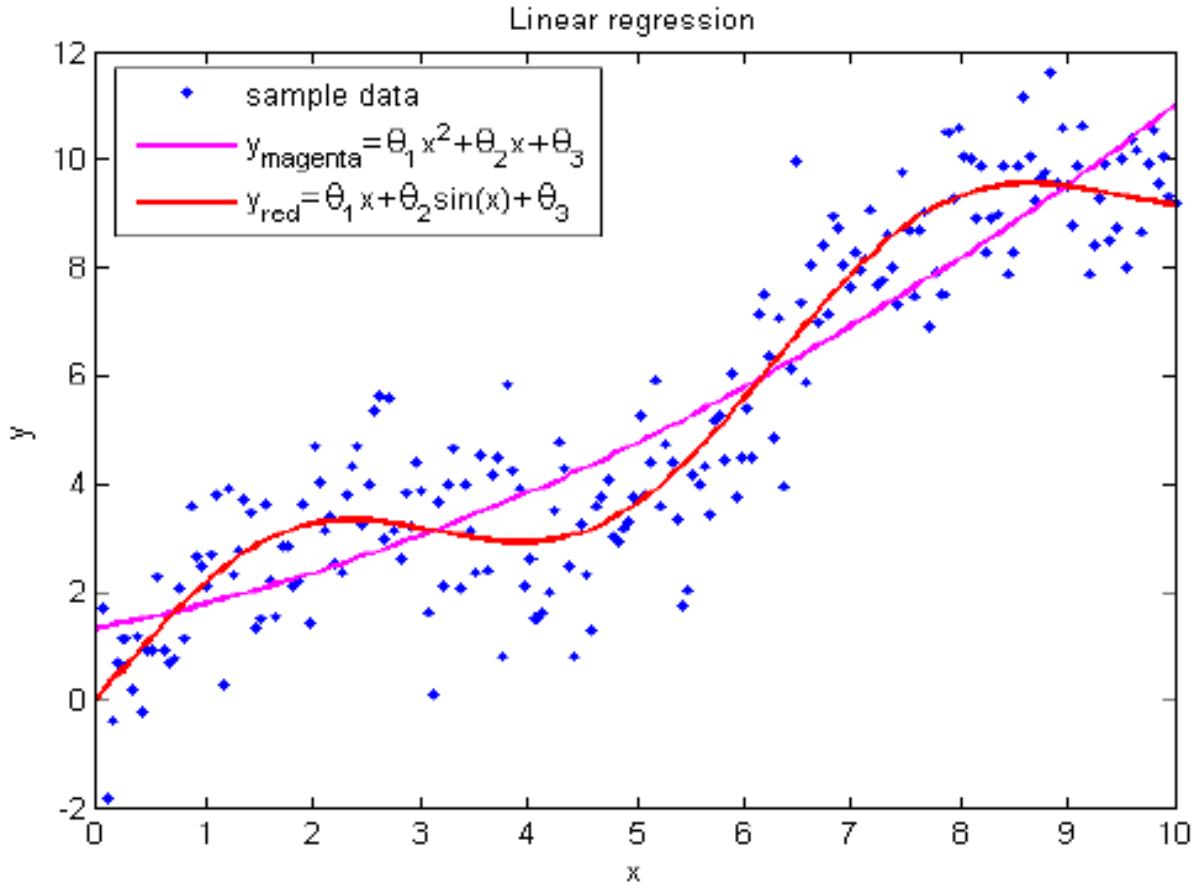
- $Y = R$ или $Y = R^n$

ПРИМЕРЫ ЗАДАЧ РЕГРЕССИИ

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

ЗАДАЧА РЕГРЕССИИ

$X = Y = \mathbb{R}$, $\ell = 200$, $n = 3$ признака: $\{x, x^2, 1\}$ или $\{x, \sin x, 1\}$



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

- $Y = R$ или $Y = R^n$

Ранжирование

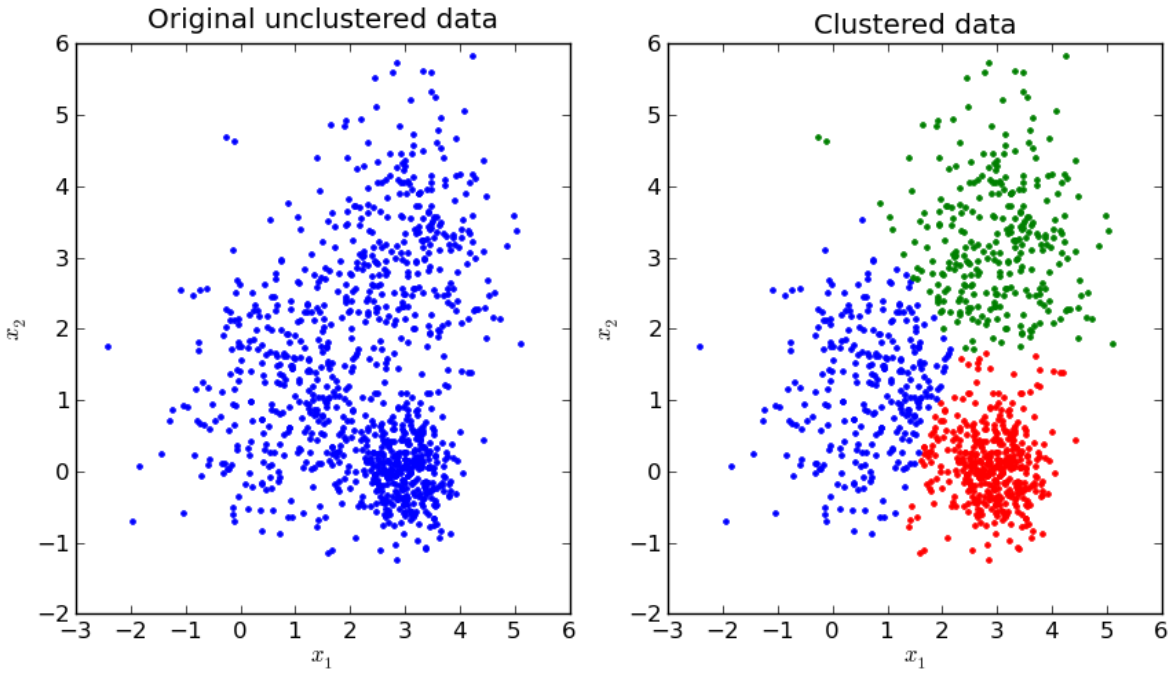
- Y – конечное упорядоченное множество

ПРИМЕРЫ ЗАДАЧ РАНЖИРОВАНИЯ

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

ЗАДАЧИ, В КОТОРЫХ НЕТ ЦЕЛЕВОЙ ПЕРЕМЕННОЙ

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



ПРИМЕРЫ ЗАДАЧ КЛАСТЕРИЗАЦИИ

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

ЗАДАЧИ, В КОТОРЫХ НЕТ ЦЕЛЕВОЙ ПЕРЕМЕННОЙ

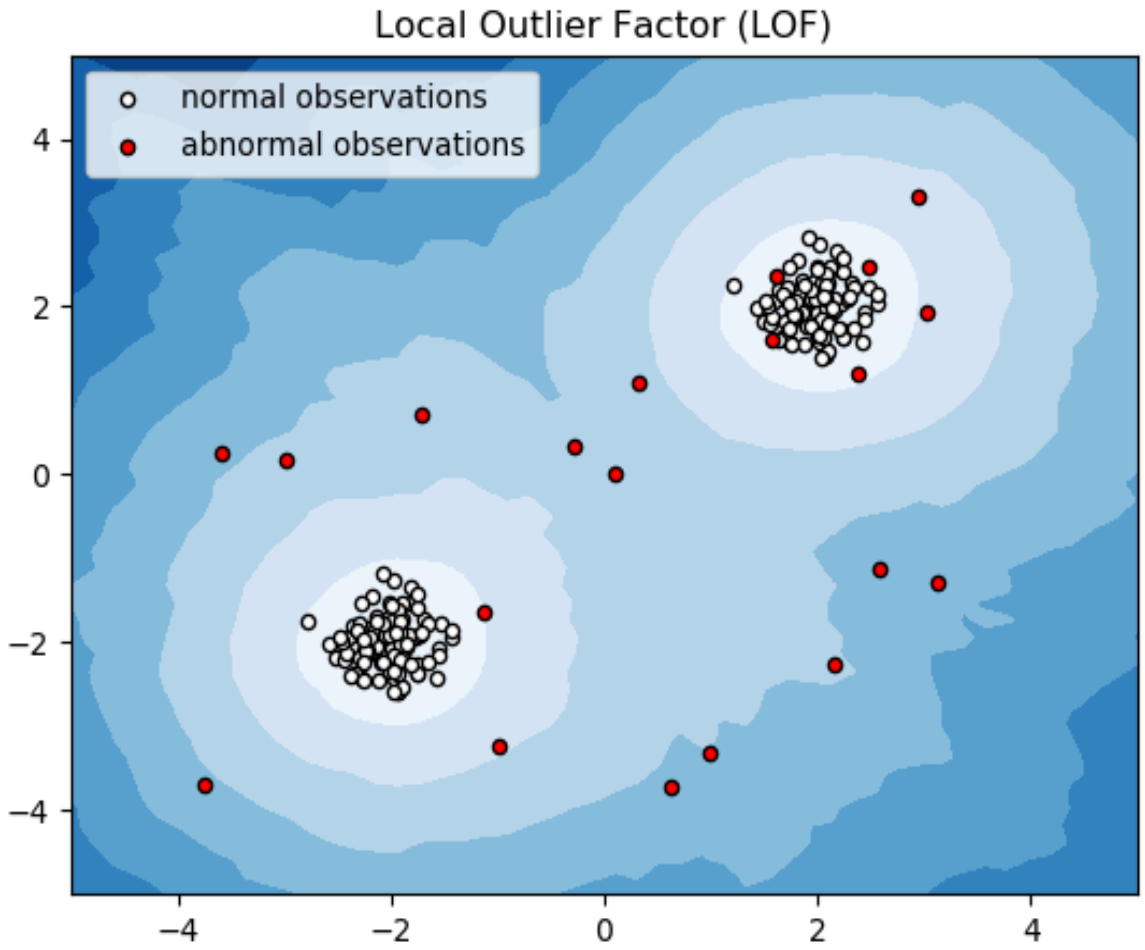
- **Кластеризация –** задача разделения объектов на группы, при этом где целевые переменные для объектов неизвестны (или не существуют). Разделение происходит только на основе признаковых описаний объектов.
- **Понижение размерности –** задача генерации новых признаков (их число меньше, чем число старых), так, что с их помощью задача решается не хуже, чем с исходными.

ЗАДАЧИ, В КОТОРЫХ НЕТ ЦЕЛЕВОЙ ПЕРЕМЕННОЙ

- **Кластеризация –** задача разделения объектов на группы, при этом где целевые переменные для объектов неизвестны (или не существуют). Разделение происходит только на основе признаковых описаний объектов.
- **Понижение размерности –** задача генерации новых признаков (их число меньше, чем число старых), так, что с их помощью задача решается не хуже, чем с исходными.
- **Оценивание плотности –** задача приближения распределения объектов.

ПРИМЕР ОЦЕНИВАНИЯ ПЛОТНОСТИ

- Поиск аномалий с помощью оценивания плотностей



ЗАДАЧИ, В КОТОРЫХ НЕТ ЦЕЛЕВОЙ ПЕРЕМЕННОЙ

- **Кластеризация –** задача разделения объектов на группы, при этом где целевые переменные для объектов неизвестны (или не существуют). Разделение происходит только на основе признаковых описаний объектов.
- **Понижение размерности –** задача генерации новых признаков (их число меньше, чем число старых), так, что с их помощью задача решается не хуже, чем с исходными.
- **Оценивание плотности –** задача приближения распределения объектов.
- **Визуализация –** задача изображения многомерных объектов в 2х или 3хмерном пространстве с сохранением зависимостей между ними.

ТИПЫ ЗАДАЧ МАШИННОГО ОБУЧЕНИЯ

- Если нам известны значения целевой переменной, то есть алгоритм обучается так, чтобы правильно предсказывать целевую переменную – это **обучение с учителем**. Сюда относят классификацию, регрессию и ранжирование.

ТИПЫ ЗАДАЧ МАШИННОГО ОБУЧЕНИЯ

- Если нам известны значения целевой переменной, то есть алгоритм обучается так, чтобы правильно предсказывать целевую переменную – это **обучение с учителем**. Сюда относят классификацию, регрессию и ранжирование.
- Если нам неизвестны значения целевой переменной или целевая переменная вообще отсутствует, то есть алгоритм обучается только по признакам объектов, то это **обучение без учителя**. Примерами обучения с учителем являются кластеризация, понижение размерности и др.

ОБУЧЕНИЕ АЛГОРИТМА, ОЦЕНКА ПРЕДСКАЗАТЕЛЬНОЙ СПОСОБНОСТИ

ОБУЧЕНИЕ АЛГОРИТМА

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

ОБУЧЕНИЕ АЛГОРИТМА

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

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

ОБУЧЕНИЕ АЛГОРИТМА

Предположим, что мы хотим предсказать *стоимость дома* 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:

$$\boldsymbol{a(x) = w_0 + w_1 x_1 + w_2 x_2},$$

где w_0, w_1, w_2 -

параметры модели (*веса*).

ОБУЧЕНИЕ АЛГОРИТМА

Предположим, что мы хотим предсказать *стоимость дома* 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:

$$\boldsymbol{a(x) = w_0 + w_1 x_1 + w_2 x_2},$$

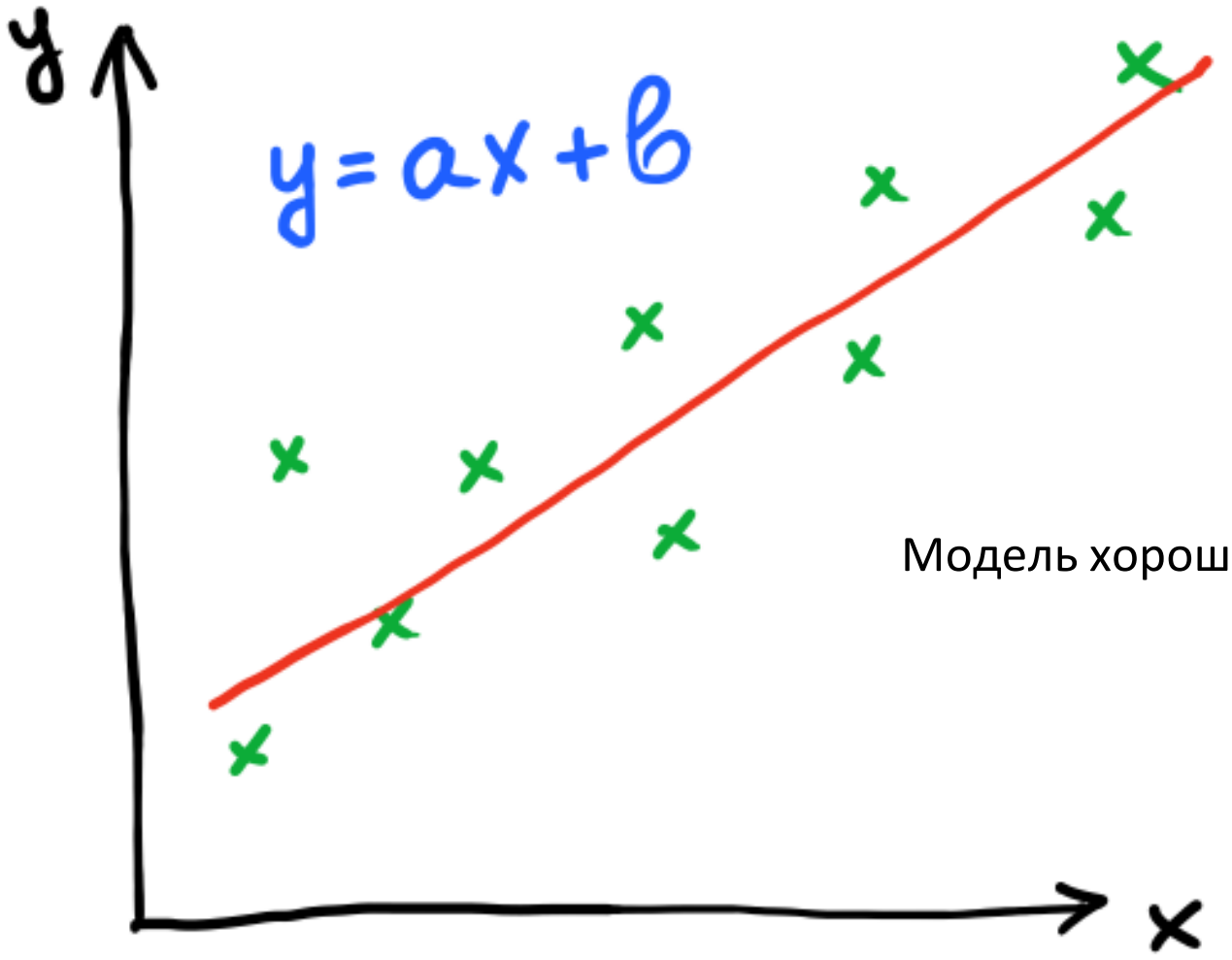
где w_0, w_1, w_2 -

параметры модели (*веса*).

Общий вид линейных моделей:

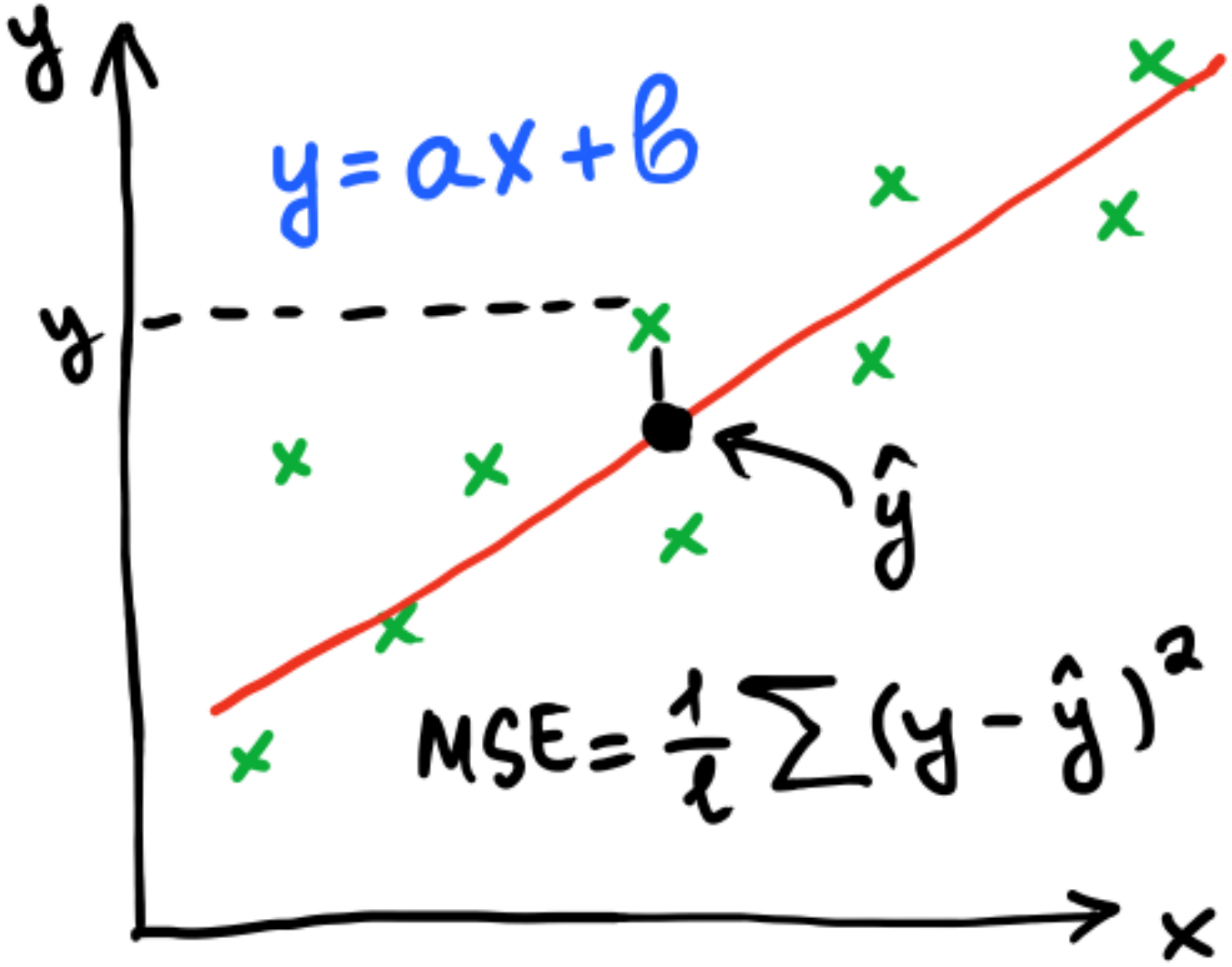
$$\boldsymbol{A = \{a(x) = w_0 + w_1 x_1 + \cdots + w_d x_d | w_0, w_1, \ldots, w_d \in \mathbb{R}\}}$$

ОБУЧЕНИЕ АЛГОРИТМА



Модель хорошая?

ОБУЧЕНИЕ АЛГОРИТМА



ФУНКЦИОНАЛ ОШИБКИ

- Как измерить ошибку алгоритма на всех объектах выборки?

Функционал ошибки – функционал, измеряющий качество работы алгоритма.

Пример (среднеквадратичная ошибка, MSE):

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

X – объекты, l – количество объектов

a – алгоритм, $a(x_i)$ – ответ алгоритма на объекте x_i

y_i - истинные ответы

ФУНКЦИОНАЛ ОШИБКИ

Функционал ошибки – функционал, измеряющий качество работы алгоритма.

Пример (среднеквадратичная ошибка, MSE):

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

X – объекты, l – количество объектов

a – алгоритм, $a(x_i)$ – ответ алгоритма на объекте x_i

y_i - истинные ответы

При обучении алгоритма мы минимизируем функционал ошибки.

ОБУЧЕНИЕ АЛГОРИТМА

Пример (семейство линейных моделей):

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

Функционал ошибки:

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

Функционал ошибки для линейной модели стоимости дома:

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

ОБУЧЕНИЕ АЛГОРИТМА

Параметры w_0, w_1, w_2 подбираются так, чтобы на них достигался минимум функции потерь (на обучающей выборке):

Функционал ошибки для линейной модели стоимости дома:

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

ОБУЧЕНИЕ АЛГОРИТМА (ОБЩИЙ ВИД ЛИНЕЙНОЙ РЕГРЕССИИ)

Параметры $w_0, \dots, w_n$ подбираются так, чтобы на них достигался минимум функции потерь (на обучающей выборке):

Функционал ошибки для линейной модели:

$$Q(a, X) = \frac{1}{l} \sum_{i=1}^{l} \left(w_0 + \sum_{j=1}^{d} w_j x_{ij} - y_i \right)^2 \to \min_{w_0, \dots, w_d}$$

ОБУЧЕНИЕ АЛГОРИТМА

Процесс поиска оптимального алгоритма (оптимального набора параметров или *весов*) называется **обучением**.

ОЦЕНКА КАЧЕСТВА МОДЕЛЕЙ

В задачах машинного обучения для оценки качества моделей и сравнения различных алгоритмов используются ***метрики качества***.

МЕТРИКИ КАЧЕСТВА

В задачах машинного обучения для оценки качества моделей и сравнения различных алгоритмов используются ***метрики качества***.

Примеры:

- Корень из среднеквадратичной ошибки – для регрессии

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

МЕТРИКИ КАЧЕСТВА

В задачах машинного обучения для оценки качества моделей и сравнения различных алгоритмов используются ***метрики качества*.**

Примеры:

- Корень из среднеквадратичной ошибки **–** для регрессии
- Доля правильных ответов **–** для классификации

$$accuracy(a, X) = \frac{1}{l}\sum_{i=1}^{l}[a(x_i) = y_i]$$

АЛГОРИТМ РЕШЕНИЯ ЗАДАЧИ АНАЛИЗА ДАННЫХ

1. Постановка задачи

АЛГОРИТМ РЕШЕНИЯ ЗАДАЧИ АНАЛИЗА ДАННЫХ

1. Постановка задачи
2. Выделение признаков

АЛГОРИТМ РЕШЕНИЯ ЗАДАЧИ АНАЛИЗА ДАННЫХ

1. Постановка задачи
2. Выделение признаков
3. Формирование выборки

АЛГОРИТМ РЕШЕНИЯ ЗАДАЧИ АНАЛИЗА ДАННЫХ

1. Постановка задачи
2. Выделение признаков
3. Формирование выборки
4. Выбор функции потерь и метрики качества

АЛГОРИТМ РЕШЕНИЯ ЗАДАЧИ АНАЛИЗА ДАННЫХ

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

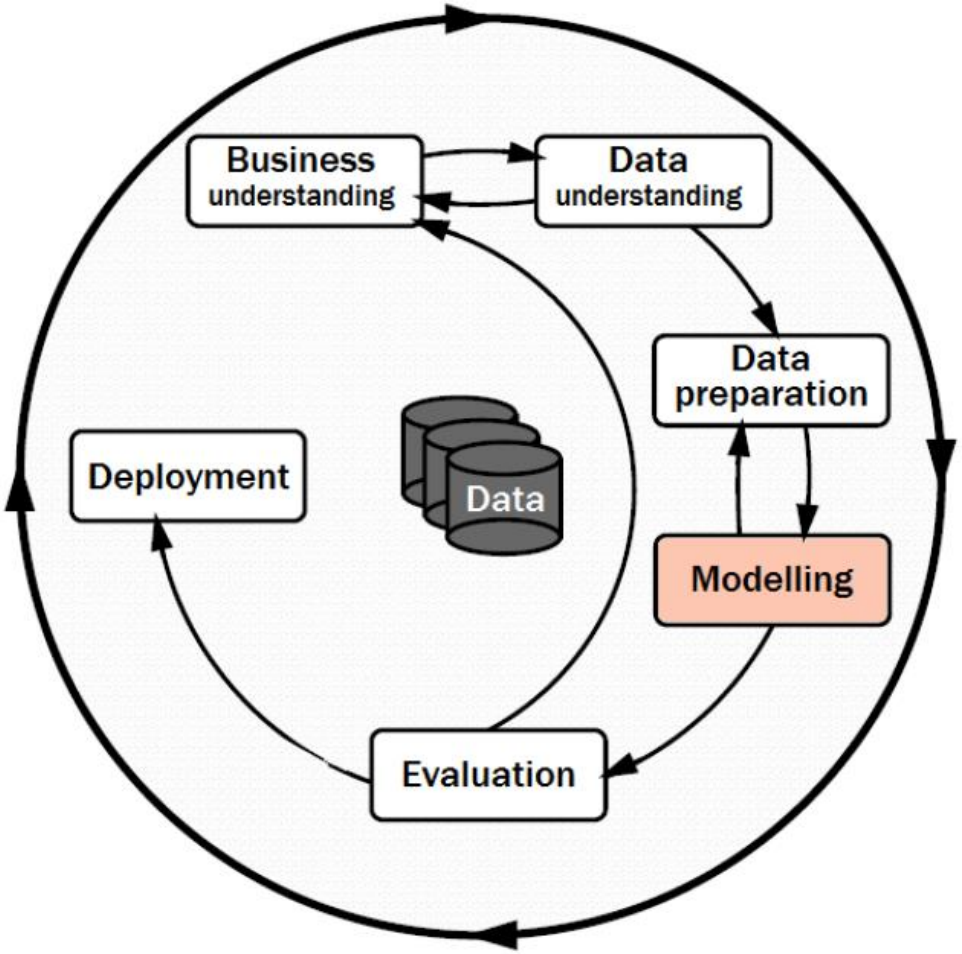
АЛГОРИТМ РЕШЕНИЯ ЗАДАЧИ АНАЛИЗА ДАННЫХ

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

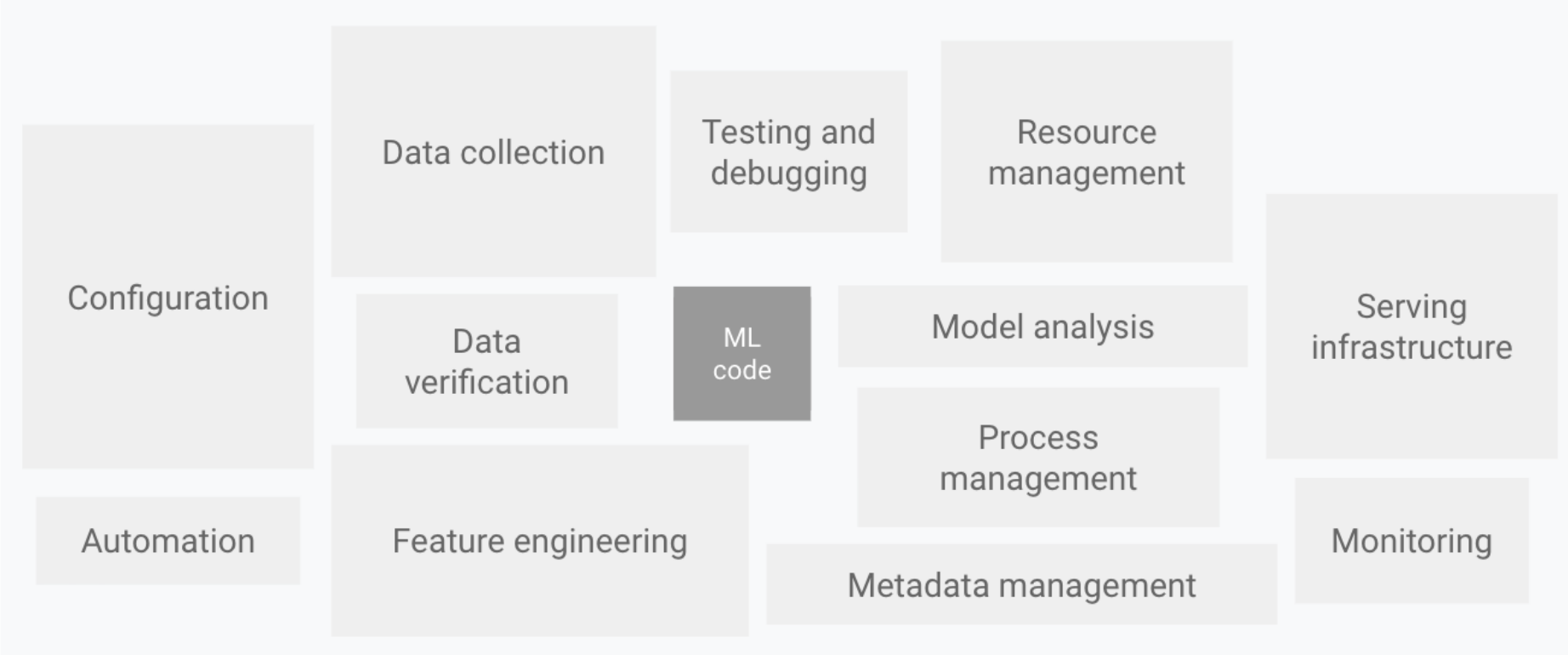
АЛГОРИТМ РЕШЕНИЯ ЗАДАЧИ АНАЛИЗА ДАННЫХ

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

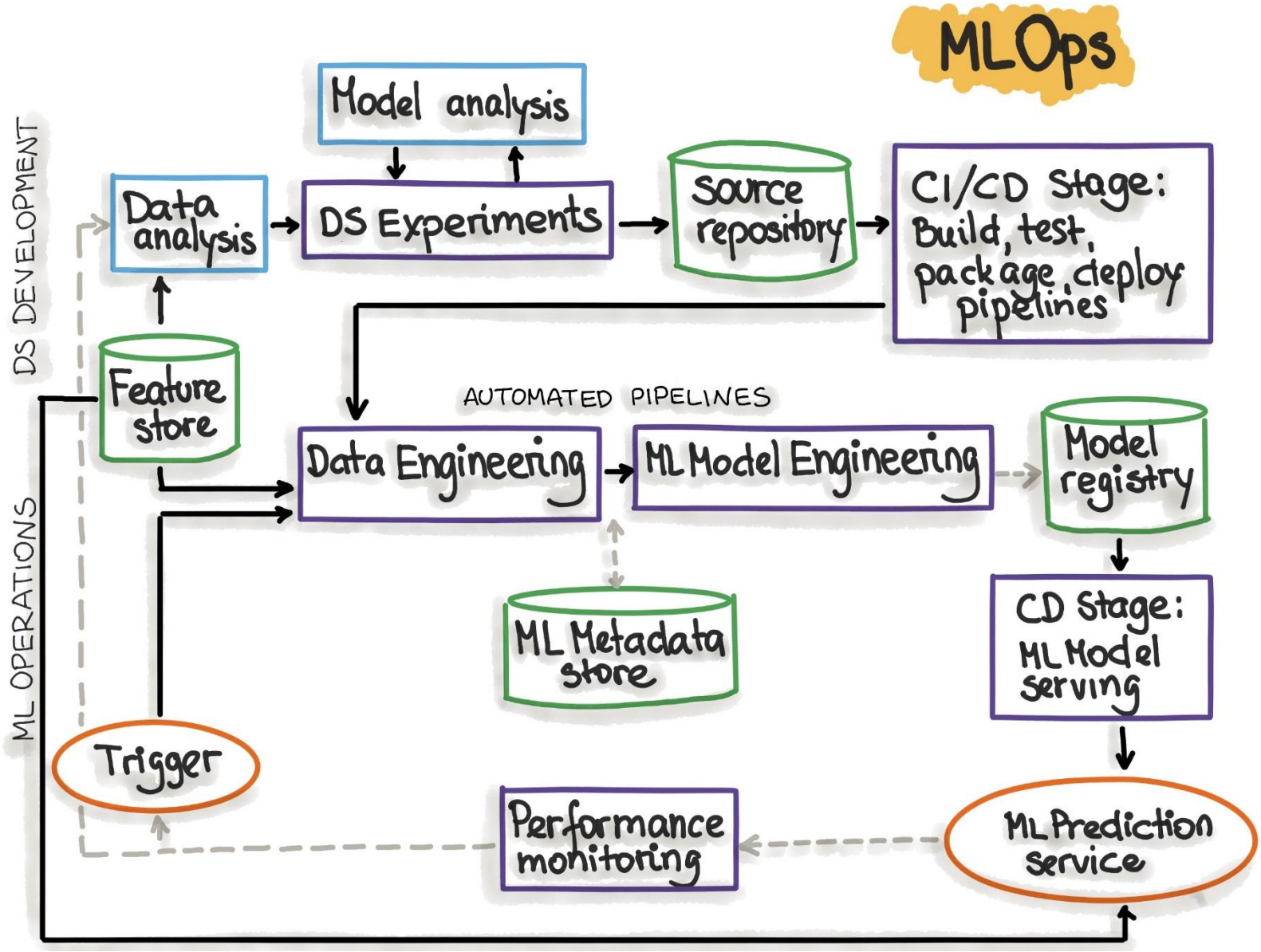
СТАДИИ РАЗРАБОТКИ МОДЕЛИ МАШИННОГО ОБУЧЕНИЯ (CRISP)



ПРОЕКТ ПО ML ЭТО НЕ ТОЛЬКО ОБУЧЕНИЕ МОДЕЛИ...



MLOPS


MLOps
Model analysis
Data analysis
DS Experiments
Source repository
CI/CD Stage: Build, test, package, deploy pipelines
DS DEVELOPMENT
Feature store
AUTOMATED PIPELINES
Data Engineering
ML Model Engineering
Model registry
ML OPERATIONS
ML Metadata store
CD Stage: ML Model serving
Trigger
Performance monitoring
ML Prediction service

ПРОФЕССИИ, СВЯЗАННЫЕ С ML

- Data Analyst → 180000 рублей в месяц
- Data Scientist → 250000 рублей в месяц
- Data Engineer → 270000 рублей в месяц
- ML Engineer → 300000 рублей в месяц

* Средняя зарплата по данным за июнь-июль 2022 года по вакансиям в интернете с указанной зарплатой.

МАТЕМАТИКА ДЛЯ ML

- Линейная алгебра
- Математический анализ
- Теория вероятностей и математическая статистика

ЛИНЕЙНАЯ АЛГЕБРА В ML

Линейная алгебра применяется для решения всех основных задач Data Science, таких как создание и обучение моделей.

Примеры алгоритмов, в которых используется линейная алгебра:

- Линейная регрессия
- Логистическая регрессия (классификатор)
- Метод главных компонент
- Матричные разложения (например, в задаче построения рекомендаций)
- Нейронные сети

ЧТО НУЖНО ЗНАТЬ ИЗ ЛИНЕЙНОЙ АЛГЕБРЫ

- векторы
- матрицы
- транспонирование матрицы
- обратная матрица
- определитель матрицы
- след матрицы
- скалярное произведение
- собственные значения
- собственные векторы

МАТЕМАТИЧЕСКИЙ АНАЛИЗ В ML

Методы математического анализа нужны при решении задач оптимизации в ML:

- Минимизация функции потерь при обучении алгоритма
- Метод обратного распространения ошибки при обучении нейронных сетей и др.

Необходимые навыки из математического анализа:

- Дифференцирование (производные, частные производные функций одной и многих переменных)
- Интегрирование

ВЕКТОРНОЕ И МАТРИЧНОЕ ДИФФЕРЕНЦИРОВАНИЕ

- Абсолютное большинство задач оптимизации в ML можно записать и решить в матричном виде (операции с матрицами и векторами осуществляются в Python в разы быстрее, чем поэлементные)

ВЕКТОРНОЕ И МАТРИЧНОЕ ДИФФЕРЕНЦИРОВАНИЕ

- Например, задачу минимизации функции потерь линейной модели

$$Q(w, X) = \sum_{i=1}^{l} \left(w_0 + \sum_{j=1}^{d} w_j x_{ij} - y_i \right)^2 \to \min_{w_0, \ldots, w_d}$$

Можно переписать в векторном виде так:

$$Q(w, X) = (y - Xw)^T (y - Xw) \to \min_{w}$$

ВЕКТОРНОЕ И МАТРИЧНОЕ ДИФФЕРЕНЦИРОВАНИЕ

- Например, задачу минимизации функции потерь линейной модели

$$Q(w, X) = \sum_{i=1}^{l} \left(w_0 + \sum_{j=1}^{d} w_j x_{ij} - y_i \right)^2 \to \min_{w_0, \ldots, w_d}$$

Можно переписать в векторном виде так:

$$Q(w, X) = (y - Xw)^T (y - Xw) \to \min_{w}$$

- Для решения этой задачи нам нужно уметь дифференцировать функцию по вектору, а в других задачах – и функцию по матрице. Здесь появляется *векторное и матричное дифференцирование*.