Введение в машинное обучение

Паточенко Евгений

НИУ ВШЭ

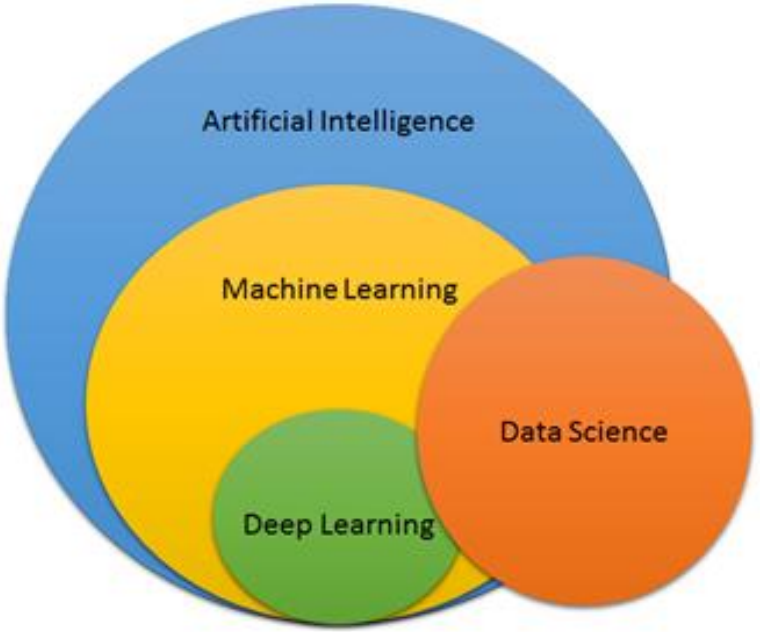
План занятия

- Немного истории
- Основные понятия
- Типы задач в машинном обучении
- Разведочный анализ данных (EDA)
- Обучение и валидация модели

Машинное обучение

Машинное обучение:

- Наука, изучающая способы извлечения закономерностей из ограниченного количества примеров (Е. Соколов)
- Наука, изучающая алгоритмы, автоматически улучшающиеся благодаря опыту (хендбук Яндекса)
- Процесс, который дает возможность компьютерам обучаться выполнять что-то без явного написания кода (А. Л. Самуэль)



Немного истории

1947 г. Британский математик и программист Алан Тьюринг высказывается об идее создания «интеллектуальных машин», которые должны изменять свое внутреннее состояние исходя из полученного опыта. В 1950 создает *тест Тьюринга* для оценки искусственного интеллекта компьютера.

1956 г. Дартмутская конференция. Американским информатиком Джоном Маккарти предложен термин «искусственный интеллект»

1960 г. Создан первый работающий пример алгоритма, моделирующий работу мозгового нейрона — перцептрон Розенблатта, на основе специально созданной машины Mark-1.

Немного истории

1966 г. Создается диалоговая система Eliza, которая моделирует разговор с психотерапевтом (обрабатывает естественный язык!)

Алгоритм подставляет значимые слова в шаблонную фразу, во многом перефразирует фразы пациента:

— Пациент: «У меня болит голова»

— Доктор: «Почему вы говорите, что у вас болит голова?»

— Пациент: «Мой отец меня ненавидит»

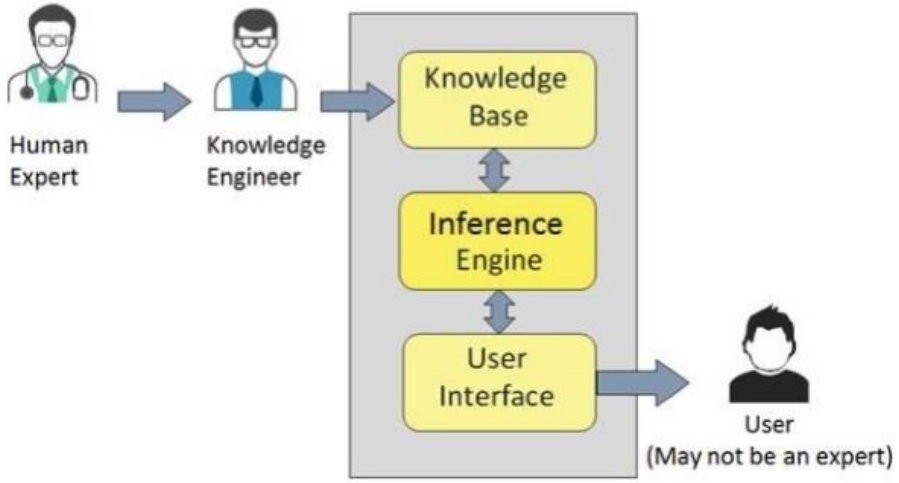
— Доктор: «Кто еще из семьи вас ненавидит?»

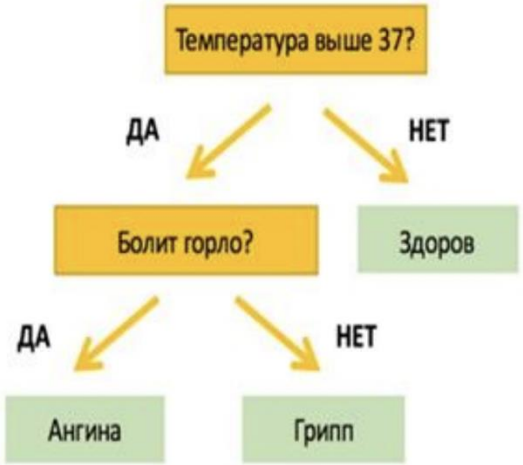
В непонятных ситуациях отвечала «Понятно».

Немного истории

1970-е — 1980-е гг. Появляются экспертные системы. В Стэнфорде разрабатывают систему MYCIN для помощи врачам в диагностировании и лечении серьезных бактериальных заболеваний. Происходит расцвет подходов на основе правил (rule-based).

1984 г. Разрабатывается алгоритм автоматического построения решающего дерева.



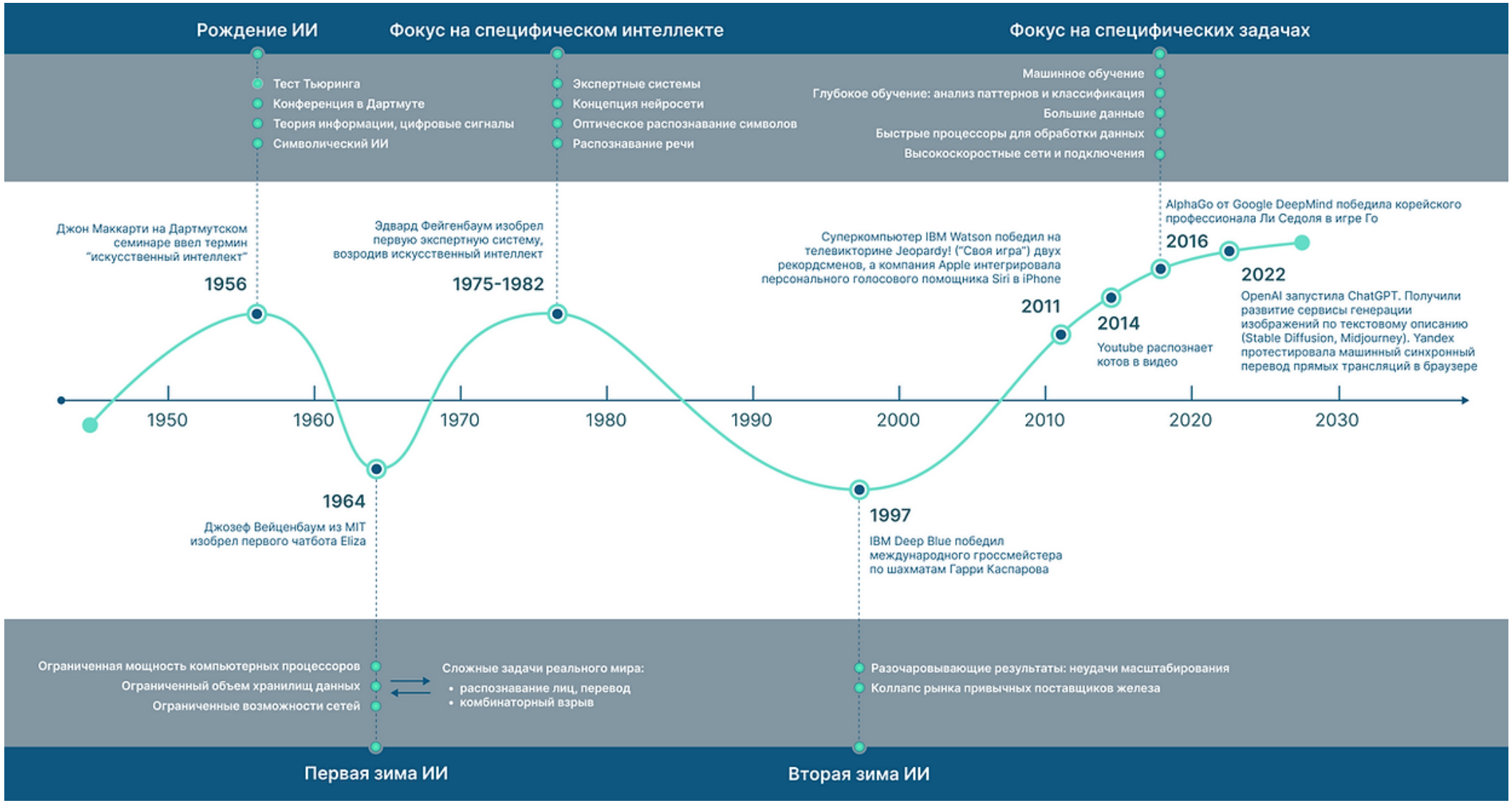


Немного истории

2016 г. Начинает использоваться нейросетевой машинный перевод. Глубокие нейронные сети обучаются на параллельных корпусах текстов. Качество перевода становится значительно лучше, чем у перевода на основе правил.

2017 г. Выходит статья Attention Is All You Need, про механизм внимания, который ложится в основу SoTA-решений в ИИ, используемых по сей день.

Немного истории



Основные понятия

Объекты — сущности, для которых хотим сделать предсказание

Признаки — Характеристики объектов. Бывают численные, категориальные, бинарные

Обучающая выборка — набор объектов, для которых известны правильные ответы

Целевая переменная (ответ) — величина, которую хотим предсказать

Модель (алгоритм) — функция, отображающая объекты в предсказания (алгоритм, который учится находить закономерности в данных и делает предсказания)

Основные понятия

Формальная постановка задачи

X — множество объектов

Y — множество ответов

Дано:

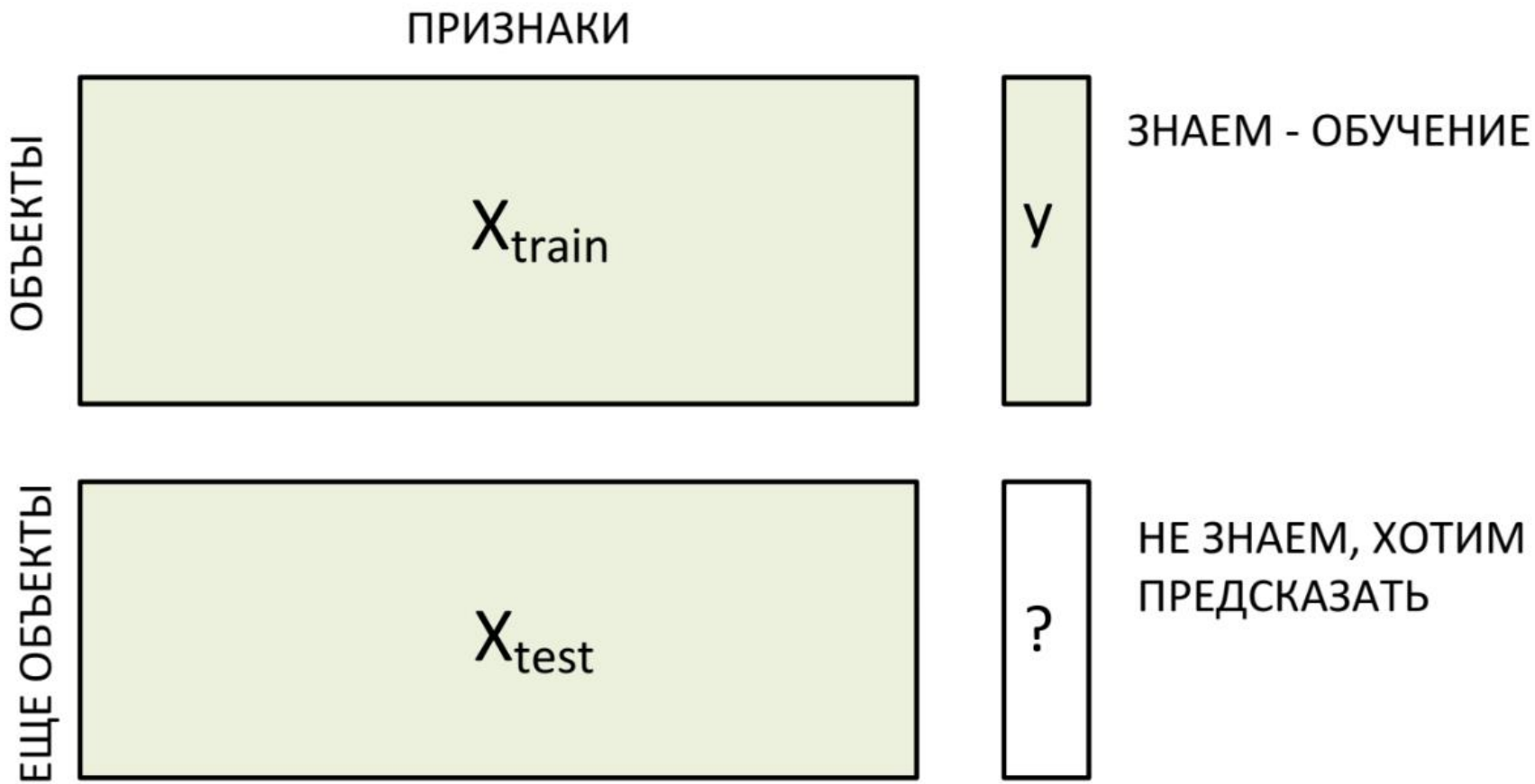
$\{x_1, \dots, x_n\} \subset X$ — обучающая выборка

$\{y_1, \dots, y_n\}, y_i = y(x_i)$ — известные ответы

Найти:

$a: X \longrightarrow Y$ — алгоритм, приближающий y на всем множестве X

Основные понятия



Основные понятия

Пример: задача скоринга

Задача: по характеристикам клиента (пол, возраст, средний доход, рейтинг кредитной истории, семейное положение и прочие) предсказать, вернет ли клиент кредит.

Основные понятия

Пример: задача скоринга

Задача: по характеристикам клиента (пол, возраст, средний доход, рейтинг кредитной истории, семейное положение и прочие) предсказать, вернет ли клиент кредит.

Целевая переменная —

Признаки —

Объекты —

Основные понятия

Пример: задача скоринга

Задача: по характеристикам клиента (пол, возраст, средний доход, рейтинг кредитной истории, семейное положение и прочие) предсказать, вернет ли клиент кредит.

Целевая переменная — число (1 — если вернет кредит и 0 — если нет)

Признаки —

Объекты —

Основные понятия

Пример: задача скоринга

Задача: по характеристикам клиента (пол, возраст, средний доход, рейтинг кредитной истории, семейное положение и прочие) предсказать, вернет ли клиент кредит.

Целевая переменная — число (1 — если вернет кредит и 0 — если нет)

Признаки — характеристики клиента (пол, возраст и так далее)

Объекты —

Основные понятия

Пример: задача скоринга

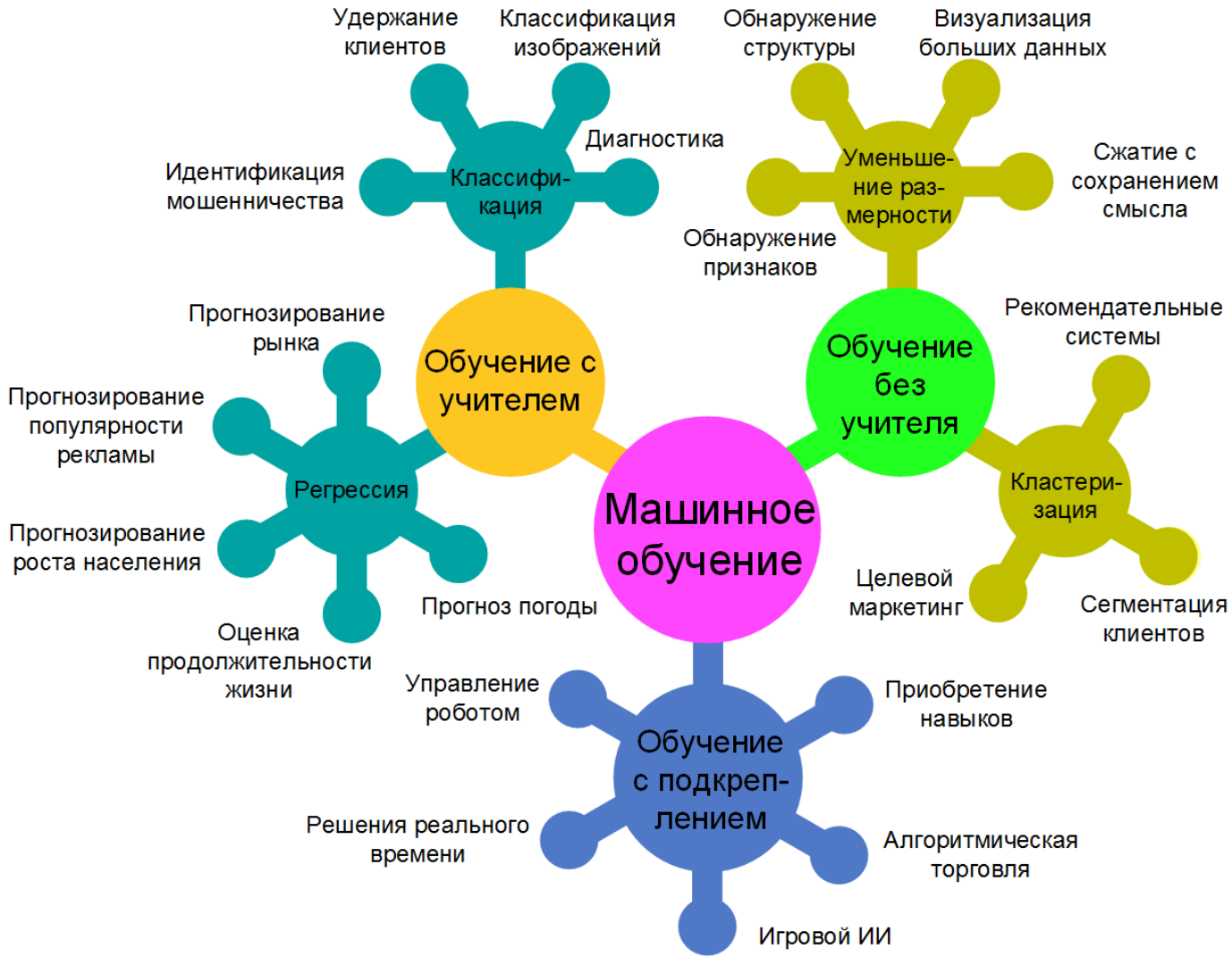
Задача: по характеристикам клиента (пол, возраст, средний доход, рейтинг кредитной истории, семейное положение и прочие) предсказать, вернет ли клиент кредит.

Целевая переменная — число (1 — если вернет кредит и 0 — если нет)

Признаки — характеристики клиента (пол, возраст и так далее)

Объекты — имеющиеся в датасете клиенты, для которых есть признаковое описание

Типы задач в машинном обучении



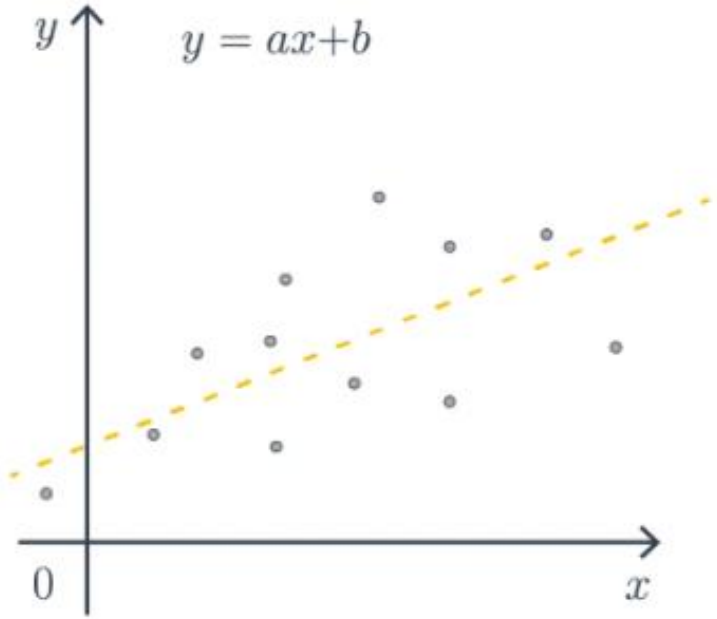
Типы задач в машинном обучении

Регрессия

• Предсказание стоимости недвижимости (стоимость квартиры в Москве)

• Предсказание прибыли ресторана

• Предсказание зарплаты выпускника вуза по его оценкам



x — (единственный) признак

y — таргет

Типы задач в машинном обучении

Регрессия

Базовая модель машинного обучения и статистики, используемая для оценки зависимости между одной зависимой переменной (целевой) и одной или несколькими независимыми переменными (признаками), при этом целевая переменная — непрерывная величина

Общий вид уравнения линейной регрессии

$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n + \varepsilon$, где

- y — зависимая (целевая) переменная
- $x_1, x_2, \ldots x_n$ — независимые переменные (признаки)
- β_0 — свободный член (intercept)
- $\beta_1, \beta_2, \ldots \beta_n$ — коэффициенты (веса) модели
- ε — ошибка модели (ошибка прогноза)

Типы задач в машинном обучении

Классификация

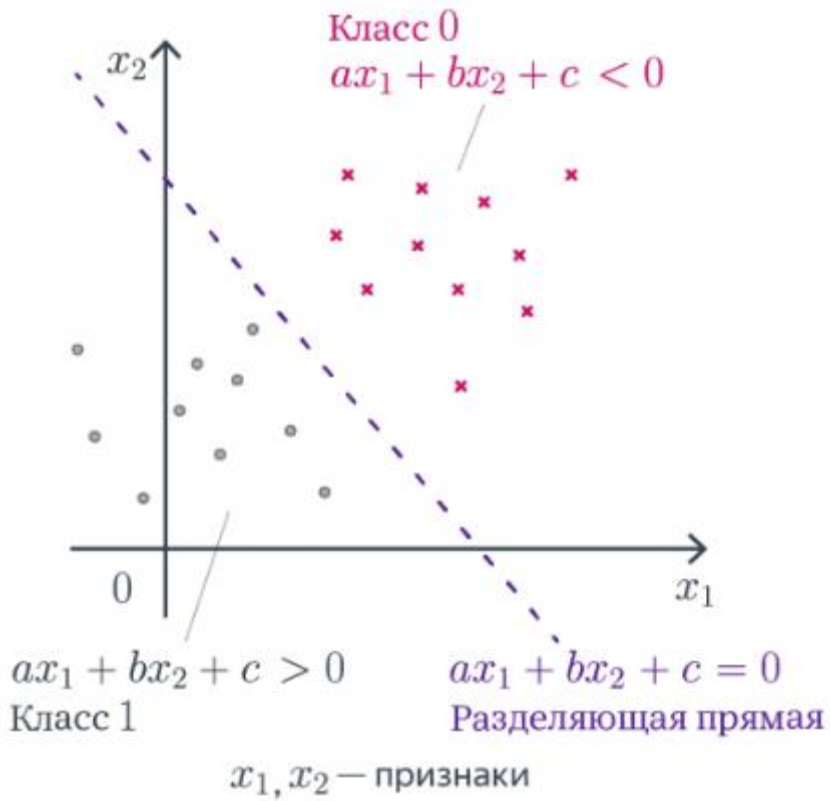
• Задачи медицинской диагностики (пациент здоров или болен)

• Задачи кредитного скоринга (вернет ли клиент кредит)

• Задача предсказания оттока клиентов (уйдет ли клиент в следующем месяце)

• Предсказание поведения пользователя (кликнет ли пользователь по баннеру)

• Классификация изображений (кошка/собака/мышь/енот)



Типы задач в машинном обучении

Классификация

Модель машинного обучения, используемая для прогнозирования категориальной (дискретной) целевой переменной на основе одной или нескольких независимых переменных (признаков). Целевая переменная принимает конечное число классов или меток.

Может быть:

— Бинарной (классификация на два класса) $Y = \{0,1\}$

— Многоклассовой (классификация на M непересекающихся классов) $Y = \{1, \dots, M\}$

— Многоклассовой (классификация на M классов, которые могут пересекаться) $Y = \{0,1\}^M$

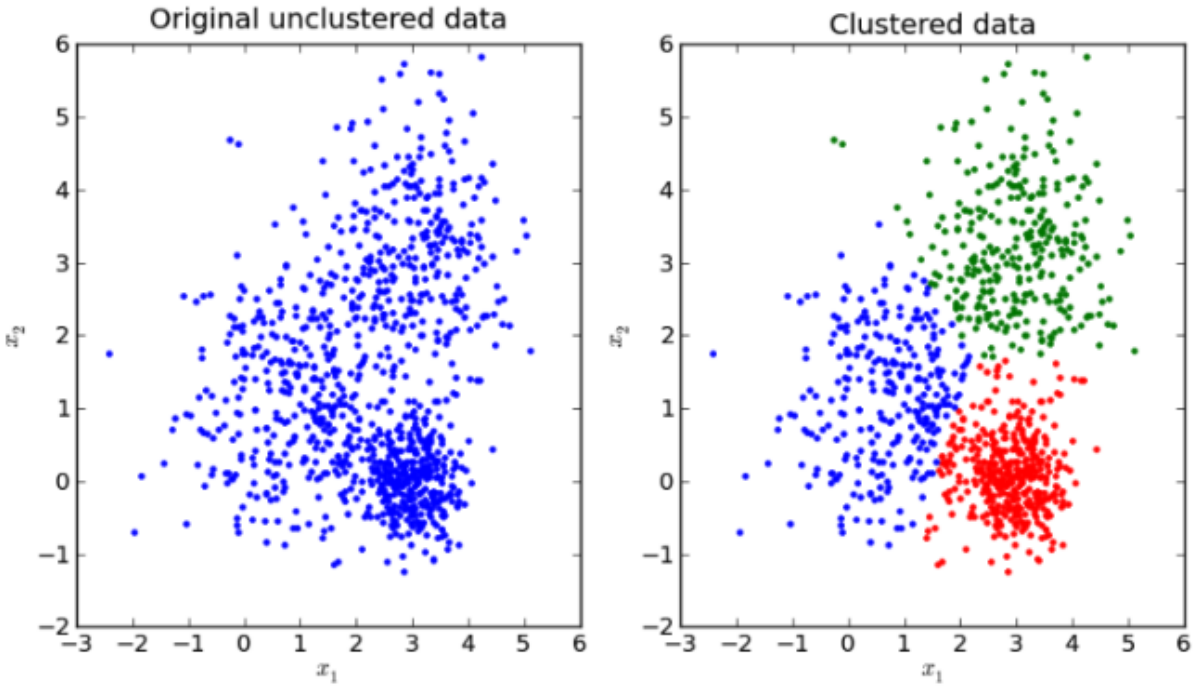
Типы задач в машинном обучении

Кластеризация

• Сегментация клиентов (например, для таргетирования рекламы)

• Группирование текстов по темам (например, новости по категориям)

• Группировка изображений (например, в фотогалереях)



Типы задач в машинном обучении

Кластеризация

Метод машинного обучения без учителя, используемый для разделения объектов на группы (кластеры) таким образом, чтобы объекты внутри одного кластера были максимально похожи друг на друга, а объекты из разных кластеров — как можно более различны. При кластеризации целевая переменная отсутствует, а группы формируются на основе структуры данных.

В отличие от классификации, у нас нет обучающей выборки, а классы — не определены заранее.

Типы задач в машинном обучении

Другие задачи машинного обучения:

- Ранжирование
- Рекомендации
- Снижение размерности
- Поиск аномалий
- Прогнозирование временных рядов
- Обнаружение и отслеживание объектов

и т.д.

Процесс машинного обучения

Этапы решения задачи машинного обучения

Получение данных

Сбор данных — процесс, нацеленный на получение значимой информации с целью построения согласованного и полного датасета для конкретной цели бизнеса, например, для принятия решений, ответов на исследовательские вопросы или стратегического планирования.

- Определить, какую информацию вам нужно собирать
- Найти источники релевантных данных
- Выбрать способы и инструменты сбора данных
- Решить какой объем данных будет достаточным
- Подготовить технологию хранения данных

Разведочный анализ данных (EDA)

EDA — анализ основных свойств данных, нахождение в них общих закономерностей, распределений и аномалий.

Разведочный анализ данных (EDA)

Описание датасета

Прежде всего необходимо понять как можно больше про данные, с которыми будем работать:

- Посмотреть объем и полноту
- Проверить характеристики данных
- Определить распределение
- Узнать есть ли зависимость между признаками и признаков с целевой переменной

Разведочный анализ данных (EDA)

Заполнение пропусков

- Заполнение простыми статистиками (среднее, медиана, мода)
- Заполнение специальными значениями (константа, групповая статистика)
- Заполнение с помощью машинного обучения (регрессия, KNN, Random Forest/ XGBoost)
- Удаление пропущенных значений:
  - если пропусков мало, можно удалить строки
  - если пропусков слишком много, стоит рассмотреть удаление признака

Разведочный анализ данных (EDA)

Типы признаков

- **Бинарные** — принимают одно из двух возможных значений и чаще всего выражаются комбинацией 0 и 1 или -1 и 1

Разведочный анализ данных (EDA)

Типы признаков

- **Бинарные** — принимают одно из двух возможных значений и чаще всего выражаются комбинацией 0 и 1 или -1 и 1
  Примеры:
  мужчина/женщина, купил/не купил, имеется что-либо / не имеется, здоров / болен и т. д.

Разведочный анализ данных (EDA)

Типы признаков

- **Бинарные** — принимают одно из двух возможных значений и чаще всего выражаются комбинацией 0 и 1 или -1 и 1
  Примеры:
  мужчина/женщина, купил/не купил, имеется что-либо / не имеется, здоров / болен и т. д.

- **Категориальные** — принимают одно из некоторого конечного множества значений. Для дальнейшего использования в модели должны быть переведены в числовой формат (закодированы). Могут быть номинальными (категории не сравнимы: например цветовая палитра) или ранговыми (в категориях подразумевается порядок: например, воинские звания)

Разведочный анализ данных (EDA)

Типы признаков

- **Бинарные** — принимают одно из двух возможных значений и чаще всего выражаются комбинацией 0 и 1 или -1 и 1
  Примеры:
  мужчина/женщина, купил/не купил, имеется что-либо / не имеется, здоров / болен и т. д.

- **Категориальные** — принимают одно из некоторого конечного множества значений. Для дальнейшего использования в модели должны быть переведены в числовой формат (закодированы). Могут быть номинальными (категории не сравнимы: например цветовая палитра) или ранговыми (в категориях подразумевается порядок: например, воинские звания)
  Примеры:
  оценки на экзамене, палитра цветов, город рождения, класс объекта недвижимости и т. д.

Разведочный анализ данных (EDA)

Типы признаков

- **Непрерывные** — принимают одно из непрерывного подмножества множества действительных чисел. Могут быть целочисленными (тип int) или вещественными (float)

Разведочный анализ данных (EDA)

Типы признаков

- **Непрерывные** — принимают одно из непрерывного подмножества множества действительных чисел. Могут быть целочисленными (тип int) или вещественными (float)
  Примеры:
  цена, вес, рост, температура, длина, высота, влажность, угол наклона, доход и т. д.

Разведочный анализ данных (EDA)

Типы признаков

- **Непрерывные** — принимают одно из непрерывного подмножества множества действительных чисел. Могут быть целочисленными (тип int) или вещественными (float)
  Примеры:
  цена, вес, рост, температура, длина, высота, влажность, угол наклона, доход и т. д.

- **Географические данные** — данные формата широта-долгота на поверхности Земли

Разведочный анализ данных (EDA)

Типы признаков

- **Непрерывные** — принимают одно из непрерывного подмножества множества действительных чисел. Могут быть целочисленными (тип int) или вещественными (float)
  Примеры:
  цена, вес, рост, температура, длина, высота, влажность, угол наклона, доход и т. д.

- **Географические данные** — данные формата широта-долгота на поверхности Земли

- **Данные о дате и времени** — специальный формат для описания времени наступления некоторого события или его длительности (datetime в pandas)

Разведочный анализ данных (EDA)

Типы признаков

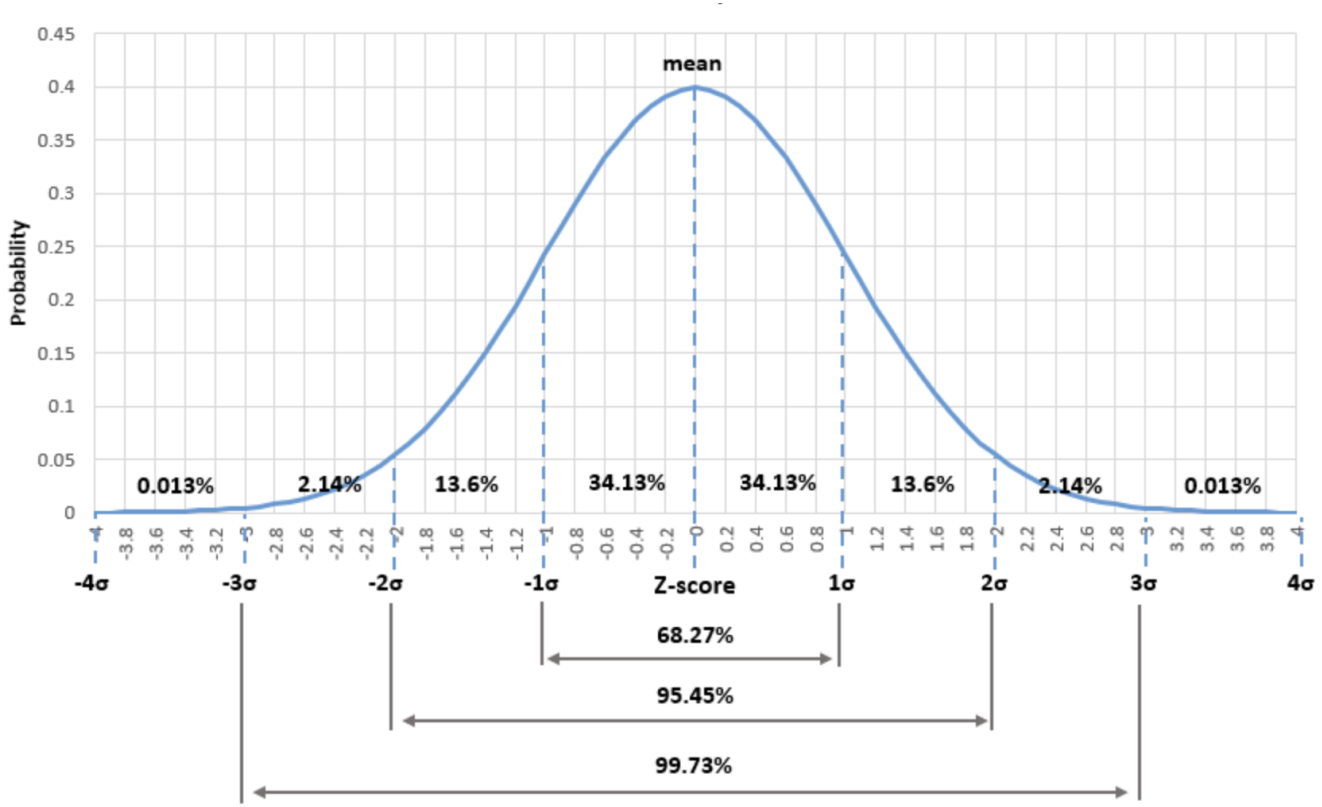
Пример датасета с информацией о пассажирах «Титаника».
Какие типы признаков здесь присутствуют?

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

Разведочный анализ данных (EDA)

Обработка выбросов

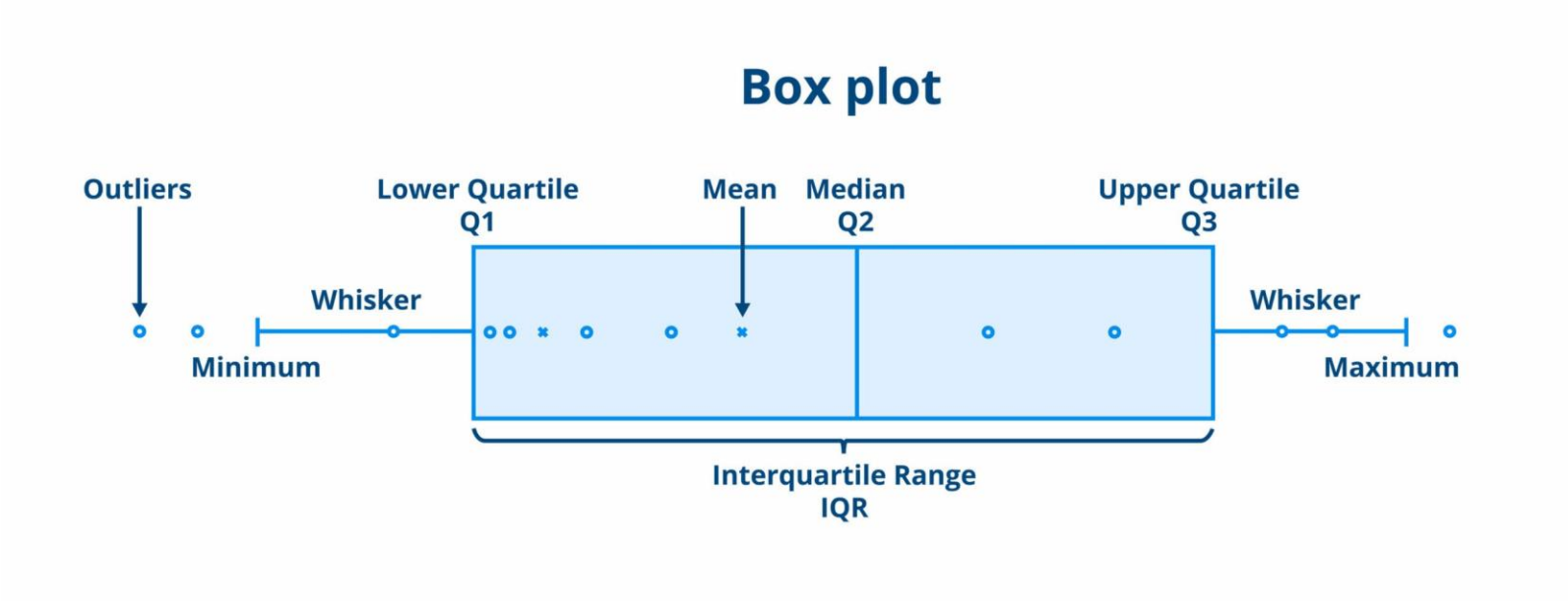
- **Статистика**
  - Z-оценка: если значение лежит дальше, чем ± 3 стандартных отклонения от среднего, это выброс



Разведочный анализ данных (EDA)

Обработка выбросов

- **Статистика**
  - Межквартильный размах (IQR)



Разведочный анализ данных (EDA)

Обработка выбросов

- **Статистика**

Критерий	Z-score	IQR
Основан на	Среднем и стандартном отклонении	Квартили (медиана, Q1, Q3)
Подходит для	**Нормально распределенных данных**	**Ненормальных / скошенных распределений**
Простота реализации	Прост в реализации	Прост в реализации
Надежность	Может дать сбои при наличии сильных выбросов	Надежнее для «реальных» грязных данных

Разведочный анализ данных (EDA)

Обработка выбросов

- **Машинное обучение**
  - Isolation Forest
    Специальный алгоритм для обнаружения выбросов
  - One-Class SVM
    Лучше подходит для высокоразмерных данных
  - Local Outlier Factor (LOF)
    Находит аномалии на основе плотности соседей

Разведочный анализ данных (EDA)

Обработка выбросов

- **Визуализация**
  - Boxplot
    Простая и наглядная визуализация межквартильного размаха
  - Scatter plot / Pairplot
    Помогает увидеть выбросы в мультипризнаковом пространстве
  - Time series plot
    Для временных рядов

Разведочный анализ данных (EDA)

Обработка выбросов

- **Оставить как есть!**

Иногда выбросы — это не ошибки в данных, а реальные значения, представляющие важность для работы алгоритма, и их важно сохранить (например, при анализе технических сбоев или выявлении мошенничества)

Разведочный анализ данных (EDA)

Feature Engineering

- **Извлечение признаков (Feature Extraction)** — конструирование новых признаков из исходных данных

Разведочный анализ данных (EDA)

Feature Engineering

- **Извлечение признаков (Feature Extraction)** — конструирование новых признаков из исходных данных
- **Трансформация признаков (Feature Transformation)** — изменение существующих признаков для повышения степени их пригодности для моделирования (нормализация, стандартизация, кодирование, логарифмирование и т. д.)

Разведочный анализ данных (EDA)

Feature Engineering

- **Извлечение признаков (Feature Extraction)** — конструирование новых признаков из исходных данных
- **Трансформация признаков (Feature Transformation)** — изменение существующих признаков для повышения степени их пригодности для моделирования (нормализация, стандартизация, кодирование, логарифмирование и т. д.)
- **Отбор признаков (Feature Selection)** — выделение наиболее значимых и релевантных признаков из данных для снижения размерности датасета и повышения эффективности последующего построения модели

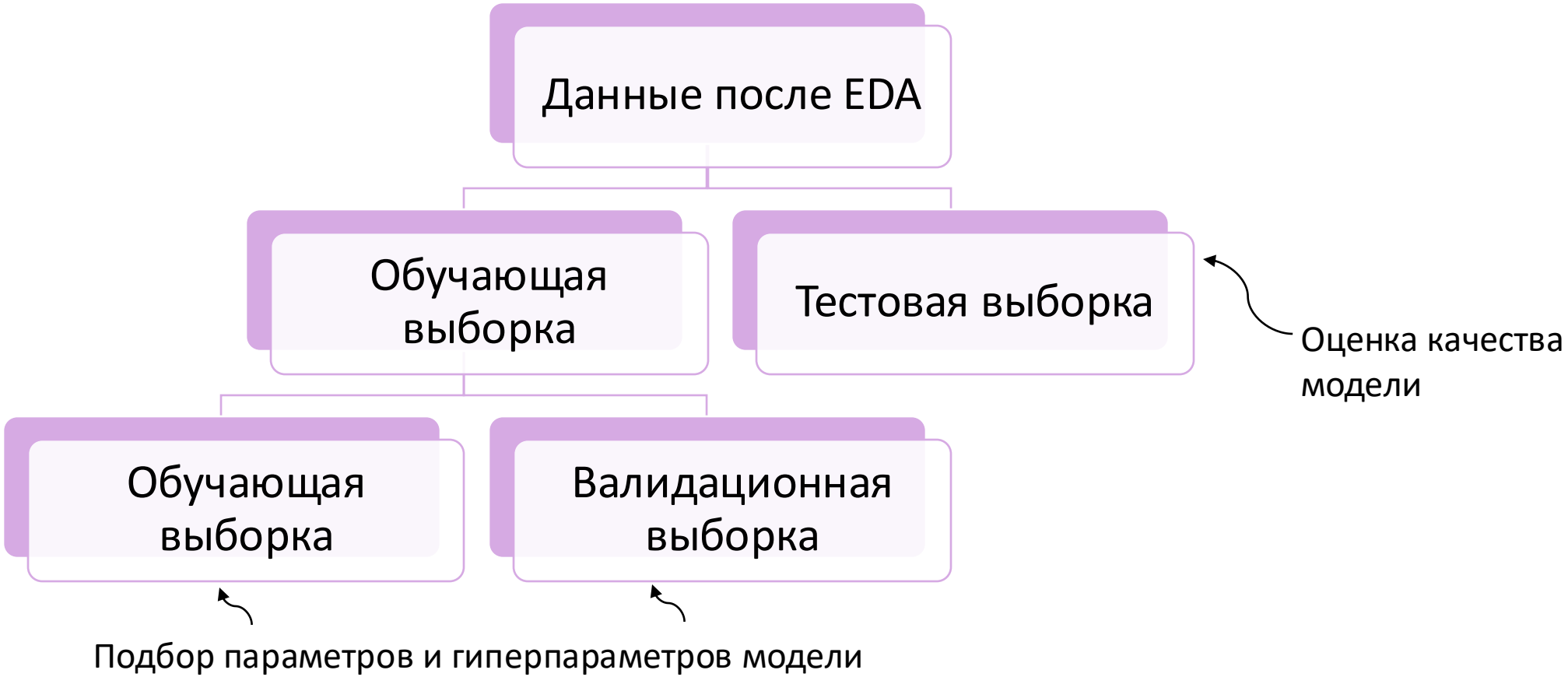
Обучение и валидация модели

Процесс

1. Выбор модели (линейные модели, деревья, бустинги, нейронные сети)
2. Обучение модели
3. Валидация модели (оценка качества модели на тестовых данных)
4. Подбор гиперпараметров модели
5. Выбор наилучшей модели

Обучение и валидация модели

Разбиение данных на обучающую и тестовую выборку



Обучение и валидация модели

Параметры и гиперпараметры

Гиперпараметры настраиваются и фиксируются до этапа настройки параметров на обучающей выборке.

Примеры:
шаг градиентного спуска, значение силы регуляризации в линейной модели, глубина решающего дерева и т. д.

Параметры настраиваются в процессе обучения модели на данных.

Примеры:
веса в линейной регрессии, нейросетях, структура решающего дерева

Обучение и валидация модели

Функции потерь и метрики качества

Функция потерь — функция, измеряющая качество работы алгоритма (ошибку предсказания модели) на объектах обучающей выборки.

Пример: MSE (среднеквадратичная ошибка, разница между предсказанным значением и истинным, возведенная в квадрат)

Метрика качества — функция, измеряющая качество работы алгоритма (ошибку предсказания модели) на данных, которые модель еще не видела, и использующаяся для сравнения моделей.

Примеры: MAE (абсолютная ошибка, модуль разницы между предсказанным значением и истинным), RMSE (корень из MSE)

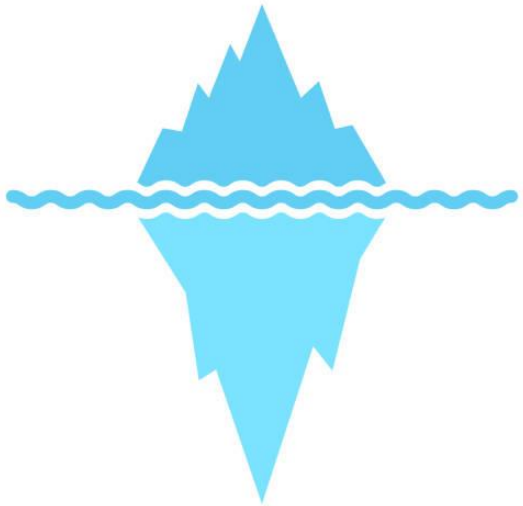
Обучение и валидация модели

Обучение и валидация

Обучение — это по сути процесс подбора весов или параметров модели так, чтобы оптимизировать (в большинстве случаев — минимизировать) функцию потерь

Валидация — этап оценки модели после обучения: вычисляются метрики качества, подбираются гиперпараметры и выбирается наилучшая модель

Что после обучения?



Что после обучения?

Пошаговая сборка пайплайна ML-проекта

1. Jupyter Notebook и csv-файл

Обучение модели,
верхушка айсберга

Что после обучения?

Пошаговая сборка пайплайна ML-проекта

1. Jupyter Notebook и csv-файл

Обучение модели,
верхушка айсберга

Что после обучения?

Пошаговая сборка пайплайна ML-проекта

1. Jupyter Notebook и csv-файл
2. Git, DVC
3. Airflow и DAG
4. Логирование метрик и параметров. MLFlow
5. Model Registry
6. Внедрение мониторинга с алертами
7. Feature Store
8. Переобучение модели
9. Внедрение мониторинга

Обучение модели, верхушка айсберга

Айсберг :)

Компоненты


Оркестрация / управление ML процессами
Хранилище кода VCS
Инструментарий CI/CD
Мониторинг и Журналирование
Ноутбуки / IDE
Управление экспериментами
Реестр моделей
Инструментарий среды исполнения / (Распределенные) Вычисления
Реестр артефактов
Реестр докер-образов
Системы хранения данных

Инструментарий

- Система контроля версий
  – Git, GitHub, GitLab, DVC
- CI/CD
  – GitLab CI, Jenkins
- Логирование
  – MLflow
  – Weights & Biases, Neptune
- Оркестрация
  – Airflow, Kubernetes, KubeFlow, Argo Workflows
- Реестр моделей
  – MLflow, DVC
- Деплой
  – Docker, Kubernetes
  – Amazon SageMaker, Azure ML
- Мониторинг
  – Grafana, Prometheus
  – Evidently, Great Expectations
- Feature store
  – Feast, Tecton
  – Amazon SageMaker Feature Store

Квиз

Мы решаем задачу определения вида животного по фотографии. Что является целевой переменной?

a) Одна фотография

b) Вид животного (кошка, собака, енот…)

c) Наличие ушей на фотографии, количество лап, цвет шерсти

d) Невозможно определить

Квиз

Что является целевой переменной в задаче регрессии?

a) Класс (например, кошка или собака)

b) Числовая непрерывная величина

c) Группа объектов (кластер)

d) Нет верного ответа

Квиз

К какому типу относится задача определения тональности отзыва на фильм (положительный или отрицательный)?

a) Классификация

b) Регрессия

c) Кластеризация

d) Невозможно определить

Квиз

Какой пример относится к обучению с учителем (supervised learning)?

a) Прогноз цены квартиры

b) Поиск аномалий

c) Кластеризация изображений

d) Нет верного ответа

Рекомендуемая литература

- Владимир Савельев – Статистика и котики [easy]
- Хендбук по машинному обучению от Яндекс Образования [medium]
- Сергей Николенко – Машинное обучение: основы [hard]
- Михаил Лагутин – Наглядная математическая статистика [medium+]
- Джеймс Г. и др. – Введение в статистическое обучение с примерами на Python [hard]

*Ссылки активны