Понимание задачи

Вопрос: какой тип задачи?

- Регрессия: $Y \in \mathbb{R}$
  - Сколько денег потеряет банк, выдав кредит клиенту?
- Классификация: $Y = \{1, \ldots, K\}$
  - Сможет ли клиент вернуть кредит?
- Кластеризация: $Y = ?$
  - Найти группы клиентов банка, имеющих схожее поведение

Вопрос: как измеряется качество решения?

- В регрессии:
  - MSE: $\frac{1}{\ell}\sum_{i=1}^{\ell}(a(x_i) - y_i)^2$
  - MAE: $\frac{1}{\ell}\sum_{i=1}^{\ell}|a(x_i) - y_i|$
- В классификации:
  - Доля верных ответов: $\frac{1}{\ell}\sum_{i=1}^{\ell}[a(x_i) = y_i]$
  - Точность и полнота
  - AUC-ROC, AUC-PRC
- В кластеризации:
  - Зависит от конкретной задачи

Вопрос: как измеряется качество решения?

Бизнес-метрики:

- Прибыль интернет-магазина после внедрения рекомендательной системы
  - Алгоритм предсказывает, какой товар купит пользователь
  - Оптимизация прибыли не эквивалентна оптимизации числа верных ответов!
- Количество кликов в письмах с рекомендациями отелей
  - Алгоритм предсказывает, кликнет ли пользователь по рекламе отеля
  - Сколько отелей рекомендовать — один, три, шесть?
  - Качество фотографий отеля важнее, чем качество модели?

Данные

Вопрос: на основе каких данных будем решать задачу?

- Числовые признаки: возраст, доход, ...
- Категориальные признаки: образование, цвет, идентификатор пользователя
- Текстовые признаки
- Изображения, сигналы
- Координаты

Формирование признаков (feature engineering)

Задача: получить матрицу «объекты-признаки»

- Преобразование числовых признаков
- Извлечение числовых признаков из сырых данных
  - Категориальные признаки
  - Текстовые признаки

Предобработка данных

Данные могут быть «грязными»:

- Выбросы
- Шумы в признаках
- Пропущенные значения

Мусор на входе — мусор на выходе

Построение алгоритма

Основные семейства в задачах обучения с учителем:

- Линейные модели
- Композиции деревьев (градиентный бустинг, случайный лес)
- Нейронные сети (глубокое обучение — много слоев, сложная архитектура)

А также:

- Отбор признаков
- Понижение размерности

Оценивание качества

Как оценить качество алгоритма и настроить гиперпараметры?

- Отложенная выборка
- Кросс-валидация
  - Сколько блоков?
- Как разбивать данные на блоки?

Заключение

Основные этапы анализа данных:

1. Понимание задачи: постановка и мера качества
2. Понимание данных
3. Формирование признаков
4. Предобработка данных
5. Построение алгоритма
6. Оценивание качества