Лекция 1. Введение

Основы интеллектуального анализа данных

Полузёров Т. Д.

БГУ ФПМИ

Структура лекции

Цели анализа данных

Data mining - процесс извлечения знаний из различных источников данных (базы данных, текст, картинки). Полученные знания должны быть достоверными, полезными, интерпретируемыми.

Моделирование - процесс построения модели, хорошо описывающей закономерности, которые порождают данные.

Подходы к построению моделей:

- статистический
- на основе машинного обучения
- вычислительный

Типы задач

Основные две группы задач:

- Обучение с учителем:
  - Регрессия - прогноз численного значения
  - Классификация - определения класса объекта
  - Ранжирование - упорядочивание объектов
- Обучение без учителя:
  - Кластеризация - выделение семейств, групп в данных
  - Поиск ассоциативных правил - поиск зависимых событий
  - Понижение размерности - сжатие данных при разумной потере информации

Этапы решения задач

Классическая схема решения задачи состоит из этапов:

1. Определение задачи которую нужно решить
2. Сбор и подготовка данных
3. Определение используемых инструментов, моделей
4. **Построение модели**
5. **Первичная оценка качества модели (offline evaluation)**
6. Внедрение или доработка модели
7. Оценка результатов работы в продакшене (online evaluation)

Постановка задачи. Обучение с учителем

$\mathbb{X}$ - множество объектов

$\mathbb{Y}$ - множество ответов

$y^* : \mathbb{X} \to \mathbb{Y}$ -неизвестная зависимость (target function)

Дано:

$X = \{x_1, ..., x_\ell\} \subset \mathbb{X}$ - обучающая выборка (samples)

$Y = \{y_1, ..., y_\ell\} = \{y^*(x_i), i = 1...\ell\} \subset \mathbb{Y}$ - известные ответы (targets)

Необходимо:

Найти алгоритм (решающую функцию, модель) $a : \mathbb{X} \to \mathbb{Y}$ приближающую y^* **на всём** множестве $\mathbb{X}$

Признаковое описание объектов

Отображения $f_j : \mathbb{X} \to D_j, j = 1, ..., n$ - признаки объекта (features), измерение некоторох характеристик объекта
Вектор $(f_1(x), ..., f_n(x))$ - признаковое описание объекта x.

Матрица "объекты-признаки":

$$F = (f(x_{ij}))_{\ell \times n} = \begin{pmatrix} f_1(x_1) & ... & f_n(x_1) \\ ... & ... & ... \\ f_1(x_\ell) & ... & f_n(x_\ell) \end{pmatrix}$$

Далее будем отождествлять признаковое описание объекта с самим объектом:
$x := (f_1(x), ..., f_n(x))$, т.е. $X := F$

Типы признаков

Основные типы признаков:

- $D_j = \{0, 1\}$ - **бинарный** признак f_j
- $|D_j| < \infty$ и определена только операция сравнения на равенство - **категориальный** признакf_j
- $|D_j| < \infty$ f_j и определены операции сравнения больше, меньше, равенство - **порядковый** (ранговый) признак f_j
- $D_j \subseteq \mathbb{R}$ - **количественный** признак f_j

Примеры:

- Цвет - категориальный признак, нельзя сказать "Красный"> "Синий"
- Офицерские звания - пример порядкового признака, можно ортировать категории
- Время/дата - может проявлять свойства непрерывных, категориальных, циклических типов

Форма множества ответов - определяет тип задачи

Задача классификации:

- $\mathbb{Y} = \{0, 1\}$ - бинарная классификация
- $\mathbb{Y} = \{1, ...M\}$ - на M непересекающихся классов (multiclass)
- $\mathbb{Y} = \{0, 1\}^M$ - на M классов, которые могут пересекаться (multilabel)

Восстановления регрессии:

- $\mathbb{Y} = \mathbb{R}$
- $\mathbb{Y} = \mathbb{R}^m$

В задачах "обучения без учителя" - множество $\mathbb{Y}$ не определено

Модель - как семейство параметризованных функций

Модель - параметрическое семейство функций

$$\mathbb{A} = \{a(x, \theta) | \theta \in \Theta\}$$

где $a : \mathbb{X} \times \Theta \to \mathbb{Y}$ - фиксированная функция, Θ - множество допустимых значений θ

Пример:

- $\{a(x) = \sum_{j=1}^{n} \omega_j x_j \ | \omega_j \in \mathbb{R}\}$ - семейство линейных моделей для задачи регрессии, $\mathbb{Y} = \mathbb{R}$
- $\{a(x) = [(\sum_{j=1}^{n} \omega_j x_j) > 0] | \omega_j \in \mathbb{R}\}$ - семейство линейных моделей для бинарной классификации, $\mathbb{Y} = \{0, 1\}$

Метод обучения

Метод обучения (learning algorithm) - это отображение вида

$$\mu : (\mathbb{X} \times \mathbb{Y}) \to \mathbb{A}$$

которое произвольной конечной выборке $(X \times Y) = \{(x_i, y_i\}_{i=1}^{\ell}$ ставит в соответствие некоторый алгоритм $a \in \mathbb{A}$

Обучить модель (fit) - значит с помощью метода обучения μ определить конкретные значения параметров для модели из выбранного семейства.

Функционалы качества

$\mathcal{L}(a, x)$ - **функция потерь** (loss function) - неотрицательная функция пропорциональная величине ошибки алгоритма $a \in \mathbb{A}$ на объекте $x \in \mathbb{X}$, если верный ответ есть $y \in \mathbb{Y}$

Функции потерь для задач классификации:

- $\mathcal{L}(a, x) = [a(x) \neq y]$ - индикатор ошибки

Функции потерь для задач регрессии:

- $\mathcal{L}(a, x) = |a(x) - y|$ - абслолютное значение ошибки
- $\mathcal{L}(a, x) = (a(x) - y)^2$ - квадрат ошибки

Эмпирический риск - функционал качества алгоритма a на конечной выборке $X \subset \mathbb{X}$

$$Q(a, X) = \frac{1}{\ell} \sum_{i=1}^{\ell} \mathcal{L}(a, x_i)$$

Основной метод обучения

Метод **минимизации эмпирического риска**:

$$\mu(X) = arg \min_{a \in A} Q(a, X)$$

Пример: метод наименьших квадратов, $Y = \mathbb{R}, \mathcal{L}$ - квадратична

$$\mu(X) = arg \min_{a \in A} \sum_{i=1}^{\ell} (a(x_i, \theta) - y_i)^2$$

Два этапа модели:

- Этап обучения (fit): по имеющейся выборке X с помощью метода обучения μ построить a.
- Этап применения обученной модели (predict): $\hat{y}_i = a(x_i')$

Обобщающая способность

Если минимум функционала $Q(a, X)$ достигается на алгоритме a, то это еще не гарантирует, что a будет хорошо приближать целевую зависимость на произвольной контрольной выборке $X' \in \mathbb{X}$

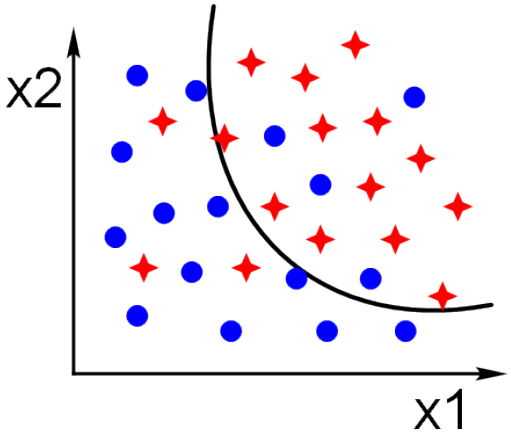
Обобщающая способность метода μ характеризуется величиной $Q(\mu(X), X')$, где X и X' получены из одного и того же неизвестного распределения $\mathbb{X}$

Крайние ситуации при обучении:

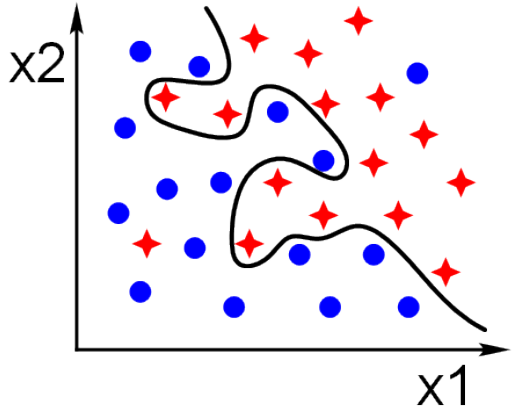
- **Недообучение** - ситуация, когда качество плохое и на X, и на X'
- **Переобучение** - качество на X хорошее, но на X' существенно хуже

Пример недо- и переобучения

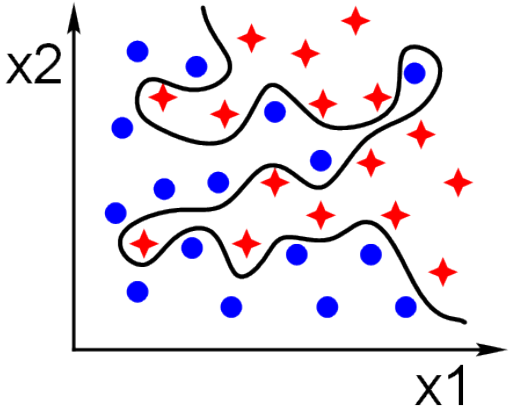
Недообучение



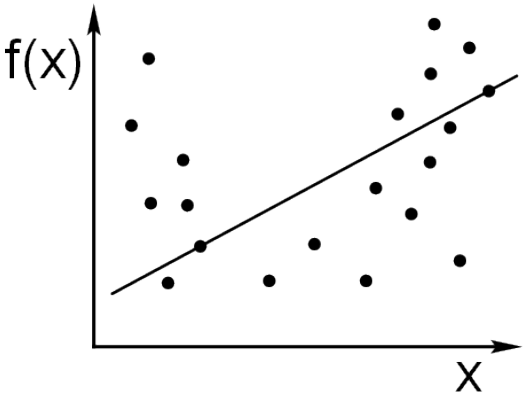
Оптимум



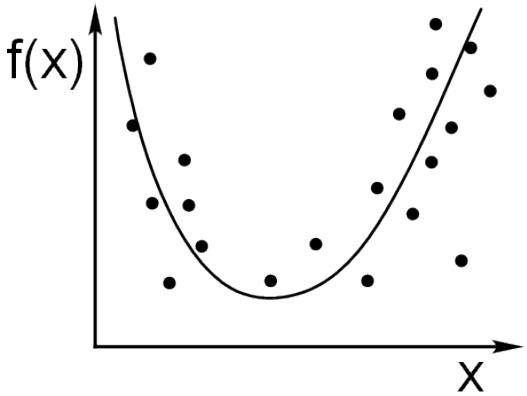
Переобучение



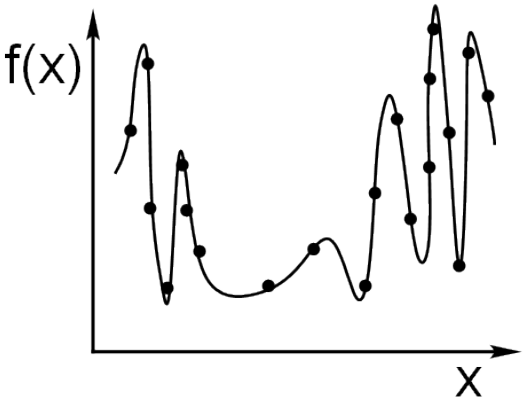
Недообучение



Оптимум



Переобучение



Переобучение - проблема обобщающей способности

Из-за чего возникает переобучение? Избыточная сложность пространства параметров Θ позволяет черезмерно точно подстроиться под обучающую выборку. Переобучение есть всегда, когда оптимизация идет по конечной выборке

Избавиться нельзя. Как минимизировать?

- Использовать класс более "простых"моделей
- Накладывать ограничение на параметры модели - регуляризация
- Увеличить обучающую выборку

Эмпирические оценки обобщающей способности

- Отложенная выборка (hold-out), $X = X_{train} \sqcup X_{test}$

$$HO(\mu, X_{train}, X_{test}) = Q(\mu(X_{train}), X_{test}) \to min$$

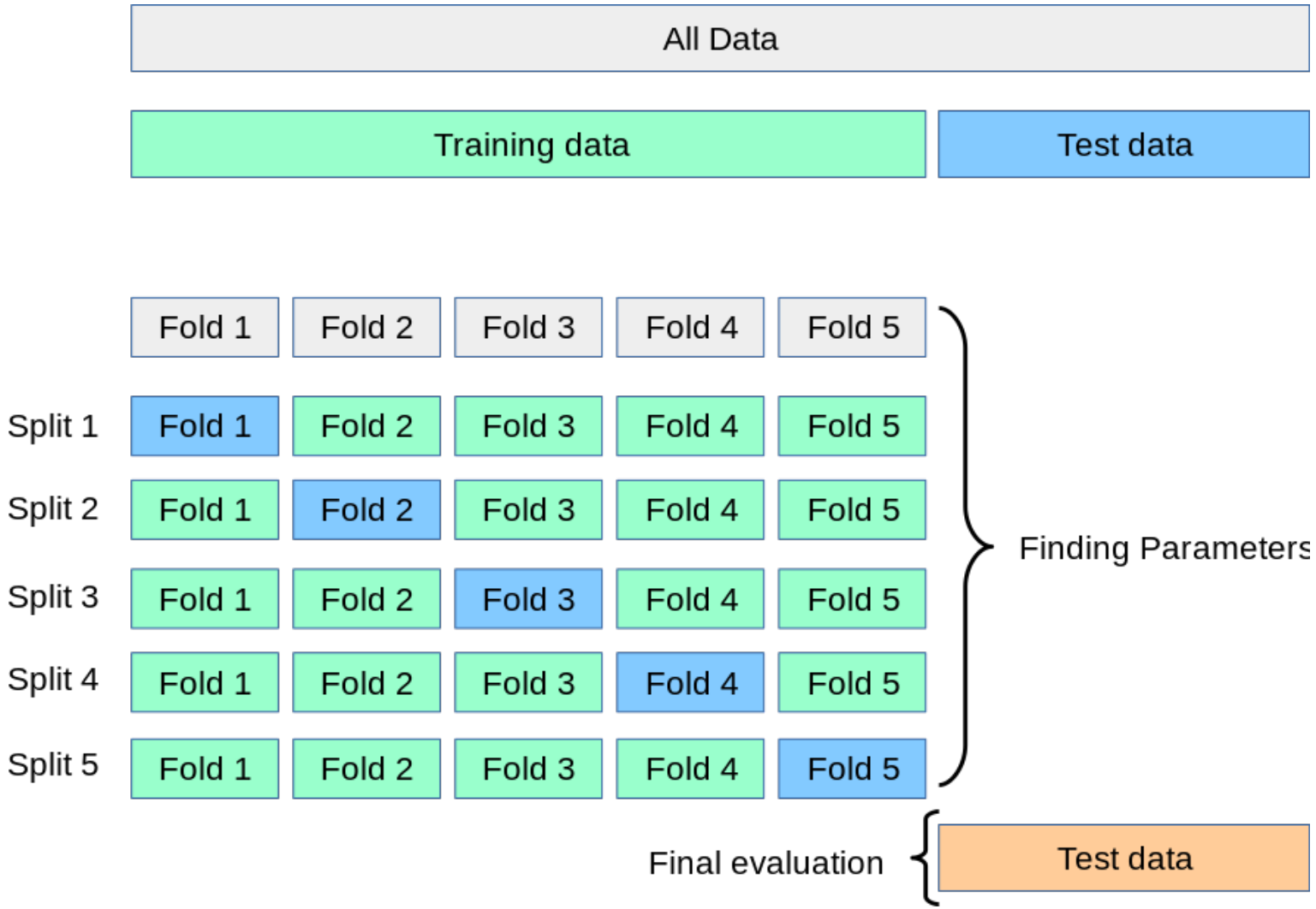
- Скользящий контроль (leave-one-out):

$$LOO(\mu, X) = \frac{1}{L} \sum_{i=1}^{L} Q(\mu(X \setminus \{x_i\}), x_i) \to min$$

- Кросс-проверка (cross-validation), $X = X_1 \sqcup X_2 \sqcup ... \sqcup X_k$

$$CV(\mu, X^L) = \frac{1}{k} \sum_{i=1}^{k} Q(\mu(X \setminus X_k), X_k) \to min$$

Кросс-валидация



Кредитный скоринг

Объекты - заявки клиентов на кредит
Цель - одобрить или отклонить заявку

Признаки:

- Бинарные: пол, наличие авто, имеет ли действующие кредиты
- Непрерывные: зарплата, сумма кредита
- Порядковые: образование, должность
- Категориальные: тип кредита, семейный статус

Особенности задачи:

- Дисбаланс классов: очень мало дефолтных
- Требование оценки вероятности дефолта
- Интерпретируемость модели

Отток клиентов

Объекты - абонент в определенный момент времени
Цель - распознавать риск ухода клиента

Признаки:

- Бинарные: пол, подключался ли во время акций
- Непрерывные: месячный расход трафика, число подключенных функций
- Категориальные: источник привлечения клиента

Особенности задачи:

- Оценивание вероятностей
- Сверхбольшие выборки
- Сложность в формировании признакового описания объектов

Биометрическая идентификация

Объекты - образцы отпечатков пальцев
Цель - идентифицировать человека

Особенности задачи:

- Нетривиальное преобразование входных данных в информативные признаки
- Требование сверхвысокой точности

Оценка стоимости недвижимости

Объекты - описание объекта недвижимости
Цель - оценить стоимость
Признаки:

- Бинарные: комерческая ли, наличие балкона, лифта, мусоропровода
- Непрерывные: площадь, год постройки дома
- Категориальные: район города

Особенности задачи:

- Данные могут быть очень разнородной
- Стоимость меняется со временем: зависимость непостоянна
- Влияние внешних экономических факторов

Прогнозирование объемов продаж

Объекты - тройка (товар, магазин, день)
Цель - прогноз числа продаж

Особенности задачи:

- Разреженные данные
- Функция потерь сильно не симметрична

Ранжирование поисковой выдачи

Объекты - поисковой запрос
Цель - формирование выдачи по убыванию релевантности

Особенности задачи:

- Очень много данных,
- Требование быстрой обработки запросов
- Сложность формирования размеченной выборки

Резюме

- Основные понятия: объект, признак, модель, функция потерь, метод обучения, эмпирический риск, обобщающая способность
- Модель - функция, заданная с точностью до параметров. Обучить модель - найти оптимальный набор параметров
- Проблема описывается математически $\rightarrow$ сводится к задаче оптимизации $\rightarrow$ решается