ИНТЕЛЛЕКТУАЛЬНОЕ МОДЕЛИРОВАНИЕ

Образовательная программа «6B06103 – Компьютерная инженерия»

Лекция 4

Описательная статистика и предобработка данных

Неделя 4 · Tan, Steinbach, Kumar: главы 2.2, 2.3, 3.2

Цель и план лекции

Цель: описывать данные статистическими показателями и готовить их к построению моделей

1. Меры центра, разброса и положения
2. Пропуски, дубликаты, выбросы
3. Методы предобработки по учебнику Tan
4. Практика в pandas и scikit-learn

Описательная статистика

Сводит массив чисел к нескольким понятным показателям: центр, разброс, форма распределения

- **Частоты и мода** — для категориальных признаков
- **Центр** — среднее, медиана, усечённое среднее
- **Разброс** — размах, дисперсия, стандартное отклонение, IQR
- **Положение** — квартили и процентили

Среднее, медиана, усечённое среднее

```
{1, 2, 3, 4, 5, 90}
```

Показатель	Значение	К выбросам
Среднее	17,5	Чувствительно
Медиана	3,5	Устойчива
Усечённое среднее (40 %)	3,5	Устойчиво

Пример учебника Tan: отбрасываем 1 и 90 → (2 + 3 + 4 + 5) / 4

Меры разброса

$$s^2 = \frac{\sum_{i=1}^{m} (x_i - \bar{x})^2}{m-1} \qquad s = \sqrt{s^2}$$

- **Размах** $= \max - \min$ — один выброс его резко меняет
- **Дисперсия** и **стандартное отклонение** — разброс вокруг среднего (pandas делит на m – 1)
- Устойчивые альтернативы: **MAD** и **межквартильный размах**

Квартили и межквартильный размах

- **Q1** — 25 % наблюдений ниже
- **Q2** — медиана, 50 %
- **Q3** — 75 % наблюдений ниже

$\mathrm{IQR} = Q_3 - Q_1$ — разброс центральных 50 % данных

Не зависит от крайних значений → основа поиска выбросов

Пропуски и дубликаты

- **Удалить** строки или столбцы — если пропусков мало
- **Заполнить** — средним, медианой, модой, по группе, моделью
- **Игнорировать** — некоторые методы работают с пропусками
- **Дубликаты** завышают вес объектов — находим и удаляем

Цель Data Mining — найти закономерности, несмотря на проблемы качества

Выбросы: два метода

IQR: выброс вне [Q1 – 1,5·IQR; Q3 + 1,5·IQR]

z-оценка: $z = \frac{x - \bar{x}}{s}$, выброс при $|z| > 3$

- IQR устойчивее; z-оценка предполагает нормальное распределение
- Не удаляйте выброс, не поняв причину: это может быть мошенничество

Пример: зарплаты, тенге

```
500 000, 550 000, 600 000, 750 000,
800 000, 1 000 000, 2 000 000
```

Величина	Значение
Q1 / Q3	575 000 / 900 000
IQR	325 000
Верхняя граница	1 387 500

2 000 000 > 1 387 500 → выброс

Методы предобработки (Tan, 2.3)

Метод	Что делает
Агрегация	Объединяет объекты: дни → месяцы
Выборка	Берёт репрезентативную часть данных
Отбор признаков	Убирает избыточные и лишние признаки
Создание признаков	Месяц из даты, ИМТ $= \frac{\text{вес}}{\text{рост}^2}$
Дискретизация	Число → интервалы
Нормализация	Единый масштаб признаков

Выборка и отбор признаков

- **Простая случайная** выборка — с возвращением или без
- **Стратифицированная** — доля из каждой группы, редкие группы не теряются
- **Отбор признаков**: встроенный (в алгоритме), фильтр (до обучения), обёрточный (перебор подмножеств)
- **PCA** строит новые признаки с наибольшей изменчивостью

Дискретизация и бинаризация

- **Равная ширина** — интервалы одной длины (pd.cut)
- **Равная частота** — поровну объектов в интервале (pd.qcut)
- **One-hot-кодирование** — категория → набор столбцов 0/1 (pd.get_dummies)

Модели scikit-learn принимают только числа, поэтому категории кодируют

Нормализация и преобразования

Min-max: $x' = \dfrac{x - \min}{\max - \min} \to [0, 1]$

Стандартизация: $x' = \dfrac{x - \bar{x}}{s} \to$ среднее 0, $s = 1$

- Нужна методам на расстояниях и с регуляризацией
- Логарифм сжимает скошенные признаки: доход, цены

Статистика в pandas

```
df.describe()                # count, mean, std, квартили
x.mean(), x.median(), x.std()
stats.trim_mean(x, 0.2)  # усечённое
x.quantile(0.75) - x.quantile(0.25)
city.value_counts(normalize=True)
city.mode()[0]
```

trim_mean(x, 0.2) отбрасывает по 20 % с каждой стороны

Очистка и предобработка в коде

```
df = df.drop_duplicates()
df.isna().sum()
df['age'] = df['age'].fillna(df['age'].median())
df = df[df['salary'].between(low, high)]
pd.qcut(ages, q=3)
pd.get_dummies(df['city'])
StandardScaler().fit_transform(X)
```

Выводы

- Среднее и s чувствительны к выбросам, медиана и IQR — устойчивы
- Очистка: пропуски, дубликаты, выбросы, несогласованность
- Предобработка: агрегация, выборка, отбор и создание признаков, дискретизация, нормализация

Далее: визуализация данных в matplotlib

Контрольные вопросы

1. Какие показатели допустимы для номинальных, порядковых и количественных признаков?
2. Вычислите среднее, медиану и усечённое среднее ряда 3, 5, 6, 7, 8, 9, 60.
3. Найдите границы выбросов по IQR для ряда 10, 12, 13, 15, 16, 18, 45.
4. Когда уместна каждая стратегия обработки пропусков?
5. Сравните дискретизацию равной ширины и равной частоты, min-max и стандартизацию.
6. Напишите программу очистки DataFrame с пропусками, дубликатами и выбросами.

Литература

1. Tan P.-N., Steinbach M., Kumar V. Introduction to Data Mining. — Pearson, 2014. — 2.2, 2.3, 3.2.
2. McKinney W. Python for Data Analysis. — O’Reilly. — Chapter 7.
3. Грас Дж. Data Science. Наука о данных с нуля. — 2020. — Глава 5.
4. Документация pandas: Working with missing data.
5. Документация scikit-learn: Preprocessing data.