ИНТЕЛЛЕКТУАЛЬНОЕ МОДЕЛИРОВАНИЕ

Образовательная программа «6B06103 – Компьютерная инженерия»

Лекция 8

Повторение материала. Сквозной пример

Неделя 8

Цель и план лекции

Цель: систематизировать лекции 1–7 и закрепить анализ данных от исходной таблицы до оценки модели

1. Data Mining, данные и типы признаков
2. pandas, статистика и предобработка
3. Визуализация и разведочный анализ
4. Линейная регрессия и подготовка признаков
5. Типовые задачи и сквозной пример

Логика первой половины курса

Лекция	Главный вопрос	Этап CRISP-DM
1	Что такое Data Mining и какие задачи он решает?	Понимание бизнеса
2	Из чего состоят данные?	Понимание данных
3	Как работать с таблицей в Python?	Понимание и подготовка
4	Как описать и подготовить данные?	Подготовка данных
5	Как показать данные на графике?	Понимание данных
6	Какие закономерности есть в данных?	Понимание данных
7	Как построить первую модель?	Моделирование и оценка

Data Mining и процесс KDD

- **Data Mining** — автоматический поиск полезных и ранее неизвестных закономерностей в больших данных
- Поиск записи по номеру договора — не Data Mining

```
Данные → Предобработка → Data Mining →
Постобработка → Знание
```

- Самый трудоёмкий этап — **предобработка**
- **CRISP-DM** — шесть этапов с возможными возвратами

Основные задачи Data Mining

Задача	Тип обучения	Результат	Пример
Классификация	С учителем	Категория	Вернёт ли заёмщик кредит
Регрессия	С учителем	Число	Цена квартиры
Ассоциации	Без учителя	Правила	Товары, покупаемые вместе
Кластеризация	Без учителя	Группы	Сегменты клиентов
Аномалии	Чаще без учителя	Необычные объекты	Мошенническая операция

Есть ли целевая переменная? Если есть, она категориальная или числовая?

Типы признаков

Тип	Операции	Примеры	Статистики
Номинальный	=, ≠	Цвет глаз, ИИН, город	Мода, частоты
Порядковый	+ <, >	Оценка, размер S/M/L	+ медиана, процентили
Интервальный	+ +, −	°C, календарная дата	+ среднее, ст. отклонение
Относительный	+ ×, /	Возраст, масса, доход	+ среднее геометрическое

- Средний почтовый индекс бессмыслен, хотя он записан числом
- Асимметричный признак: важно только ненулевое значение

Меры близости

Данные	Мера	Формула
Числовые, один масштаб	Евклидово	$d = \sqrt{\sum_k (x_k - y_k)^2}$
Покоординатные отклонения	Манхэттенское	$d = \sum_k \lvert x_k - y_k \rvert$
Симметричные бинарные	SMC	$\frac{f_{11} + f_{00}}{f_{00} + f_{01} + f_{10} + f_{11}}$
Асимметричные бинарные	Жаккара	$\frac{f_{11}}{f_{01} + f_{10} + f_{11}}$
Разреженные векторы частот	Косинусное	$\cos(x, y) = \frac{x \cdot y}{\lVert x \rVert \, \lVert y \rVert}$

Перед вычислением расстояний признаки масштабируют

pandas на маленькой таблице

```
df = pd.DataFrame({'name': ['Aru', 'Bek', 'Dana', 'Erlan'],
                   'group': ['A', 'B', 'A', 'B'],
                   'score': [85, 62, 91, 70]})

df.loc[0:2, 'name']     # 3 строки: метка 2 включена
df.iloc[0:2, 0]         # 2 строки: позиция 2 не включена
df.loc[df['score'] > 65, ['name', 'score']]
df.groupby('group')['score'].mean()   # A: 88.0, B: 66.0
```

- df['name'] — Series, df[['name']] — DataFrame

pandas: на что обратить внимание

Что сделать	Как записать	Важно
Несколько столбцов	df[['a', 'b']]	Двойные скобки
По меткам	df.loc[строки, столбцы]	Граница включается
По позициям	df.iloc[строки, столбцы]	Граница не включается
Объединить условия	(усл1) & (усл2)	Скобки, а не and
Изменить значение	df.loc[0, 'a'] = 1	Не df['a'][0] = 1
Соединить по ключу	pd.merge(a, b, on='id')	inner — только совпавшие

Описательная статистика

$$s^2 = \frac{1}{m-1}\sum_{i=1}^{m}(x_i - \bar{x})^2, \qquad \mathrm{IQR} = Q_3 - Q_1$$

$$[Q_1 - 1{,}5 \cdot \mathrm{IQR};\ Q_3 + 1{,}5 \cdot \mathrm{IQR}]$$

- **Чувствительны к выбросам:** среднее, размах, стандартное отклонение
- **Устойчивы:** медиана, IQR
- Выброс нельзя удалять, не выяснив причину

Нормализация признаков

$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}}, \qquad z = \frac{x - \bar{x}}{s}$$

- **Min-max** приводит значения к диапазону [0, 1]
- **Стандартизация** даёт среднее 0 и стандартное отклонение 1
- Нужна для методов на основе расстояний и регуляризации

Очистка и предобработка

Проблема	Что сделать	Инструмент
Дубликаты	Найти и удалить	drop_duplicates
Пропуски	Удалить или заполнить	fillna, SimpleImputer
Невозможные значения	Заменить на пропуск	Условие в loc
Скошенное распределение	Логарифмировать	np.log1p
Разный масштаб	Нормализовать	StandardScaler
Номинальный признак	One-hot-кодирование	OneHotEncoder
Нужны категории	Дискретизировать	pd.cut, pd.qcut

Визуализация: какой график выбрать

Задача	График	matplotlib
Изменение во времени	Линейный	plot
Сравнение категорий	Столбчатая диаграмма	bar, barh
Распределение признака	Гистограмма, ECDF	hist, step
Сравнение групп	Ящики с усами	boxplot
Связь двух признаков	Диаграмма рассеяния	scatter
Доли целого	Круговая диаграмма	pie

Проверка готового графика — принципы **ACCENT**; у гистограммы по оси X интервалы, а не категории

Разведочный анализ данных

Этапы: обзор → качество → один признак → пары → многомерный анализ → выводы

$$r_{xy} = \frac{\operatorname{cov}(x, y)}{s_x s_y}, \qquad -1 \le r_{xy} \le 1$$

- Корреляция измеряет только **линейную** связь
- Она **не доказывает** причинность
- Смешение групп может изменить знак: в Iris −0,43, а внутри видов связь положительная

Линейная регрессия

$$\hat{y} = ax + b, \quad a = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2}, \quad b = \bar{y} - a\bar{x}$$

$$R^2 = 1 - \frac{\mathrm{SSE}}{\mathrm{SST}}, \quad \mathrm{SSE} = \sum_i (y_i - \hat{y}_i)^2, \quad \mathrm{SST} = \sum_i (y_i - \bar{y})^2$$

- МНК минимизирует сумму квадратов остатков
- Хорошие остатки выглядят как случайный шум
- Дуга — нелинейность; растущий разброс — логарифм

Подготовка признаков для модели

Проблема	Решение
Категории	One-hot-кодирование, drop='first'
Скошенная целевая переменная	Логарифмирование
Нелинейность	Полиномиальные признаки
Мультиколлинеарность	Корреляционная матрица, удаление признака
Утечка данных	Параметры — только по обучающей выборке, Pipeline

Задача 1. Статистика и выбросы

Дан ряд: 4, 8, 6, 5, 3, 7, 40

1. Среднее = 73 / 7 ≈ 10,43; медиана = 6
2. Q1 = 4,5; Q3 = 7,5; IQR = 3
3. Значение 40 выходит за верхнюю границу

$$4{,}5 - 1{,}5 \cdot 3 = 0, \qquad 7{,}5 + 1{,}5 \cdot 3 = 12$$

Без выброса среднее и медиана равны 5,5

Задача 2. Меры близости

x = (1, 1, 0, 0, 1, 0), y = (1, 0, 0, 0, 1, 1)

- $f_{11} = 2$, $f_{10} = 1$, $f_{01} = 1$, $f_{00} = 2$
- SMC = 4/6 ≈ 0,67; **Жаккар** = 2/4 = 0,5 — он и подходит для корзин

$$d_E = \sqrt{9 + 16} = 5, \qquad d_M = 3 + 4 = 7$$

Точки (1, 2) и (4, 6): E — евклидово, M — манхэттенское расстояние

Задача 3. Нормализация

Значения 10, 20, 30, 40

- Min-max: 0; 0,33; 0,67; 1
- Среднее 25, стандартное отклонение s ≈ 12,91

$$z(40) = \frac{40 - 25}{12{,}91} \approx 1{,}16$$

Задача 4. Простая регрессия

x	y	$x - \bar{x}$	$y - \bar{y}$	$(x - \bar{x})(y - \bar{y})$	$(x - \bar{x})^2$	$\hat{y}$	Остаток
1	3	−2	−1	2	4	2,6	0,4
2	4	−1	0	0	1	3,3	0,7
3	2	0	−2	0	0	4,0	−2,0
4	5	1	1	1	1	4,7	0,3
5	6	2	2	4	4	5,4	0,6
Σ				7	10		0

$$a = \frac{7}{10} = 0{,}7, \quad b = 4 - 0{,}7 \cdot 3 = 1{,}9, \quad \hat{y} = 0{,}7x + 1{,}9, \quad R^2 = 1 - \frac{5{,}1}{10} = 0{,}49 = r^2$$

Сквозной пример: данные

200 студентов: hours, attendance, group, score

```
df = pd.DataFrame({
    'hours': rng.uniform(1, 12, n).round(1),
    'attendance': rng.uniform(40, 100, n).round(0),
    'group': rng.choice(['A', 'B', 'C'], n)})
```

- Внесены проблемы: **10 пропусков**, **3 дубликата**, посещаемость **250 %**
- В группе C балл в среднем выше: заложена надбавка

Шаг 1–2. Обзор и очистка

```
print(df.shape)                # (203, 4)
print(df.isna().sum())         # hours: 10
print(df.duplicated().sum()) # 3

df = df.drop_duplicates()    # (200, 4)
df.loc[df['attendance'] > 100, 'attendance'] = np.nan
```

- describe показывает максимум посещаемости 250 %
- Пропуски заполним в конвейере: медиана — только по обучающей выборке

Шаг 3. Разведочный анализ

Группа	Студентов	Средний балл
A	69	62,4
B	66	62,0
C	65	68,4

Пара признаков	r
hours — score	0,86
attendance — score	0,26
hours — attendance	−0,06

- Сильнее всего балл связан с часами подготовки
- Предикторы почти не коррелируют: мультиколлинеарности нет
- Группа C набирает больше — включаем её в модель

Шаг 4. Конвейер и модель

```
numeric = make_pipeline(SimpleImputer(strategy='median'),
                        StandardScaler())
prep = ColumnTransformer([
    ('num', numeric, ['hours', 'attendance']),
    ('cat', OneHotEncoder(drop='first'), ['group'])])
model = Pipeline([('prep', prep),
                  ('reg', LinearRegression())])
model.fit(X_train, y_train)
```

Предобработка внутри Pipeline — **без утечки данных**

Шаг 5. Результат

Показатель на тесте	Модель	Предсказание средним
MAE, баллов	4,49	13,14
R^2	0,856	—

Признак	Коэффициент	Смысл
hours	11,14	Самый влиятельный числовой признак
attendance	4,23	Влияние слабее
group_B	2,27	Надбавка к группе A
group_C	4,79	Надбавка к группе A

Частые ошибки: данные

Ошибка	Как правильно
Считать среднее для кодов категорий	Использовать моду и частоты
Сравнивать корзины через SMC	Использовать коэффициент Жаккара
Считать расстояния без масштабирования	Стандартизировать признаки
Путать срезы loc и iloc	Проверять размер результата
Удалять выбросы автоматически	Сначала выяснить причину

Частые ошибки: анализ и модель

Ошибка	Как правильно
Ограничиваться статистиками	Смотреть на графики
Считать корреляцию причинностью	Проверять по группам и знаниям о предмете
Кодировать районы числами 1, 2, 3	Применять one-hot-кодирование
Масштабировать до разделения	Использовать Pipeline
Оценивать модель на обучающих данных	Использовать тестовую выборку

Выводы

- Лекции 1–7 образуют единый процесс: от задачи к модели
- Модель зависит от подготовки данных
- Сначала данные и графики, потом модель
- Оценка — только на данных, которых модель не видела
- Модель сравнивают с простым базовым предсказанием

Далее: вероятностные методы классификации

Контрольные вопросы

1. Опишите KDD и CRISP-DM; определите задачу: цена квартиры, сегменты клиентов, мошенничество.
2. Сравните четыре шкалы; определите тип ИИН, размера одежды, °C и дохода.
3. Чем отличаются df['a'] и df[['a']], loc и iloc?
4. Для ряда 2, 4, 5, 7, 9, 11, 30 найдите среднее, медиану, IQR и границы выбросов.
5. Опишите этапы EDA и назовите три ограничения корреляции.
6. Для x = (1, 2, 3, 4), y = (2, 4, 5, 9) найдите a, b, SSE, SST, R^2; что такое утечка данных?

Литература

1. Tan P.-N., Steinbach M., Kumar V. Introduction to Data Mining. — Pearson, 2014. — Chapters 1–3; Appendix D.
2. McKinney W. Python for Data Analysis. — O’Reilly. — Chapters 5, 7, 8, 9.
3. VanderPlas J. Python Data Science Handbook. — O’Reilly, 2016. — Chapters 3–5.
4. Грас Дж. Data Science. Наука о данных с нуля. — 2020. — Главы 3, 5, 14, 15.
5. Мюллер А., Гвидо С. Введение в машинное обучение с помощью Python. — 2017. — Глава 4.
6. Chapman P. et al. CRISP-DM 1.0. — SPSS Inc., 2000.
7. Документация scikit-learn: Pipelines and composite estimators.