ИНТЕЛЛЕКТУАЛЬНОЕ МОДЕЛИРОВАНИЕ

Образовательная программа «6B06103 – Компьютерная инженерия»

Лекция 7

Линейная регрессия и подготовка признаков

Неделя 7 · Tan, Steinbach, Kumar: приложение D

Цель и план лекции

Цель: строить, интерпретировать и оценивать линейную регрессию с правильно подготовленными признаками

1. Простая регрессия и метод наименьших квадратов
2. Ручной расчёт и коэффициент детерминации
3. Множественная регрессия
4. Подготовка признаков и утечка данных
5. Конвейер в scikit-learn

Задача регрессии

Предсказать **числовую** целевую переменную по признакам: цену квартиры, выручку, балл

- **Зависимая** переменная Y — что предсказываем
- **Независимые** X — предикторы
- Один предиктор — простая регрессия, несколько — множественная
- Обучение с учителем: ответы известны на обучающих данных

Простая линейная регрессия

$$\hat{y} = a \cdot x + b$$

- **a** — на сколько в среднем меняется Y при росте X на 1
- b — значение Y при $X = 0$
- Пример: балл $= 10 \cdot$ часы $+ 30$ — каждый час подготовки даёт +10 баллов

Метод наименьших квадратов

$$\mathrm{SSE} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \to \min$$

$$a = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2} \qquad b = \bar{y} - a \cdot \bar{x}$$

- Квадраты: ошибки разных знаков не компенсируются, большие штрафуются сильнее
- Знак наклона совпадает со знаком корреляции

Ручной расчёт

x	y	x − x̄	y − ȳ	произведение	$(x-\bar{x})^2$
1	2	−2	−2	4	4
2	4	−1	0	0	1
3	5	0	1	0	0
4	4	1	0	0	1
5	5	2	1	2	4
$\sum$				6	10

$$\bar{x}=3,\ \bar{y}=4 \to a=\frac{6}{10}=\mathbf{0{,}6},\ b=4-0{,}6\cdot 3=\mathbf{2{,}2} \to \hat{y}=0{,}6x+2{,}2$$

Коэффициент детерминации R²

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

В примере: $\text{SSE} = 2{,}4$, $\text{SST} = 6 \to R^2 = 1 - \frac{2{,}4}{6} = \mathbf{0{,}6}$ — модель объясняет 60 % изменчивости y

- $R^2 = 1$ — идеально, $R^2 = 0$ — не лучше среднего
- Остатки хорошей модели выглядят как случайный шум

Множественная регрессия

$$\hat{y} = a_1 x_1 + a_2 x_2 + \ \ldots \ + a_k x_k + b$$

- a_j — влияние x_j **при неизменных остальных** признаках
- Решение МНК: $w = (X^T X)^{-1} X^T y$ — нормальные уравнения
- Линейно зависимые признаки → $X^T X$ необратима

Подготовка признаков

Проблема	Решение
Номинальный признак	One-hot, drop='first'
Разный масштаб	StandardScaler
Скошенная цель (цены)	Логарифм
Нелинейная связь	PolynomialFeatures: x^2, $x_1 \cdot x_2$
Пропуски	SimpleImputer

Район кодируем не числами 1, 2, 3, а столбцами 0/1

Две ловушки

Мультиколлинеарность

Сильно связанные предикторы → неустойчивые коэффициенты с неожиданным знаком. Решение: убрать дубль или регуляризация

Утечка данных

Масштаб и заполнение пропусков считают только по обучающей выборке. Решение: Pipeline

Регрессия в scikit-learn

```
from sklearn.linear_model import LinearRegression
X = np.array([1, 2, 3, 4, 5]).reshape(-1, 1)
y = np.array([2, 4, 5, 4, 5])
model = LinearRegression().fit(X, y)
model.coef_, model.intercept_  # 0.6, 2.2
model.score(X, y)                  # R2 = 0.6
```

Совпадает с ручным расчётом. fit — обучить, predict — предсказать

Конвейер: цены квартир

```
prep = ColumnTransformer([
    ('num', StandardScaler(), ['area', 'rooms', 'floor']),
    ('cat', OneHotEncoder(drop='first'), ['district'])])
model = Pipeline([('prep', prep),
                  ('reg', LinearRegression())])
model.fit(X_train, y_train)
model.predict(X_test)
```

Предобработка обучается только на train — утечки нет

Результат модели

0,93

R^2 на тестовой выборке

3,3

MAE, млн тенге

- Наибольший коэффициент у площади (после стандартизации)
- Медеуский район: +12,7 млн к Алатаускому (заложено 12)

Выводы

- Коэффициенты находит МНК; они понятно интерпретируются
- R^2 — доля объяснённой изменчивости, остатки — проверка формы модели
- Качество зависит от подготовки признаков
- Pipeline защищает от утечки данных

Далее (после Midterm): классификация, наивный Байес

Контрольные вопросы

1. Дайте определение линейной регрессии; сравните простую и множественную.
2. Опишите МНК и выведите формулы a и b.
3. Вычислите a, b, SSE, SST и R^2 для $x = (1, 2, 3, 4)$, $y = (2, 3, 5, 6)$.
4. Как кодировать категориальные признаки и что такое «ловушка фиктивных переменных»?
5. Что такое мультиколлинеарность и утечка данных, как их предотвратить?
6. Создайте конвейер: масштабирование, кодирование, регрессия и R^2 на тесте.

Литература

1. Tan P.-N., Steinbach M., Kumar V. Introduction to Data Mining. — Pearson, 2014. — Appendix D. Regression.
2. VanderPlas J. Python Data Science Handbook. — O’Reilly, 2016. — Chapter 5.
3. Грас Дж. Data Science. Наука о данных с нуля. — 2020. — Главы 14–15.
4. Мюллер А., Гвидо С. Введение в машинное обучение с помощью Python. — 2017. — Глава 4.
5. Документация scikit-learn: Linear Models; Pipelines.