ИНТЕЛЛЕКТУАЛЬНОЕ МОДЕЛИРОВАНИЕ

Образовательная программа «6B06103 – Компьютерная инженерия»

Лекция 6

# Разведочный анализ данных (EDA) с pandas и seaborn

Неделя 6 · Tan, Steinbach, Kumar: глава 3

# Цель и план лекции

**Цель:** проводить разведочный анализ и строить статистические графики в pandas и seaborn

1. Цели и этапы разведочного анализа
2. Ковариация и корреляция
3. Графики pandas и seaborn
4. Многомерные данные: pairplot, параллельные координаты
5. Итоги анализа набора Iris

# Что такое разведочный анализ

**EDA** — знакомство с данными до моделирования: структура, закономерности, аномалии, гипотезы

- Подход Джона Тьюки (1977): сначала смотреть, потом проверять
- Какие признаки и типы? Есть ли пропуски и выбросы?
- Как признаки связаны друг с другом и с целевой переменной?

# Этапы разведочного анализа

1. **Обзор**: shape, info, head
2. **Качество**: пропуски, дубликаты, невозможные значения
3. **Один признак**: статистики, гистограммы, ящики
4. **Пары признаков**: рассеяние, корреляция, группы
5. **Многомерный обзор**: pairplot, тепловая карта
6. **Выводы**: план предобработки и гипотезы

# Набор данных Iris

150 цветков, 3 вида по 50, 4 признака в см: длина и ширина чашелистика и лепестка

| Вид | Длина лепестка | Ширина лепестка |
|---|---|---|
| setosa | 1,46 | 0,25 |
| versicolor | 4,26 | 1,33 |
| virginica | 5,55 | 2,03 |

Средние по видам: df.groupby('species').mean()

# Ковариация и корреляция

$$r = \frac{\operatorname{cov}(x_i, x_j)}{s_i \cdot s_j}, \quad \text{от } -1 \text{ до } 1$$

- Ковариация зависит от единиц, корреляция — нет
- $r \approx 1$ — сильная прямая связь, $r \approx -1$ — обратная, $r \approx 0$ — нет **линейной** связи
- Корреляция **не доказывает причинность**
- Для порядков и монотонных связей — корреляция Спирмена

# Корреляции в Iris

**0,96**

Длина и ширина лепестка почти дублируют друг друга

**−0,43**

Ширина чашелистика и длина лепестка в целом

Внутри каждого вида эта связь положительна (0,18; 0,56; 0,40). Смешение групп меняет знак — проверяйте корреляции по группам

# Графики прямо из DataFrame

```
num.hist(bins=15, figsize=(10, 6))  # все признаки
num.plot.box()
df.plot.scatter(x='petal_length', y='petal_width')
means.sort_values().plot.barh()
```

- Метод plot вызывает matplotlib «под капотом»
- df.hist() — быстрый первый взгляд на все распределения

# Библиотека seaborn

```
import seaborn as sns
sns.set_theme(style='whitegrid')
```

- Красивое оформление по умолчанию
- Принимает DataFrame и имена столбцов
- Параметр **hue** раскрашивает по категории
- Плотности, «скрипки», регрессия, тепловые карты из коробки

# Распределения и группы

```
sns.histplot(data=df, x='petal_length',
             hue='species', kde=True)
sns.kdeplot(data=df, x='sepal_width', hue='species')
sns.boxplot(data=df, x='species', y='petal_length')
sns.violinplot(data=df, x='species', y='sepal_width')
```

KDE — гладкая гистограмма; «скрипка» = ящик + плотность

# Связи между признаками

```
sns.scatterplot(data=df, x='petal_length',
                y='petal_width', hue='species')
sns.regplot(data=df, x='petal_length', y='petal_width')
sns.heatmap(num.corr(), annot=True, fmt='.2f',
            cmap='coolwarm', vmin=-1, vmax=1)
```

Тепловая карта корреляции находит связанные признаки с одного взгляда

# Многомерные данные

```
sns.pairplot(df, hue='species', corner=True)


from pandas.plotting import parallel_coordinates
parallel_coordinates(df, 'species')
```

- **Матрица рассеяния** — график для каждой пары признаков
- **Параллельные координаты** — объект = ломаная по осям признаков; классы образуют «пучки»

# Итоги разведочного анализа Iris

- Пропусков нет, одна повторяющаяся строка
- Классы сбалансированы — по 50 объектов
- Признаки лепестка бимодальны и сильно связаны ($r = 0{,}96$)
- Setosa линейно отделим; versicolor и virginica перекрываются
- Единицы одинаковые, но для методов на расстояниях масштабирование полезно

# Выводы

- EDA — обязательный шаг перед моделированием
- Корреляция — только линейная связь и не причинность
- pandas — быстрые графики, seaborn — статистические с группами
- Для многих признаков — pairplot и параллельные координаты

Далее: линейная регрессия

# Контрольные вопросы

1. Дайте определение EDA и перечислите его этапы.
2. Запишите формулы ковариации и корреляции; назовите ограничения корреляции.
3. Почему корреляция ширины чашелистика и длины лепестка в Iris отрицательна?
4. Сравните возможности pandas и seaborn для графиков.
5. Опишите матрицу рассеяния и параллельные координаты.
6. Для набора tips постройте ящики чаевых по дням, тепловую карту корреляции и назовите три самые связанные пары.

# Литература


1. Tan P.-N., Steinbach M., Kumar V. Introduction to Data Mining. — Pearson, 2014. — 3.1, 3.2.5, 3.3.4.
2. McKinney W. Python for Data Analysis. — O’Reilly. — Chapter 9.2.
3. VanderPlas J. Python Data Science Handbook. — O’Reilly, 2016. — Chapter 4.
4. Tukey J. W. Exploratory Data Analysis. — Addison-Wesley, 1977.
5. Документация seaborn: seaborn.pydata.org