ИНТЕЛЛЕКТУАЛЬНОЕ МОДЕЛИРОВАНИЕ

Образовательная программа «6B06103 – Компьютерная инженерия»

Лекция 1

Введение в интеллектуальное моделирование и Data Mining

Лектор к.ф.-м.н., ассоциированный профессор Уалиева Ирина Маратовна

Tan, Steinbach, Kumar. Introduction to Data Mining, глава 1

Цель и план лекции

ЦЕЛЬ

Сформировать целостное представление о задачах, этапах и инструментах интеллектуального анализа данных

1. Что такое Data Mining
2. Процесс KDD и методология CRISP-DM
3. Причины появления Data Mining
4. Четыре основные задачи
5. Обучение с учителем и без учителя
6. Первый анализ данных на Python

Что такое Data Mining

Data Mining — процесс автоматического обнаружения полезной, ранее неизвестной информации в больших хранилищах данных

Информационный поиск

Найти запись клиента по номеру договора, найти страницы по ключевым словам — ответ заранее понятен

Data Mining

Найти скрытые зависимости, группы похожих объектов, необычные наблюдения, сделать прогноз

Процесс извлечения знаний (KDD)



- **Предобработка** — самый трудоёмкий этап: объединение источников, очистка, отбор признаков, нормализация
- **Постобработка** — фильтрация и визуализация шаблонов, статистическая проверка
- «Замыкание цикла» — результаты встраиваются в принятие решений

Методология CRISP-DM

1

Понимание бизнеса

Цель и критерии успеха

2

Понимание данных

Сбор, обзор, качество

3

Подготовка данных

Очистка, признаки

4

Моделирование

Выбор и настройка алгоритмов

5

Оценка

Качество и соответствие цели

6

Внедрение

Использование и сопровождение

Между этапами возможны возвраты. Курс повторяет эту логику: данные (нед. 2–6) → модели → оценка (нед. 11–12)

Почему возник Data Mining

Трудность	Суть
Масштабируемость	Гигабайты и терабайты данных
Высокая размерность	Сотни и тысячи признаков
Сложные данные	Тексты, графы, временные ряды, ДНК
Распределённость	Данные у разных владельцев и в разных местах
Нетрадиционный анализ	Тысячи гипотез, данные собраны «попутно»

Data Mining объединяет статистику, машинное обучение, базы данных и параллельные вычисления

Два типа задач Data Mining

Предсказательные

Предсказать значение **целевой переменной** по значениям **объясняющих переменных** (предикторов)

Классификация, регрессия

Описательные

Найти закономерности, кратко описывающие данные: связи, группы, тренды, аномалии

Ассоциации, кластеры, аномалии

Четыре основные задачи

Классификация и регрессия

Вернёт ли заёмщик кредит?

Какой будет выручка?

Анализ ассоциаций

{подгузники} → {молоко}: что покупают вместе

Кластерный анализ

Статьи делятся на «экономику» и «здравоохранение»

Обнаружение аномалий

Операция по карте не похожа на профиль клиента

Обучение с учителем и без учителя

Задача	Тип обучения	Алгоритмы курса
Классификация	С учителем	Наивный Байес, логистическая регрессия, дерево
Регрессия	С учителем	Линейная регрессия, Ridge, Lasso
Кластеризация	Без учителя	k-средних
Ассоциации	Без учителя	Apriori, FP-Growth
Аномалии	Чаще без учителя	Метод IQR, z-оценка

С учителем — правильные ответы известны при обучении. **Без учителя** — алгоритм ищет структуру в самих данных.

Первый анализ: набор данных Iris

```
from sklearn.datasets import load_iris
iris = load_iris(as_frame=True)
df = iris.frame
df['species'] = iris.target_names[
    iris.target]
print(df.shape)
print(df.groupby('species')
    ['petal length (cm)'].mean())
```

150 цветков, 3 вида по 50

Вид	Лепесток, см
setosa	1,46
versicolor	4,26
virginica	5,55

ПРОСТЕЙШАЯ МОДЕЛЬ — ПРАВИЛО

Длина лепестка < 2,5 см → Setosa

Правило выполняется ровно для 50 цветков, и все они — Setosa: ни одной ошибки. Дальше научимся строить такие модели автоматически.

Инструменты курса: Python

Библиотека	Назначение	Недели
NumPy	Вычисления с массивами	1–3
pandas	Табличные данные	3–4
matplotlib	Графики	5
seaborn	Статистическая визуализация	6
scikit-learn	Машинное обучение, оценка моделей	7–13
mlxtend	Ассоциативные правила	14

Среда: Anaconda + Jupyter Notebook или Google Colab

Этика и «знай свои данные»

- **Конфиденциальность** — Закон РК «О персональных данных и их защите»
- **Справедливость** — модель может воспроизводить дискриминацию
- **Корреляция ≠ причинность**
- **Прозрачность** решений модели

ПРИМЕР ИЗ УЧЕБНИКА

«Поле 1» оказалось лучшим предиктором. Это был ID, присвоенный после сортировки по целевой переменной: связь реальная, но бессмысленная.

Выводы

- Data Mining — автоматический поиск полезных, ранее неизвестных закономерностей
- Центральный этап процесса KDD; предобработка — самый трудоёмкий шаг
- Задачи: классификация и регрессия, ассоциации, кластеры, аномалии
- Инструменты: Python, pandas, matplotlib, seaborn, scikit-learn

Далее: из чего состоят данные — типы признаков и наборов данных

Контрольные вопросы

1. Дайте определение Data Mining; чем он отличается от информационного поиска?
2. Опишите этапы KDD; почему предобработка — самый трудоёмкий этап?
3. Перечислите этапы CRISP-DM и сопоставьте их с этапами KDD.
4. Сравните предсказательные и описательные задачи, приведите примеры.
5. Объясните различие классификации и регрессии, обучения с учителем и без.
6. Напишите программу: число объектов каждого вида Iris и средняя ширина лепестка по видам.

Литература

1. Tan P.-N., Steinbach M., Kumar V. Introduction to Data Mining. — Pearson, 2014. — Chapter 1. Introduction.
2. Грас Дж. Data Science. Наука о данных с нуля. — СПб.: БХВ-Петербург, 2020. — Глава 1.
3. VanderPlas J. Python Data Science Handbook. — O’Reilly, 2016. — Chapter 5.
4. Chapman P. et al. CRISP-DM 1.0: Step-by-step data mining guide. — SPSS Inc., 2000.
5. Документация scikit-learn: scikit-learn.org