ИНТЕЛЛЕКТУАЛЬНОЕ МОДЕЛИРОВАНИЕ

Образовательная программа «6B06103 – Компьютерная инженерия»

Лекция 2

Данные: типы признаков и структура наборов данных

Неделя 2 · Tan, Steinbach, Kumar. Introduction to Data Mining, глава 2

Цель и план лекции

Цель: научиться определять типы признаков и структуру набора данных и выбирать подходящие операции и меры близости

1. Объекты и признаки
2. Четыре типа признаков и допустимые операции
3. Виды наборов данных и их характеристики
4. Качество данных
5. Меры сходства и различия
6. Типы данных в pandas

Объекты и признаки

- **Объект** — запись, наблюдение, пример (строка таблицы)
- **Признак** — свойство объекта, которое меняется от объекта к объекту или во времени (столбец)
- **Шкала измерения** — правило, сопоставляющее признаку число или символ

ID студента	Курс	Средний балл
1034262	4	3,24
1052663	2	3,51

Четыре типа признаков

Тип	Операции	Примеры	Статистики
Номинальный	= ≠	Индекс, ИИН, цвет глаз	Мода
Порядковый	< >	Оценка, размер S/M/L	Медиана
Интервальный	+ −	Дата, температура в °C	Среднее
Относительный	× /	Возраст, деньги, масса	Все, в т. ч. отношения

- Номинальный + порядковый = **категориальные**
- Интервальный + относительный = **количественные**

Пример: температура

- По Цельсию — **интервальная** шкала: ноль выбран условно
- «20 °C вдвое теплее 10 °C» — **неверно**
- «На 10 градусов теплее» — **верно**
- По Кельвину — **относительная** шкала: 2 K вдвое больше 1 K

Допустимые преобразования: $°F = 1{,}8 \cdot °C + 32$ (интервальная), метры ↔ футы (относительная)

Дискретные, непрерывные, асимметричные

- **Дискретный** — конечное или счётное число значений: число детей, индекс
- **Бинарный** — два значения: да/нет, 0/1
- **Непрерывный** — вещественные значения: рост, вес, доход
- **Асимметричный** — важна только «1»: купил товар, прошёл курс

Асимметричные признаки — основа анализа покупательских корзин (неделя 14)

Характеристики наборов данных

- **Размерность** — число признаков. Много признаков → **проклятие размерности**: расстояния выравниваются, алгоритмы теряют эффективность
- **Разреженность** — доля нулей; часто ненулевых меньше 1 %
- **Разрешение** — уровень детализации: почасовые продажи показывают суточные колебания, помесячные — сезонность

Виды наборов данных

Группа	Примеры
Записи	Таблицы, транзакции (чеки), матрица данных, матрица «документ — термин»
Графы	Веб-страницы и ссылки, социальные сети, молекулы
Упорядоченные	История покупок, последовательности ДНК, временные ряды, пространственные данные

Временные и пространственные данные: близкие по времени или месту значения похожи (**автокорреляция**)

Качество данных

Проблема	Пример
Шум	Ошибки датчика, помехи при записи
Выбросы	Доход 500 млн тенге среди студентов
Пропуски	Клиент не указал возраст
Несогласованность	Индекс не соответствует городу
Дубликаты	Клиент записан дважды

Точность ≠ смещение: весы, стабильно «добавляющие» 200 г, точны, но смещены

Сходство и различие

- **Сходство** — чем больше, тем объекты похожее; обычно от 0 до 1
- **Различие** — чем меньше, тем объекты ближе; частный случай — **расстояние**
- На близости основаны кластеризация, ближайшие соседи, поиск аномалий
- Признаки в разных единицах сначала **масштабируют**

Расстояния между числовыми объектами

Евклидово: $d(x,y) = \sqrt{\sum_{k=1}^{n} (x_k - y_k)^2}$

- **Манхэттенское:** $d = \sum_{k=1}^{n} |x_k - y_k|$
- **Чебышёва:** $d = \max_k |x_k - y_k|$
- Все три — частные случаи расстояния Минковского ($r = 1,\ 2,\ \infty$)

Пример: $p_1 = (0, 2)$, $p_2 = (2, 0)$ → евклидово $\sqrt{8} \approx$ **2,83**, манхэттенское **4**

Бинарные векторы: SMC и Жаккар

- $\text{SMC} = \dfrac{f_{11} + f_{00}}{\text{все признаки}}$ — учитывает совпадающие нули
- **Жаккар** $J = \dfrac{f_{11}}{f_{01} + f_{10} + f_{11}}$ — нули игнорирует

```
x = 1 0 0 0 0 0 0 0 0 0
y = 0 0 0 0 0 0 1 0 0 1
```

$\text{SMC} = \frac{7}{10} = \mathbf{0{,}7}$, $J = \frac{0}{3} = \mathbf{0}$. Для покупательских корзин верен Жаккар: общих покупок нет

Косинусное сходство: тексты

$$\cos(x,y) = \frac{x \cdot y}{\|x\| \cdot \|y\|}$$

```
d1 = 3 2 0 5 0 0 0 2 0 0
d2 = 1 0 0 0 0 0 0 1 0 2
```

$d_1 \cdot d_2 = 5,\ \|d_1\| = \sqrt{42} \approx 6{,}48,\ \|d_2\| = \sqrt{6} \approx 2{,}45 \rightarrow \cos \approx$ **0,315**

Не учитывает нули и не зависит от длины документа

Типы признаков в pandas

```
df['gender'] = df['gender'].astype('category')
order = pd.CategoricalDtype(
    ['удовл', 'хор', 'отл'], ordered=True)
df['grade'] = df['grade'].astype(order)
df[df['grade'] ⩾ 'хор']

from scipy.spatial.distance import pdist, jaccard
```

- Номинальный → category; порядковый → category с ordered=True
- Числовые → int64 / float64; даты → datetime64

Выводы

- Тип признака определяет допустимые операции и статистики
- Наборы данных: записи, графы, упорядоченные; важны размерность, разреженность, разрешение
- Мера близости выбирается по типу данных: евклидово, Жаккар, косинус

Далее: работа с таблицами в pandas

Контрольные вопросы

1. Дайте определения объекта, признака и шкалы измерения.
2. Сравните четыре шкалы: какие операции допустимы для каждой?
3. Почему температура в °C интервальная, а в K — относительная?
4. Объясните размерность, разреженность, разрешение и проклятие размерности.
5. Сравните SMC и коэффициент Жаккара; какой подходит для корзин?
6. Создайте DataFrame с номинальным, порядковым и числовым признаками и вычислите евклидово расстояние между двумя объектами.

Литература

1. Tan P.-N., Steinbach M., Kumar V. Introduction to Data Mining. — Pearson, 2014. — Chapter 2: 2.1, 2.2, 2.4.
2. McKinney W. Python for Data Analysis. — O’Reilly. — Chapter 7.5. Categorical Data.
3. Грас Дж. Data Science. Наука о данных с нуля. — СПб.: БХВ-Петербург, 2020. — Глава 4.
4. Документация pandas: Categorical data.
5. Документация SciPy: Distance computations.