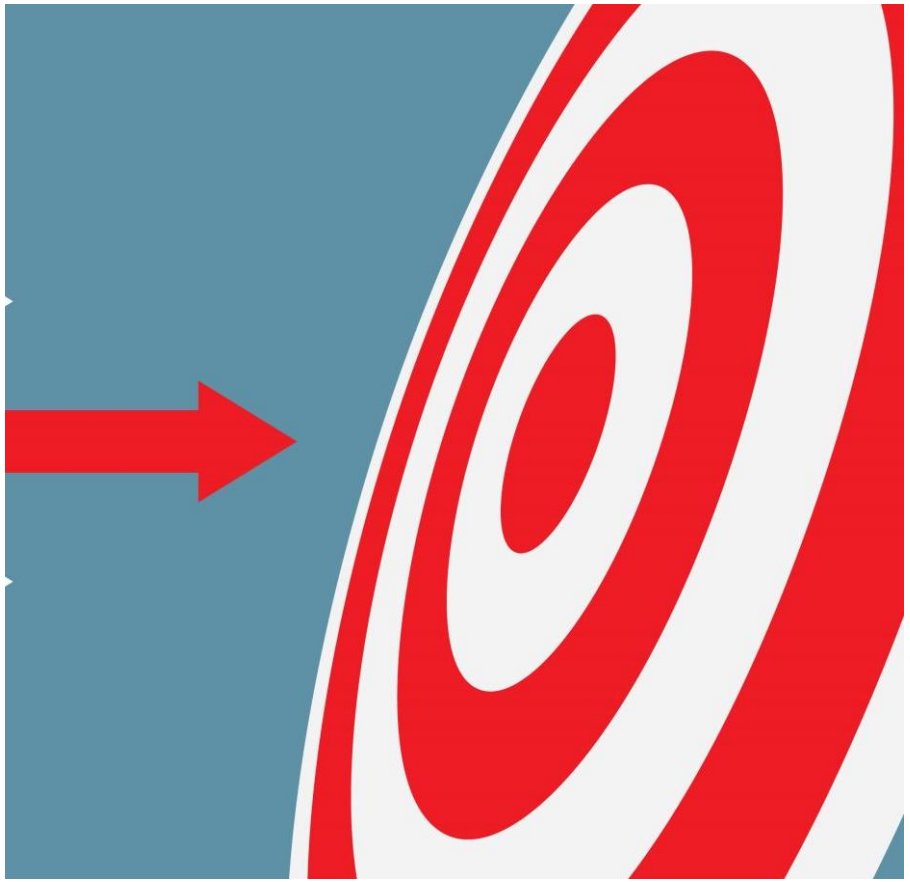


ИССЛЕДОВАНИЕ ДАННЫХ И МЕРЫ СХОДСТВА

ДИСЦИПЛИНА «ИНТЕЛЛЕКТУАЛЬНЫЙ АНАЛИЗ ДАННЫХ»

ЦЕЛЬ И РЕЗУЛЬТАТЫ ОБУЧЕНИЯ



Понять роль исследования данных перед применением Data Mining.

Различать понятия «сходство» (similarity) и «расстояние» (distance).

Рассчитывать основные меры расстояния и сходства.

Выбирать меру в зависимости от типа данных и задачи.

Интерпретировать результаты и применять их в Data Mining.

ПЛАН ЛЕКЦИИ

1. Исследование данных и пространство признаков

2. Сходство и расстояние между объектами

3. Euclidean, Manhattan, Minkowski, Chebyshev

4. Масштабирование признаков

5. Cosine Similarity

6. Бинарные данные: Hamming и Jaccard

7. Смешанные данные и Gower

8. Практические примеры и Python

ЧТО ТАКОЕ ИССЛЕДОВАНИЕ ДАННЫХ?

Исследование данных — изучение структуры, качества, распределения и взаимосвязей в наборе данных.

Проверяем: количество объектов и признаков, типы данных, пропуски, дубликаты и выбросы.

Изучаем распределения и взаимосвязи признаков.

Определяем, какие методы Data Mining подходят для задачи.

ОБЪЕКТЫ И ПРИЗНАКИ

Объект — единица наблюдения: клиент, пациент, транзакция, документ, товар.

Признак — характеристика объекта.

Пример клиента: возраст, доход, количество покупок.

Клиент	Возраст	Доход	Покупки в месяц
A	25	300	4
B	27	350	5
C	45	900	12
D	47	850	11

Каждый объект можно представить в виде вектора:

$$A = (25, 300, 4)$$
$$B = (27, 350, 5)$$
$$C = (45, 900, 12)$$

ЗАЧЕМ ИЗМЕРЯТЬ СХОДСТВО?

Поиск похожих клиентов или товаров.

Кластеризация объектов.

Алгоритмы k-NN.

Поиск похожих документов.

Рекомендательные системы.

Выявление аномалий.

Поиск дубликатов.

СХОДСТВО И РАССТОЯНИЕ

Сходство (Similarity) показывает степень похожести двух объектов.

Расстояние (Distance) показывает степень различия между объектами.

Чем выше сходство — тем объекты ближе по выбранному критерию.

Чем меньше расстояние — тем объекты ближе.

Конкретный смысл зависит от выбранной меры.

Например, возьмём:

$$A = (1, 0)$$
$$B = (0.95, 0.312)$$

Тогда:

$$Sim(A, B) = \frac{1(0.95) + 0(0.312)}{\sqrt{1^2 + 0^2}\sqrt{0.95^2 + 0.312^2}}$$
$$Sim(A, B) = 0.95$$

означает высокую степень сходства.

РАССТОЯНИЕ

Distance показывает, насколько объекты отличаются друг от друга.

Чем меньше расстояние, тем ближе объекты.

Например:

$$d(A, B) = 2$$

означает, что A и B находятся ближе друг к другу, чем:

$$d(A, C) = 10.$$

ПРОСТРАНСТВО ПРИЗНАКОВ

Каждый объект представляется точкой или вектором в пространстве признаков.

Расстояние между точками можно использовать как меру различия.

Для двух признаков объекты можно представить на координатной плоскости.

Для n признаков получаем n-мерное пространство.

ГЕОМЕТРИЧЕСКАЯ ИНТЕРПРЕТАЦИЯ

Рассмотрим два признака:

- x_1— возраст;
- x_2— количество покупок.

Объект A:

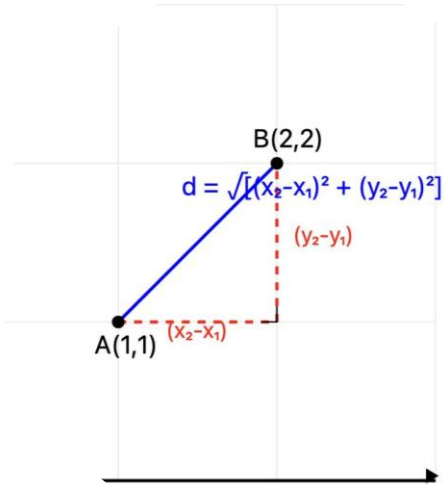
$$A = (25, 4)$$

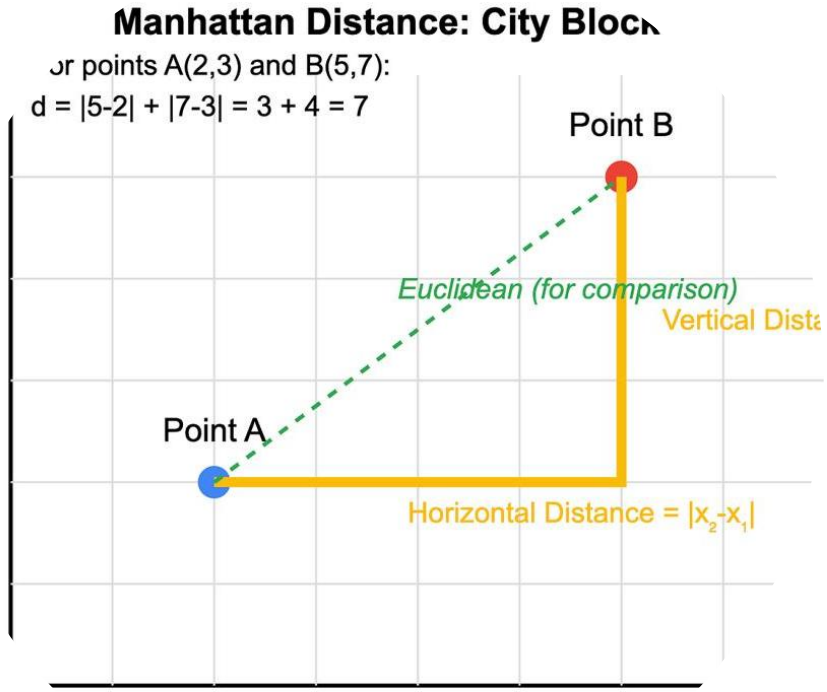
Объект B:

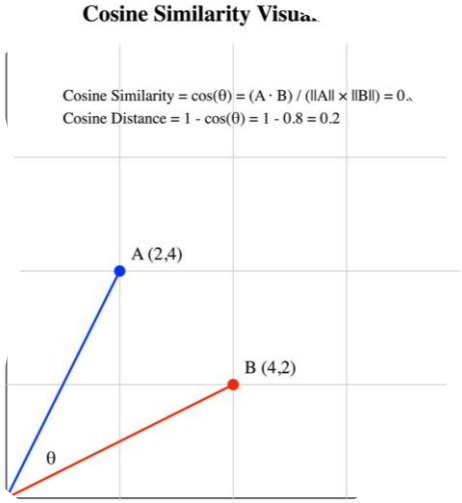
$$B = (27, 5)$$

Их можно представить точками на плоскости.

Расстояние между точками показывает степень различия объектов.


B(2,2)
d = √[(x₂-x₁)² + (y₂-y₁)²]
(y₂-y₁)
A(1,1)
(x₂-x₁)


Manhattan Distance: City Block
or points A(2,3) and B(5,7):
d = |5-2| + |7-3| = 3 + 4 = 7
Point B
Euclidean (for comparison)
Vertical Dista
Point A
Horizontal Distance = |x₂-x₁|
anhattan Distance = |x₂-x₁| + |y₂-y₁|


Cosine Similarity Visua
Cosine Similarity = cos(θ) = (A · B) / (||A|| × ||B||) = 0.
Cosine Distance = 1 - cos(θ) = 1 - 0.8 = 0.2
A (2,4)
B (4,2)
θ

ЕВКЛИДОВО РАССТОЯНИЕ

Наиболее известная мера расстояния — **евклидово расстояние**.

Для двух объектов:

$X=(x_1,\ldots,x_n)$ и $Y=(y_1,\ldots,y_n)$:

формула:

$$d(X,Y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

Это геометрическое расстояние по прямой.
Часто используется для числовых признаков и алгоритмов кластеризации.

Для двух признаков:

$$d(X,Y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2}$$

ПРИМЕР: РАСЧЁТ ЕВКЛИДОВА РАССТОЯНИЯ

A = (2, 3), B = (5, 7)

Тогда:

$$d(A,B) = \sqrt{(2-5)^2 + (3-7)^2}$$
$$= \sqrt{(-3)^2 + (-4)^2}$$
$$= \sqrt{9+16}$$
$$= \sqrt{25} = 5$$

Ответ:

$$\boxed{d(A,B) = 5}$$

СРАВНЕНИЕ ТРЁХ ОБЪЕКТОВ

Пусть:

$$A = (2, 3)$$
$$B = (3, 4)$$
$$C = (8, 9)$$

Рассчитаем расстояние от A.

A и B:

$$d(A, B) = \sqrt{(2-3)^2 + (3-4)^2}$$
$$= \sqrt{1+1} = \sqrt{2} \approx 1.41$$

A и C:

$$d(A, C) = \sqrt{(2-8)^2 + (3-9)^2}$$
$$= \sqrt{36+36} = \sqrt{72} \approx 8.49$$

Следовательно:

$$d(A, B) < d(A, C)$$

B значительно ближе к A, чем C.

МАНХЭТТЕНСКОЕ РАССТОЯНИЕ

Манхэттенское расстояние определяется как сумма абсолютных разностей координат:

$$d(X,Y) = \sum_{i=1}^{n} | x_i - y_i |$$

Для двух признаков:

$$d(X,Y) = | x_1 - y_1 | + | x_2 - y_2 |$$

Пример

$$A = (2,3)$$
$$B = (5,7)$$

Тогда:

$$d(A,B) = | 2 - 5 | + | 3 - 7 |$$
$$= 3 + 4 = 7$$

Ответ:

$$\boxed{d(A,B) = 7}$$

EUCLIDEAN И MANHATTAN: СРАВНЕНИЕ

Для A=(2,3), B=(5,7):

Обратите внимание:

Евклидово расстояние = 5

Манхэттенское расстояние = 7

Это разные геометрические способы измерения различия.

Обе меры сравнивают объекты, но по-разному определяют «близость».

Выбор меры зависит от природы задачи.

КОГДА ИСПОЛЬЗОВАТЬ МАНХЭТТЕНСКОЕ РАССТОЯНИЕ?

Манхэттенское расстояние особенно удобно, когда:

- признаки имеют смысл отдельных изменений;
- перемещение происходит по координатной сетке;
- данные содержат много признаков;
- необходимо уменьшить влияние отдельных больших отклонений по сравнению с квадратом разности в Euclidean distance.

Название связано с аналогией с улицами Манхэттена: расстояние между двумя точками определяется не прямой линией, а суммарным перемещением по направлениям.

РАССТОЯНИЕ МИНКОВСКОГО

Евклидово и Манхэттенское расстояния являются частными случаями **расстояния Минковского**:

$$d(X,Y) = \left(\sum_{i=1}^{n} | x_i - y_i |^p \right)^{1/p}$$

где p— параметр.

Если:

$$p = 1$$

получаем Manhattan:

$$d(X,Y) = \sum | x_i - y_i |$$

Если:

$$p = 2$$

получаем Euclidean:

$$d(X,Y) = \sqrt{\sum (x_i - y_i)^2}$$

РАССТОЯНИЕ ЧЕБЫШЁВА

Расстояние Чебышёва:

$$d(X,Y) = \max_i | x_i - y_i |$$

То есть рассматривается максимальная разница по одному из признаков.

Пример

$$A = (2,3,5)$$
$$B = (4,7,6)$$

Разности:

$$| 2 - 4 | = 2$$
$$| 3 - 7 | = 4$$
$$| 5 - 6 | = 1$$

Следовательно:

$$d(A,B) = \max(2,4,1) = 4$$

СРАВНЕНИЕ ОСНОВНЫХ РАССТОЯНИЙ

Для: $A = (2, 3)$ *и* $B = (5, 7)$

получаем:

Мера	Результат
Euclidean	5
Manhattan	7
Chebyshev	4

Важно понимать, что **нет одной универсально лучшей меры расстояния**.

Выбор зависит от:

- типа данных;
- масштаба признаков;
- природы задачи;
- алгоритма;
- наличия выбросов;

смысла отдельных признаков.

ПОЧЕМУ ВАЖНО МАСШТАБИРОВАНИЕ?

Признаки могут иметь очень разные диапазоны.

Например: возраст — десятки лет, доход — сотни тысяч тенге.

Большой по масштабу признак может доминировать в Euclidean Distance.

Перед расчётом расстояний часто используют Min-Max Scaling или Standardization.

РАССМОТРИМ КЛИЕНТОВ:

Клиент	Возраст	Доход
A	25	300 000
B	26	310 000
C	40	700 000

Если использовать Euclidean Distance непосредственно:

$$d(A,B) = \sqrt{(25-26)^2 + (300000-310000)^2}$$
$$= \sqrt{1 + 100000000}$$

Практически всё расстояние определяется **доходом**.

Возраст почти не влияет.

Это может быть нежелательно.

MIN-MAX НОРМАЛИЗАЦИЯ

Поэтому перед расчётом расстояний часто используют **масштабирование**.

Min-Max нормализация

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}}$$

В результате значения обычно находятся в диапазоне: $[0, 1]$

Пример

Пусть возраст находится от 20 до 60 лет.

Для человека 40 лет:

$$x' = \frac{40 - 20}{60 - 20} = \frac{20}{40} = 0.5$$

СТАНДАРТИЗАЦИЯ Z-SCORE

Формула: $z = \frac{x-\mu}{\sigma}$

где:

μ— среднее значение;

σ— стандартное отклонение.

После стандартизации признаки имеют приблизительно:

$$\mu = 0$$

и

$$\sigma = 1.$$

После стандартизации признаки получают сопоставимый масштаб.

Особенно полезна для методов, чувствительных к масштабу.

КОГДА МАСШТАБИРОВАНИЕ ОСОБЕННО ВАЖНО?

Если алгоритм использует расстояния, например:

- k-NN;
- K-means;
- DBSCAN;
- некоторые методы поиска похожих объектов,

то необходимо проверить:

Не доминирует ли один признак над остальными из-за масштаба?

Это одна из наиболее распространённых ошибок при Data Mining.

КОСИНУСНОЕ СХОДСТВО

Для некоторых данных абсолютное расстояние не является главным.

Например, при анализе текстов нас интересует:

Насколько похожи направления двух векторов?

Для этого используется **Косинусное сходство**.

$$Sim(X, Y) = \frac{X \cdot Y}{\| X \| \| Y \|}$$

где:

$$X \cdot Y = \sum x_i \, y_i$$

ПРИМЕР КОСИНУСНОГО СХОДСТВА

Пусть: $A = (1, 2, 3)$ и $B = (2, 4, 6)$
Скалярное произведение:

$$A \cdot B = 1 \cdot 2 + 2 \cdot 4 + 3 \cdot 6 = 2 + 8 + 18 = 28$$

Длина A:

$$\| A \| = \sqrt{1^2 + 2^2 + 3^2} = \sqrt{14}$$

Длина B:

$$\| B \| = \sqrt{2^2 + 4^2 + 6^2} = \sqrt{56}$$

Тогда:

$$Sim(A, B) = \frac{28}{\sqrt{14}\sqrt{56}} = 1$$

То есть:

$$\boxed{Sim(A, B) = 1}$$

Векторы имеют одинаковое направление.

ПОЧЕМУ КОСИНУСНОЕ СХОДСТВО ПОЛЕЗНО ДЛЯ ТЕКСТА?

Рассмотрим два документа.

Документ A:

Data Mining Data Mining Machine Learning

Документ B:

Data Mining Machine Learning

Можно представить документы в виде векторов частот слов.

Например:

Слово	A	B
Data	2	1
Mining	2	1
Machine	1	1
Learning	1	1

Косинусное сходство позволяет оценить **сходство содержания**, не придавая такого значения абсолютному размеру документа, как обычное евклидово расстояние.

Поэтому косинусное сходство широко используется при:

- поиске похожих документов;
- информационном поиске;
- рекомендательных системах;
- анализе текстов;
- сравнении профилей пользователей.

БИНАРНЫЕ ДАННЫЕ

Признаки принимают значения 0 и 1.

Пример: купил товар = 1, не купил = 0.

Пример: подписан на сервис = 1, не подписан = 0.

Для бинарных данных используются специальные меры сходства и расстояния.

Клиент	Купил ноутбук	Купил телефон	Купил планшет	Использует кредит
A	1	1	0	1
B	1	0	0	1

РАССТОЯНИЕ ХЭММИНГА

Расстояние Хэмминга— число позиций, в которых два вектора различаются.

Пусть:

A = (1,1,0,1,0)

B = (1,0,0,1,1)

Всего различий: 2

Следовательно:

Сравниваем позиции: $\boxed{d_H(A,B) = 2}$

Позиция	A	B	Различие
1	1	1	Нет
2	1	0	Да
3	0	0	Нет
4	1	1	Нет
5	0	1	Да

КОЭФФИЦИЕНТ ЖАККАРА

Для бинарных данных часто используется **Коэффициент Жаккара**.

$$J(A,B) = \frac{M_{11}}{M_{11} + M_{10} + M_{01}}$$

где:

- M_{11}— количество случаев, когда оба объекта имеют 1;
- M_{10}— A=1, B=0;
- M_{01}— A=0, B=1.

Совпадения 0–0 не учитываются.

Пример Коэффициент Жаккара

Пусть:

$$A = (1, 1, 0, 1, 0)$$
$$B = (1, 0, 0, 1, 1)$$

Получаем:

$$M_{11} = 2$$

позиции 1 и 4.

$$M_{10} = 1$$

позиция 2.

$$M_{01} = 1$$

позиция 5.

Тогда: $J(A, B) = \frac{2}{2+1+1} = \frac{2}{4} = 0.5$

Ответ: $\boxed{J(A, B) = 0.5}$

То есть Jaccard Similarity составляет **50%**.

СМЕШАННЫЕ ДАННЫЕ И GOWER

В реальных наборах данных признаки часто имеют разные типы.

Например:

Клиент	Возраст	Доход	Пол	Категория клиента
A	25	300000	Ж	Стандарт
B	27	350000	Ж	Стандарт

В реальных наборах одновременно встречаются числовые, категориальные и бинарные признаки.

Обычное Смешанные данные не подходит напрямую для всех таких признаков.

Gower Distance объединяет вклад признаков разных типов.

МЕРА GOWER

Идея состоит в том, что сходство рассчитывается отдельно для каждого признака, а затем объединяется.

Для числового признака:

$$s_{ij} = 1 - \frac{|\, x_i - y_i \,|}{range}$$

Для категориального признака:

$$s_{ij} = \begin{cases} 1, & x_i = y_i \\ 0, & x_i \neq y_i \end{cases}$$

Затем рассчитывается общее сходство.

Это особенно полезно для:

- клиентских профилей;
- медицинских данных;
- социальных исследований;
- анкет;
- маркетинговых данных.

ИССЛЕДОВАНИЕ ДАННЫХ ПЕРЕД РАСЧЁТОМ РАССТОЯНИЯ

Перед выбором меры сходства необходимо задать несколько вопросов.

Вопрос 1. Какой тип данных?

- числовые;
- категориальные;
- бинарные;
- текстовые;
- временные;
- смешанные.

Вопрос 2. Имеют ли признаки одинаковый масштаб?

Если нет — возможно, потребуется:

- Min-Max Scaling;
- Standardization.

Вопрос 3. Есть ли выбросы?

Если есть, необходимо учитывать их влияние на расстояние.

Вопрос 4. Что означает «похожесть» в данной предметной области?

Это принципиально важно.

МАТРИЦА РАССТОЯНИЙ

Если имеется несколько объектов, удобно представить расстояния в виде матрицы.
Например:

	A	B	C
A	0	1.41	8.49
B	1.41	0	7.07
C	8.49	7.07	0

На диагонали:

$$d(A, A) = 0$$

Матрица симметрична:

$$d(A, B) = d(B, A)$$

Такие матрицы используются в:

- кластеризации;
- nearest-neighbor методах;
- иерархической кластеризации;
- поиске похожих объектов.

ПРАКТИЧЕСКАЯ ЗАДАЧА ДЛЯ СТУДЕНТОВ

Даны три объекта:

$$A = (2, 4)$$
$$B = (5, 8)$$
$$C = (3, 5)$$

Необходимо:

1. Рассчитать:

- Euclidean Distance;
- Manhattan Distance;
- Chebyshev Distance.

2. Определить, какой объект ближе к A.

3. Сравнить результаты трёх мер.

Решение

A и B

Euclidean:

$$d(A,B) = \sqrt{(2-5)^2 + (4-8)^2}$$
$$= \sqrt{9+16} = 5$$

Manhattan:

$$d(A,B) = 3 + 4 = 7$$

Chebyshev:

$$d(A,B) = \max(3,4) = 4$$

A и C

Euclidean:

$$d(A,C) = \sqrt{(2-3)^2 + (4-5)^2} = \sqrt{2} \approx 1.41$$

Manhattan:

$$d(A,C) = 1 + 1 = 2$$

Chebyshev:

$$d(A,C) = 1$$

Таким образом, по всем трём мерам:

$\boxed{C \text{ ближе к } A}$

ВЫБОР МЕРЫ: ПРАКТИЧЕСКАЯ СХЕМА

Числовые признаки → Euclidean / Manhattan / другие метрики.

Текстовые или векторные данные → Cosine Similarity.

Бинарные последовательности → Hamming.

Множества и положительные события → Jaccard.

Смешанные типы → Gower или специализированный подход.

Всегда учитывайте масштаб, выбросы и смысл признаков.

СВЯЗЬ С МЕТОДАМИ DATA MINING

k-NN — поиск ближайших объектов.

K-means — расстояние до центроидов.

DBSCAN — соседство объектов по расстоянию ε.

Иерархическая кластеризация — объединение близких объектов/кластеров.

Recommendation Systems — поиск похожих пользователей и товаров.

Text Mining — сравнение документов.

ПРАКТИЧЕСКИЙ ПРИМЕР: СЕГМЕНТАЦИЯ КЛИЕНТОВ

Данные: возраст, доход, количество покупок.

1. Проверить качество данных.
2. Выполнить масштабирование.
3. Выбрать меру расстояния.
4. Рассчитать матрицу расстояний.
5. Выполнить кластеризацию.
6. Интерпретировать полученные группы клиентов.

PYTHON: EUCLIDEAN DISTANCE

```
import numpy as np
A = np.array([2, 3])
B = np.array([5, 7])
distance = np.linalg.norm(A - B)
print(distance)
```

Результат: 5.0

PYTHON: MANHATTAN И COSINE

```
from scipy.spatial.distance import cityblock
from sklearn.metrics.pairwise import cosine_similarity
A = [2, 3]; B = [5, 7]
cityblock(A, B)  # 7
X = [[1,2,3]]; Y = [[2,4,6]]
cosine_similarity(X, Y)  # [[1.]]
```

ПРАКТИЧЕСКАЯ ЗАДАЧА ДЛЯ СТУДЕНТОВ

Даны A=(2,4), B=(5,8), C=(3,5).

1. Рассчитать Euclidean Distance.
2. Рассчитать Manhattan Distance.
3. Рассчитать Chebyshev Distance.
4. Определить ближайший к A объект.
5. Сравнить результаты трёх мер.
6. Повторить расчёты в Python.

КОНТРОЛЬНЫЕ ВОПРОСЫ

1. Что такое объект и признак?
2. Чем similarity отличается от distance?
3. Запишите формулу Euclidean Distance.
4. Почему необходимо масштабирование?
5. Для чего используется Cosine Similarity?
6. Что измеряет Hamming Distance?
7. Когда применяется Jaccard?
8. Какие меры подходят для смешанных данных?

ТИПИЧНЫЕ ОШИБКИ

Использование расстояния без проверки масштаба признаков.

Применение Euclidean к категориальным признакам без корректного кодирования.

Игнорирование выбросов.

Использование одной меры для всех типов данных.

Неверная интерпретация значения расстояния как процента различия.

ОСНОВНЫЕ ВЫВОДЫ

Расстояние измеряет различие, сходство — степень близости.

Выбор меры зависит от типа данных и предметной задачи.

Euclidean, Manhattan и Chebyshev применяются преимущественно для числовых данных.

Cosine полезна для векторных и текстовых данных.

Hamming и Jaccard подходят для бинарных данных, но решают разные задачи.

Для смешанных данных могут применяться специальные меры, например Gower.

Правильная мера сходства — важная часть решения Data Mining.

САМОСТОЯТЕЛЬНАЯ РАБОТА

Тема: «Сравнение объектов с использованием различных мер сходства и расстояния».

A=(2,4,6), B=(3,5,7), C=(10,2,4).

Рассчитать Euclidean, Manhattan и Chebyshev.

Построить матрицы расстояний.

Определить наиболее близкие пары.

Реализовать расчёты в Python.

Сделать вывод о выборе меры.

СПАСИБО ЗА ВНИМАНИЕ!

Ключевой вопрос:

«Какая мера сходства или расстояния лучше всего отражает смысл именно вашей задачи?»

Ответ зависит от типа данных, масштаба признаков и предметной области.