

КЛАСТЕРИЗАЦИЯ ДАННЫХ

ЛЕКЦИЯ 7

ЦЕЛЬ ЛЕКЦИИ:

- изучить основные методы кластеризации данных, принципы их работы, критерии выбора алгоритма и применение кластеризации для решения практических задач анализа данных.

ОЖИДАЕМЫЕ РЕЗУЛЬТАТЫ ОБУЧЕНИЯ.

После изучения темы студент сможет:

- Объяснять понятие кластеризации и её отличие от классификации.
- Выбирать метрики расстояния для сравнения объектов.
- Применять алгоритмы K-means, иерархическую кластеризацию и DBSCAN.
- Оценивать качество кластеризации.
- Интерпретировать результаты кластеризации на практических примерах.

ПОНЯТИЕ КЛАСТЕРИЗАЦИИ ДАННЫХ

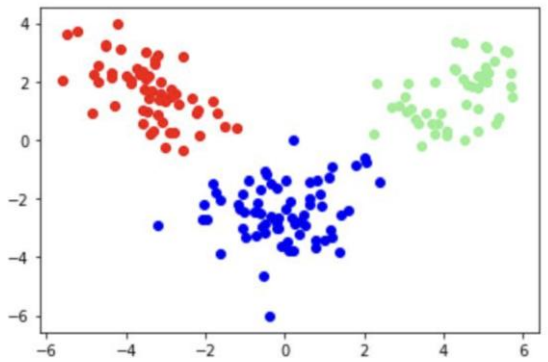
- **Кластеризация (clustering)** — это метод интеллектуального анализа данных, предназначенный для автоматического разделения объектов на группы, называемые кластерами, таким образом, чтобы объекты внутри одного кластера были максимально похожи друг на друга, а объекты из разных кластеров — как можно более различными.

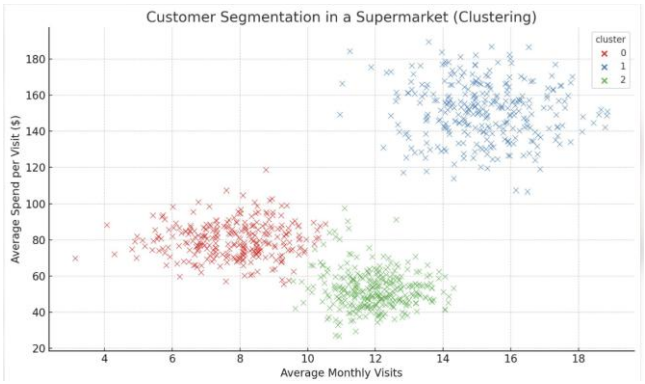
Кластеризация относится к методам обучения без учителя (Unsupervised Learning). Это означает, что для каждого объекта заранее не указана правильная категория или метка класса.

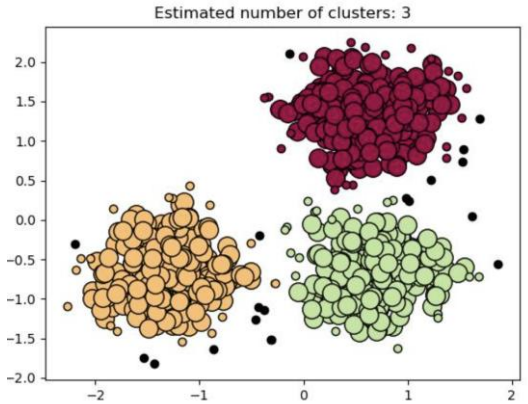
Например, университет может проанализировать данные студентов и выделить группы:

- студенты с высокой успеваемостью;
- студенты со средней успеваемостью;
- студенты, которым требуется дополнительная поддержка.

При этом алгоритму заранее не сообщают, какие студенты относятся к каждой группе. Он выявляет структуру данных на основе сходства между объектами.







ОТЛИЧИЕ КЛАСТЕРИЗАЦИИ ОТ КЛАССИФИКАЦИИ

Критерий	Кластеризация	Классификация
Тип обучения	Без учителя	С учителем
Исходные метки классов	Не требуются	Требуются
Основная задача	Найти естественные группы	Отнести объект к известному классу
Пример	Сегментация клиентов	Определение спам-сообщения
Результат	Кластеры, обнаруженные в данных	Предсказанные классы

Главное отличие: при классификации алгоритм учится на заранее размеченных примерах, а при кластеризации самостоятельно обнаруживает группы в данных.

ОБЛАСТИ ПРИМЕНЕНИЯ КЛАСТЕРИЗАЦИИ

Кластеризация применяется в различных областях:

- Маркетинг: сегментация клиентов по покупательскому поведению.
- Медицина: группировка пациентов по результатам обследований.
- Финансы: выявление групп транзакций с похожими характеристиками.
- Образование: анализ учебной активности студентов.
- Кибербезопасность: поиск необычных сетевых соединений и групп похожих атак.
- Обработка изображений: сегментация изображений на области.
- Биология: группировка генов и биологических объектов.
- Рекомендательные системы: поиск пользователей со сходными интересами.

ОСНОВНЫЕ ПОНЯТИЯ КЛАСТЕРИЗАЦИИ

- Объект — элемент набора данных, который необходимо сгруппировать. Например, клиент, студент или транзакция.
- Признак — характеристика объекта. Для клиента это могут быть возраст, доход, количество заказов и средний чек.
- Кластер — группа объектов, обладающих сходными характеристиками.
- Центроид — среднее положение объектов кластера в пространстве признаков.
- Расстояние — численная мера различия между двумя объектами.
- Шум — объекты, которые не относятся к достаточно плотным группам согласно выбранному алгоритму, например DBSCAN.

Качество кластеризации зависит не только от алгоритма, но и от того, какие признаки выбраны, как они масштабированы и какую метрику расстояния используют.

ПОДГОТОВКА ДАННЫХ ПЕРЕД КЛАСТЕРИЗАЦИЕЙ

Перед применением алгоритмов необходимо подготовить данные. Ошибки на этом этапе могут привести к неправильному формированию кластеров.

1. ВЫБОР ПРИЗНАКОВ

Необходимо выбрать характеристики, которые отражают сходство объектов.

Например, для сегментации клиентов можно использовать:

- возраст;
- ежемесячный доход;
- количество покупок за месяц;
- среднюю сумму покупки.

Если включить признаки, не связанные с задачей, качество кластеризации может ухудшиться.

2. ОБРАБОТКА ПРОПУЩЕННЫХ ЗНАЧЕНИЙ

Пропущенные данные можно обработать следующими способами:

- удалить объекты с большим количеством пропусков;
- заменить пропуски средним или медианным значением;
- использовать специальные методы восстановления данных.

3. МАСШТАБИРОВАНИЕ ПРИЗНАКОВ

Многие алгоритмы кластеризации чувствительны к масштабу признаков. Например, если доход измеряется тысячами, а возраст — десятками, доход может чрезмерно влиять на расстояние между объектами.

Один из распространённых методов масштабирования — стандартизация:

$$z = \frac{x - \mu}{\sigma}$$

где:

- x — исходное значение признака;
- μ — среднее значение;
- σ — стандартное отклонение;
- z — стандартизированное значение.

После стандартизации среднее значение признака становится равным нулю, а стандартное отклонение — единице, если оно изначально ненулевое.

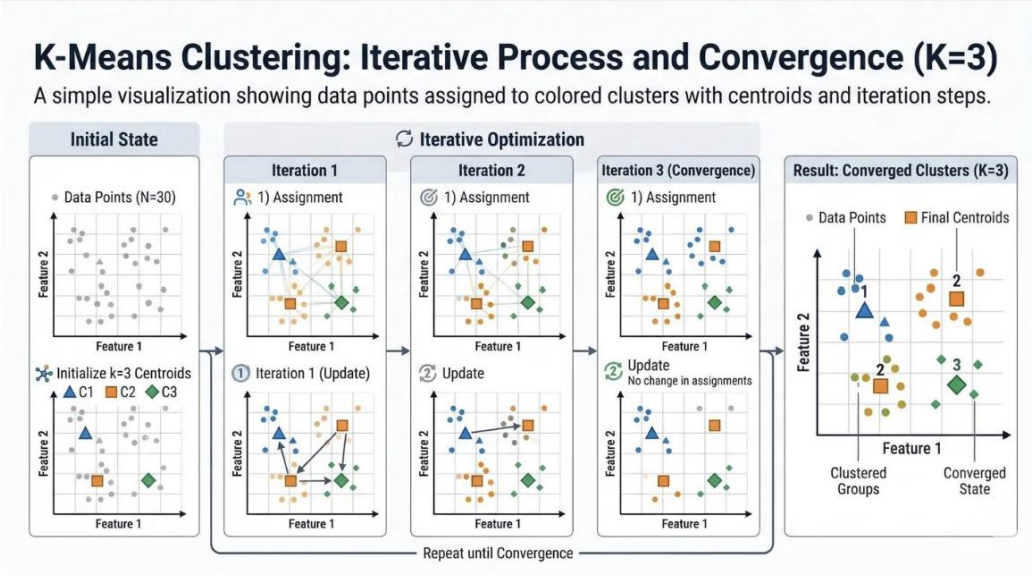
- Другой метод — нормализация Min-Max:

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

Она обычно преобразует значения в диапазон от 0 до 1.
Важно: масштабирование следует выполнять после разделения данных, если задача предполагает последующую оценку на отложенной выборке, чтобы избежать утечки информации.

АЛГОРИТМ K-MEANS

K-means (метод k-средних) — один из наиболее распространённых алгоритмов кластеризации. Он разделяет данные на K кластеров, стремясь минимизировать разброс объектов внутри каждого кластера.
Алгоритм подходит для числовых данных, когда кластеры относительно компактны и хорошо разделимы.



ПРИНЦИП РАБОТЫ K-MEANS

- **Шаг 1. Выбрать количество кластеров K**

Например, K=2.

- **Шаг 2. Инициализировать центроиды**

Выбрать начальные центры кластеров. На практике часто применяют K-means++.

- **Шаг 3. Назначить объекты кластерам**

Каждый объект относится к ближайшему центроиду по выбранной метрике.

- **Шаг 4. Пересчитать центроиды**

Для каждого кластера вычислить среднее значение по каждому признаку.

- **Шаг 5. Проверить условие остановки**

Повторять назначение и пересчёт, пока центроиды не перестанут существенно изменяться либо не будет достигнут лимит итераций.

ЦЕЛЕВАЯ ФУНКЦИЯ K-MEANS

Алгоритм минимизирует сумму квадратов расстояний от объектов до центроидов соответствующих кластеров:

$$J = \sum_{j=1}^{K} \sum_{x_i \in C_j} \|x_i - \mu_j\|^2$$

где:

- K — количество кластеров;
- C_j — множество объектов кластера j;
- x_j — отдельный объект;
- μ_j — центроид кластера j.

Чем меньше значение J, тем компактнее кластеры относительно своих центроидов. Однако само по себе минимальное значение не означает, что выбранное число кластеров оптимально.

ПРИМЕР ВЫПОЛНЕНИЯ K-MEANS

- Рассмотрим шесть объектов с двумя признаками:

Объект	x	y
A	1	1
B	1,5	2
C	3	4
D	6	8
E	3,5	5
F	4,5	5

Пусть K=2, а начальные центроиды выбраны в точках A(1, 1) и D(6, 8).

ШАГ 1. ОПРЕДЕЛЯЕМ РАССТОЯНИЯ.

- Для объекта B(1,5; 2) до центроида A:

$$d(B, A) = \sqrt{(1{,}5 - 1)^2 + (2 - 1)^2}$$

$$d(B, A) = \sqrt{0{,}25 + 1} \approx 1{,}118$$

До центроида D:

$$d(B, D) = \sqrt{(1{,}5 - 6)^2 + (2 - 8)^2}$$

$$d(B, D) = \sqrt{20{,}25 + 36} \approx 7{,}50$$

Следовательно, объект B относится к первому кластеру.
Аналогично рассчитываем расстояния для остальных объектов.

ШАГ 2. РАСПРЕДЕЛЯЕМ ОБЪЕКТЫ.

После первой итерации получаем:

- кластер 1: A, B, C;
- кластер 2: D, E, F.

ШАГ 3. ПЕРЕСЧИТЫВАЕМ ЦЕНТРОИДЫ.

- Для первого кластера:

$$\mu_1 = \left(\frac{1 + 1{,}5 + 3}{3}, \frac{1 + 2 + 4}{3} \right)$$

$$\mu_1 \approx (1{,}833;\ 2{,}333)$$

Для второго кластера:

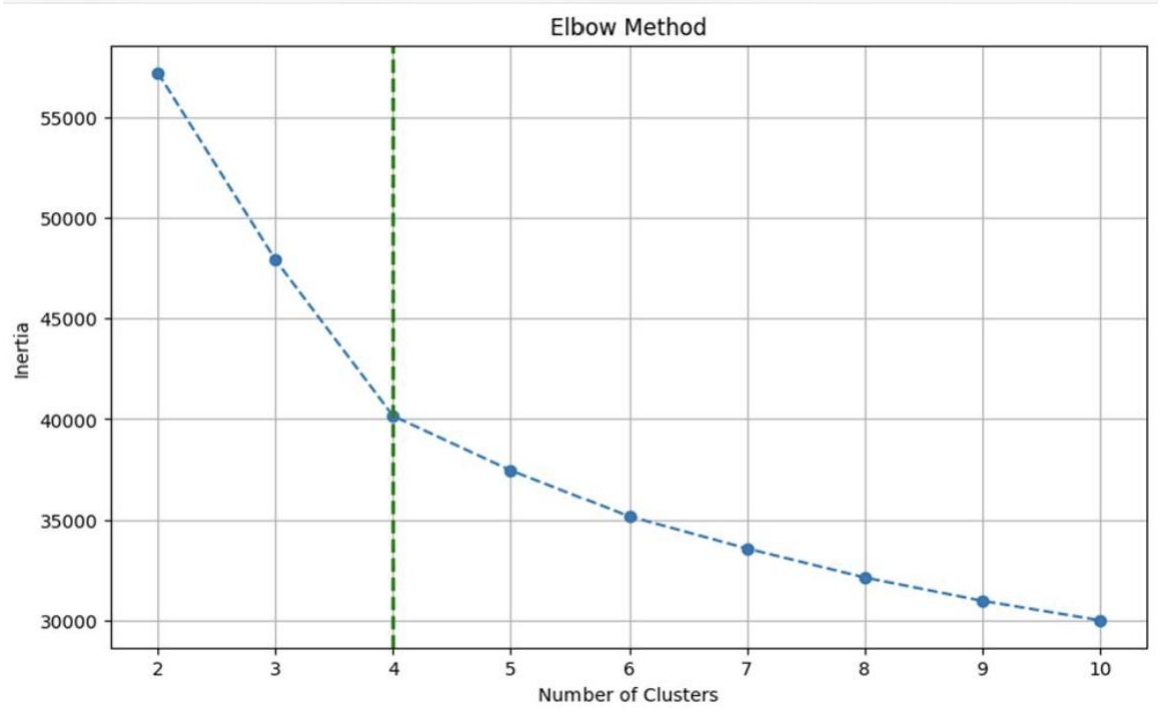
$$\mu_2 = \left(\frac{6 + 3{,}5 + 4{,}5}{3}, \frac{8 + 5 + 5}{3} \right)$$

$$\mu_2 \approx (4{,}667;\ 6)$$

После пересчёта нужно повторно проверить принадлежность объектов кластерам.
Если состав кластеров не меняется, алгоритм может остановиться.

КАК ВЫБРАТЬ КОЛИЧЕСТВО КЛАСТЕРОВ?

- Количество K обычно задаётся до запуска K-means. Для его выбора можно применять несколько подходов.
- **Метод локтя (Elbow Method).** Рассчитывают сумму квадратов расстояний внутри кластеров для разных значений K. Затем строят график и ищут точку, после которой дальнейшее увеличение количества кластеров даёт значительно меньший выигрыш.



- **Метод силуэта.** Сравнивают среднее качество группировки для разных вариантов разбиения. Чем выше средний коэффициент силуэта, тем лучше объекты в среднем отделены от соседних кластеров.
- Ни один из методов не гарантирует единственно правильного значения K. Результаты необходимо интерпретировать с учётом предметной области.

ПРЕИМУЩЕСТВА И НЕДОСТАТКИ K-MEANS

Преимущества:

- простота реализации;
- относительно высокая скорость;
- хорошая работа с большими наборами числовых данных;
- понятная интерпретация центроидов.

Недостатки:

- требуется заранее выбрать K;
- чувствителен к выбросам;
- зависит от начальных центроидов;
- плохо работает с кластерами сложной формы и разной плотности;
- требует осмысленной подготовки и масштабирования признаков.

ИЕРАРХИЧЕСКАЯ КЛАСТЕРИЗАЦИЯ

Иерархическая кластеризация формирует систему вложенных кластеров. Результат часто изображают в виде дендрограммы — дерева, показывающего, как объединяются или разделяются группы объектов.

Существует два основных подхода.

- **Агломеративный:** каждый объект сначала является отдельным кластером, после чего наиболее похожие кластеры постепенно объединяются.
- **Дивизивный:** все объекты начинают в одном кластере, который последовательно разделяется на более мелкие группы.

Агломеративный метод используется особенно часто.

1. ЭТАПЫ АГЛОМЕРАТИВНОЙ КЛАСТЕРИЗАЦИИ

1. Создать отдельный кластер для каждого объекта.
2. Рассчитать расстояния между объектами или кластерами.
3. Найти два наиболее близких кластера.
4. Объединить их.
5. Пересчитать расстояния между новыми кластерами.
6. Повторять до достижения нужного числа кластеров либо объединения всех объектов.

2. МЕТОДЫ ОПРЕДЕЛЕНИЯ РАССТОЯНИЯ МЕЖДУ КЛАСТЕРАМИ

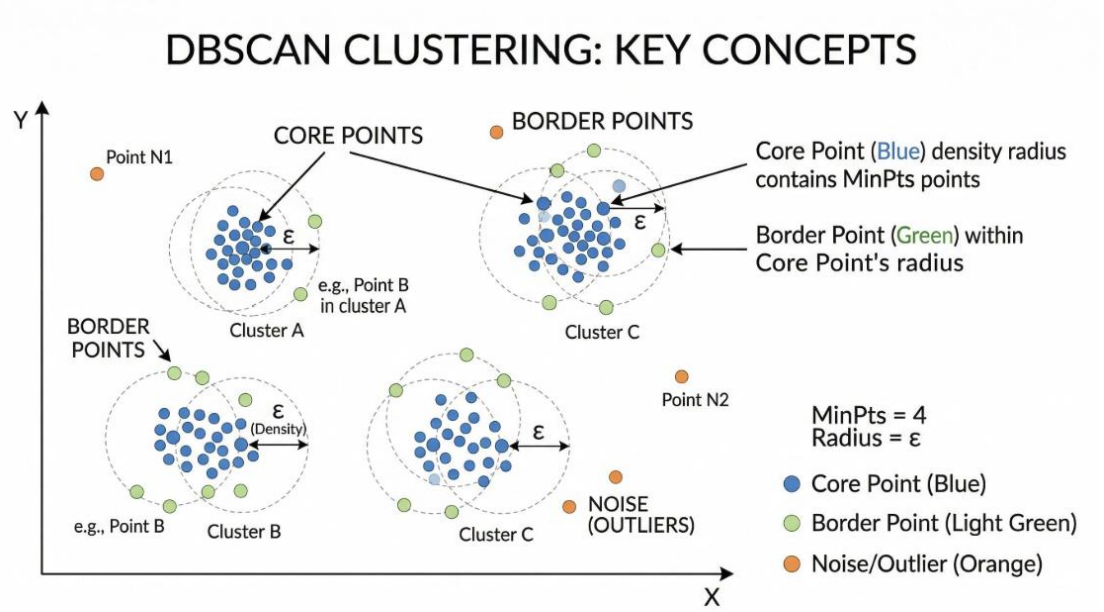
Метод	Принцип
Single linkage	Минимальное расстояние между объектами двух кластеров
Complete linkage	Максимальное расстояние между объектами двух кластеров
Average linkage	Среднее расстояние между парами объектов двух кластеров
Ward linkage	Объединение, минимизирующее увеличение внутрикластерной суммы квадратов отклонений

Выбор метода влияет на форму и структуру получаемых кластеров.

Преимущества: не обязательно заранее задавать число кластеров; можно исследовать структуру данных на разных уровнях.
Недостатки: вычислительные и затратные по памяти процедуры на больших наборах данных; решения об объединении обычно необратимы.

АЛГОРИТМ DBSCAN

- DBSCAN (Density-Based Spatial Clustering of Applications with Noise) — алгоритм, который формирует кластеры на основе плотности объектов.
- В отличие от K-means, DBSCAN не требует заранее задавать количество кластеров и может обнаруживать группы сложной формы. Он также способен выделять шумовые объекты.



1. ОСНОВНЫЕ ПАРАМЕТРЫ DBSCAN

- ε(eps) — радиус окрестности объекта.
- MinPts — минимальное количество объектов в окрестности, необходимое для определения плотного ядра.

В распространённой реализации самого объекта также считают при подсчёте точек в его окрестности.

2. ТИПЫ ОБЪЕКТОВ

- **Ядровая точка (core point)** — точка, в окрестности которой находится достаточное количество объектов.
- **Граничная точка (border point)** — точка, которая не является ядровой, но попадает в окрестность ядровой точки.
- **Шумовая точка (noise point)** — точка, которую алгоритм не смог отнести к кластеру при заданных параметрах.

3. ПРИНЦИП РАБОТЫ DBSCAN

1. Выбирается ещё не обработанная точка.
2. Рассматриваются все точки в радиусе varepsilon.
3. Если соседей достаточно, создаётся кластер и он расширяется через достижимые ядровые точки.
4. Если соседей недостаточно, точка временно помечается как шумовая.
5. Такая точка позднее может оказаться граничной, если попадёт в окрестность ядровой точки другого кластера.
6. Процесс продолжается, пока все точки не будут обработаны.

4. ПРЕИМУЩЕСТВА И НЕДОСТАТКИ DBSCAN

Преимущества:

- не нужно заранее определять количество кластеров;
- обнаруживает кластеры сложной формы;
- выделяет шум и отдельные объекты.

Недостатки:

- результат зависит от параметров eps и MinPts;
- трудно подобрать единые параметры для кластеров с существенно разной плотностью;
- чувствителен к масштабу признаков;
- в пространствах высокой размерности расстояния могут терять информативность.

СРАВНЕНИЕ АЛГОРИТМОВ КЛАСТЕРИЗАЦИИ

Характеристика	K-means	Иерархическая	DBSCAN
Нужно задавать число кластеров	Да	Не обязательно заранее	Нет
Обнаружение шума	Ограниченное	Зависит от метода	Да
Сложная форма кластеров	Обычно плохо	Зависит от метода	Хорошо при подходящей плотности
Чувствительность к масштабу	Высокая	Зависит от метрики	Высокая
Работа с большими данными	Обычно хорошая	Часто затратная	Зависит от реализации и индексации
Основные параметры	K инициализация	Метрика, linkage, порог	eps, MinPts

РЕКОМЕНДАЦИЯ ПО ВЫБОРУ:

- K-means — если нужны компактные группы и данные преимущественно числовые.
- Иерархическая кластеризация — если важно изучить отношения между группами и их вложенную структуру.
- DBSCAN — если ожидаются кластеры сложной формы и необходимо выявлять шум.

ОЦЕНКА КАЧЕСТВА КЛАСТЕРИЗАЦИИ

- В кластеризации часто нет заранее известных правильных меток, поэтому качество оценивают с помощью внутренних метрик, устойчивости результатов и экспертной интерпретации.

КОЭФФИЦИЕНТ СИЛУЭТА (SILHOUETTE SCORE)

- Для объекта i:

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$

где:

- $a(i)$ — среднее расстояние до объектов своего кластера;
- $b(i)$ — минимальное среднее расстояние до объектов другого кластера, рассчитанное по альтернативным кластерам.

Значение находится в диапазоне от −1 до 1.

Значение	Интерпретация
Близко к 1	Объект хорошо соответствует своему кластеру
Близко к 0	Объект находится около границы кластеров
Меньше 0	Возможно, объект ближе к другому кластеру

Средний силуэт помогает сравнивать варианты кластеризации, но его следует использовать с учётом формы и структуры данных.

2. ВНУТРИКЛАСТЕРНАЯ СУММА КВАДРАТОВ

- Для K-means применяется показатель inertia — сумма квадратов расстояний объектов до центроидов. С увеличением K этот показатель обычно не увеличивается, поскольку каждый объект может оказаться ближе к своему центру.
- Поэтому нельзя просто выбрать максимальное количество кластеров: необходимо учитывать компромисс между компактностью, сложностью модели и полезностью групп.

3. ДРУГИЕ ПОКАЗАТЕЛИ

- Davies–Bouldin Index: меньшие значения обычно указывают на более компактные и хорошо разделённые кластеры.
- Calinski–Harabasz Index: большие значения обычно указывают на лучшее разделение относительно разброса внутри кластеров.
- ARI и NMI: применяются, когда для проверки имеются эталонные метки, хотя обучение выполняется без учителя.

ПРАКТИЧЕСКОЕ ЗАДАНИЕ ДЛЯ СТУДЕНТОВ

Тема: сегментация клиентов методом кластеризации.

Цель — разделить клиентов интернет-магазина на группы на основе поведения при покупках.

Исходные признаки:

- возраст клиента;
- ежемесячный доход;
- количество покупок за месяц;
- средний чек.

Задания:

1. Создать или загрузить набор данных минимум из 100 клиентов.
2. Проверить пропущенные значения и обработать их.
3. Выбрать признаки, подходящие для сегментации.
4. Выполнить стандартизацию.
5. Применить K-means с несколькими значениями K, например от 2 до 6.
6. Построить график метода локтя.
7. Рассчитать коэффициент силуэта для нескольких вариантов.
8. Выбрать и обосновать итоговое количество кластеров.
9. Сравнить результат K-means с DBSCAN.
10. Дать содержательное описание каждого полученного сегмента и предложить возможные бизнес-решения.

КОНТРОЛЬНЫЕ ВОПРОСЫ

1. Что такое кластеризация данных?
2. Почему кластеризация относится к обучению без учителя?
3. Чем кластеризация отличается от классификации?
4. Для чего выполняют стандартизацию признаков?
5. Как рассчитывается евклидово расстояние?
6. Какие этапы включает алгоритм K-means?
7. Что такое центроид и как он вычисляется?
8. Как работает метод локтя?
9. В чём особенности иерархической кластеризации?
10. Что означают параметры eps и MinPts в DBSCAN?
11. Чем ядровая точка отличается от граничной?
12. Что показывает коэффициент силуэта?
13. Почему K-means чувствителен к выбросам?
14. В каких случаях предпочтительнее DBSCAN?
15. Какие трудности возникают при кластеризации данных высокой размерности?

РЕКОМЕНДУЕМАЯ ЛИТЕРАТУРА

1. Han, J., Kamber, M., Pei, J. Data Mining: Concepts and Techniques. 3rd ed. Morgan Kaufmann, 2011.
2. Géron, A. Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow. 3rd ed. O'Reilly, 2022.
3. Müller, A. C., Guido, S. Introduction to Machine Learning with Python. 2nd ed. O'Reilly, 2022.
4. Документация Scikit-learn: Clustering.
5. Документация Scikit-learn: Clustering performance evaluation.