ЛЕКЦИЯ 7

МЕТОДЫ КЛАСТЕРИЗАЦИИ И СНИЖЕНИЯ РАЗМЕРНОСТИ ДАННЫХ

Основные понятия • определения • схемы • таблицы • графики • примеры

Преподаватель: проф.Ахмедиярова А.Т.

Структура и основные разделы

01 Обучение без учителя

02 Понятие кластеризации

03 Метрики расстояния и масштабирование

04 Алгоритм K-Means

05 Выбор числа кластеров K

06 Преимущества и ограничения K-Means

07 Иерархическая кластеризация

08 Алгоритм DBSCAN

09 Сравнение методов кластеризации

10 Оценка качества кластеризации

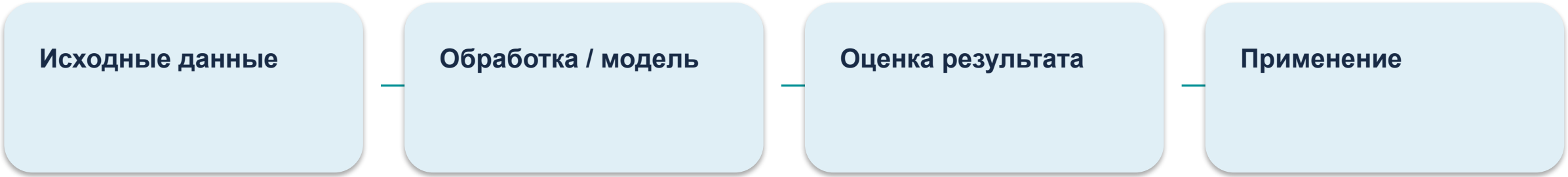
11 Кластеризация в информационной безопасности

12 Что такое снижение размерности

Обучение без учителя

В задачах обучения без учителя данные не содержат заранее заданной целевой переменной с правильными ответами. Алгоритм анализирует признаки объектов и пытается выявить внутреннюю структуру данных. К основным задачам относятся кластеризация, снижение размерности и поиск аномалий.

Схема: Обучение без учителя



Последовательность отражает основные этапы рассматриваемого процесса; конкретная реализация зависит от задачи.

Понятие кластеризации

Кластеризация — метод анализа данных, при котором объекты объединяются в группы, называемые кластерами. Желательно, чтобы объекты внутри одного кластера были максимально похожи друг на друга, а объекты разных кластеров — различались. В отличие от классификации, номера или названия кластеров заранее неизвестны.

Ключевая идея: Кластеризация — метод анализа данных, при котором объекты объединяются в группы, называемые кластерами. Желательно, чтобы объ

Понятие кластеризации — продолжение

Алгоритм самостоятельно формирует группы на основании структуры признакового пространства. Пример. Имеется набор событий информационной системы, описанных частотой обращений, объёмом переданных данных, временем активности и количеством неуспешных попыток входа. Кластеризация может выявить группы типичного поведения и небольшую группу необычных событий, требующую дополнительного анализа.

Сравнительная таблица: Понятие кластеризации

Понятие кластеризации	Кластеризация — метод анализа данных, при котором объекты объединяются в группы, называемые кластерами. Желательно,
Метрики расстояния и масштабирование	В отличие от классификации, номера или названия кластеров заранее неизвестны. Алгоритм самостоятельно формирует гру
Алгоритм K-Means	Пример. Имеется набор событий информационной системы, описанных частотой обращений, объёмом переданных данных, врем

Метрики расстояния и масштабирование

Многие методы кластеризации используют понятие расстояния между объектами. Одной из наиболее распространённых является евклидова метрика: $d(x, y) = \sqrt{\Sigma(x_i - y_i)^2}$ Евклидово расстояние зависит от масштаба признаков. Если один признак изменяется от 0 до 1, а другой — от 0 до 100000, второй признак будет существенно сильнее влиять на расстояние.

Ключевая идея: Многие методы кластеризации используют понятие расстояния между объектами. Одной из наиболее распространённых является евклид

Метрики расстояния и масштабирование — продолжение

Поэтому перед K-Means и рядом других методов данные обычно стандартизируют или нормализуют.

Алгоритм K-Means

K-Means — один из наиболее известных алгоритмов кластеризации. Пользователь заранее задаёт число кластеров K. Алгоритм ищет K центров — центроидов — и распределяет объекты между ними. Основные шаги алгоритма: 1. Выбрать число кластеров K. 2. Инициализировать K центроидов. 3. Для каждого объекта вычислить расстояние до центроидов и назначить объект ближайшему кластеру. 4.

Ключевая идея: K-Means — один из наиболее известных алгоритмов кластеризации. Пользователь заранее задаёт число кластеров K. Алгоритм ищет K

Алгоритм K-Means — продолжение

Пересчитать каждый центроид как среднее положение объектов его кластера. 5. Повторять назначение объектов и пересчёт центроидов до стабилизации решения. Цель K-Means — уменьшить суммарное внутрикластерное расстояние объектов до соответствующих центроидов. В литературе эта величина часто называется inertia или within-cluster sum of squares. Пример.

Алгоритм K-Means — продолжение

Для анализа активности пользователей можно использовать признаки: число входов, средняя длительность сессии, объём переданных данных и количество ошибок аутентификации. K-Means может разделить пользователей на несколько поведенческих групп.

Схема: Алгоритм K-Means

K-Means — один из наиболее известных алгоритмов

Основные шаги алгоритма:

1. Выбрать число кластеров K.

2. Инициализировать K центроидов.

Последовательность отражает основные этапы рассматриваемого процесса; конкретная реализация зависит от задачи.

Выбор числа кластеров K

Одной из главных особенностей K-Means является необходимость заранее определить K. На практике используют сочетание количественных критериев и содержательной интерпретации. Нельзя выбирать K только по красивой визуализации. Результат должен быть устойчивым и интерпретируемым в контексте конкретной задачи.

Ключевая идея: Одной из главных особенностей K-Means является необходимость заранее определить K. На практике используют сочетание количеств

Иерархическая кластеризация

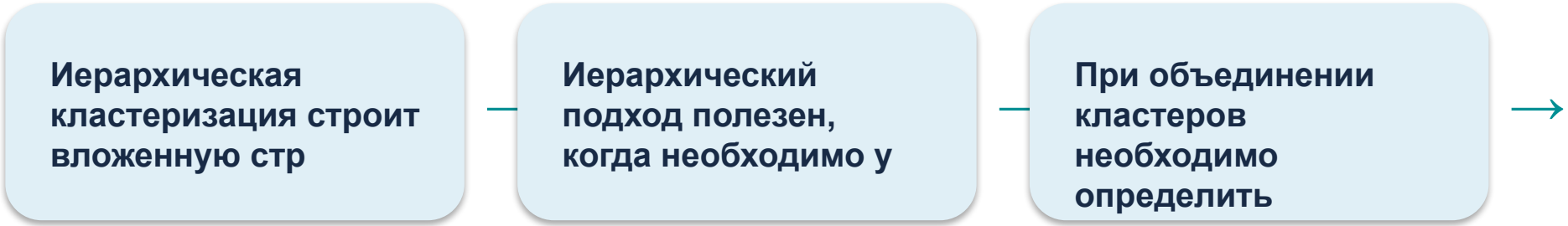
Иерархическая кластеризация строит вложенную структуру групп. При агломеративном подходе каждый объект сначала рассматривается как отдельный кластер, после чего наиболее близкие кластеры последовательно объединяются. Результат можно представить в виде дендрограммы.

Ключевая идея: Иерархическая кластеризация строит вложенную структуру групп. При агломеративном подходе каждый объект сначала рассматриваетс

Иерархическая кластеризация — продолжение

Иерархический подход полезен, когда необходимо увидеть структуру данных на нескольких уровнях детализации и исследовать, какие объекты объединяются раньше, а какие — только на более высоком уровне. При объединении кластеров необходимо определить linkage — способ вычисления расстояния между группами. Используются single linkage, complete linkage, average linkage и метод Ward.

Схема: Иерархическая кластеризация



Последовательность отражает основные этапы рассматриваемого процесса; конкретная реализация зависит от задачи.

Алгоритм DBSCAN

DBSCAN — плотностный метод кластеризации. В отличие от K-Means он не требует заранее задавать число кластеров и способен выделять области произвольной формы. Алгоритм рассматривает плотность объектов в окрестности каждой точки.

Ключевая идея: DBSCAN — плотностный метод кластеризации. В отличие от K-Means он не требует заранее задавать число кластеров и способен выде

Алгоритм DBSCAN — продолжение

Ключевые параметры DBSCAN — eps, определяющий радиус окрестности, и min_samples, задающий минимальное количество объектов, необходимое для формирования плотной области. Способность помечать отдельные объекты как шум делает DBSCAN интересным для первичного поиска необычных событий.

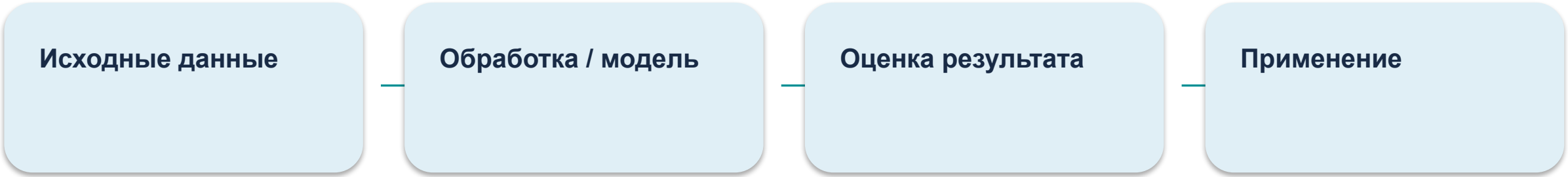
Алгоритм DBSCAN — продолжение

Однако обнаруженный шум нельзя автоматически считать атакой: это только объект, не вписавшийся в плотностную структуру данных.

Оценка качества кластеризации

В кластеризации правильные ответы часто отсутствуют, поэтому оценка отличается от классификации. Используются внутренние критерии, основанные на компактности и разделимости групп, а также экспертная интерпретация.

Схема: Оценка качества кластеризации



Последовательность отражает основные этапы рассматриваемого процесса; конкретная реализация зависит от задачи.

Сравнительная таблица: Оценка качества кластеризации

Оценка качества кластеризации	В кластеризации правильные ответы часто отсутствуют, поэтому оценка отличается от классификации. Используются внутр

Кластеризация в информационной безопасности

В информационной безопасности кластеризация может применяться для анализа поведения пользователей, группировки сетевых потоков, исследования событий SIEM, сегментации вредоносных образцов и поиска необычных паттернов. Например, сетевые потоки можно представить признаками duration, packet_count, bytes_in, bytes_out, average_packet_size и connection_rate.

Ключевая идея: В информационной безопасности кластеризация может применяться для анализа поведения пользователей, группировки сетевых потоко

Кластеризация в информационной безопасности — продолжение

После стандартизации K-Means формирует группы похожих потоков. Затем аналитик изучает характеристики кластеров и определяет, соответствуют ли они типичному поведению или требуют дополнительной проверки. В материалах по информационной безопасности также используются K-Means и DBSCAN как методы аномалийно-ориентированного анализа сетевых и хостовых метрик.

Кластеризация в информационной безопасности — продолжение

Их преимущество состоит в возможности работать без заранее подготовленных меток, а ограничением является риск ложных срабатываний и необходимость корректно описывать нормальное поведение.

Что такое снижение размерности

Современные наборы данных могут содержать десятки, сотни и тысячи признаков. Большая размерность увеличивает вычислительные затраты, затрудняет визуализацию и может приводить к избыточности признаков. Снижение размерности преобразует исходное пространство в пространство меньшей размерности с сохранением наиболее существенной информации.

Ключевая идея: Современные наборы данных могут содержать десятки, сотни и тысячи признаков. Большая размерность увеличивает вычислительные з

Что такое снижение размерности — продолжение

Основные цели снижения размерности: визуализация многомерных данных; уменьшение вычислительной сложности; устранение избыточности и коррелированных признаков; подготовка компактного представления для последующего анализа.

Метод главных компонент PCA

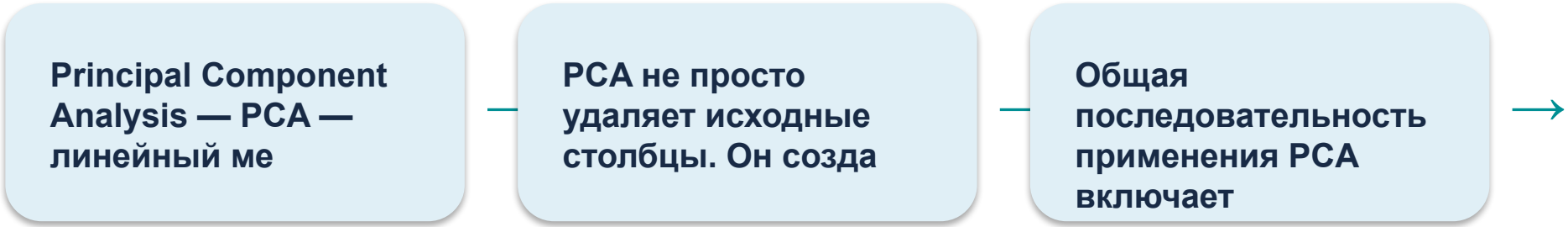
Principal Component Analysis — PCA — линейный метод снижения размерности. Он преобразует исходные признаки в новые взаимно ортогональные направления — главные компоненты. Первая компонента описывает максимально возможную долю вариации данных, следующая — максимальную оставшуюся вариацию при условии ортогональности предыдущей и т. д. PCA не просто удаляет исходные столбцы.

Ключевая идея: Principal Component Analysis — PCA — линейный метод снижения размерности. Он преобразует исходные признаки в новые взаимно ор

Метод главных компонент PCA — продолжение

Он создаёт новые признаки как линейные комбинации исходных переменных. Общая последовательность применения PCA включает стандартизацию признаков, построение ковариационной структуры, определение главных направлений и проекцию объектов на выбранное число компонент.

Схема: Метод главных компонент PCA



Последовательность отражает основные этапы рассматриваемого процесса; конкретная реализация зависит от задачи.

Объяснённая дисперсия

При использовании PCA необходимо определить, сколько компонент оставить. Для этого анализируется explained variance ratio — доля дисперсии, объясняемая каждой компонентой. Суммарная объяснённая дисперсия показывает, какая часть вариативности исходных данных сохраняется после преобразования.

Ключевая идея: При использовании PCA необходимо определить, сколько компонент оставить. Для этого анализируется explained variance ratio — д

Объяснённая дисперсия — продолжение

Например, если первые 10 компонент сохраняют большую часть объяснённой дисперсии исходных 100 признаков, их можно использовать как более компактное представление. Конкретный порог сохранения информации выбирается в зависимости от задачи, а не как универсальное значение.

Сравнительная таблица: Объяснённая дисперсия

Объяснённая дисперсия	При использовании PCA необходимо определить, сколько компонент оставить. Для этого анализируется explained variance
PCA для текстовых и высокоразмерных признаков	Например, если первые 10 компонент сохраняют большую часть объяснённой дисперсии исходных 100 признаков, их можно и

PCA для текстовых и высокоразмерных признаков

В загруженных материалах PCA применяется к числовому представлению текстов TF-IDF. Каждая строка такой матрицы соответствует документу, а столбцы — терминам. Число признаков может быть очень большим, поэтому снижение размерности позволяет сформировать компактные латентные компоненты.

Ключевая идея: В загруженных материалах PCA применяется к числовому представлению текстов TF-IDF. Каждая строка такой матрицы соответствует

PCA для текстовых и высокоразмерных признаков — продолжение

После преобразования анализируется explained_variance_ratio_, чтобы оценить объём информации, сохранённый в сокращённом пространстве. Это позволяет выбирать число компонент как компромисс между компактностью и сохранением структуры данных.

Метод t-SNE

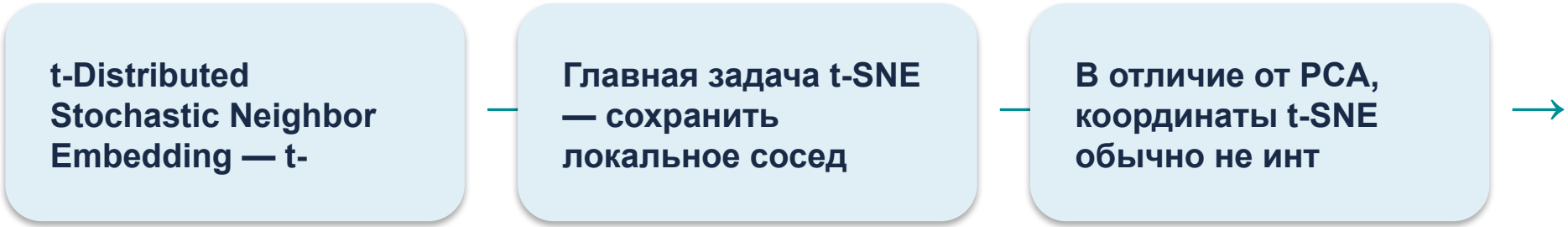
t-Distributed Stochastic Neighbor Embedding — t-SNE — нелинейный метод снижения размерности, часто применяемый для визуализации многомерных данных в двух- или трёхмерном пространстве. Главная задача t-SNE — сохранить локальное соседство: объекты, близкие в исходном многомерном пространстве, стремятся располагаться близко и на визуализации.

Ключевая идея: t-Distributed Stochastic Neighbor Embedding — t-SNE — нелинейный метод снижения размерности, часто применяемый для визуализац

Метод t-SNE — продолжение

Поэтому t-SNE удобен для исследовательского анализа возможных групп. В отличие от PCA, координаты t-SNE обычно не интерпретируют как отдельные содержательные признаки. Расстояния между далёкими группами и размеры визуальных кластеров также не следует трактовать слишком буквально.

Схема: Метод t-SNE



Последовательность отражает основные этапы рассматриваемого процесса; конкретная реализация зависит от задачи.

Связь кластеризации и снижения размерности

Эти методы часто используются совместно. Сначала данные могут быть стандартизированы и преобразованы PCA, затем в компактном пространстве выполняется кластеризация. Другой вариант — сначала получить кластеры, а затем использовать PCA или t-SNE для двумерной визуализации результата.

Ключевая идея: Эти методы часто используются совместно. Сначала данные могут быть стандартизированы и преобразованы PCA, затем в компактном

Связь кластеризации и снижения размерности — продолжение

Важно помнить, что визуальная картинка после t-SNE сама по себе не доказывает существование реальных кластеров. Кластеризация и визуализация должны оцениваться отдельными критериями.

Связь кластеризации и снижения размерности — продолжение

Пример K-Means в Python from sklearn.preprocessing import StandardScaler from sklearn.cluster import KMeans from sklearn.metrics import silhouette_score X = data[['duration', 'packets', 'bytes', 'failed_logins']] scaler = StandardScaler() X_scaled = scaler.fit_transform(X) model = KMeans(n_clusters=3, random_state=42, n_init='auto') labels = model.fit_predict(X_scaled) print('Silhouette:', silhouette_score(X_scaled, labels)) print('Centers:', model.cluster_centers_) Пример DBSCAN в Python from sklearn.preprocessing import StandardScaler from sklearn.cluster import DBSCAN X_scaled = StandardScaler().fit_transform(X) dbscan = DBSCAN(eps=0.6, min_samples=5) labels = dbscan.fit_predict(X_scaled) data['cluster'] = labels print(data['cluster'].value_counts()) 21.

Связь кластеризации и снижения размерности — продолжение

Пример PCA в Python from sklearn.preprocessing import StandardScaler from sklearn.decomposition import PCA X_scaled = StandardScaler().fit_transform(X) pca = PCA(n_components=2) X_pca = pca.fit_transform(X_scaled) print('Explained variance:', pca.explained_variance_ratio_) print('Total:', pca.explained_variance_ratio_.sum())

Сравнительная таблица: Связь кластеризации и снижения размерности

Связь кластеризации и снижения размерности	Эти методы часто используются совместно. Сначала данные могут быть стандартизированы и преобразованы PCA, затем в к
Типичные ошибки	Важно помнить, что визуальная картинка после t-SNE сама по себе не доказывает существование реальных кластеров. Кла
Практическая схема выбора метода	Пример K-Means в Python
Основные выводы	from sklearn.preprocessing import StandardScaler from sklearn.cluster import KMeans from sklearn.metrics import sil

Типичные ошибки

Применять K-Means к немасштабированным признакам с сильно различающимися диапазонами. Выбирать число кластеров только по визуальному впечатлению. Считать номера кластеров готовыми смысловыми классами. Автоматически интерпретировать все точки DBSCAN с меткой noise как атаки. Использовать t-SNE как доказательство наличия кластеров.

Ключевая идея: Применять K-Means к немасштабированным признакам с сильно различающимися диапазонами.

Типичные ошибки — продолжение

Выбирать число компонент PCA без анализа объяснённой дисперсии. Оценивать кластеризацию только одним числовым критерием без анализа содержимого групп.

Схема: Типичные ошибки



Последовательность отражает основные этапы рассматриваемого процесса; конкретная реализация зависит от задачи.

Основные выводы

Кластеризация относится к обучению без учителя и выявляет группы без заранее заданных меток. K-Means формирует K кластеров вокруг центроидов и требует выбора числа кластеров. DBSCAN выделяет плотные области и шум и не требует заранее задавать число кластеров. Результат кластеризации зависит от признаков, масштаба данных, метрики расстояния и параметров алгоритма.

Ключевая идея: Кластеризация относится к обучению без учителя и выявляет группы без заранее заданных меток.

Основные выводы — продолжение

PCA формирует новые главные компоненты и позволяет уменьшить размерность с контролем объяснённой дисперсии. t-SNE прежде всего применяется для визуального исследования локальной структуры многомерных данных. В задачах информационной безопасности методы без учителя полезны для исследования поведения и поиска необычных паттернов, но требуют экспертной интерпретации.

Контрольные вопросы

1. Что понимается под обучением без учителя? 2. Чем кластеризация отличается от классификации? 3. Что такое кластер? 4. Какую роль играет метрика расстояния? 5. Почему перед K-Means важно масштабировать признаки? 6. Каковы основные шаги K-Means? 7. Что такое центроид? 8. Почему в K-Means необходимо задавать K? 9. В чём состоит метод локтя? 10. Что показывает silhouette score? 11.

Ключевая идея: 1. Что понимается под обучением без учителя?

Контрольные вопросы — продолжение

Какие ограничения имеет K-Means? 12. Как работает иерархическая кластеризация? 13. Что показывает дендрограмма? 14. Что такое linkage? 15. В чём состоит идея DBSCAN? 16. Что означают параметры eps и min_samples? 17. Какие точки DBSCAN считает шумом? 18. Почему шум DBSCAN нельзя автоматически считать атакой? 19. Как оценивать качество кластеризации без известных меток? 20.

Контрольные вопросы — продолжение

Как кластеризация применяется в информационной безопасности? 21. Что понимается под снижением размерности? 22. Какие задачи решает PCA? 23. Что такое главная компонента? 24. Что показывает explained variance ratio? 25. Почему PCA полезен для высокоразмерных данных? 26. Для чего используется t-SNE? 27. Чем PCA отличается от t-SNE? 28.

Контрольные вопросы — продолжение

Почему визуализация t-SNE не доказывает наличие кластеров? 29. Как совместно применять PCA и кластеризацию? 30. Как выбрать метод для конкретной задачи? Список источников: 1. Силлабус дисциплины «Искусственный интеллект» для образовательной программы 7M06301 «Комплексное обеспечение информационной безопасности». Тема недели 7 — «Методы кластеризации и снижения размерности данных». 2.

Контрольные вопросы — продолжение

Учебное пособие «Основы искусственного интеллекта». Глава 7 «Кластеризация и снижение размерности»: K-Means, DBSCAN, PCA, t-SNE. 3. Материалы по применению TF-IDF, PCA и t-SNE для снижения размерности текстовых признаков и анализа объяснённой дисперсии. 4.

Схема: Контрольные вопросы

1. Что понимается под обучением без учителя?

2. Чем кластеризация отличается от классификации

4. Какую роль играет метрика расстояния?

5. Почему перед K-Means важно масштабировать при

Последовательность отражает основные этапы рассматриваемого процесса; конкретная реализация зависит от задачи.

Сравнительная таблица: Контрольные вопросы

Контрольные вопросы	
	1. Что понимается под обучением без учителя?

График: иллюстрация зависимости



Иллюстративные значения (не результаты эксперимента).

Итоги и вопросы для обсуждения

1. Объясните понятие «Обучение без учителя» и приведите пример.

2. Объясните понятие «Понятие кластеризации» и приведите пример.

3. Объясните понятие «Метрики расстояния и масштабирование» и приведите пример.

4. Объясните понятие «Алгоритм K-Means» и приведите пример.

5. Объясните понятие «Выбор числа кластеров K» и приведите пример.

6. Объясните понятие «Преимущества и ограничения K-Means» и приведите пример.

7. Объясните понятие «Иерархическая кластеризация» и приведите пример.