**Л1**

1. Назовите источники появления Больших Данных.

Источники появления больших данных включают в себя: устройства Интернета вещей (IoT), социальные сети, онлайн-транзакции, цифровые медиафайлы, датчики, генераторы данных и т.д.

1. В каких областях деятельности используются большие данные, привести примеры.

Большие данные используются во многих областях, включая медицину, финансы, маркетинг, производство, логистику и многое другое. Например, анализ больших данных может помочь маркетологам оптимизировать кампании по продвижению товаров, а анализ данных из IoT-устройств может помочь предотвратить отказы оборудования в производственной среде.

1. Основные вызовы больших данных (6V).

Основные вызовы, связанные с большими данными, называются 6V: объем (volume), скорость (velocity), разнообразие (variety), правдивость (veracity), ценность (value) и визуализация (visualization).

1. Определение термина "большие данные", источники получения больших данных.

Большие данные - это объемы данных, которые невозможно эффективно обрабатывать традиционными методами. Источники получения больших данных включают в себя IoT-устройства, социальные сети, транзакции онлайн-магазинов, медиафайлы и т.д.

1. Перечислите и охарактеризуйте логические слои для работы с большими данными.

Логические слои для работы с большими данными включают: сбор и хранение данных, обработку и анализ данных, визуализацию данных и принятие решений на основе данных.

1. Данные, информация, знания – в чем отличия?

Данные - это просто факты и цифры, информация - это данные, которые были обработаны и превращены в знания, а знания - это информация, которая может быть использована для принятия решений.

1. Области деятельности, где эффективно используются БД, примеры.

Большие данные эффективно используются в таких областях, как маркетинг, финансы, производство, медицина, логистика и т.д. Например, анализ больших данных помогает компаниям оптимизировать свою цепочку поставок и сократить издержки производства, а в медицине может помочь выявить риски заболевания и определить наилучшее лечение.

**Л2**

1. Каковы основные инструменты аналитики больших данных, провести сравнительную характеристику.

Основными инструментами аналитики больших данных являются: Hadoop, Spark, Cassandra, MongoDB, R, Python, SAS и т.д. Сравнительная характеристика этих инструментов может включать в себя такие критерии, как поддержка распределенной обработки данных, скорость обработки, возможности машинного обучения, удобство использования, поддержка сообщества и т.д.

1. Сравнительная характеристика R и Python.

R и Python являются двумя из самых популярных языков программирования для работы с данными. Они имеют сходные возможности, но отличаются в некоторых аспектах. Например, R имеет более широкие возможности для статистического анализа данных, в то время как Python имеет более широкие возможности для машинного обучения и обработки естественного языка. Python также обладает более простым синтаксисом и широкой поддержкой сообщества.

1. Охарактеризовать конструкции языка R

Конструкции языка R включают в себя переменные, функции, циклы, условные операторы, операторы присваивания, операторы сравнения, операторы логических операций и т.д.

1. Перечислить типы языка R, привести примеры.

Типы языка R включают в себя: числа (integer, numeric), символы (character), логические значения (logical), факторы (factor), векторы (vector), матрицы (matrix), списки (list) и т.д. Примеры: integer - 1, 2, 3; numeric - 1.5, 2.3, 3.8; character - "hello", "world"; logical - TRUE, FALSE; factor - "yes", "no", "maybe"; vector - c(1, 2, 3); matrix - matrix(c(1, 2, 3, 4), nrow=2); list - list(name="John", age=30, sex="male").

1. Структуры и типы данных в языке R, привести примеры.

Структуры и типы данных в R включают в себя: векторы, матрицы, фреймы (data frames), списки и факторы. Примеры: векторы - c(1, 2, 3); матрицы - matrix(c(1, 2, 3, 4), nrow=2); фреймы - data.frame(name=c("John", "Kate"), age=c(30, 25), sex=c("male", "female")); списки - list(name="John", age=30, sex="male"); факторы - factor(c("yes", "no", "maybe"), levels=c("yes", "no", "maybe")).

1. Векторы, матрицы, фреймы и факторы в R. Сходство и различия, способы обработки.

Векторы, матрицы, фреймы и факторы в R используются для хранения и обработки данных. Векторы - это одномерный массив элементов одного типа данных, который может быть использован для хранения и операций над одномерными данными.

Матрицы - это двумерные массивы, содержащие элементы одного типа данных, расположенные в виде таблицы. Матрицы в R можно создать с помощью функции matrix().

Фреймы данных (data frames) - это таблицы, которые могут содержать элементы разных типов данных. Фреймы данных могут использоваться для хранения и обработки данных в виде таблицы, где каждая строка представляет отдельное наблюдение, а каждый столбец представляет отдельную переменную. Фреймы данных можно создать с помощью функции data.frame().

Факторы (factors) - это тип данных, который используется для хранения категориальных переменных. Факторы могут быть созданы с помощью функции factor().

Факторы могут использоваться для анализа категориальных переменных, таких как пол, образование, место жительства и т.д.

1. Принцип массивных (векторных и матричных вычислений в R).

Принцип массивных (векторных и матричных) вычислений является одним из основных принципов языка R и позволяет эффективно работать с большими объемами данных. Он заключается в том, что операции выполняются над целыми векторами или матрицами, а не над отдельными элементами.

**Л3**

1. Основные понятия статистики и дескриптивный анализ

Основные понятия статистики включают сбор, анализ и интерпретацию данных с целью описания и понимания явлений в области, которую изучает статистика. Дескриптивный анализ - это метод, который используется для описания данных с помощью статистических показателей и графиков.

1. Генеральная совокупность и выборка.

Генеральная совокупность - это полный набор всех объектов, которые могут быть изучены в рамках конкретного исследования. Выборка - это часть генеральной совокупности, которая выбирается для анализа.

1. Шкалы измерений.

Шкалы измерений - это классификация методов измерения переменных в статистике. Включает в себя номинальную, порядковую, интервальную и относительную шкалы.

1. Меры центральной тенденции, их сравнительный анализ.

Меры центральной тенденции используются для описания того, где находится "центр" набора данных. Среди них: среднее, медиана, мода. Их сравнительный анализ позволяет понять, какая мера лучше отображает типичное значение набора данных.

1. Виды функций распределения. Нормальное распределение. Уровень статистической достоверности.

Виды функций распределения включают равномерное, нормальное, биномиальное и другие. Нормальное распределение - это распределение вероятностей, которое характеризуется колоколообразной формой. Уровень статистической достоверности - это вероятность ошибки, которую можно совершить при принятии статистических выводов.

1. Свойства описательных статистик (Дескриптивный анализ).

Свойства описательных статистик (дескриптивного анализа) включают минимальное и максимальное значения, размах, меры центральной тенденции, меры изменчивости и форму распределения.

1. Перечислить и охарактеризовать меры изменчивости.

Меры изменчивости используются для описания разброса данных. Включают дисперсию, стандартное отклонение, межквартильный размах и коэффициент вариации.

**Л4**

1. Способы графического представления данных.

Способы графического представления данных могут включать гистограммы, столбчатые диаграммы, диаграммы рассеяния, boxplot и другие графики.

1. Примеры использования гистограммы для обработки фотографий и оценки качества изделий.

Гистограммы могут использоваться для обработки фотографий, например, для анализа распределения яркости пикселей. Они также могут использоваться для оценки качества изделий путем анализа распределения размеров или весов.

1. Виды столбчатых диаграмм и их интерпретация.

Виды столбчатых диаграмм могут включать группированные столбчатые диаграммы, столбчатые диаграммы со ступенчатой формой и другие. Они могут использоваться для сравнения значений между разными группами или категориями.

1. Boxplot и его интерпретация, связь этого графика с другими элементами анализа.

Boxplot (ящик с усами) представляет собой график, используемый для визуализации распределения данных. Он показывает медиану, квартили и выбросы в данных. Boxplot также может использоваться для определения наличия выбросов и сравнения распределений между разными группами данных.

1. Для чего нужны гипотезы в анализе данных, какие существуют приемы работы с гипотезами?

Гипотезы используются в анализе данных для проверки статистических различий между группами данных или для проверки гипотезы о значимости какого-либо фактора. Существуют различные приемы работы с гипотезами, включая определение уровня значимости, проверку на нормальность распределения данных и другие.

**Л5**

1. Опишите стандарты жизненного цикла Big Data.

Стандарты жизненного цикла Big Data включают в себя пять этапов: сбор и хранение данных, обработку и управление данными, анализ данных, визуализацию и представление данных, а также принятие решений на основе данных.

1. Что называется когнитивным анализом данных?

Когнитивный анализ данных - это метод анализа данных, который использует искусственный интеллект для извлечения знаний из больших объемов неструктурированных данных, таких как тексты, изображения, аудио и видеофайлы.

1. Назовите этапы интеллектуального анализа данных.

Этапы интеллектуального анализа данных включают в себя: сбор и подготовку данных, исследовательский анализ данных, моделирование, оценку результатов, принятие решений.

1. Что такое статистическое обучение?

Статистическое обучение - это процесс создания математических моделей, которые могут обучаться на основе данных, и использования этих моделей для предсказания будущих результатов на основе новых данных.

1. В чем разница между описательными и предсказательными задачами DM? Какие методы анализа лучше приспособлены для описательных, а какие для предсказательных задач?

Описательные задачи DM направлены на описание и понимание данных, а предсказательные задачи DM направлены на предсказание будущих результатов на основе данных прошлых событий. Описательные задачи DM лучше решаются методами дескриптивной статистики и визуализации данных, а предсказательные задачи DM лучше решаются с использованием алгоритмов машинного обучения.

1. В каких случаях лучше использовать линейные, а когда нелинейные модели анализа данных?

Линейные модели анализа данных лучше подходят для задач, в которых существует линейная зависимость между переменными, а нелинейные модели подходят для задач, в которых зависимость между переменными нелинейная.

1. Приведите математическое выражение для параметрической модели статистического обучения, для каких задач анализа данных их лучше использовать?

Параметрическая модель статистического обучения может быть выражена как y = f(x, β), где y - целевая переменная, x - вектор независимых переменных, а β - вектор параметров модели. Параметрические модели обычно используются для задач, в которых предполагается, что данные распределены по определенному статистическому закону.

1. В чем состоят преимущества и недостатки непараметрических моделей анализа данных, как осуществить выбор между параметрической и непараметрической моделью?

Преимущества непараметрических моделей анализа данных заключаются в том, что они могут быть использованы в случаях, когда неизвестно распределение данных или когда оно не является нормальным. Недостатки непараметрических моделей заключаются в том, что они требуют больше данных для достижения той же точности, что и параметрические модели.

1. Как выполняется измерение качества модели анализа данных?

Измерение качества модели анализа данных выполняется с помощью метрик оценки точности модели, таких как средняя абсолютная ошибка (MAE), средняя квадратичная ошибка (MSE), коэффициент детерминации (R^2) и другие.

1. В чем состоит фундаментальное свойство статистического обучения?

Фундаментальное свойство статистического обучения заключается в том, что модель должна обладать способностью обобщения - то есть она должна работать не только с обучающими данными, на которых была обучена, но и с новыми данными, которые ранее не встречались. Это достигается путем использования методов, которые минимизируют ошибку на обучающих данных и контролируют переобучение модели.

**Л6**

1. Основные задачи Data Mining. Какие дисциплины охватывает Data Mining?

Основные задачи Data Mining включают в себя поиск ассоциативных правил, классификацию, кластеризацию, регрессионный анализ и прогнозирование. Data Mining охватывает дисциплины, такие как статистика, машинное обучение, искусственный интеллект, базы данных и информационные технологии.

1. Классификация методов DM.

Классификация методов Data Mining включает в себя: деревья решений, байесовские сети, нейронные сети, метод опорных векторов, регрессионный анализ, кластерный анализ, и ассоциативные правила.

1. Понятие кластерного анализа, Классификация алгоритмов кластеризации.

Кластерный анализ - это задача разбиения набора данных на группы (кластеры) схожих объектов. Алгоритмы кластеризации могут быть иерархическими или неразделительными, плоскими или на основе графов, и могут использовать различные меры близости для оценки сходства объектов.

1. Зачем нужна мера близости в кластеризации? В чем достоинства алгоритмов, построенных на основе теории графов? Перечислите виды алгоритмов кластеризации на графах.

Мера близости в кластеризации используется для оценки сходства между объектами, что является ключевым элементом большинства алгоритмов кластеризации. Алгоритмы на основе теории графов могут использоваться для кластеризации данных, представленных в виде графа, и могут использовать различные методы для определения связей между узлами, такие как алгоритмы нахождения минимального остовного дерева или алгоритмы поиска максимального потока.

1. В чем суть алгоритмов нахождения квадратичной ошибки?

Алгоритмы нахождения квадратичной ошибки используются для минимизации суммы квадратов расстояний между объектами и центроидами (средними значениями) кластеров.

1. Плоские алгоритмы кластеризации перечислить, охарактеризовать работу.

Плоские алгоритмы кластеризации включают в себя метод k-средних, методы на основе плотности, методы на основе сеток и методы на основе моделей. Каждый из этих методов использует свой подход для определения кластеров, и может быть применен к различным типам данных.

1. Поясните суть работы алгоритмов нахождения связных компонент и алгоритмов покрывающего дерева.

Алгоритмы нахождения связных компонент и покрывающего дерева - это алгоритмы графовой теории, которые позволяют найти связи между объектами в графе. Алгоритм нахождения связных компонент находит множества вершин, которые связаны друг с другом, а алгоритм покрывающего дерева находит минимальное подмножество ребер графа, которое соединяет все вершины.

1. Опишите шаги построения дендрограммы.

Шаги построения дендрограммы:

* Выбор метрики расстояния для измерения близости между объектами.
* Вычисление матрицы расстояний между объектами на основе выбранной метрики.
* Объединение двух объектов с наименьшим расстоянием в один кластер.
* Повторение шага 3 до тех пор, пока все объекты не будут объединены в один кластер.
* Построение дендрограммы, где по оси x отображаются объекты, а по оси y - расстояния между ними.

1. В чем состоит суть стандартизации и нормализации переменных, зачем они нужны?

Стандартизация и нормализация переменных используются для приведения значений различных переменных к одному масштабу. Стандартизация приводит данные к стандартному нормальному распределению с нулевым средним и единичной дисперсией, а нормализация масштабирует данные так, чтобы они находились в интервале от 0 до 1 или от -1 до 1. Это необходимо для корректной работы многих методов анализа данных, таких как кластеризация и метод главных компонент.

**Л7**

1. Что такое наивный байесовский алгоритм? Какую задачу обработки данных он выполняет?

Наивный байесовский алгоритм - это алгоритм машинного обучения, основанный на теореме Байеса, который используется для классификации и прогнозирования вероятности принадлежности объекта к определенному классу на основе его характеристик.

1. Приведите и поясните формулу теорема Байеса.

Теорема Байеса - это формула для расчета условной вероятности наступления события A при наличии события B и предположении независимости B от других событий. Формула: P(A|B) = P(B|A) \* P(A) / P(B).

1. Деревья решений, опишите процесс работы, приемы остановки работы дерева.

Деревья решений - это метод машинного обучения, который строит дерево принятия решений на основе обучающей выборки. Процесс работы заключается в построении дерева, которое разбивает выборку на более мелкие подмножества, определяя признаки, наиболее значимые для разделения классов. Дерево останавливается, когда достигается определенный критерий остановки, например, когда узел содержит объектов меньше определенного порога.

1. Какие типы деревьев решений вы знаете, какие индексы используются при работе дерева, для чего, что такое энтропия?

Существует несколько типов деревьев решений, включая ID3, C4.5, CART и CHAID. Индексы, используемые при работе дерева, включают информационную энтропию, прирост информации и Gini-индекс. Энтропия используется для измерения неопределенности системы, прирост информации - для выбора наилучшего признака для разделения выборки, а Gini-индекс - для измерения разнообразия выборки.

1. Назовите достоинства и недостатки деревьев решения.

Деревья решения могут быстро обучаться на больших объемах данных, они легко интерпретируются и применяются в задачах классификации и регрессии. Однако они могут переобучаться и не учитывать нелинейные зависимости в данных.

1. Принцип работы RandomForest.

Random Forest - это алгоритм машинного обучения, который использует множество деревьев решений, чтобы уменьшить эффект переобучения. Он создает случайные выборки данных и признаков для каждого дерева и в итоге усредняет их предсказания.

**Л8**

1. Data Mining vs. Machine Learning – в чем отличия?

Data Mining - это процесс автоматического извлечения значимых и ранее неизвестных паттернов и взаимосвязей в больших объемах данных, а Machine Learning - это область искусственного интеллекта, в которой алгоритмы позволяют компьютерам обучаться на данных и делать прогнозы.

1. Нарисуйте схему классификации методов машинного обучения.

Схема классификации методов машинного обучения может быть представлена следующим образом:

Обучение с учителем:

* Классификация
* Регрессия

Обучение без учителя:

* Кластеризация
* Снижение размерности
* Обнаружение аномалий
* Обучение с подкреплением
* Ансамбли методов машинного обучения

1. Обучение с учителем и без учителя. Приведите примеры методов для обоих вариантов.

Обучение с учителем - это процесс обучения модели на основе размеченных данных, т.е. данных, в которых уже указаны правильные ответы. Примеры методов: линейная регрессия, логистическая регрессия, метод опорных векторов (SVM), решающие деревья.

Обучение без учителя - это процесс обучения модели на основе неразмеченных данных, т.е. данных, в которых нет правильных ответов. Примеры методов: кластерный анализ, метод главных компонент (PCA), ассоциативные правила.

1. Обучение с подкреплением и ансамбли – основные разновидности и принципы работы.

Обучение с подкреплением - это процесс обучения модели на основе опыта, который заключается в последовательном взаимодействии модели с окружающей средой. Ансамбли - это методы машинного обучения, которые комбинируют несколько моделей для повышения качества прогнозов.

1. Принципы глубокого обучения в нейросетях. В чем преимущества сверточных нейронных сетей, какие задачи они решают очень хорошо?

Принципы глубокого обучения в нейросетях включают в себя использование многослойных архитектур, обучение с помощью метода обратного распространения ошибки, использование активационных функций, таких как ReLU и сигмоида, и т.д. Сверточные нейронные сети обладают способностью автоматически выделять признаки из изображений и использовать их для решения задач классификации, детектирования объектов, распознавания речи и т.д.

1. Зачем нужны рекуррентные нейросети?

Рекуррентные нейросети (RNN) используются для анализа последовательностей данных, где каждый элемент зависит от предыдущих. Они имеют память, которая позволяет им запоминать предыдущие состояния и использовать их для прогнозирования будущих состояний или для генерации новых последовательностей данных. Это делает их полезными для задач, таких как распознавание речи, машинный перевод, анализ текстов, генерация текста, музыки и т.д.

**Л9**

1. Что такое статистическая гипотеза? Какие виды гипотез вы знаете?

Статистическая гипотеза - это предположение о параметрах распределения некоторой случайной величины в генеральной совокупности на основе выборки. Она позволяет проверять статистические различия между выборками. Виды гипотез: параметрические, непараметрические, односторонние, двусторонние.

1. Как принято формулировать нулевую гипотезу?

Нулевая гипотеза формулируется так, чтобы ее можно было противопоставить альтернативной гипотезе. Она предполагает, что различия между выборками случайны и не значимы.

1. Что такое уровень значимости, как он определяется, как влияет на решение о принятие гипотезы?

Уровень значимости - это вероятность отвергнуть нулевую гипотезу, когда она верна. Обычно используются уровни значимости 0.05 или 0.01. Они определяются заранее, и если полученное значение p-уровня значимости меньше выбранного уровня значимости, то нулевая гипотеза отвергается.

1. Каков порядок обработки данных при тестировании гипотезы о равенстве, какие еще тесты при этом должны быть пройдены, какие требования к данным выдвигаются?

Порядок обработки данных при тестировании гипотезы о равенстве: формулирование нулевой и альтернативной гипотез, выбор критерия и уровня значимости, проверка условий применения критерия, вычисление тестовой статистики, нахождение p-уровня значимости, принятие решения. Для проверки гипотез могут быть использованы различные критерии, например, t-критерий Стьюдента, критерий Манна-Уитни и др.

1. Для чего вообще нужна гипотеза о равенстве средних?

Гипотеза о равенстве средних используется для проверки статистических различий между двумя выборками. Например, можно проверить, есть ли статистически значимые различия в росте мужчин и женщин.

1. Как тестируются независимые и парные выборки?

Для тестирования независимых выборок используются t-критерий Стьюдента, критерий Манна-Уитни, критерий Колмогорова-Смирнова и др. Для тестирования парных выборок используется t-критерий Стьюдента для парных выборок, критерий знаков, критерий Вилкоксона и др. При этом данные должны удовлетворять условиям применения выбранного критерия, например, должна выполняться нормальность распределения или однородность дисперсий.

**Л10**

1. Понятие корреляции, коэффициент корреляции Пирсона, Спирмена, Кендела.

Корреляция - это статистическая мера, которая показывает, насколько две переменные связаны между собой. Коэффициент корреляции Пирсона измеряет линейную связь между двумя непрерывными переменными, коэффициент корреляции Спирмена измеряет нелинейную монотонную связь между двумя переменными, а коэффициент корреляции Кендала измеряет силу зависимости между двумя переменными на основе ранговых показателей.

1. Каковы факторы, влияющие на коэффициент корреляции?

Факторы, влияющие на коэффициент корреляции, могут включать в себя выбросы, нелинейную связь между переменными, наличие пропущенных данных, смещение и разброс выборки и другие факторы.

1. Назовите виды связи между переменными при корреляции.

Виды связи между переменными при корреляции могут включать в себя положительную линейную связь, отрицательную линейную связь, нелинейную связь, отсутствие связи и другие типы связи.

1. Что такое регрессионный анализ, какие задачи DM можно проводить с его помощью?

Регрессионный анализ - это метод, используемый для изучения отношений между зависимой переменной и одной или несколькими независимыми переменными. С помощью регрессионного анализа можно предсказать значения зависимой переменной на основе значений независимых переменных, а также определить статистическую значимость отношения между переменными.

1. Какие способы визуализации корреляции были изучены в курсе Big Data?

Способы визуализации корреляции, изученные в курсе Big Data, включают матрицу корреляции, диаграмму рассеяния, коэффициент детерминации и другие методы.

**Л11**

1. Перечислите основные задачи анализа сетей на графах. Приведите примеры.

Основные задачи анализа сетей на графах: определение центральности узлов, обнаружение сообществ, анализ путей и цепочек, прогнозирование связей в сети. Примеры: анализ социальных сетей, биологические сети, транспортные сети.

1. Перечислите разновидности сложных сетей, назовите их характеристики.

Разновидности сложных сетей: малого мира, безмасштабные, сверхсети. Характеристики: наличие узлов с большим количеством связей, неравномерность распределения связей, наличие сообществ.

1. Характерные черты безмасштабных сетей, какова их связь с сетями тесного мира?

Характерные черты безмасштабных сетей: степенное распределение степеней узлов, наличие связей между узлами с разными степенями, наличие сообществ, отсутствие характерной масштабной длины. Связь с сетями тесного мира: безмасштабные сети содержат подсети, которые могут быть описаны сетями тесного мира.

1. Каковы закономерности динамики сложных сетей и законы распространения информации в них.

Закономерности динамики сложных сетей: наличие паттернов поведения в сети, появление новых связей и узлов, изменение структуры сообществ. Законы распространения информации в них: модели распространения, такие как модель IC, модель LT, модель Активации-Передачи.

1. Свойства эластичности и надежности сложных сетей.

Свойства эластичности и надежности сложных сетей: эластичность определяет способность сети сохранять свою структуру при удалении узлов или связей, а надежность - способность сохранять свою функциональность. Оба свойства зависят от структуры сети и распределения связей между узлами.

1. Понятие регрессии. Как используется этот вид анализа?

Регрессия - это статистический метод, используемый для изучения отношений между двумя или более переменными. Он позволяет определить, как одна переменная влияет на другую. Регрессионный анализ может использоваться для прогнозирования значений зависимой переменной на основе известных значений независимых переменных. Он широко применяется в экономике, финансах, маркетинге, медицине и других областях.

**Л12**

1. Дайте определение социального графа. Перечислите его типы и свойства. К какому семейству больших графов он относится?

Социальный граф - это граф, который отображает связи между людьми или социальными сущностями. Его типы могут включать в себя графы друзей в социальных сетях, профессиональные графы, графы родственников и т.д. Свойства социального графа включают в себя длинные хвосты, модульность, связи на основе интересов, влиятельных лидеров и т.д. Он относится к семейству графов с тяжелым хвостом.

1. Какие алгоритмы лежат в основе методов выделения сообществ? Дайте общее описание шагов выполнения этих алгоритмов.

В основе методов выделения сообществ могут лежать алгоритмы, такие как метод модулярности, алгоритм Лувена, метод максимальной связности и т.д. Общие шаги выполнения включают определение меры качества кластеризации, разбиение графа на кластеры и улучшение качества кластеризации путем оптимизации меры качества.

1. Перечислите основные метрики больших графов.

Основные метрики больших графов могут включать в себя степень вершин, коэффициент кластеризации, центральность вершин, центральность по посредничеству, модулярность, ассортативность и т.д.

1. Назовите критерии качества кластеризации и поясните их значение и когда они используются.

Критерии качества кластеризации включают меры внутрикластерной связности, меры межкластерной связности, сбалансированность размеров кластеров и степень подобия между оригинальной структурой графа и структурой кластеризации. Они используются для определения эффективности кластеризации, улучшения результатов и выбора наилучшего алгоритма.

1. Приведите примеры задач, которые могут быть решены с помощью больших графов.

Задачи, которые могут быть решены с помощью больших графов, включают в себя анализ социальных сетей, поиск сообществ, анализ связей в интернете и т.д.

1. Дайте сравнительный анализ алгоритмов кластеризации.

Сравнительный анализ алгоритмов кластеризации зависит от ряда факторов, таких как тип данных, размер выборки, количество и размерность признаков и т.д. Однако, в общих чертах можно выделить следующие основные алгоритмы и их характеристики:

**K-средних:**

* Простота и эффективность в реализации.
* Предполагает, что кластеры имеют сферическую форму и одинаковую дисперсию.
* Чувствительность к выбросам.
* Иерархическая кластеризация:
* Позволяет получить иерархическую структуру кластеров.
* Не требует задания числа кластеров.
* Затратен по времени и памяти.

**DBSCAN:**

* Позволяет находить кластеры произвольной формы.
* Устойчив к выбросам и шумам.
* Требует настройки нескольких параметров.

**Spectral clustering:**

* Подходит для данных с нелинейной структурой.
* Может обнаруживать кластеры произвольной формы.
* Требует настройки параметров и вычисления собственных значений и собственных векторов.

**Affinity propagation:**

* Позволяет обнаруживать несколько центров кластеров.
* Может работать с выбросами.
* Затратен по времени и памяти.