**Лабораторная работа №2**

1. **Название работы: ”Кластеризация данных с использованием алгоритма K-средних”**
2. **Цель работы:**
3. Изучение и применение алгоритма k-средних для кластеризации данных.
4. **Основные задачи работы:**
5. Изучение теории кластеризации и алгоритма K-средних.
6. Реализация алгоритма k-средних на языке Python.
7. Применение алгоритма k-средних на реальных данных для выявления групп схожих объектов.
8. Оценка и интерпретация результатов кластеризации.
9. **Теоретическая часть работы:**

**Кластеризация** – это задача разделения набора данных на группы, называемые кластерами, таким образом, чтобы объекты внутри одного кластера были максимально схожи между собой, а объекты из разных кластеров отличались друг от друга. Алгоритм k-средних является одним из наиболее распространенных и простых методов кластеризации. Он основан на итеративном разделении объектов на k кластеров, где k - заданное число кластеров. Процесс кластеризации осуществляется путем минимизации суммарного квадратичного отклонения (SSE) между объектами и центроидами кластеров. Подробный обзор шагов алгоритма k-средних:

Шаг 1: Инициализация центроидов

Выбираются случайные k объектов из набора данных в качестве начальных центроидов.

Центроиды представляют центральные точки каждого кластера.

Шаг 2: Назначение объектов к ближайшим центроидам

Каждый объект назначается к ближайшему центроиду на основе некоторой метрики расстояния (обычно евклидово расстояние).

Объекты, находящиеся ближе к одному центроиду, считаются более похожими друг на друга, чем на объекты, находящиеся ближе к другим центроидам.

Шаг 3: Перерасчет центроидов

Вычисляются новые центроиды путем вычисления среднего значения всех объектов, назначенных к каждому кластеру.

Это обновление центроидов на основе текущего разделения объектов на кластеры.

Шаг 4: Повторение шагов 2 и 3

Процесс назначения объектов к ближайшим центроидам и перерасчета центроидов повторяется до сходимости.

Сходимость достигается, когда изменение центроидов становится незначительным или достигается максимальное число итераций.

Шаг 5: Завершение алгоритма

В результате работы алгоритма каждый объект будет назначен к одному из k кластеров.

Кластеры могут быть интерпретированы как группы объектов, близких друг к другу, а объекты внутри каждого кластера будут более похожи друг на друга, чем на объекты из других кластеров.

Применение алгоритма k-средних может быть полезно в различных областях, например:

Маркетинговые исследования: Кластеризация клиентов позволяет выявить сегменты схожих покупателей для более эффективной маркетинговой стратегии.

Анализ социальных сетей: Кластеризация пользователей позволяет выявить группы схожих интересов или поведения в социальных сетях.

Медицинская диагностика: Кластеризация пациентов может помочь выявить группы схожих заболеваний для более точной диагностики и лечения.

**5. Порядок выполнения работы:**

Шаг 1. Подготовка данных

Загрузите выборку данных для кластеризации.

Произведите предобработку данных, включая масштабирование и/или нормализацию признаков.

Шаг 2. Реализация алгоритма k-средних

Напишите функцию на языке Python, которая реализует алгоритм k-средних.

В функции реализуйте итеративный процесс пересчета центроидов и перераспределения объектов по кластерам.

Используйте подходящую метрику для измерения расстояния между объектами, например, Евклидово расстояние.

Шаг 3. Применение алгоритма k-средних

Примените реализованный алгоритм k-средних на подготовленных данных.

Определите оптимальное число кластеров k, используя подходы, такие как метод локтя или индекс силуэта.

Получите результаты кластеризации, присваивая каждому объекту соответствующий кластер.

Шаг 4. Оценка результатов кластеризации

Визуализируйте результаты кластеризации.

Оцените качество кластеризации с использованием внутренних и внешних метрик, таких как среднее расстояние внутри кластера или коэффициент силуэта.

Проанализируйте и интерпретируйте полученные кластеры, выявив схожие характеристики или закономерности.

6. Контрольные вопросы:

   1. Что такое кластеризация и для чего она используется в машинном обучении?

   2. Как работает алгоритм k-средних?

   3. Как выбрать оптимальное число кластеров k в алгоритме k-средних?

   4. Какие метрики используются для оценки качества кластеризации?

Вариант 1

Загрузите набор данных "Iris" (https://archive.ics.uci.edu/ml/datasets/iris) с помощью библиотеки Pandas.

Произведите масштабирование данных с использованием MinMaxScaler из библиотеки scikit-learn.

Примените алгоритм K-средних с числом кластеров k=3.

Визуализируйте результаты кластеризации, используя диаграмму рассеяния с отображением цветом кластеров.

Вариант 2

Сгенерируйте m точек в пространстве R[0,10]xR[0,10] таким образом, чтобы визуально его можно было разделить на k кластеров.

Примените алгоритм k-средних с числом кластеров k=3.

Визуализируйте результаты кластеризации, используя диаграмму рассеяния с отображением цветом кластеров.