МИНОБРНАУКИ РОССИИ
САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ
ЭЛЕКТРОТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ
«ЛЭТИ» ИМ. В.И. УЛЬЯНОВА (ЛЕНИНА)
Кафедра МО ЭВМ

ОТЧЕТ

по лабораторной работе №4

по дисциплине «Машинное обучение»

Тема: Кластеризация (k-средних, иерархическая)

Студент гр. 6304 _____ Антонов С.А.

Преподаватель _____ Жангиров Т.Р.

Санкт-Петербург

2020

Цель работы:

Ознакомиться с методами ассоциативного анализа из библиотеки MLxtend.

Ход работы:

Загрузка данных

1. На данном этапе был скачан и загружен датасет в датафрейм.

```
data = pd.read_csv('iris.data', header=None)
print(data)
```

	0	1	2	3	4
0	5.1	3.5	1.4	0.2	Iris-setosa
1	4.9	3.0	1.4	0.2	Iris-setosa
2	4.7	3.2	1.3	0.2	Iris-setosa
3	4.6	3.1	1.5	0.2	Iris-setosa
4	5.0	3.6	1.4	0.2	Iris-setosa
..	...	...	...	...	...
145	6.7	3.0	5.2	2.3	Iris-virginica
146	6.3	2.5	5.0	1.9	Iris-virginica
147	6.5	3.0	5.2	2.0	Iris-virginica
148	6.2	3.4	5.4	2.3	Iris-virginica
149	5.9	3.0	5.1	1.8	Iris-virginica

Рисунок 1 Загруженный датасет

K-means:

1. Проведена кластеризация методом k-средних.

```
kmeans = KMeans(init='k-means++', n_clusters=3, n_init=15)
kmeans.fit(no_labeled_data)
```

2. Получены центры кластеров и определены какие кластеры наблюдения попали в какой кластер.

```
k_means_cluster_centers = k_means.cluster_centers_
k_means_labels = pairwise_distances_argmin(no_labeled_data,
k_means_cluster_centers)
```

3. Построены результаты классификации для 4-х признаков попарно.

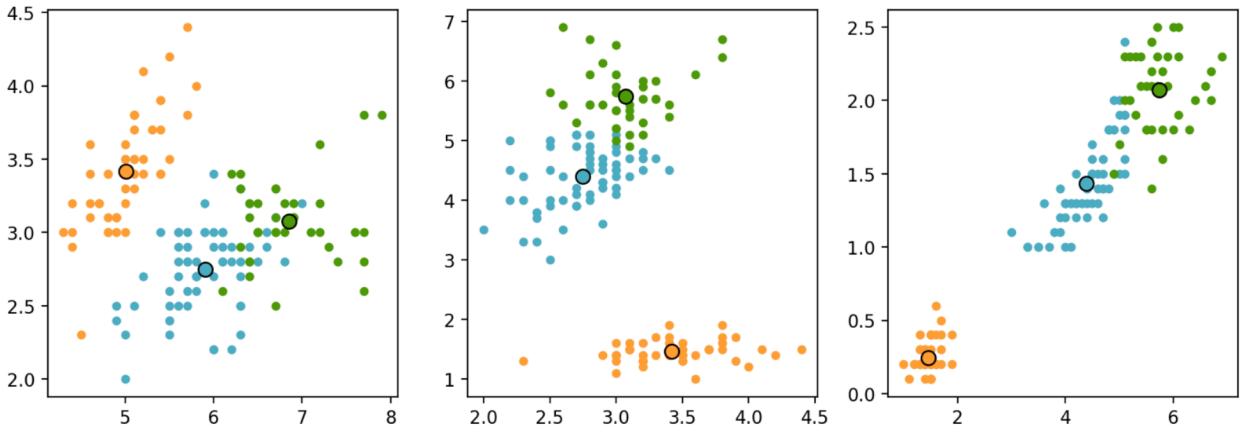

Рисунок 2 Попарные результаты

Исходя из рисунка, наилучшее разделение прошло по признакам 3 и 4. Параметр n_init не оказал видимых результатов.

4. Проведено уменьшение размерности до 2 с помощью PCA и составлена карта области значений, на которой каждый кластер занимает определенную область.

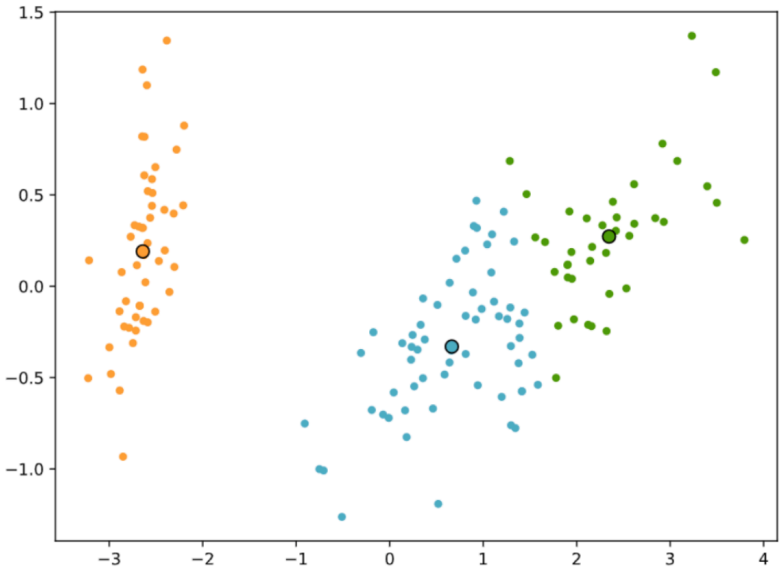

Рисунок 3 Классификация с уменьшенной размерностью данных

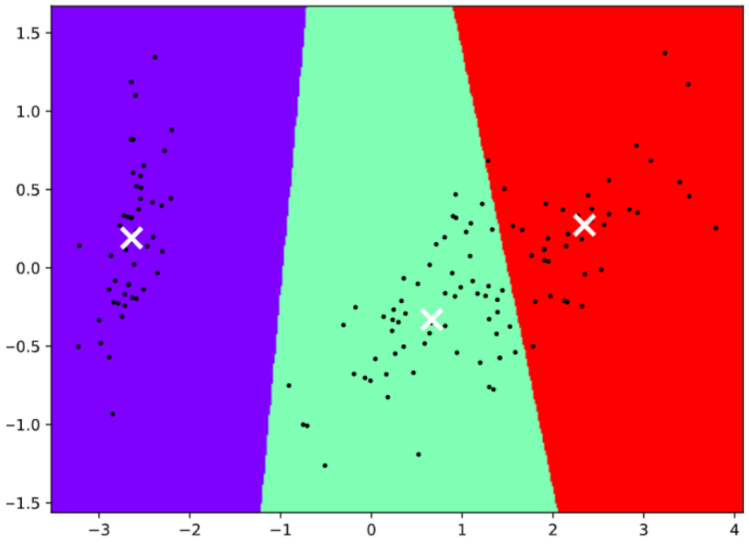

Рисунок 4 Карта области значений с уменьшенной размерностью

5. Исследована работа алгоритма при различных параметрах init. Сначала алгоритм был запущен с параметром random, затем для выбранных вручную точек.

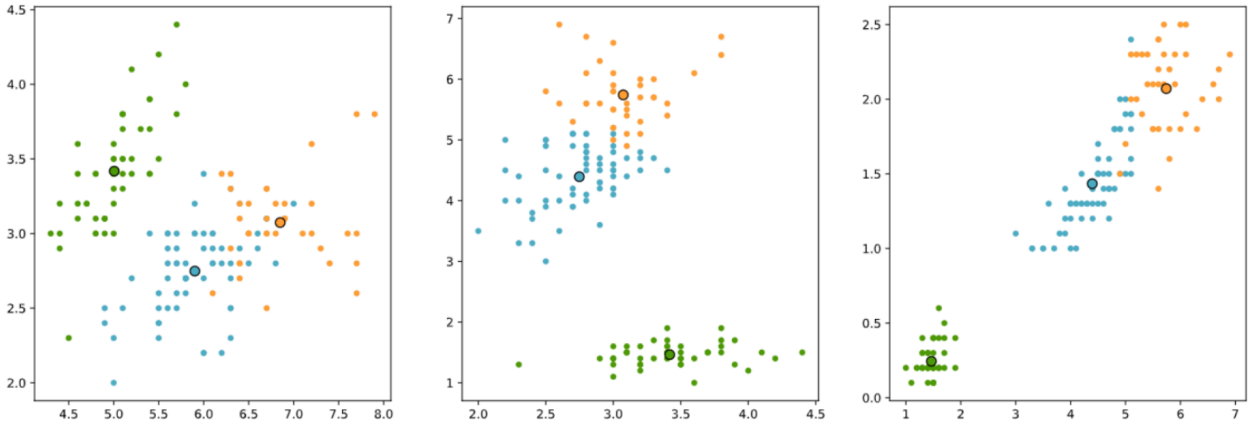

Рисунок 5 init = random, max_iter = 5

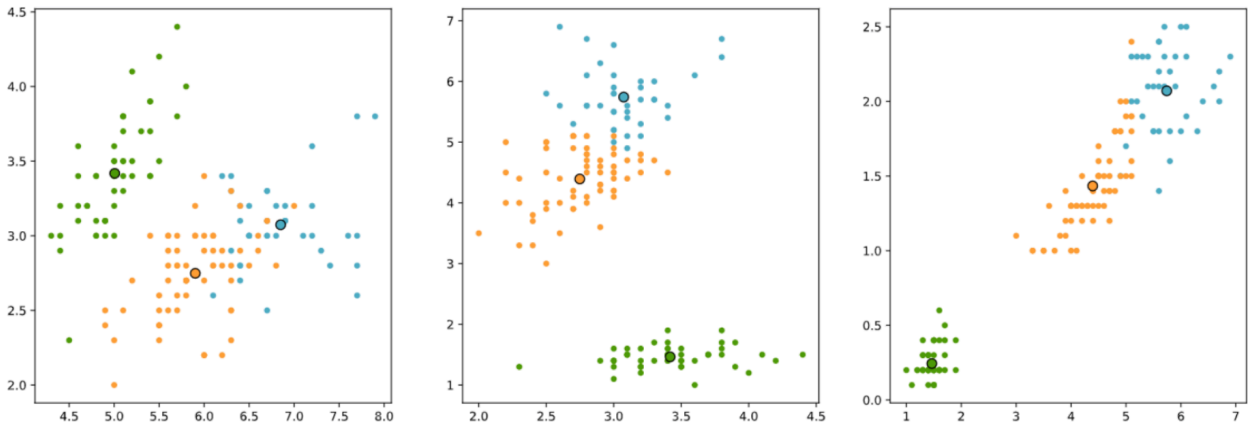

Рисунок 6 init = random, max_iter = 100

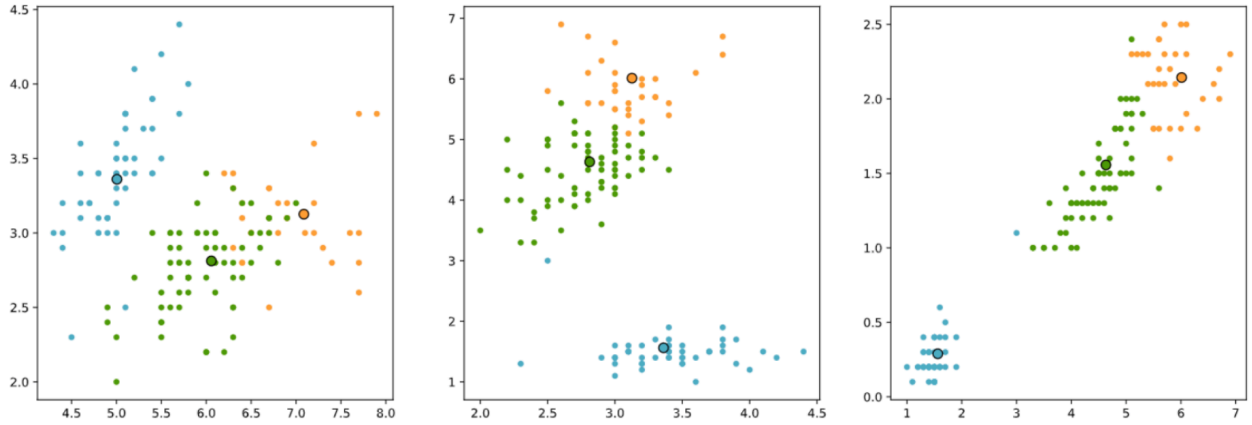

Рисунок 7 init = np.array([[0, 0, 0, 0], [0, 0, 0, 0], [0, 0, 0, 0]]), max_iter = 5

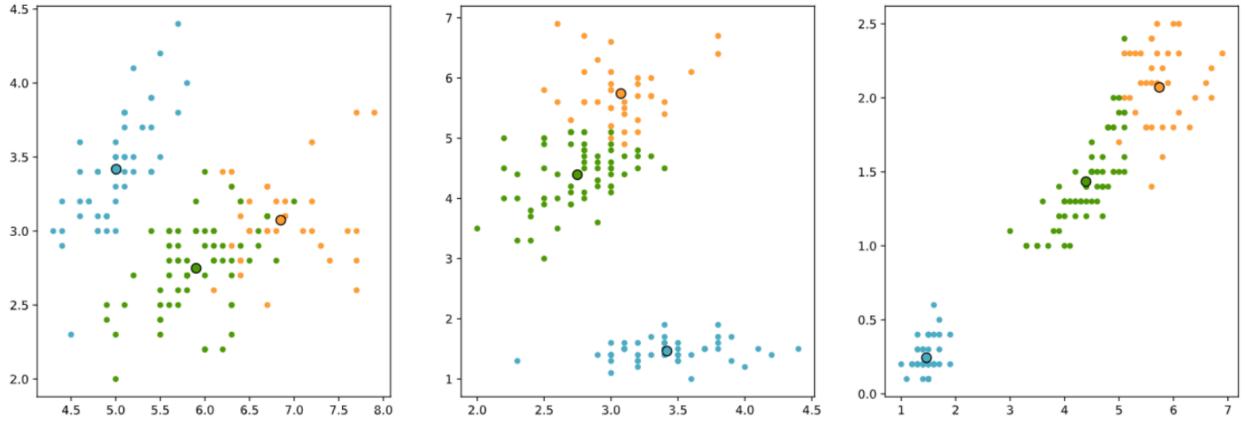

Рисунок 8 init = np.array([[0, 0, 0, 0], [0, 0, 0, 0], [0, 0, 0, 0]]), max_iter = 300

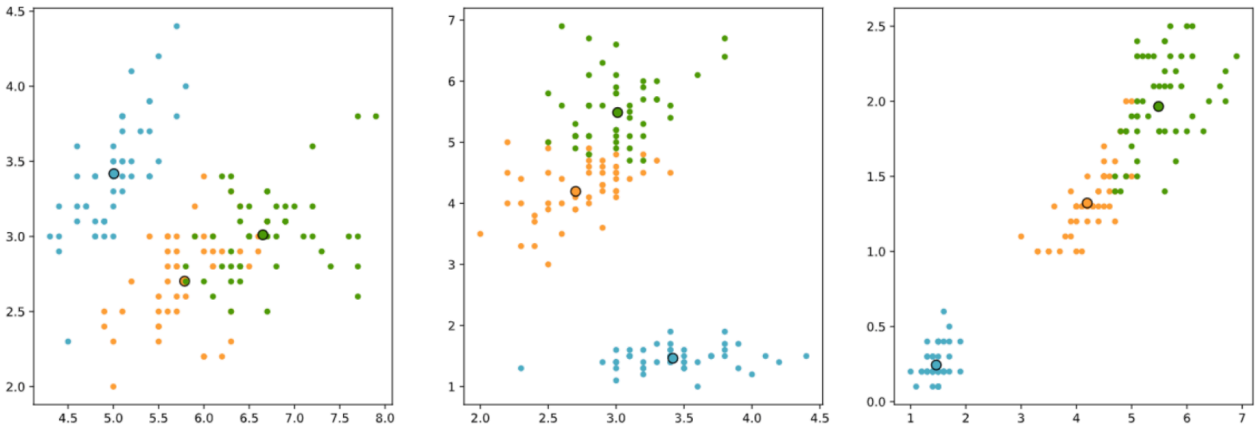

Рисунок 9 init = np.array([[5, 2, 1, 0], [5, 2, 3, 1], [6, 4, 5, 2]]), max_iter = 5

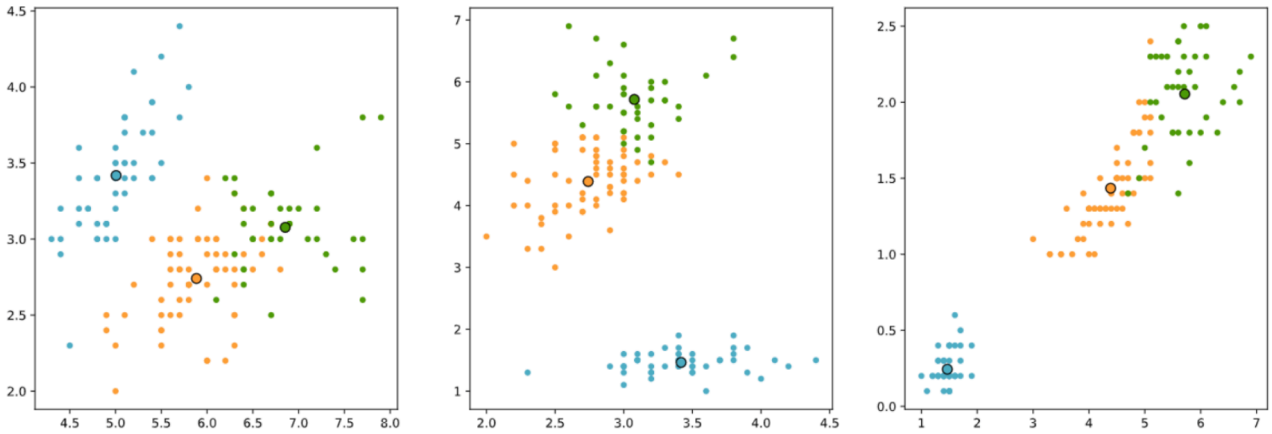

Рисунок 10 init = np.array([[5, 2, 1, 0], [5, 2, 3, 1], [6, 4, 5, 2]]), max_iter = 300

6. Методом локтя определено наилучшее количество кластеров:

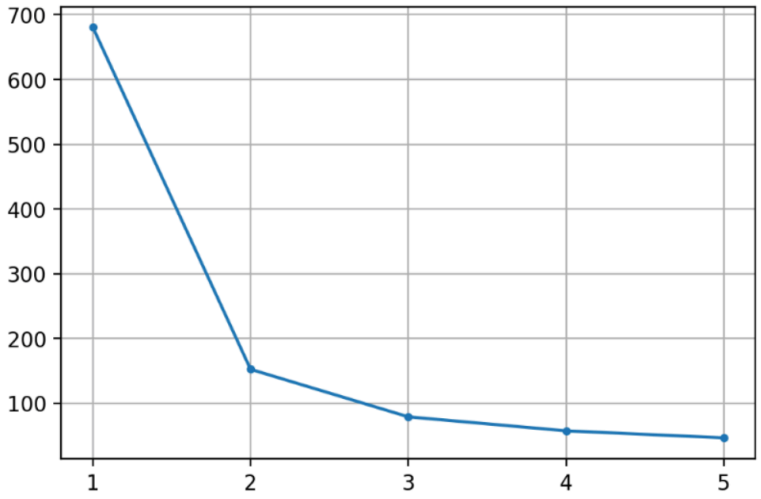

Рисунок 11 Метод локтя

Наилучшее количество кластеров – 2

7. Проведена кластеризация с использованием пакетной кластеризации k-средних. Построена диаграмма рассеяния, на которой выделены точки, которые для разных методов попали в разные кластеры. На вход MiniBatchKMeans подаются пакеты данных, а не полный набор – это увеличивает скорость работы, но снижает точность.

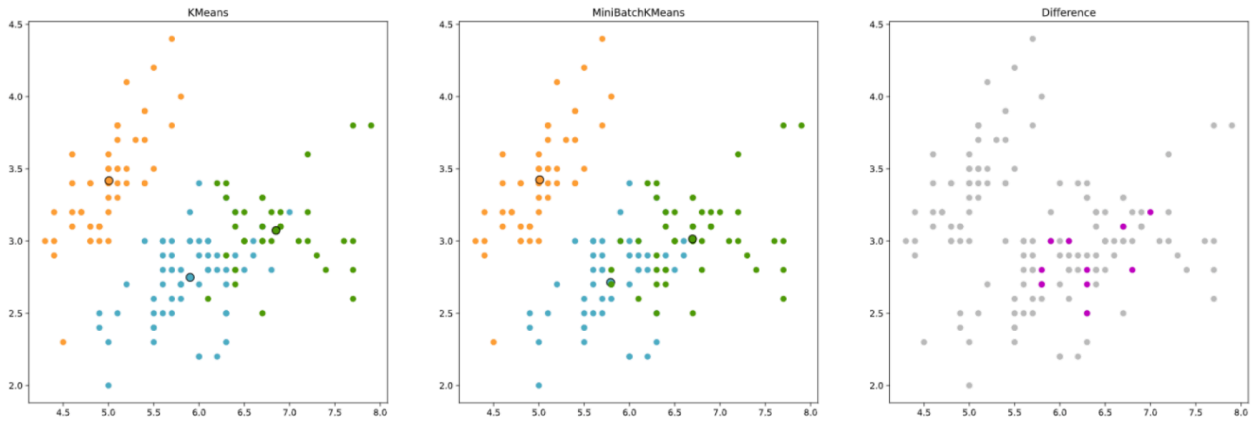

Рисунок 12 Отличие результатов KMeans и MiniBatchKMeans

Иерархическая кластеризация

1. На тех же данных была проведена иерархическая кластеризация:

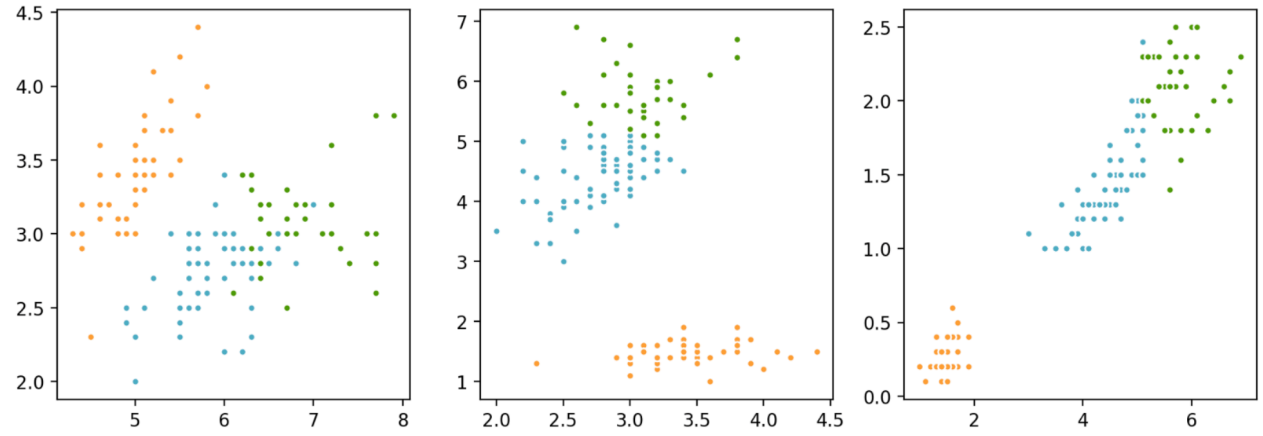

Рисунок 13 Результаты иерархической кластеризации

Отличие AgglomerativeClustering от Kmeans: изначально все точки принадлежат собственному кластеру, состоящему из одной точки. Алгоритм объединяет ближайшие кластеры на основе выбранной метрики.

2. Проведено исследование для различного количества кластеров:

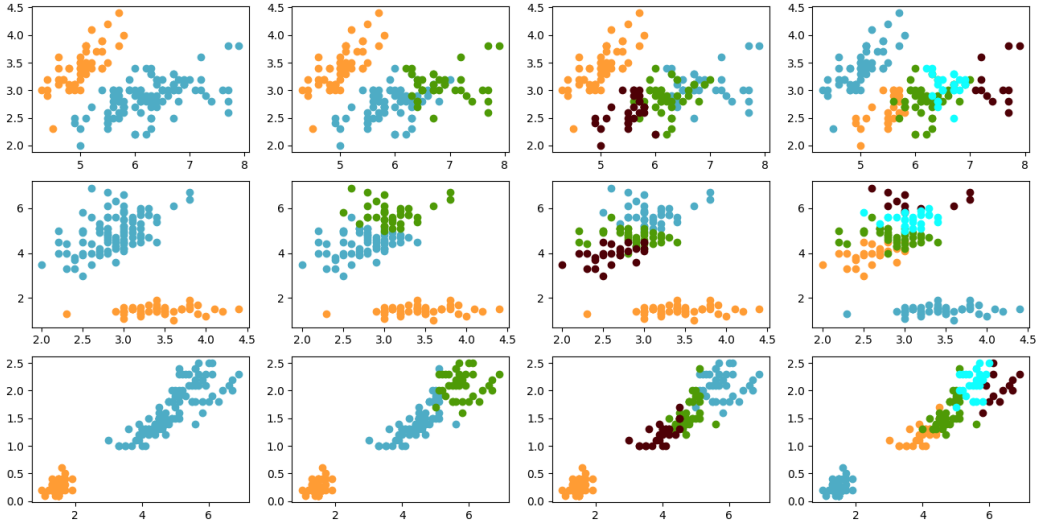

Рисунок 14 Результаты для различного количества кластеров

3. Нарисована дендограмма до уровня 6.

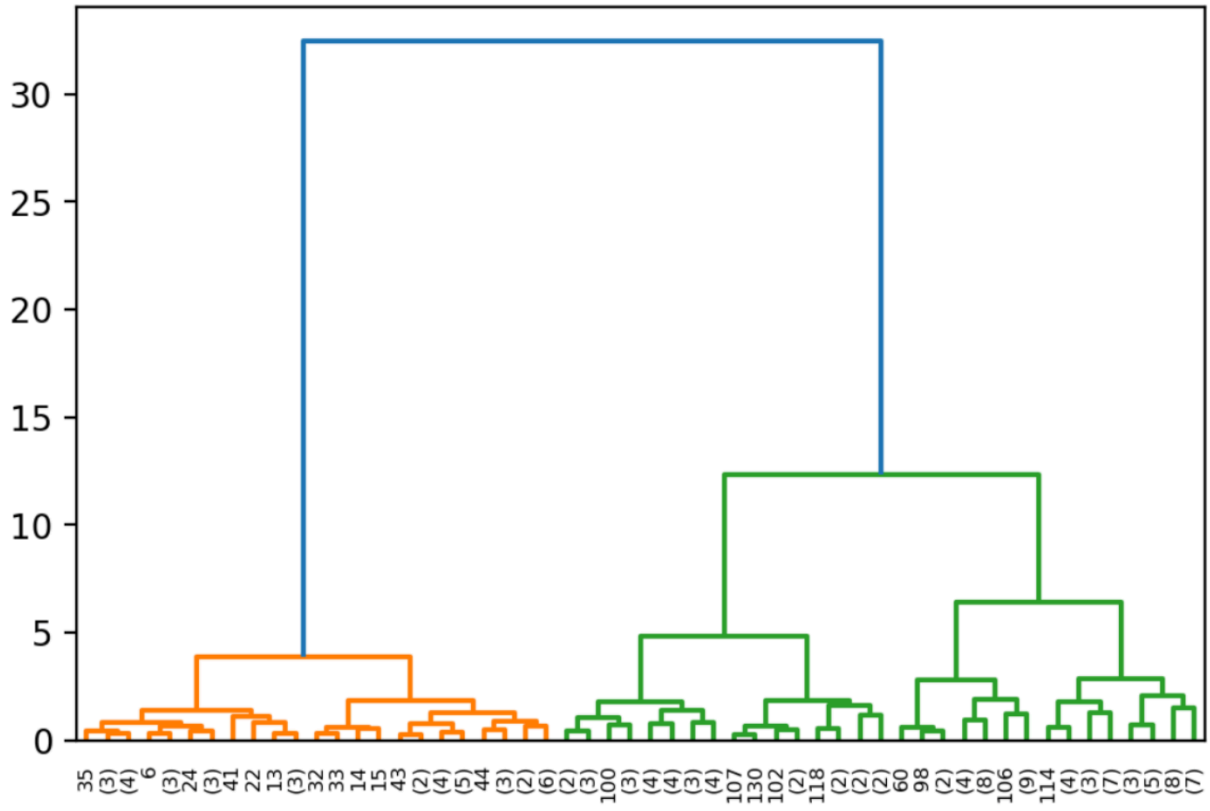

Рисунок 15 Дендограмма

4. Сгенерированы случайные данные в виде 2 колец:

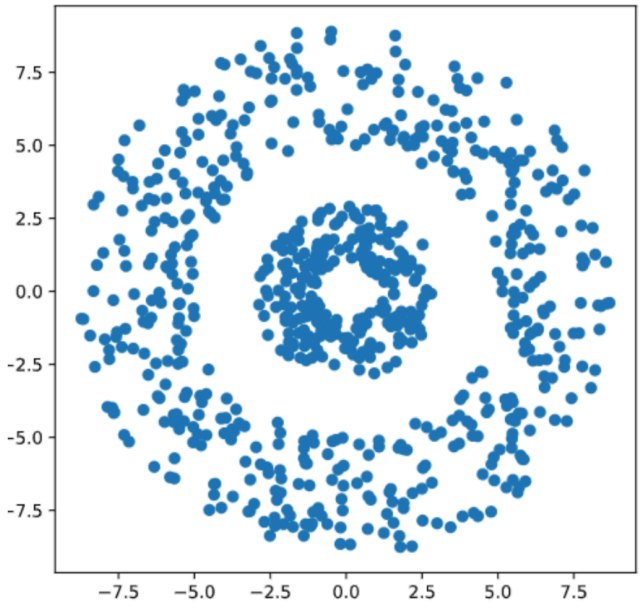

Рисунок 16 Сгенерированные данные

5. Проведена иерархическая кластеризация при использовании метрики Уорда

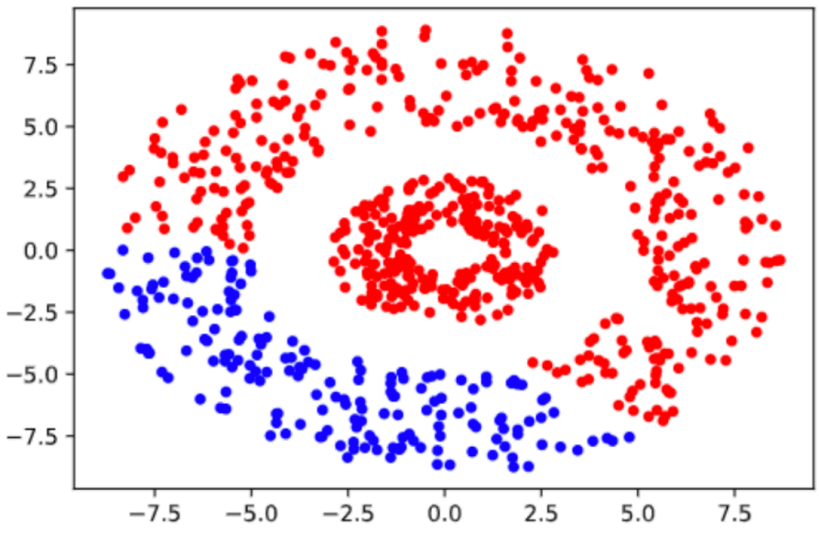

Рисунок 17 Результат иерархической кластеризации

6. Исследована кластеризация для различных параметров linkage:

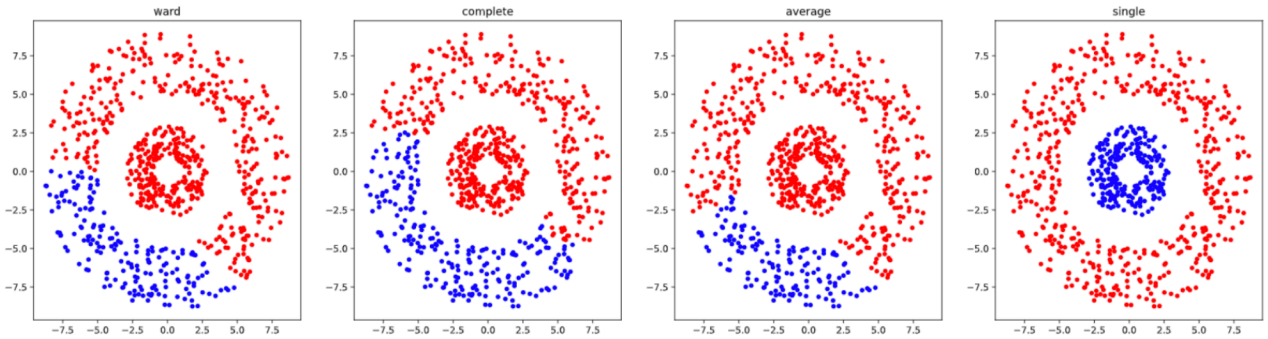

Рисунок 18 Результат иерархической кластеризации при различных параметрах linkage

- Ward – минимизация суммы квадратов разностей
- Complete – минимизация максимального расстояния
- Average – минимизация среднего расстояния
- Single – минимизация расстояния

По результатам видно, что разделение колец произошло только при использовании Single.

Выводы:

В результате выполнения лабораторной работы было проведено знакомство с методами кластеризации k-средних и иерархической кластеризации в модуле Sklearn. Использование пакетного метода k-средних приводит к небольшим изменениям результата в сравнении с полным k-средних. Метод иерархической кластеризации при правильной настройке смог определить нелинейную зависимость между синтетическими данными.

.