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

ОТЧЕТ

по лабораторной работе №5

по дисциплине «Машинное обучение»

Тема: «Кластеризация (k-средних, иерархическая)»

Студенты гр. 6307 _____ Ходос А.А.

Преподаватель _____ Жангиров Т.Р.

Санкт-Петербург
2020

Цель работы

Ознакомиться с методами кластеризации модуля Sklearn.

Ход работы

1. Загрузка данных

Был загружен датасет, представляющий собой информацию о трех классах цветов. Для дальнейшей работы с данными, в новый датасет были помещены значения первых четырех столбцов исходного датасета.

2. K-means

Была проведена кластеризация K-means, получены центры кластеров и определены, какие наблюдение в какой кластер попали. Результаты кластеризации представлены попарно для признаков на рисунке 1.

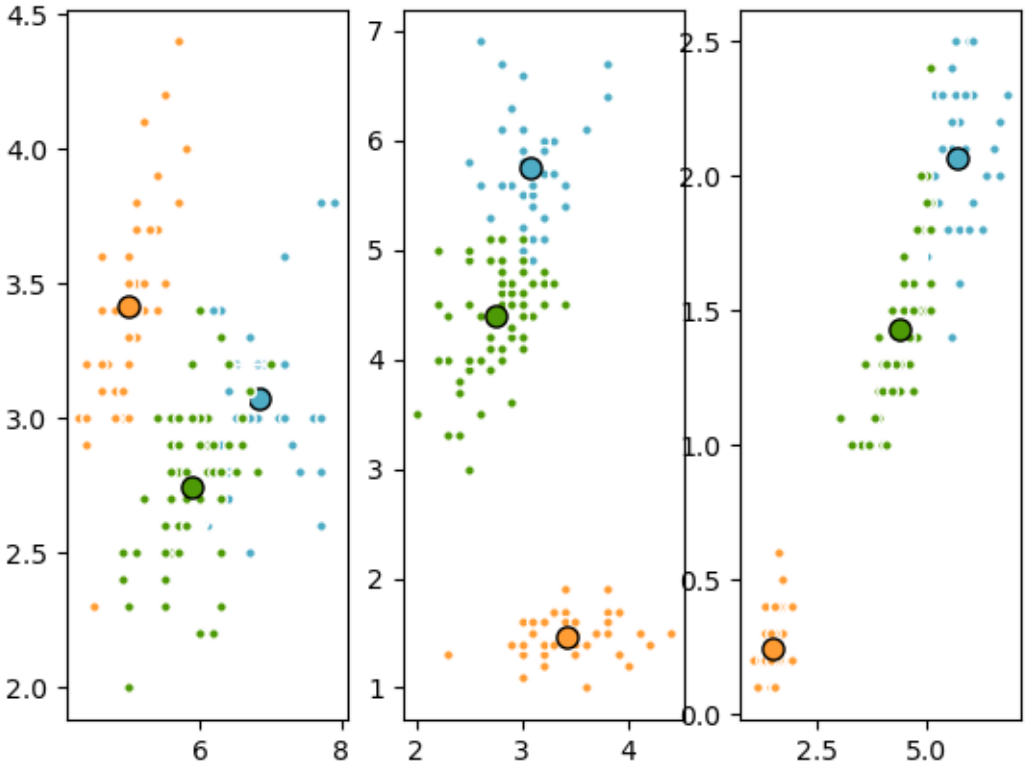

Рисунок 1 – Результаты кластеризации k-средних

Параметр n_init определяет сколько раз алгоритм выполнится с различными начальными значениями. Результат будет выбран исходя из

инерции – суммы квадратов расстояния от точек до центра ближайшего к ним кластера.

Была уменьшена размерность данных до 2 с использованием метода главных компонент. Карта области значений, на которой каждый кластер занимает определенную область со своим цветом представлена на рисунке 2.

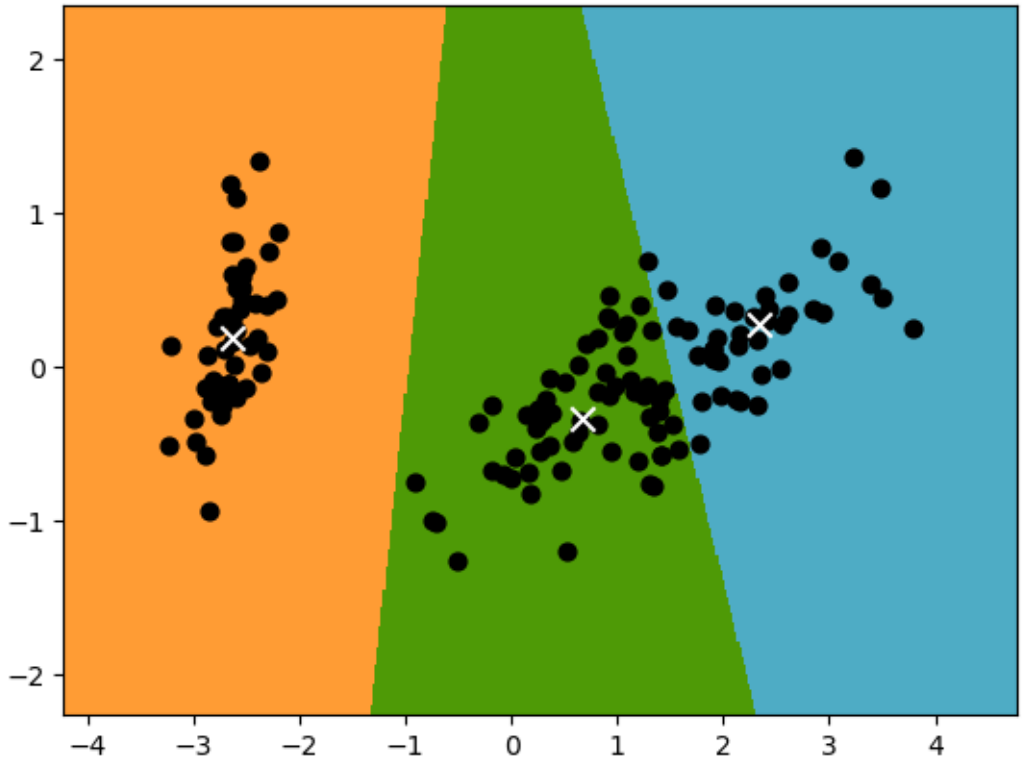

Рисунок 2 – Карта области значений кластеров

Метод локтя было определено наилучшее количество кластеров. По рисунку 3 видно, что наилучшее количество кластеров равно трём.

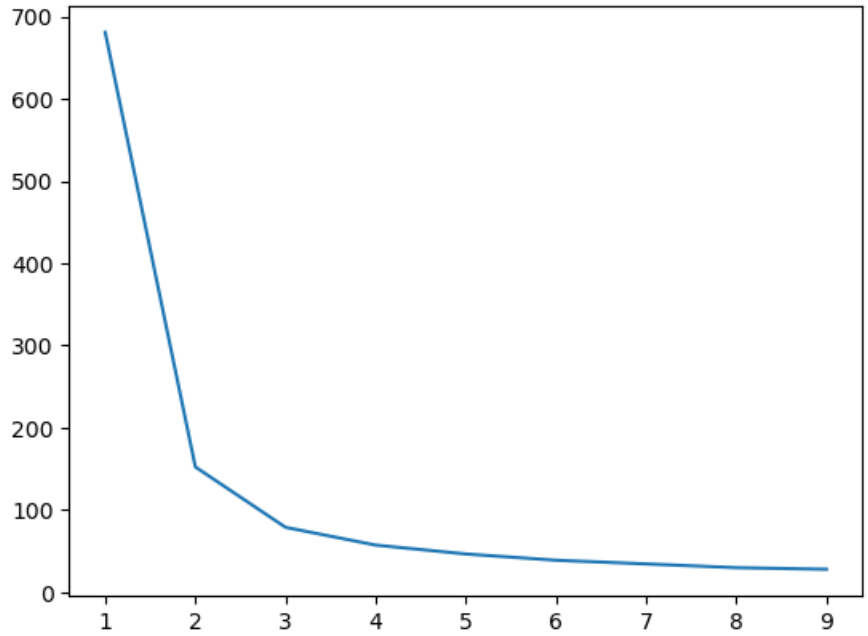

Рисунок 3 – Метод локтя

Проведена кластеризация с использованием пакетной кластеризации k-средних. Результаты представлена на рисунке 4, где выделены точки, попавшие в разные кластеры для разных методов. В отличии от k-средних, пакетная k-средних на каждой итерации использует случайные подмножества данных, а не целые наборы, за счет чего увеличивается скорость кластеризации, но снижается точность. Как видно, результаты стали лишь немного хуже.

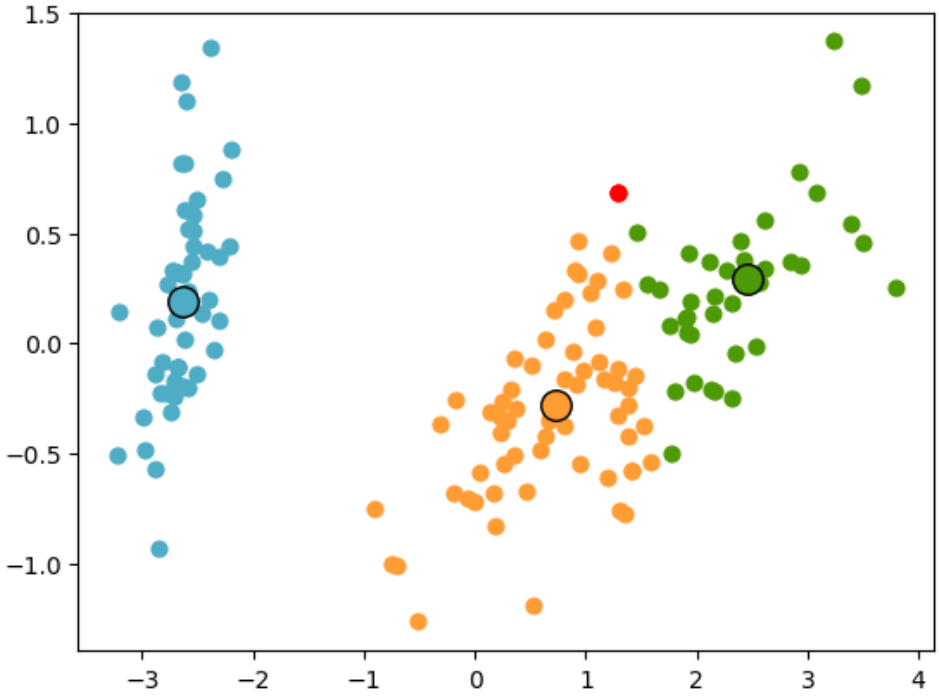

Рисунок 4 – Результаты кластеризации k-средних и пакетной k-средних

3. Иерархическая кластеризация

Для тех же данных была проведена иерархическая кластеризация. В отличии от k-средних, в иерархической кластеризации все точки изначально принадлежат кластерам размера 1, алгоритм объединяет кластеры на основе выбранной метрики. Было проведено исследование для различного количества кластеров (от 2 до 5). Результаты приведены на рисунке 5-8. Для уровня 6 была нариосва дендограмма, представленная на рисунке 9.

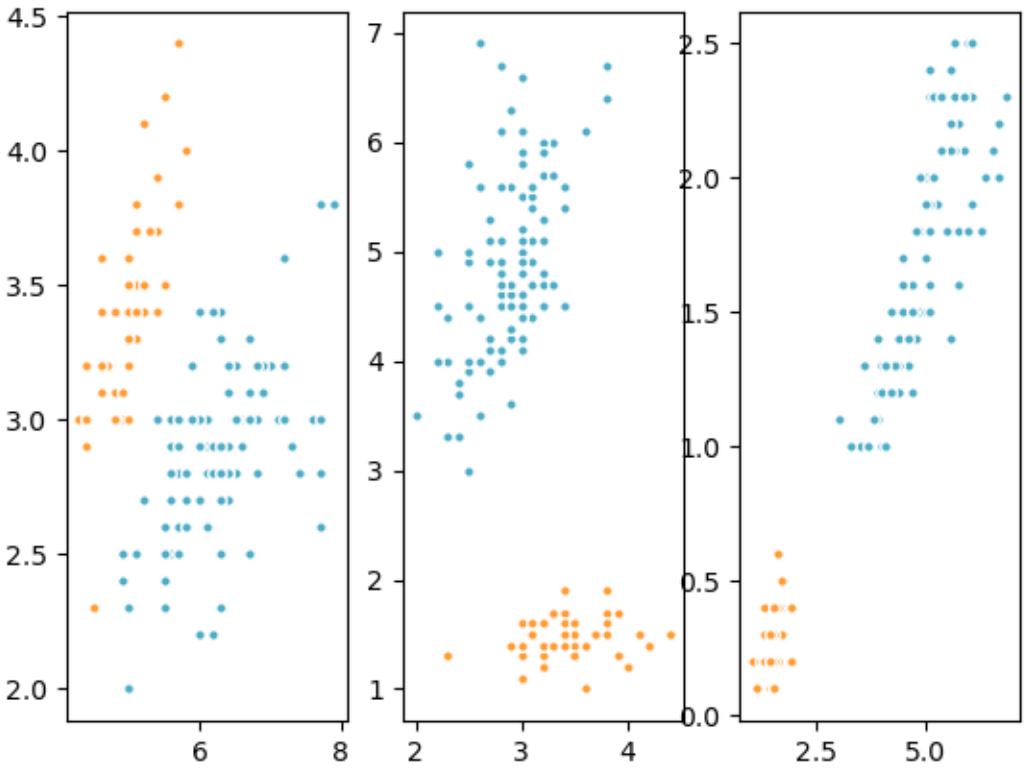

Рисунок 5 – Результаты иерархической кластеризации (количество кластеров 2)

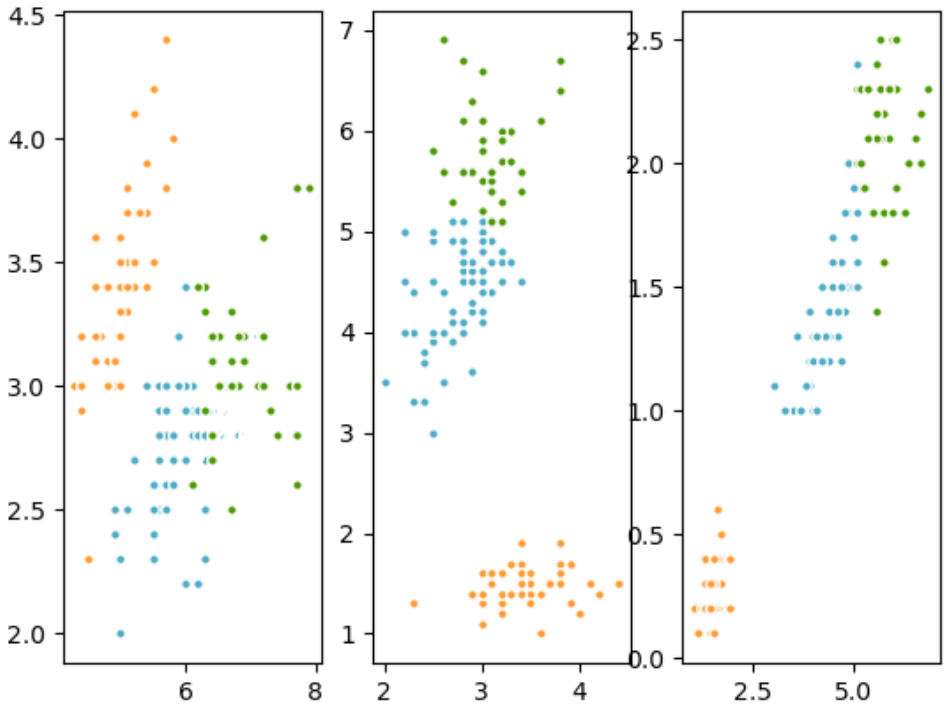

Рисунок 6 – Результаты иерархической кластеризации (количество кластеров 3)

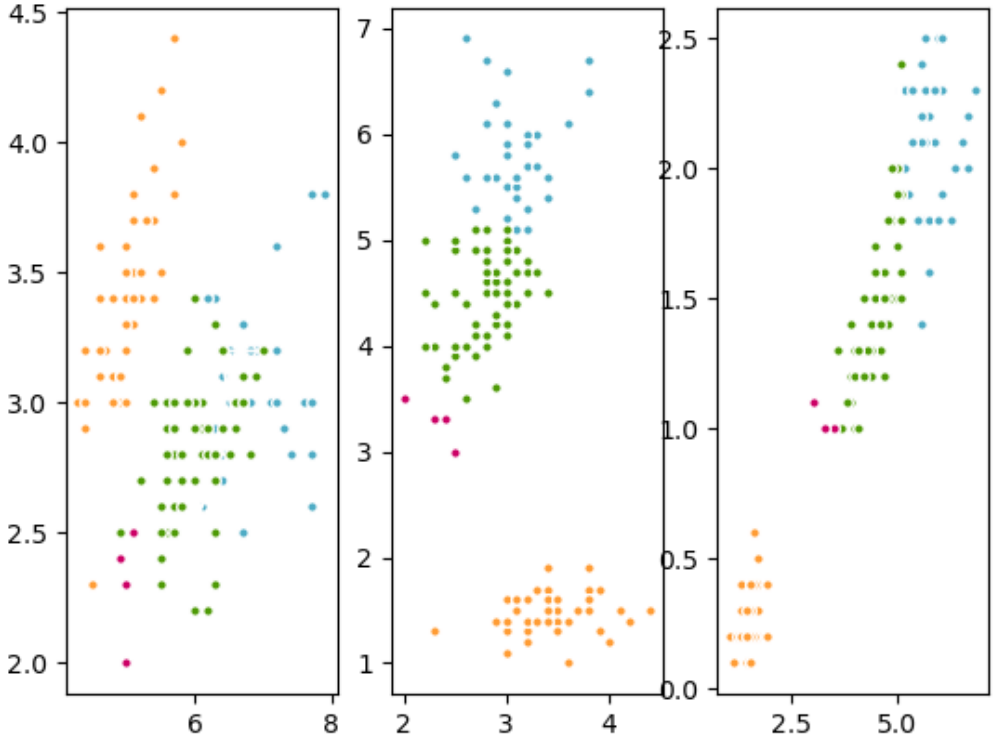

Рисунок 7 – Результаты иерархической кластеризации (количество кластеров 4)

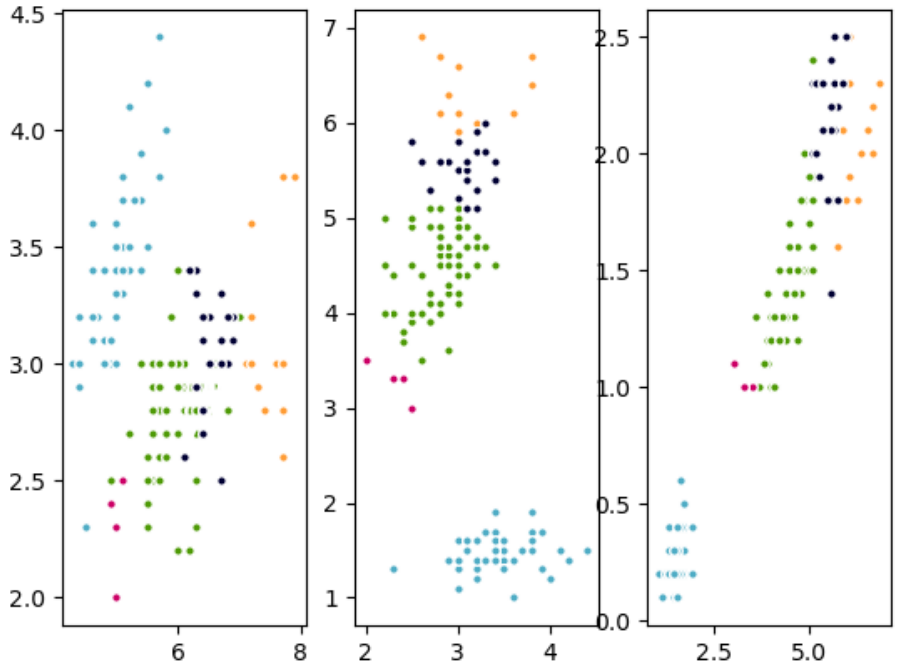

Рисунок 8 – Результаты иерархической кластеризации (количество кластеров 5)

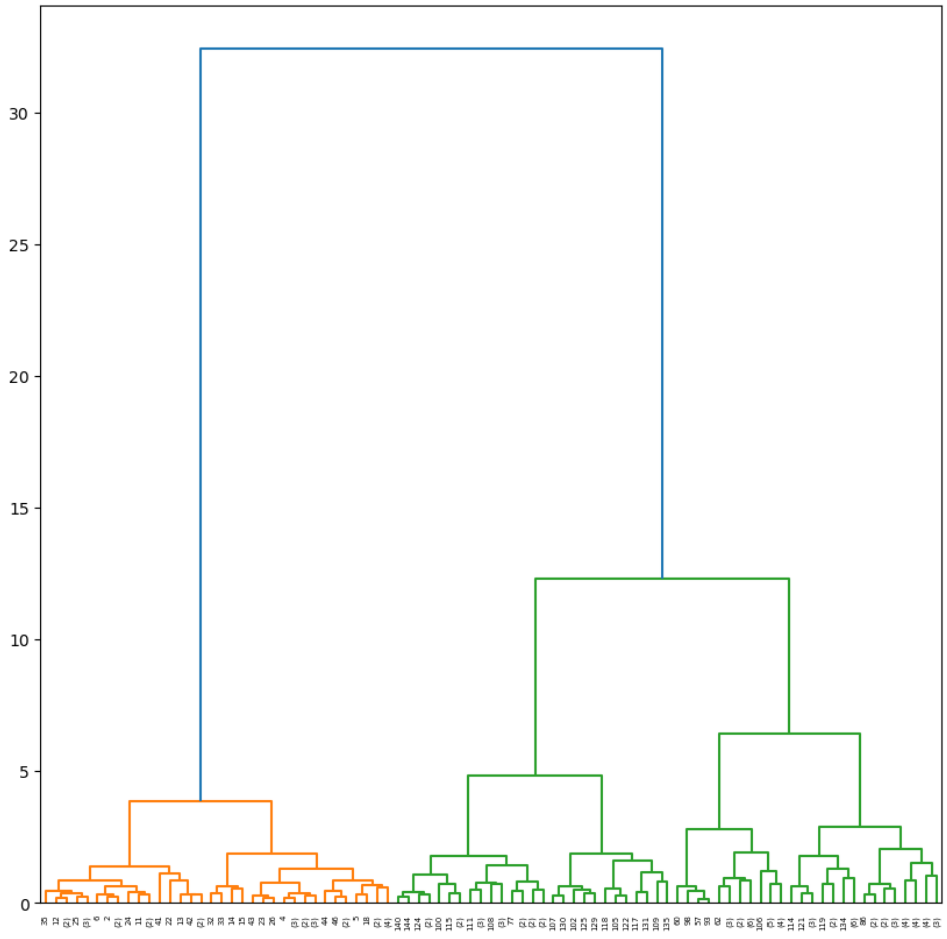

Рисунок 9 – Дендограмма для уровня 6

Были сгенерированы случайные данные в виде двух колец. Проведена иерархическая кластеризация с использованием различных linkage. Результаты представлены на рисунке 10-13.

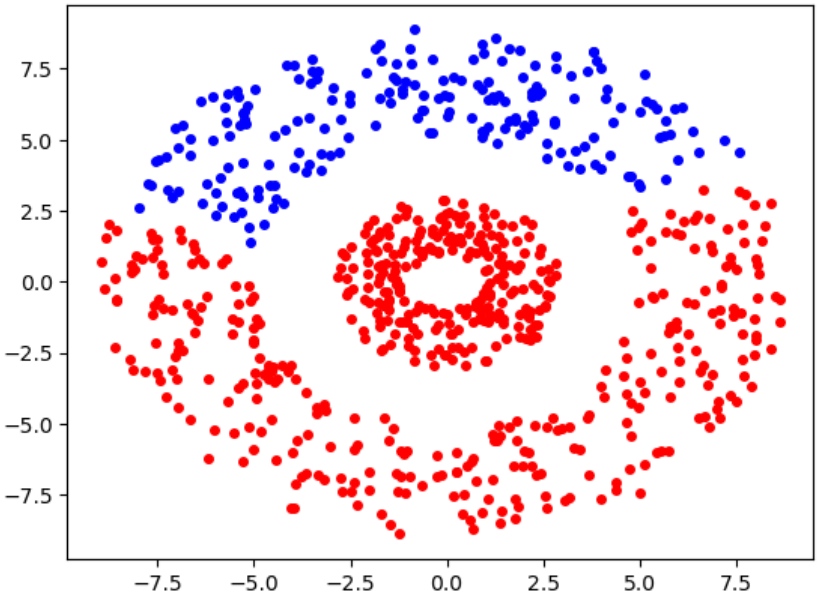

Рисунок 10 – Результаты иерархической кластеризации (linkage = ward)

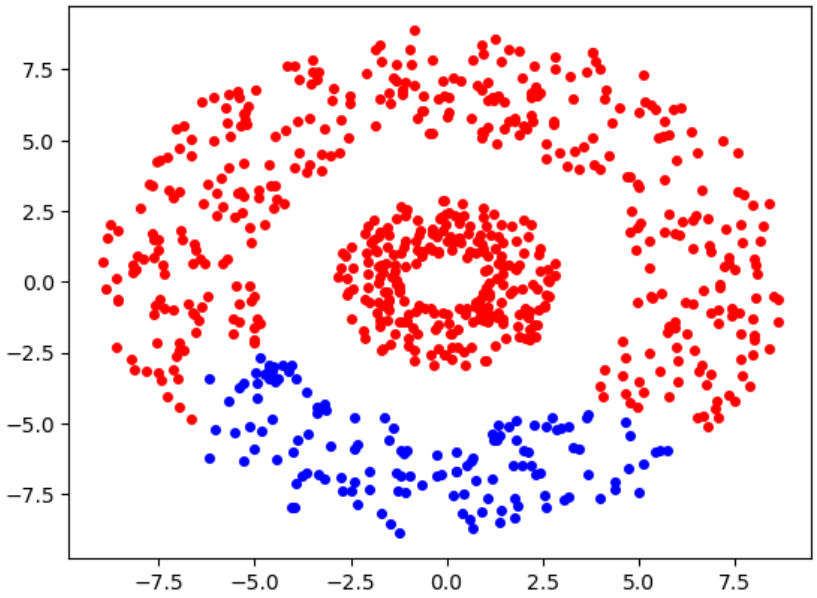

Рисунок 11 – Результаты иерархической кластеризации (linkage = complete)

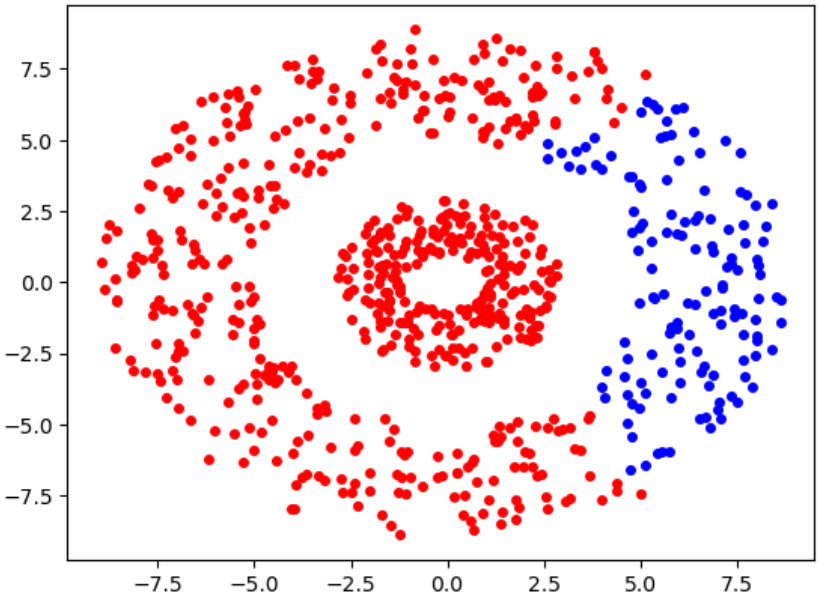

Рисунок 12 – Результаты иерархической кластеризации (linkage = average)

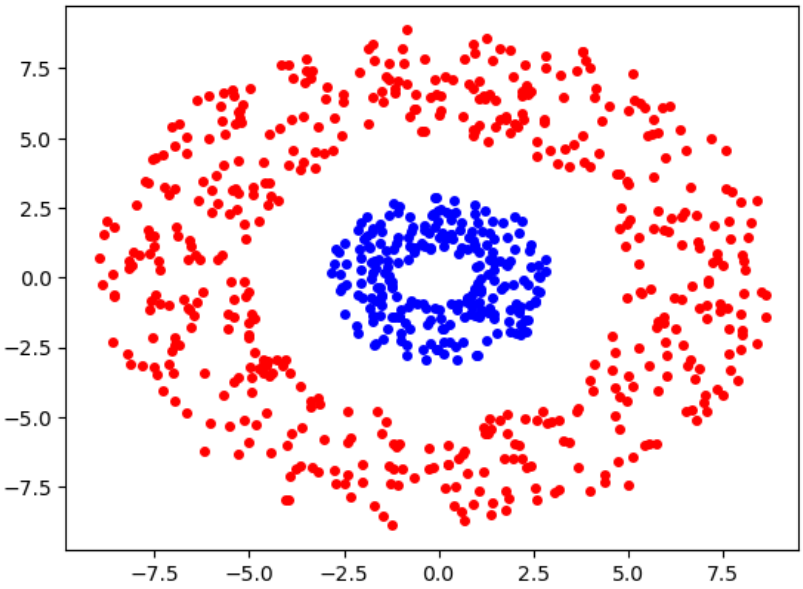

Рисунок 13 – Результаты иерархической кластеризации (linkage = single)

В зависимости от задачи, стоит применять определенный linkage. Например, если нашей задаче было разбить две окружности на кластеры, то нужно использовать linkage = single.

Вывод

Были получены навыки работы с методами кластеризации модуля Sklearn.