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

ОТЧЕТ

по лабораторной работе №5

по дисциплине «Машинное обучение»

Студенты гр. 6304 _____ Тимофеев А.А.

Преподаватель _____ Жангиров Т.Р.

Санкт-Петербург

2020

Цель работы

Ознакомиться с методами кластеризации модуля Sklearn **Ход работы**

Загрузка данных

1. Был создан датафрейм Pandas на основе загруженного датасета (https://archive.ics.uci.edu/ml/datasets/iris)

K-means

1. Была выполнена кластеризация методом K-means. Результат кластеризации представлена на рисунке 1.

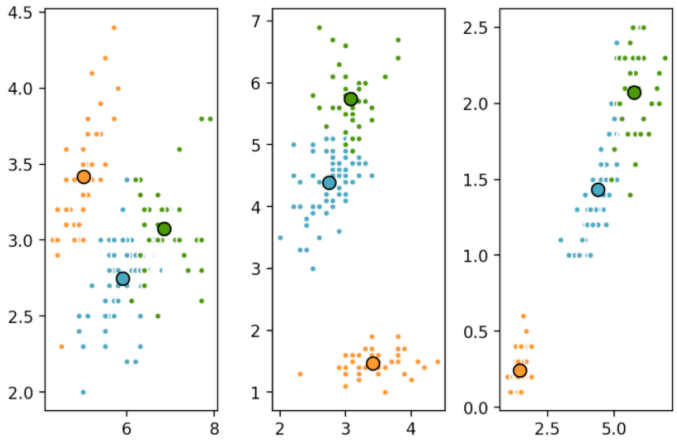

Рисунок 1 – Попарный результат кластеризации

Исходя из полученных графиков, наилучшее разбиение произошло для признаков 3 и 4 (крайний правый график). Параметр n_init отвечает за количество раз, когда алгоритм k-средних будет выполняться с разными начальными значениями центроидов.

2. Размерность выборки была уменьшена да 2 при помощи PCA. Была нарисована карта областей значений для каждого кластера после кластеризации получившихся данных. Карта представлена на рисунке 2.

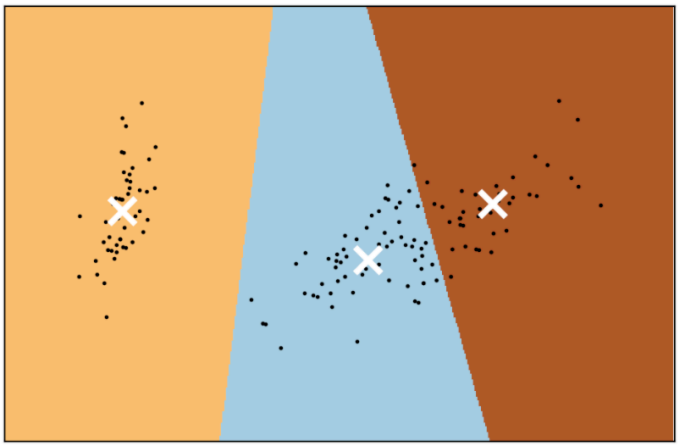

Рисунок 2 – Карта областей значений каждого кластера

3. Было выполнено 3 запуска кластеризации с значением *random* параметра *init*, а также 3 запуска для вручную заданных начальных центроидов. Результаты представлены на рисунках 3 и 4 соответственно.

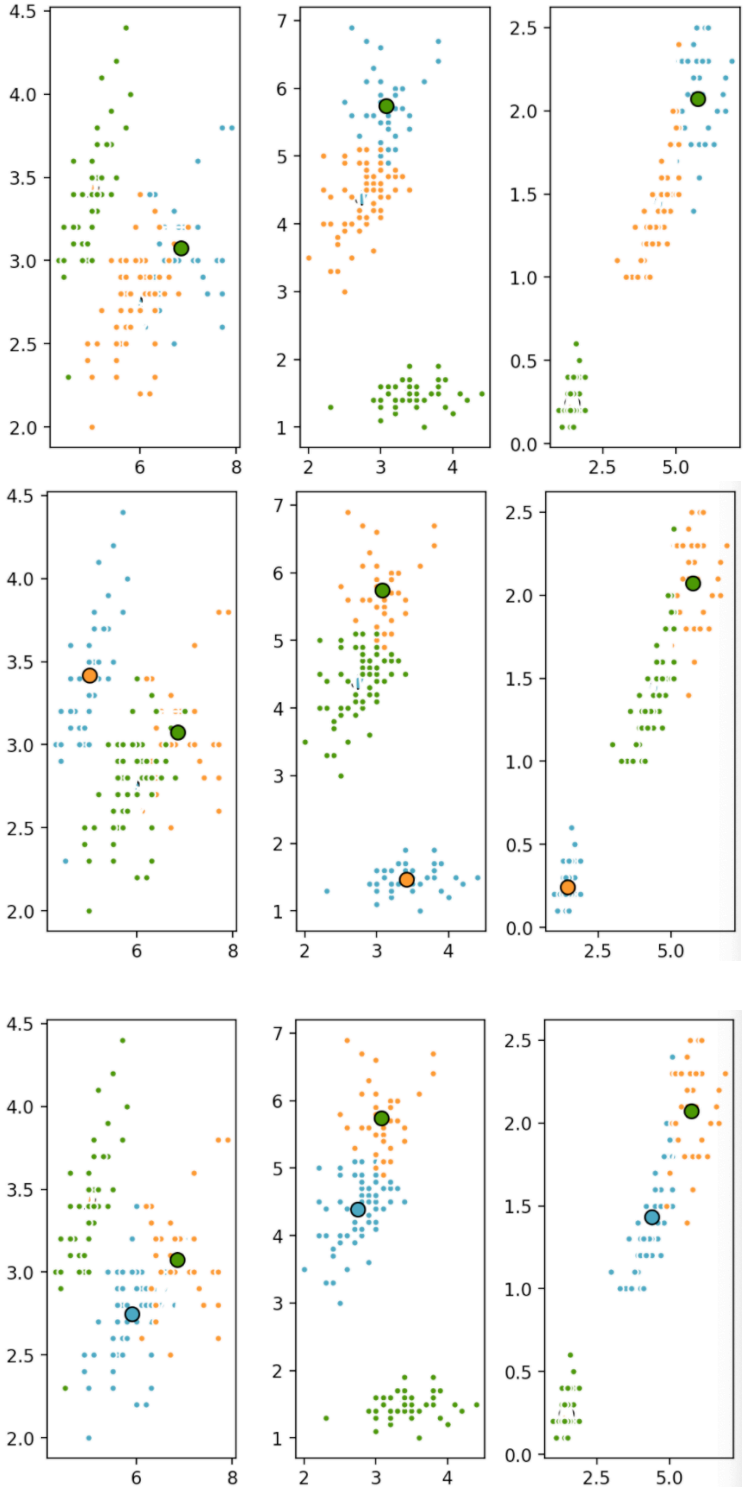

Рисунок 3 – Графики для случайного выбора данных при инициализации центроидов

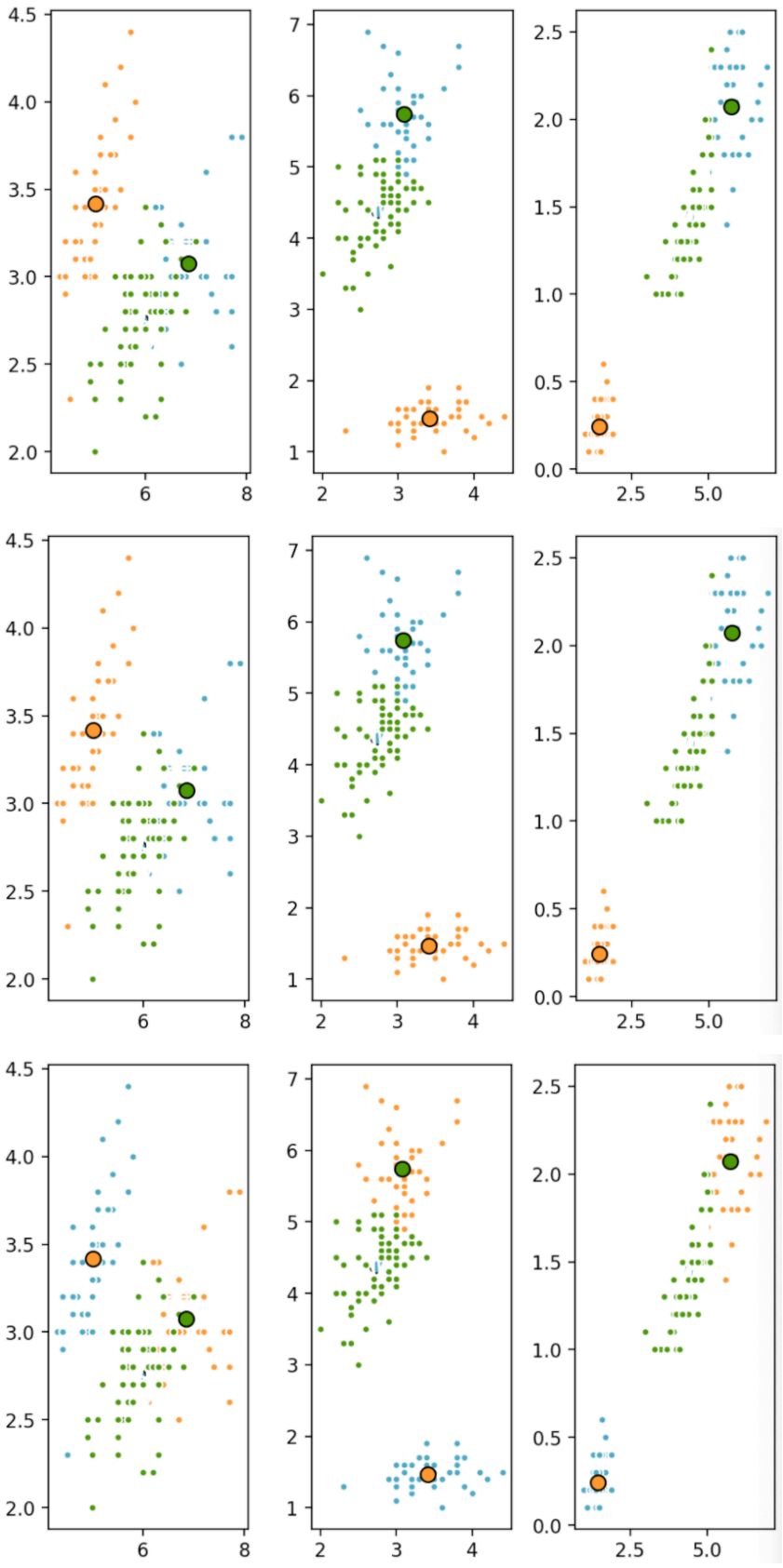

Рисунок 4 – Графики для вручную инициализированных центроидов

Результаты кластеризации идентичны во всех случаях. Возможно, это следует из небольшого размера выборки.

4. Было определено наилучшее количество кластеров при помощи метода локтя. График зависимости суммы квадратов расстояний точек то центроидов в соответствующих им кластерам от количества кластеров представлен на рисунке 5.

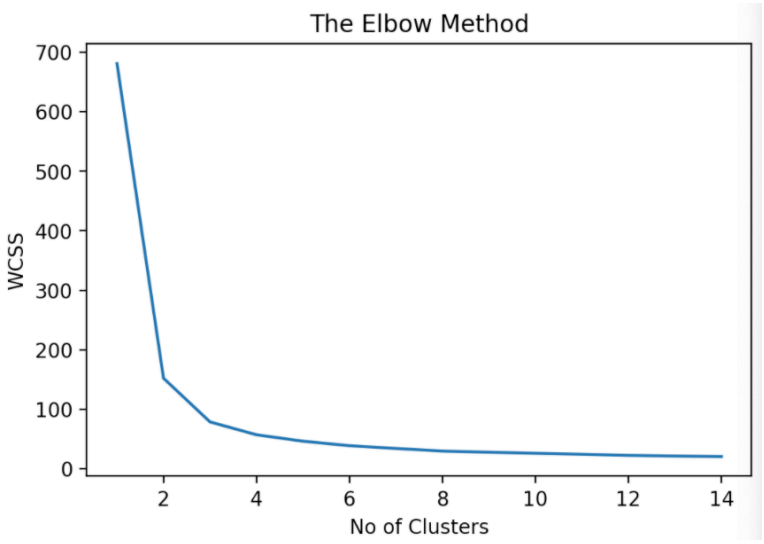

Рисунок 5 – График зависимости WCSS от числа кластеров

Из графика можно сделать вывод, что наилучшее количество кластеров – 3.

5. Была выполнена пакетная кластеризация k-средних. Отличие от обычного алгоритма заключается в том, что при кластеризации используются не все данные из выборки, а случайно выбранные из нее экземпляры. На рисунке 6 показаны отличия в результатах работы 2 версий алгоритма. Красные точки – различающиеся.

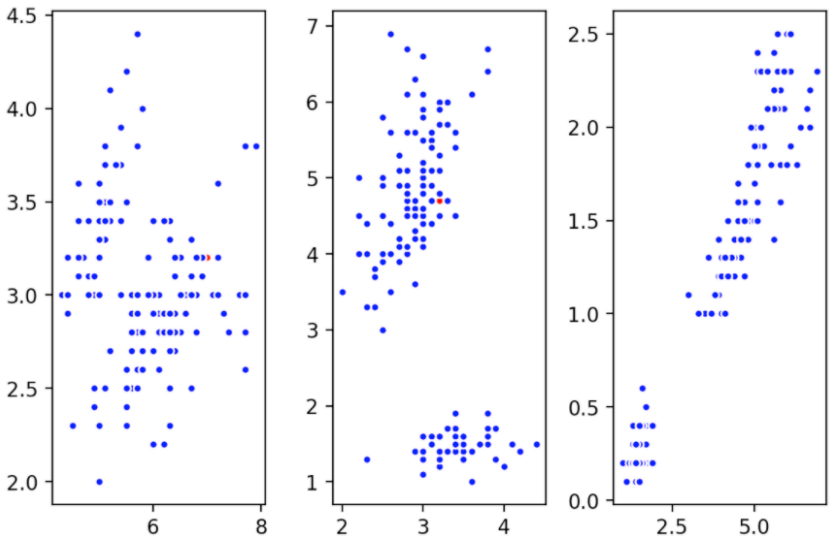

Рисунок 6 – Различия в результатах кластеризации K-means и пакетными K-means

Иерархическая кластеризация

1. Была проведена иерархическая кластеризация на тех же данных. Результаты представлены на рисунке 7.

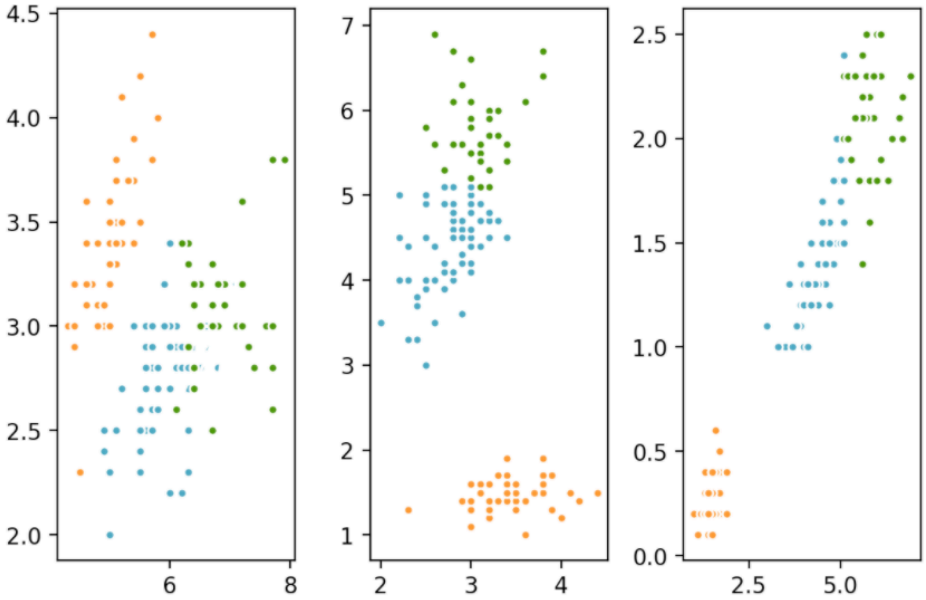

Рисунок 7 – Результаты иерархической кластеризации

В результатах кластеризации различия не наблюдаются, однако различия существуют в самих алгоритмах. В начале K-means все точки распределяются по кластерам с центроидами, проинициализированными одним из нескольких способов. Далее центроиды пересчитываются, а точки перераспределяются между кластерами. В AgglomerativeClustering каждая точка изначально принадлежит одному кластеру. На каждом шаге алгоритма два ближайших кластера объединяются в один, пока количество кластеров не станет равным заданному.

2. Были проведены кластеризации для различного размера кластеров. Результаты приведены на рисунке 8.

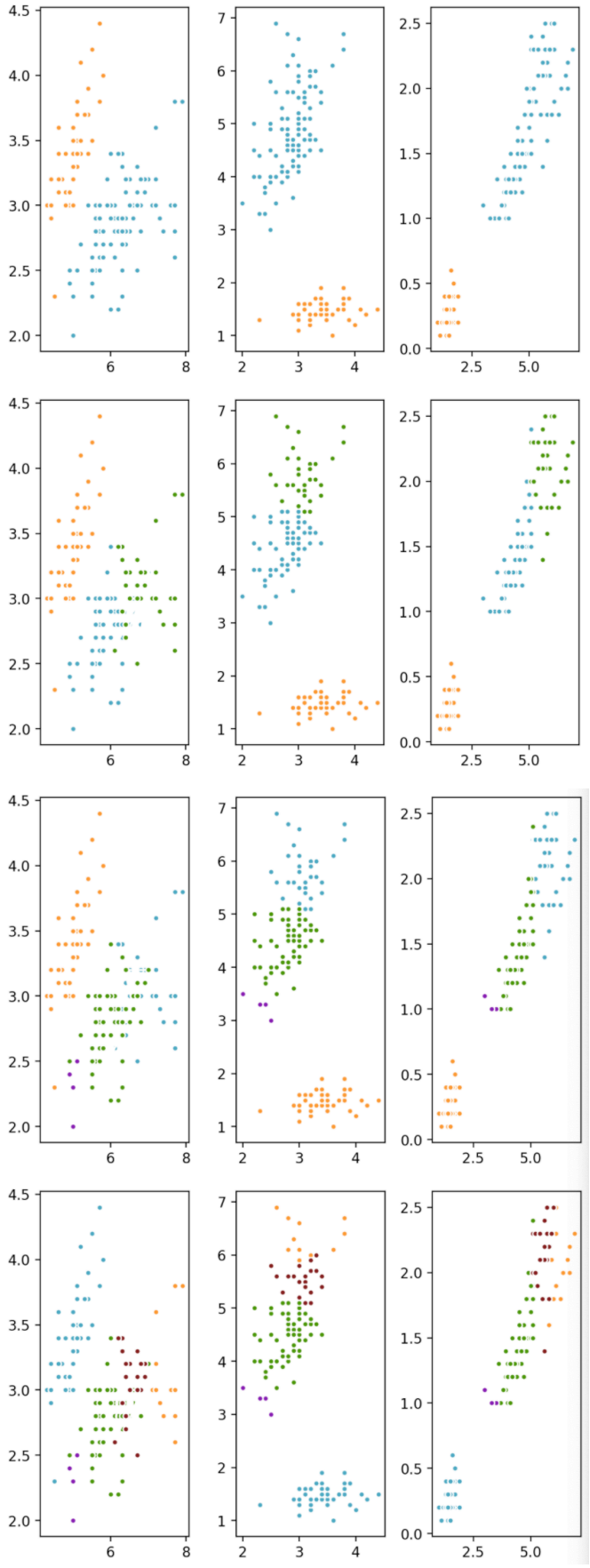

Рисунок 8 – Результаты кластеризации для различного числа кластеров

3. Была нарисована дендограмма до уровня 6. Дендограмма представлена на рисунке 9.

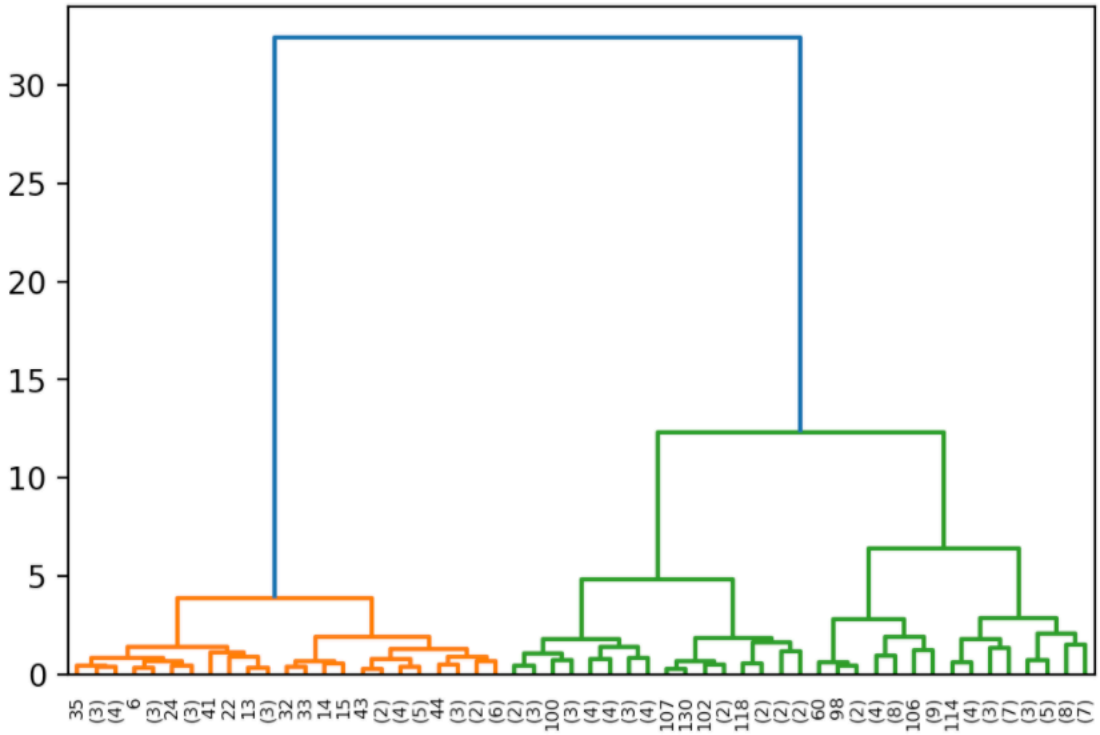

Рисунок 9 – Полученная дендограмма

4. Было проведено исследование кластеризации при различных параметрах *linkage* на данных в виде двух колец. Результаты представлены на рисунке 10.

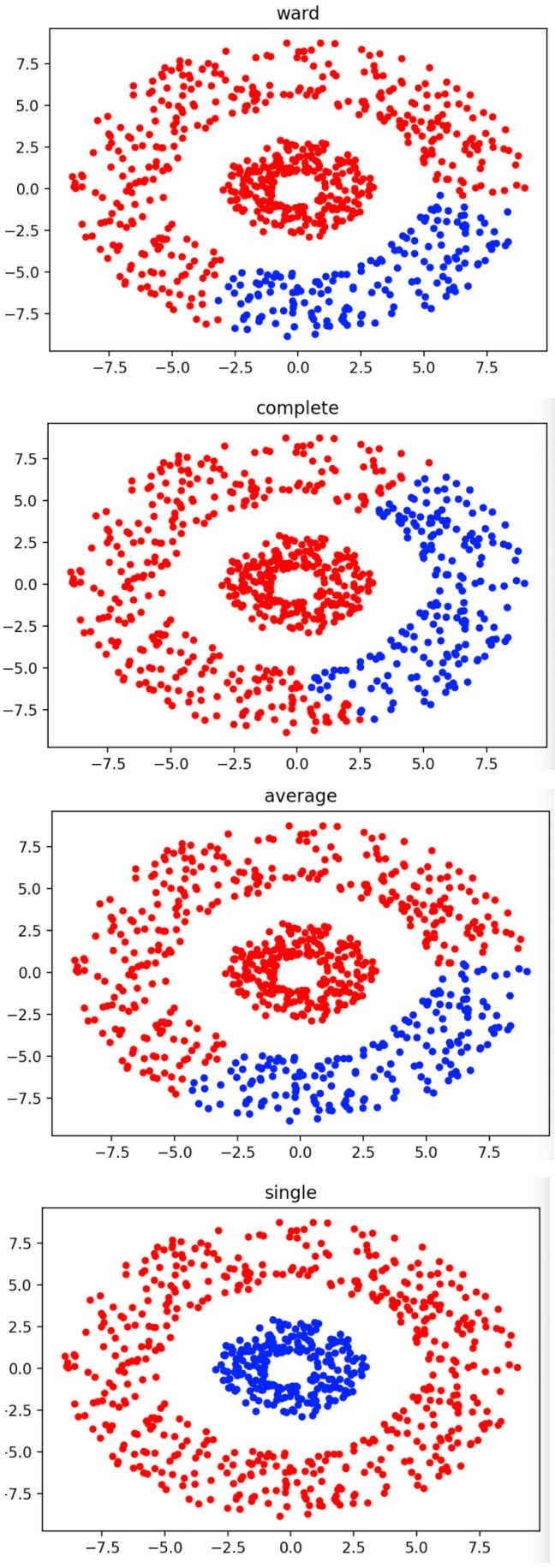

Рисунок 10 – Результаты кластеризации

Данный параметр отвечает за то, каким методом будет рассчитано расстояние между кластерами.

- Single – метод одиночной связи
- Average – метод средней связи
- Complete – метод полной связи
- Ward – метод Уорда (минимизации дисперсии)

Выводы

В ходе выполнения данной лабораторной работы было произведено знакомство с методами кластеризации модуля Sklearn. Кластеризация производилась с помощью методов K-means, пакетный K-means и AgglomerativeClustering.