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

ОТЧЕТ

по лабораторной работе №5

по дисциплине «Машинное обучение»

Тема: Кластеризация (k-средних, иерархическая)

Студент гр. 6304 _____ Виноградов К.А.

Преподаватель _____ Жангиров Т.Р.

Санкт-Петербург

2020

Загрузка данных.

Загрузим данные из csv таблицы и проанализируем их с помощью метода k-средних. Результат представлен на рис.1.

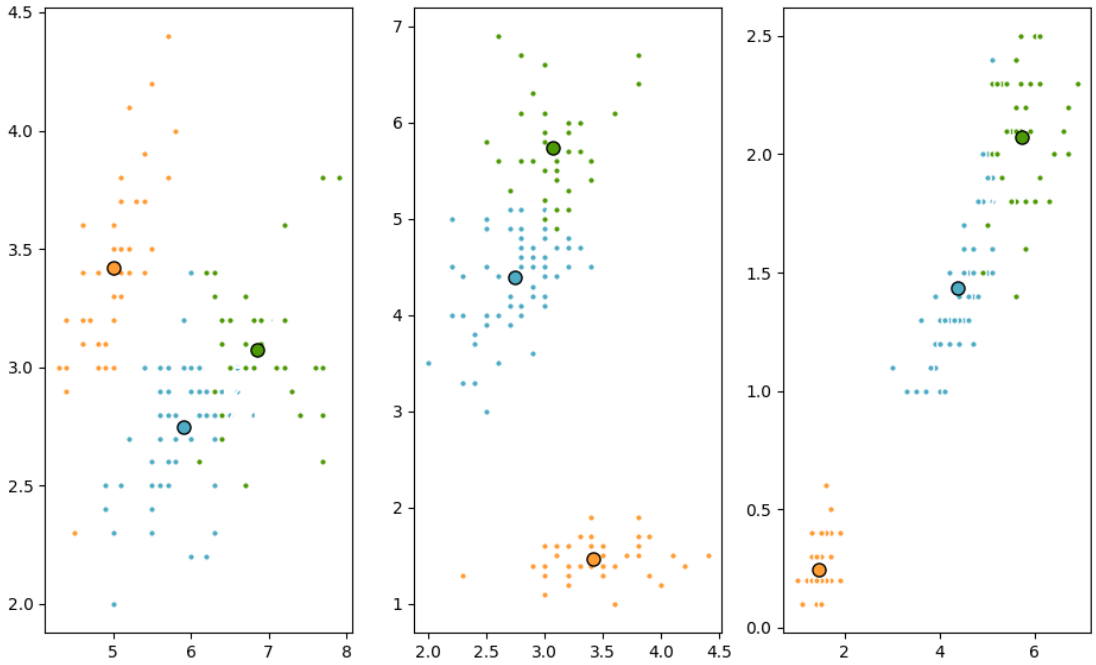

Рисунок 1 – Обработка данных методом k-средних

Разделение по кластерам во всех случая произвелось с примерно одинаковой эффективностью. Параметр n_init влияет на количество прогонов алгоритма.

Произведем понижение размерности пространства и кластериуем плоскость. Результаты на рис. 2.

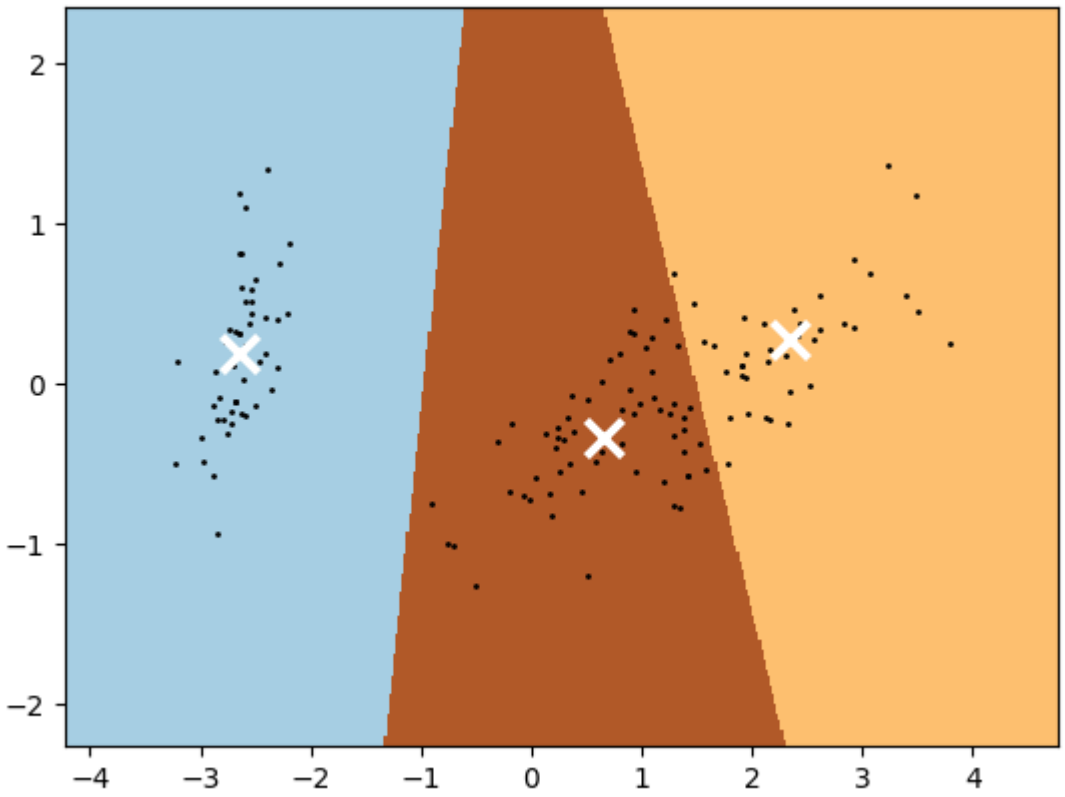

Рисунок 2 – Полученные кластеры

Изучим поведение алгоритма при выборе начальных центров случайно и вручную. Результаты на рис. 3 – 6.

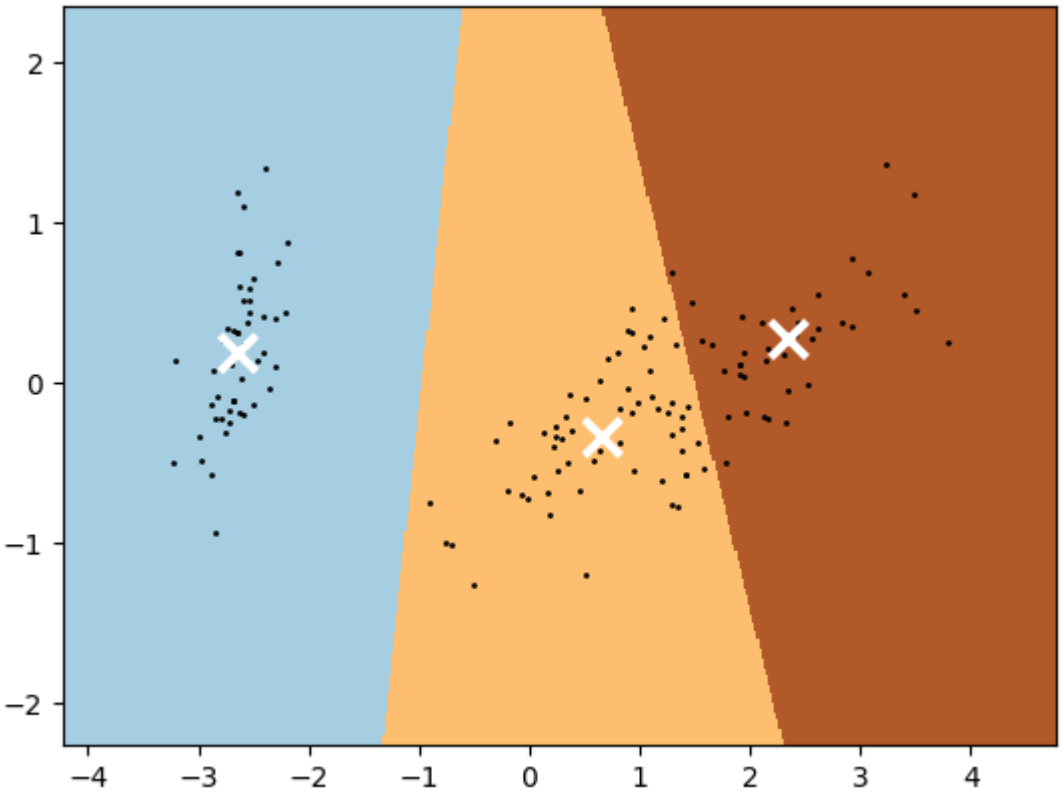

Рисунок 3 – Первый прогон со случайным выбором

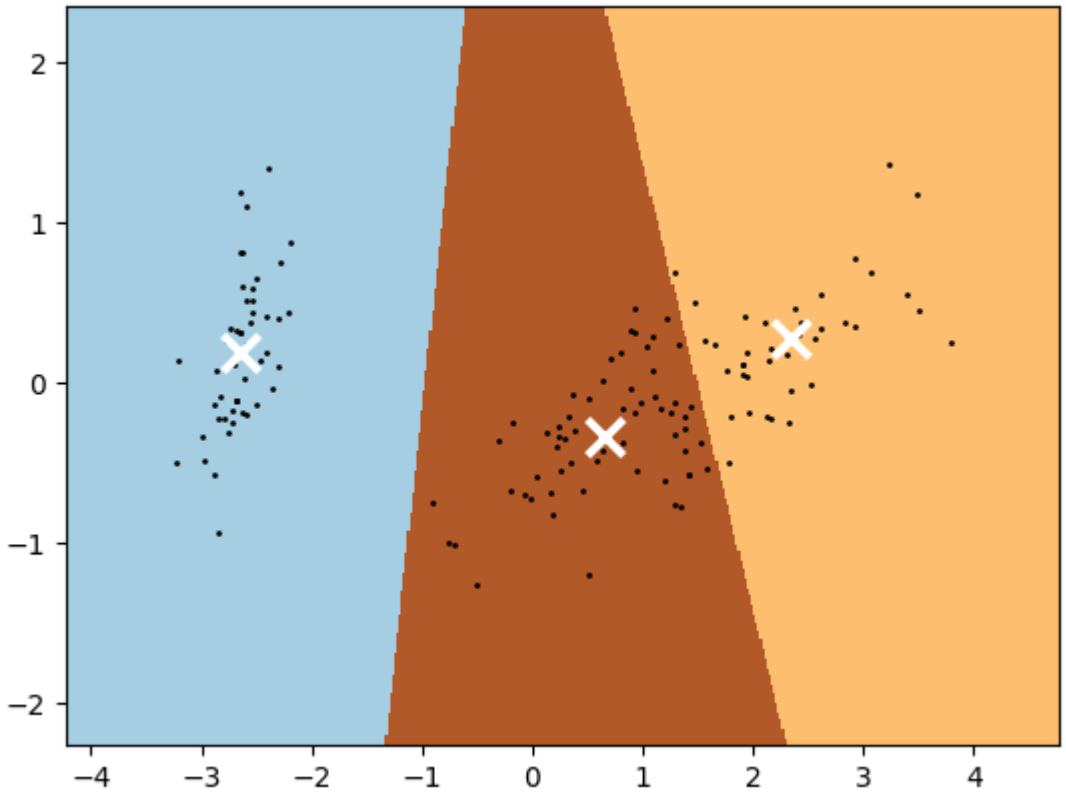

Рисунок 4 – Второй прогон со случайным выбором

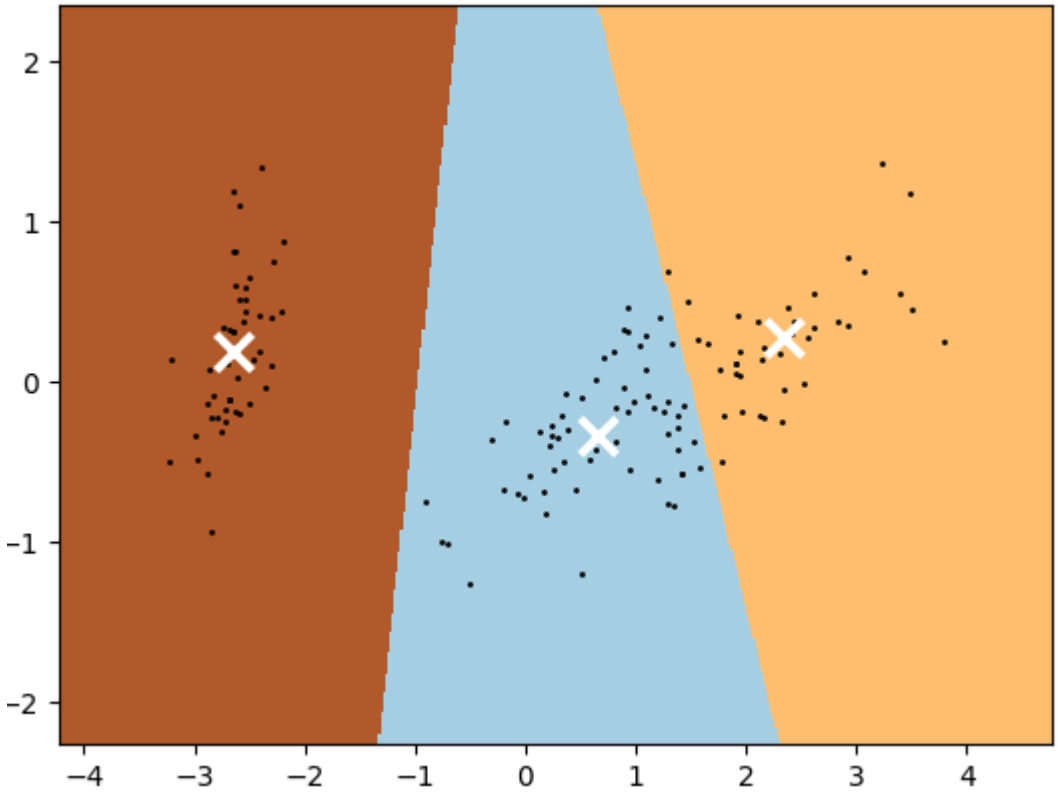

Рисунок 5 – Третий прогон со случайным выбором

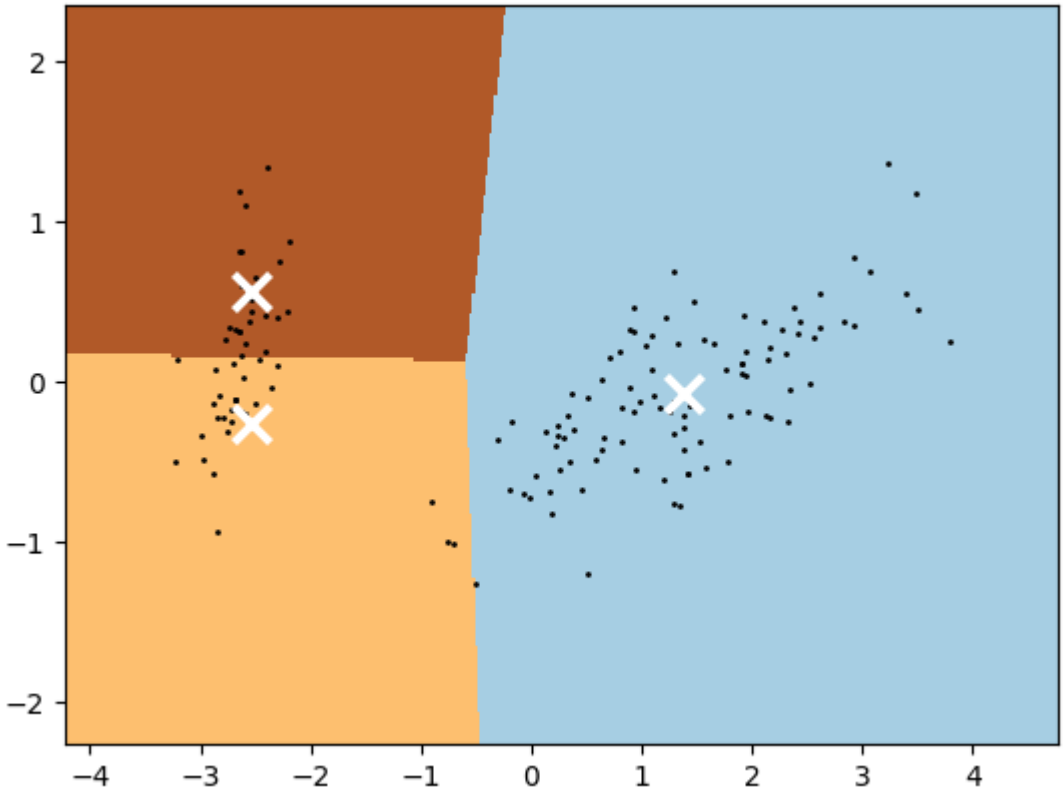

Рисунок 6 – Выбор центров вручную

Нетрудно заметить что из-за нескольких прогонов со случайными центрами результат кластеризации серьезно улучшается.

Найдем наилучшее количество кластеров методом локтя. Результат на рис. 7.

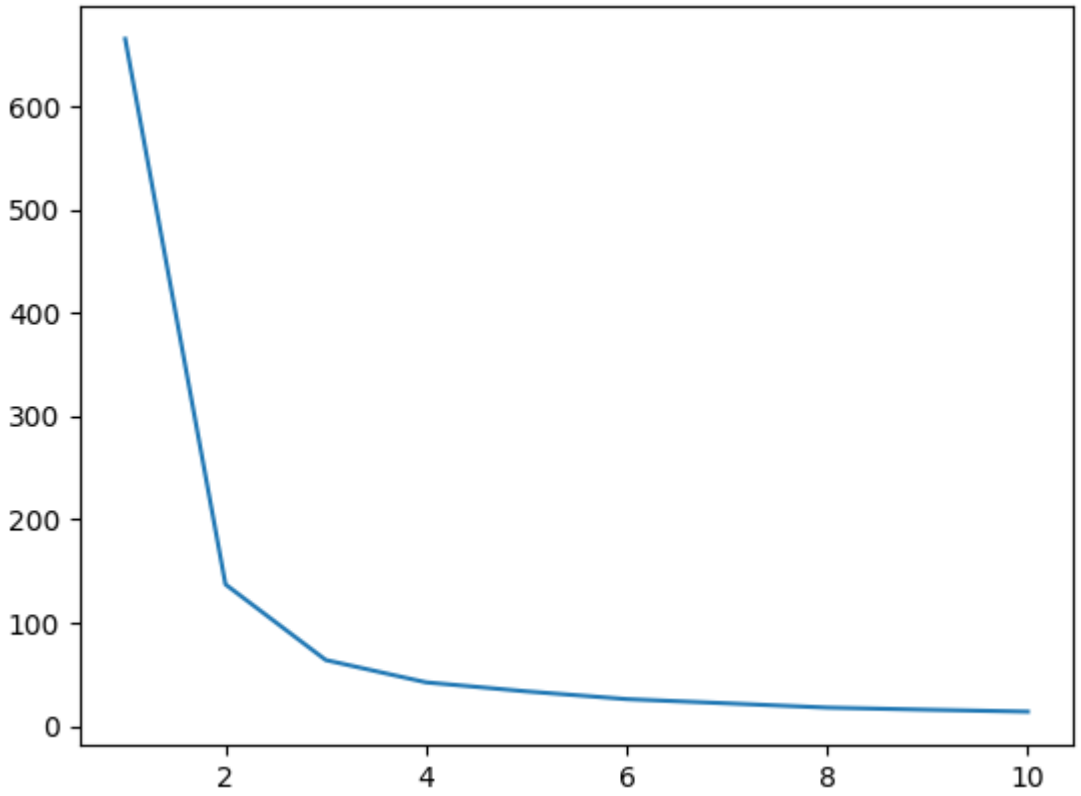

Рисунок 7 – Метод локтя

Можно заметить что после 3 кластеров результат практически не улучшается, следовательно подходящее число кластеров 3.

Попробуем распределить на элементы на кластеры с помощью пакетного k-среднего. Данный метод осуществляет разделение путем выбора только части значений и вычисления с помощью них, что повышает быстродействие но при этом точность не слишком страдает. Результат сравнения обычного и пакетного алгоритмов на рис. 8

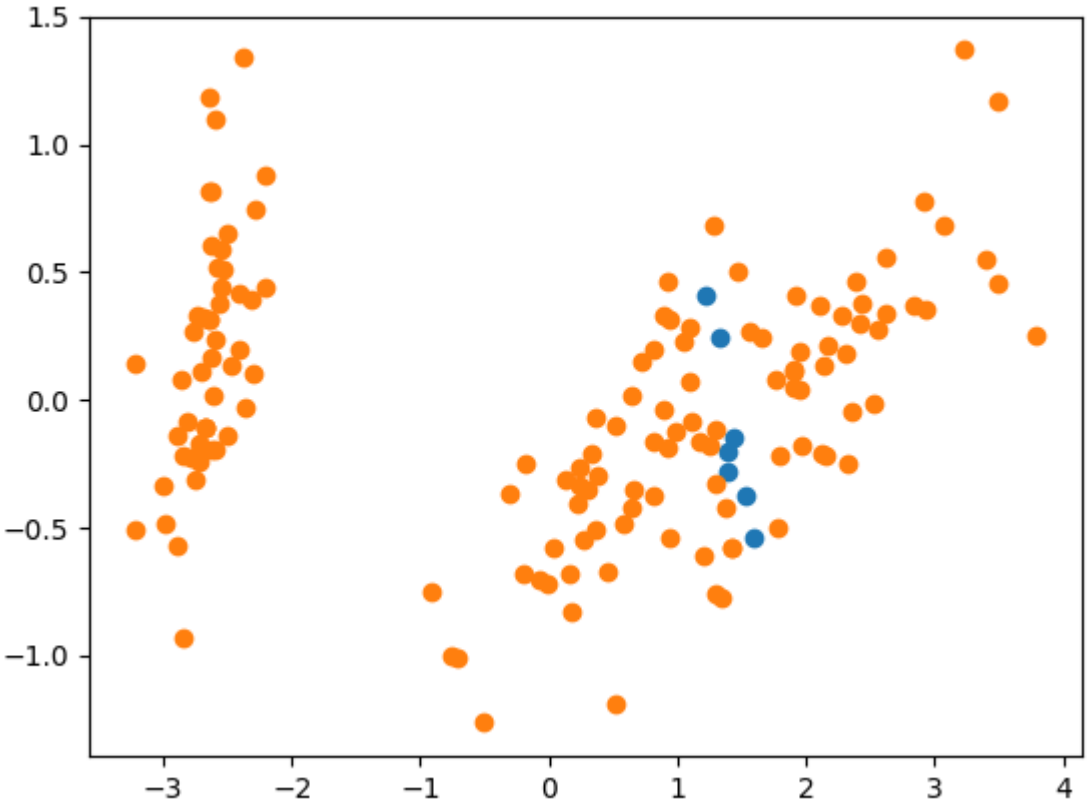

Рисунок 8 – Сравнение MiniBatch K-means и обычного K-means

Можно отметить что точки попавшие в различные кластеры при применении различных алгоритмов находятся преимущественно на стыке кластеров, однако так как точек достаточно мало результат изменяется не сильно.

Иерархический кластеринг.

В отличие от k-средних иерархический кластеринг включает в новые кластеры точки поочередно с помощью минимизации определенной характеристики и без пересчета центроид. Результаты представлены на рис. 9 – 12.

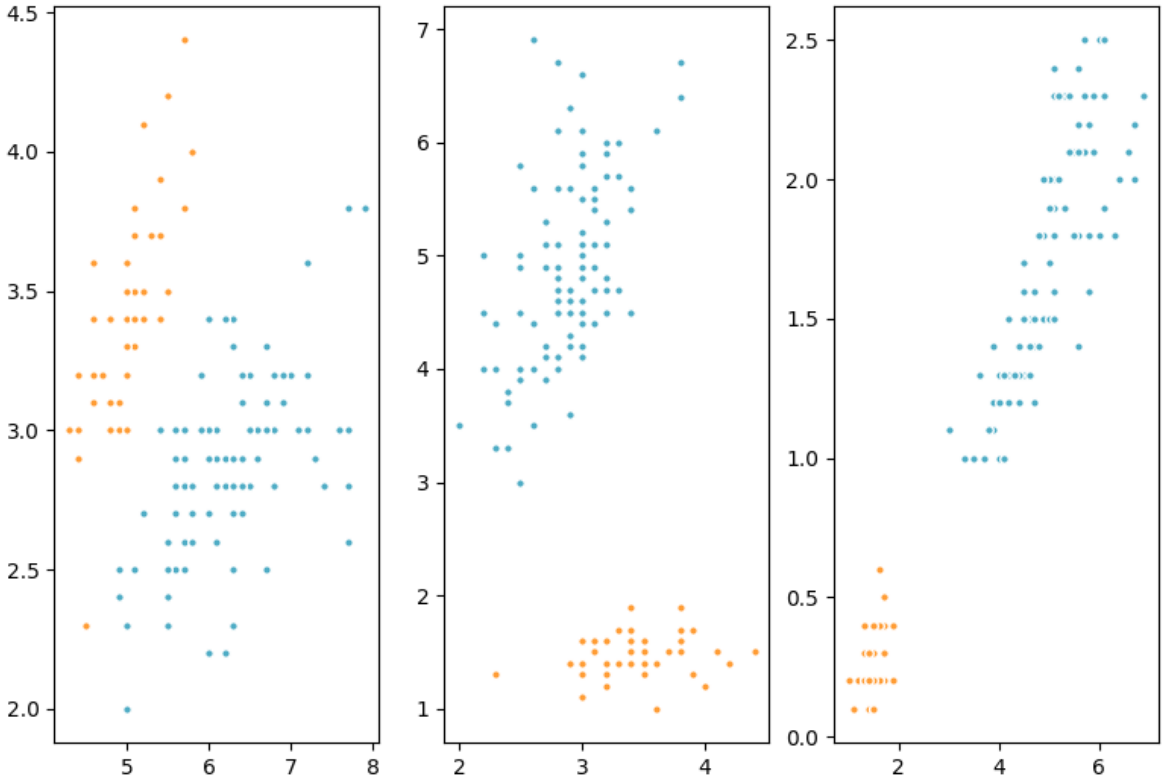

Рисунок 9 – Иерархический кластеринг с количеством кластеров 2

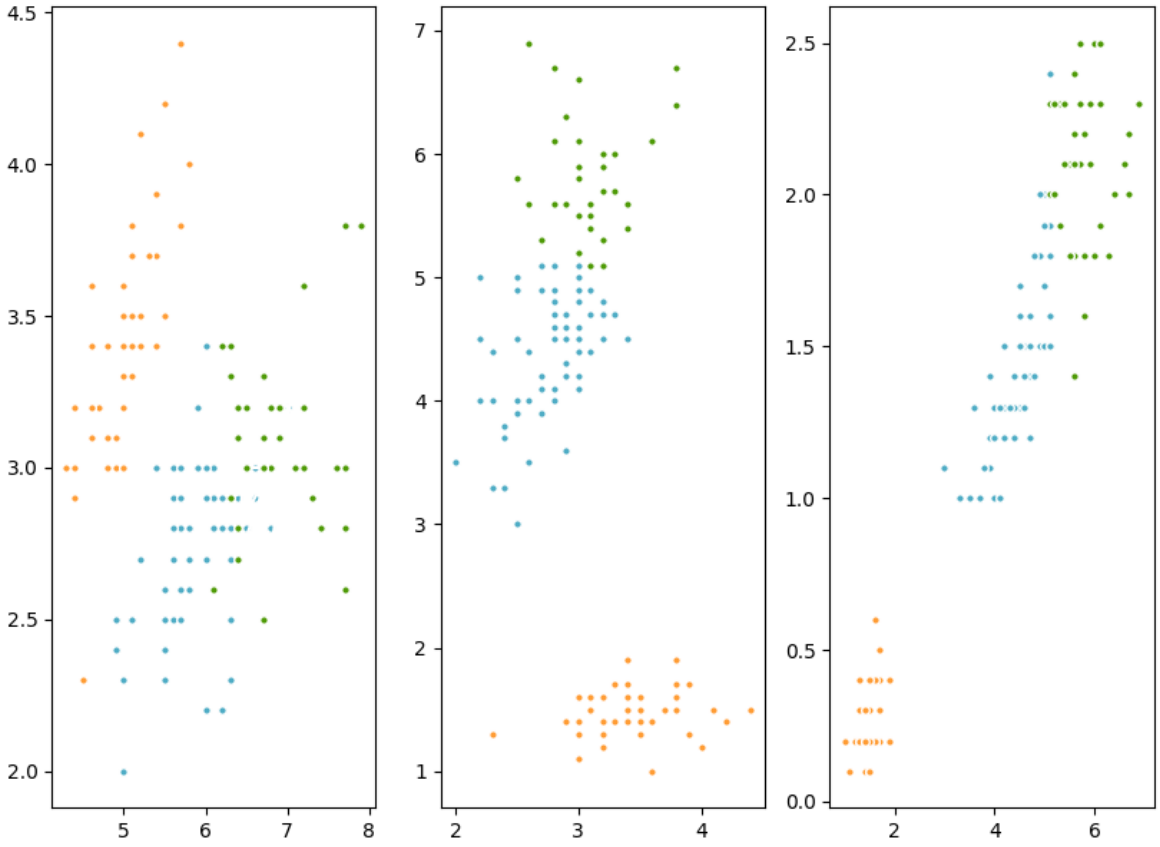

Рисунок 10 – Иерархический кластеринг с количеством кластеров 3

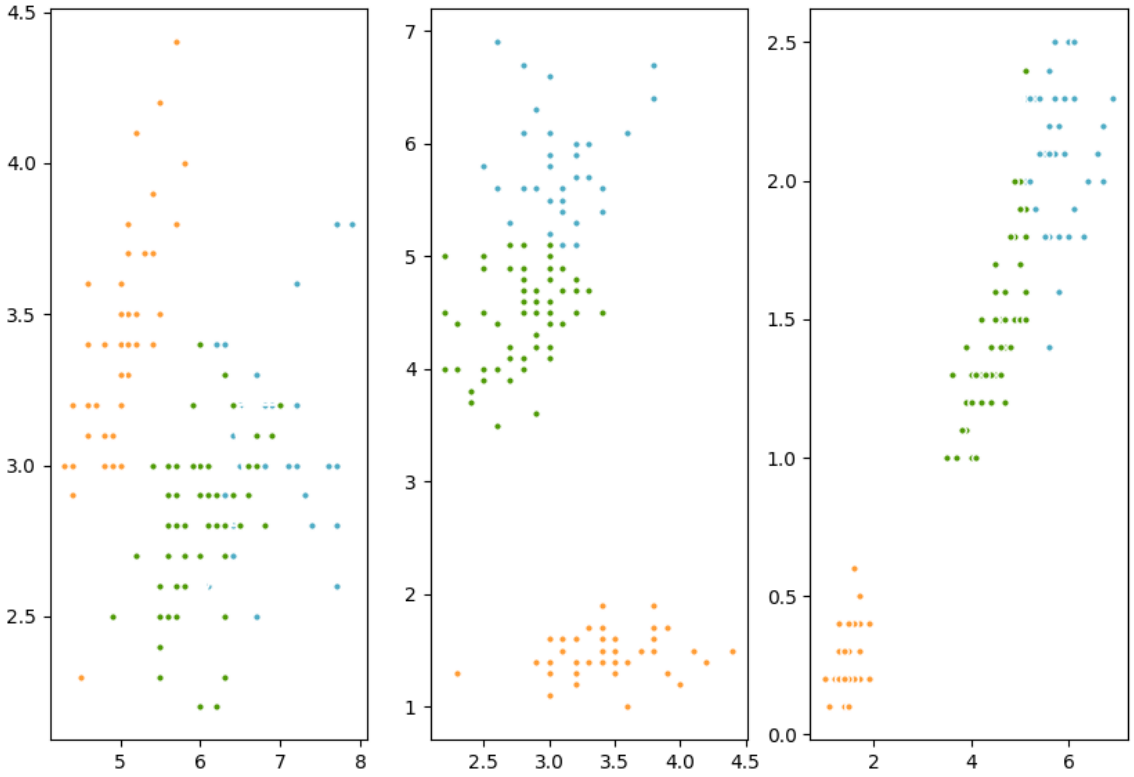

Рисунок 11 – Иерархический кластеринг с количеством кластеров 4

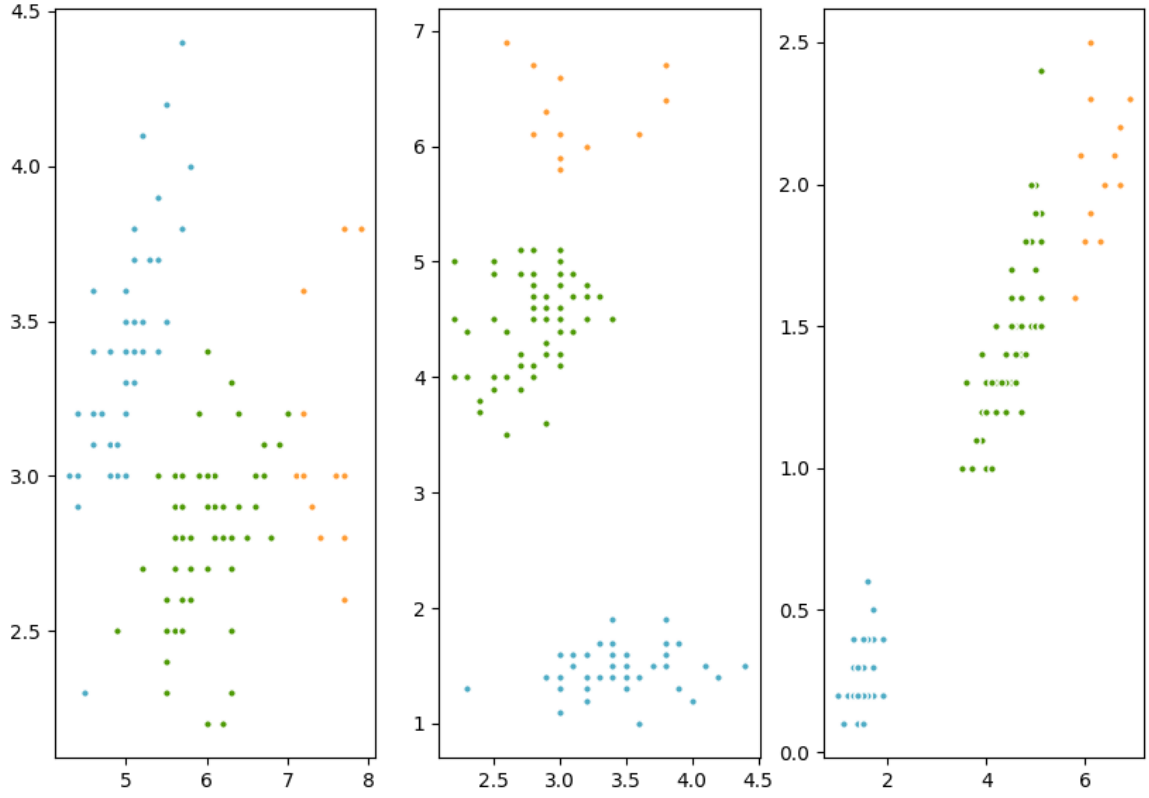

Рисунок 12 – Иерархический кластеринг с количеством кластеров 5

Нарисуем дендограмму иерархической кластеризации. Резкльтаты на рис. 13.

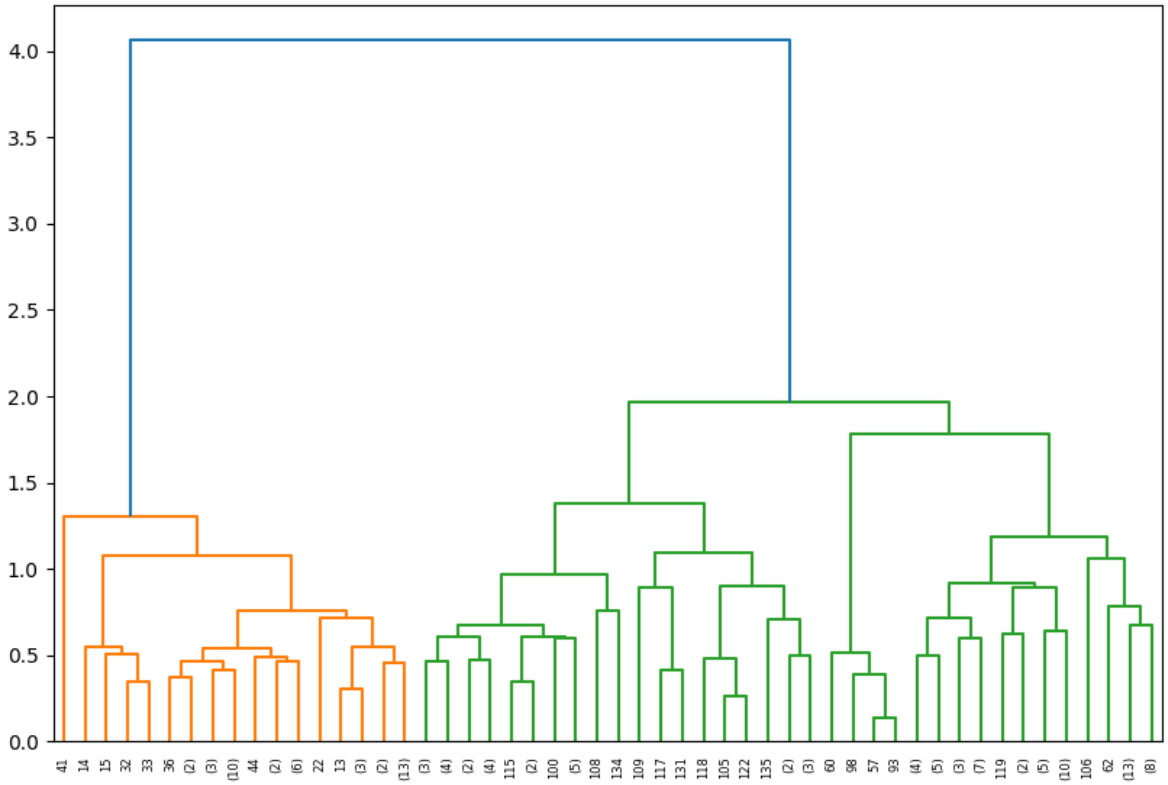

Рисунок 13 – Дендограмма

Попробуем кластеризовать набор данных состоящий из двух колец с помощью иерархической кластеризации с различными параметрами linkage. Настройки параметра linkage:

- ward минимизирует дисперсию
- average минимизирует среднее расстояние
- complete or maximum минимизирует максимальное расстояние
- single минимизирует минимальное расстояние

Результаты представлены на рис. 14 – 17.

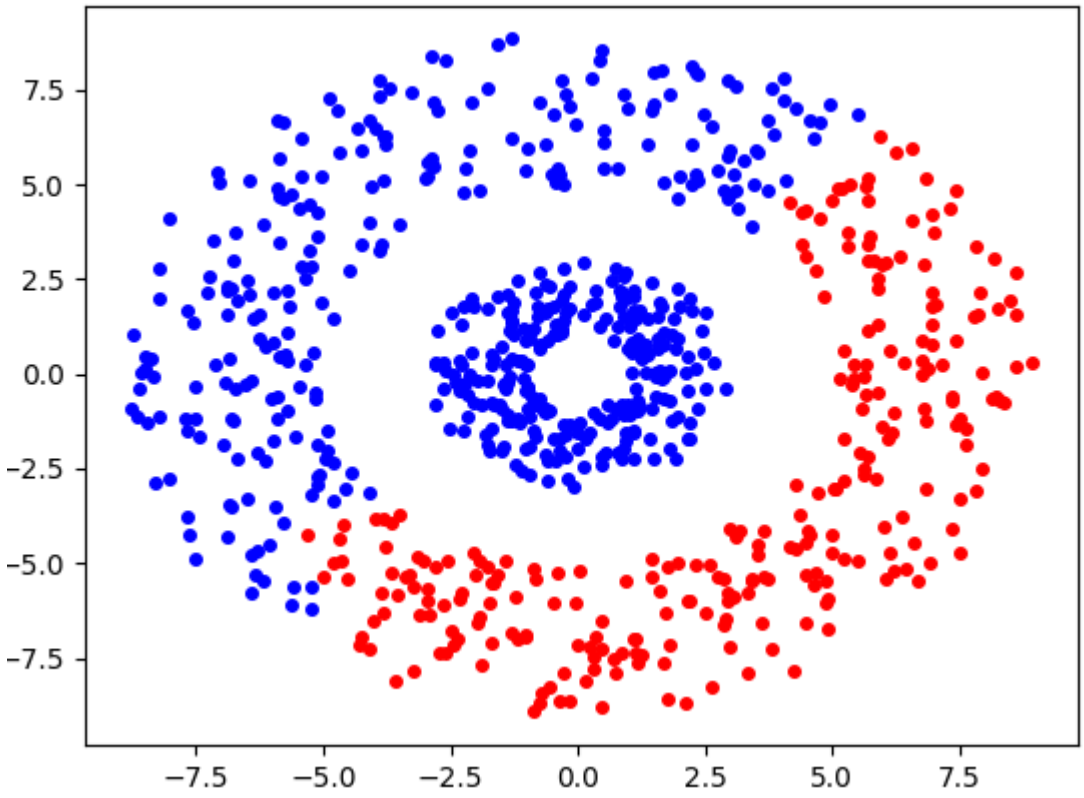

Рисунок 14 – Кластеризация с параметром ward

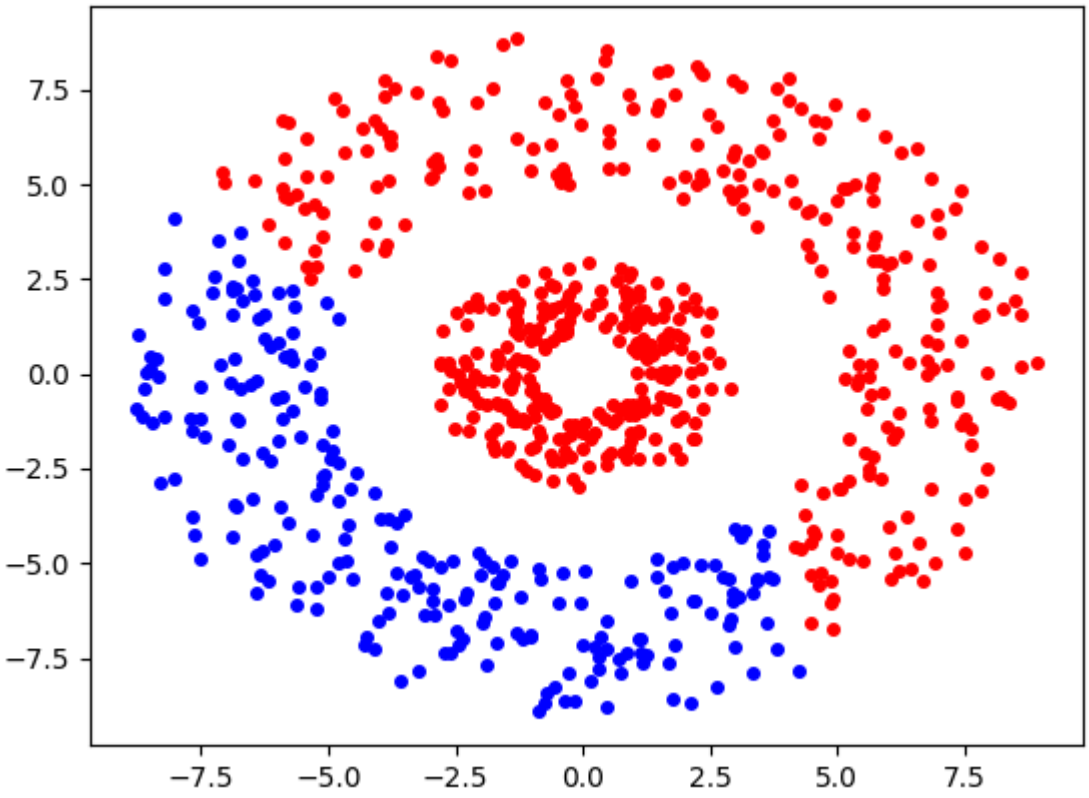

Рисунок 15 – Кластеризация с параметром complete

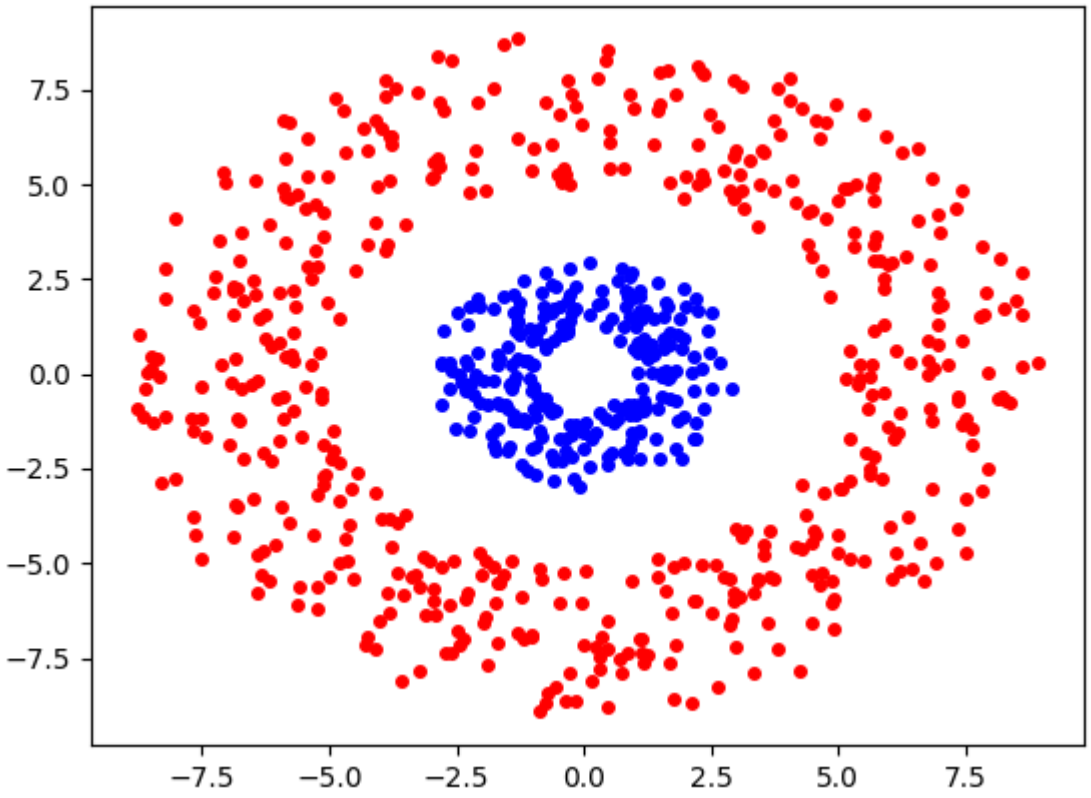

Рисунок 16 – Кластеризация с параметром single

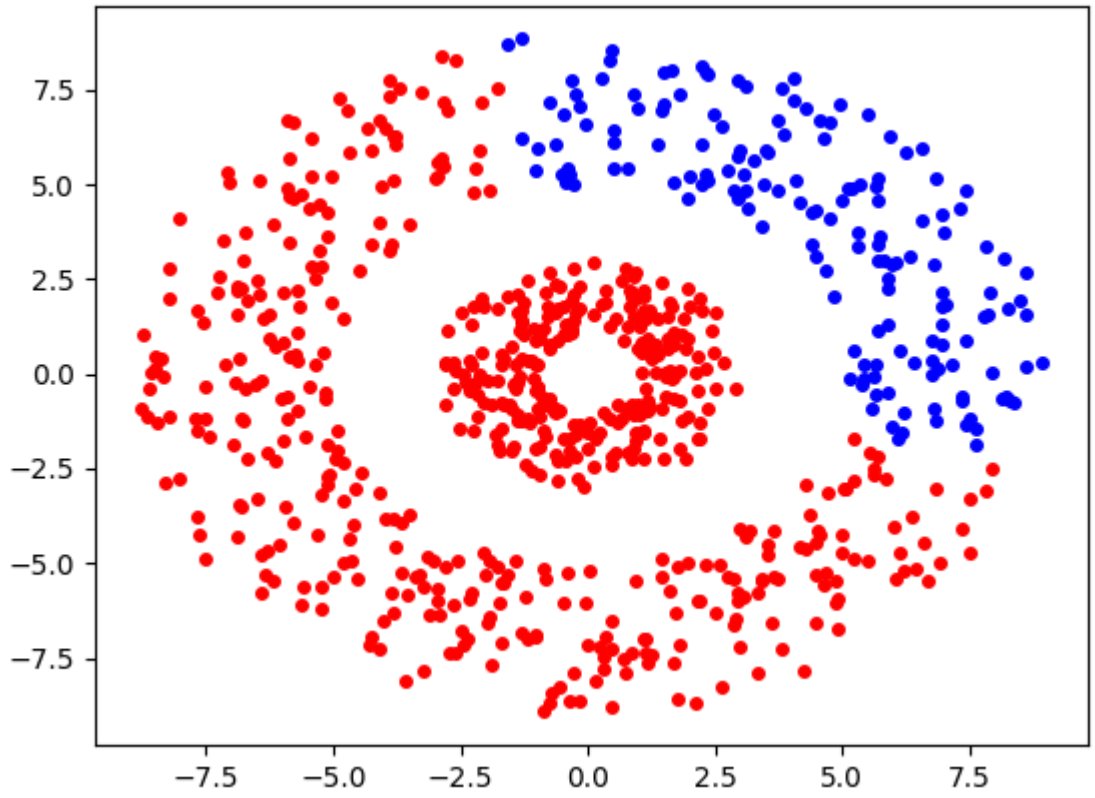

Рисунок 17 – Кластеризация с параметром average

Single может применяться на больших наборах данных и на данных с неглобулярной геометрией, однако неустойчив к выбросам.

Ward наиболее надежен однако не может менять настройки расчета расстояния.

Complete и average хорошая замена ward когда необходимо расчитать неевклидово расстояние.

Выводы

В ходе выполнения данной лабораторной работы было произведено знакомство с ассоциативным иерархической кластеризацией и кластеризацией k-means, а также их модификациями.