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)

Студентка гр. 8304 _____ Сергеев А. Д.

Преподаватель _____ Жангиров Т. Р.

Санкт-Петербург

2021

Цель работы

Ознакомиться с методами кластеризации модуля Sklearn.

Ход работы

1. Загрузка данных.

1) Был загружен по ссылке требуемый датасет.

2. K-means.

1) Была проведена кластеризация методом k-средних.

2) Были получены центры кластеров и определено, какие наблюдения попали в какой кластер.

```
Cluster centers: [[5.9016129  2.7483871  4.39354839 1.43387097]
 [5.006      3.418      1.464      0.244     ]
 [6.85       3.07368421 5.74210526 2.07105263]]
Cluster labels: [1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 1 1 1 1 1 1 1 1 1 1 1 1 0 0 2 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 0 0 0 2 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 2 0 2 2 2 2 0 2 2 2 2
 2 2 0 0 2 2 2 2 0 2 0 2 0 2 2 0 0 2 2 2 2 2 0 2 2 2 2 0 2 2 2 0 2 2 2 0 2
 2 0]
```

Рисунок 1 - Центры и значения кластеров

3) Были построены результаты классификации для признаков попарно (1 и 2, 2 и 3, 3 и 4). Кластеризация для каждого случая произведена одинаково эффективно. Параметр n_init влияет на количество итераций алгоритма.

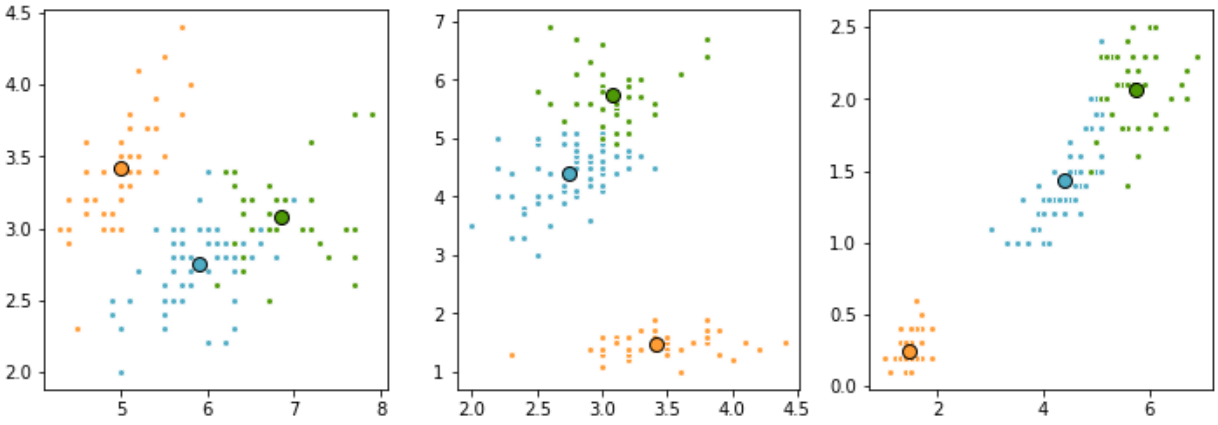

Рисунок 2 - Обработка данных методом k-средних

4) Размерность данных была уменьшена до 2 с использованием метода главных компонент, была нарисована карта для всей области значений, на которой каждый кластер занимает определенную область со своим цветом.

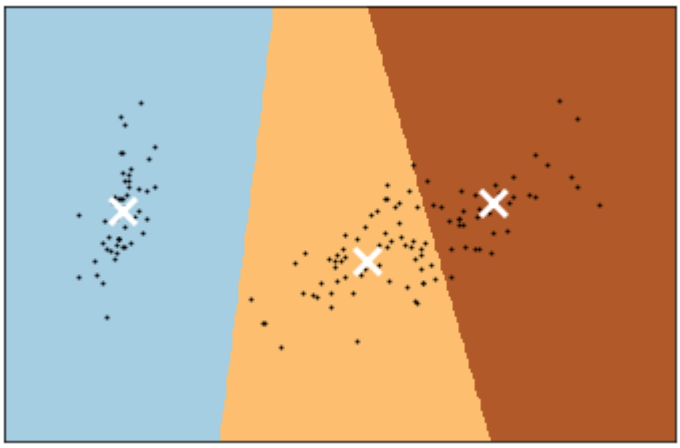

Рисунок 3 - Полученные кластеры

5) Была исследована работа алгоритма k-средних при различных параметрах init. Сначала он был выполнен несколько раз с параметром 'random', затем вручную для выбранных точек.

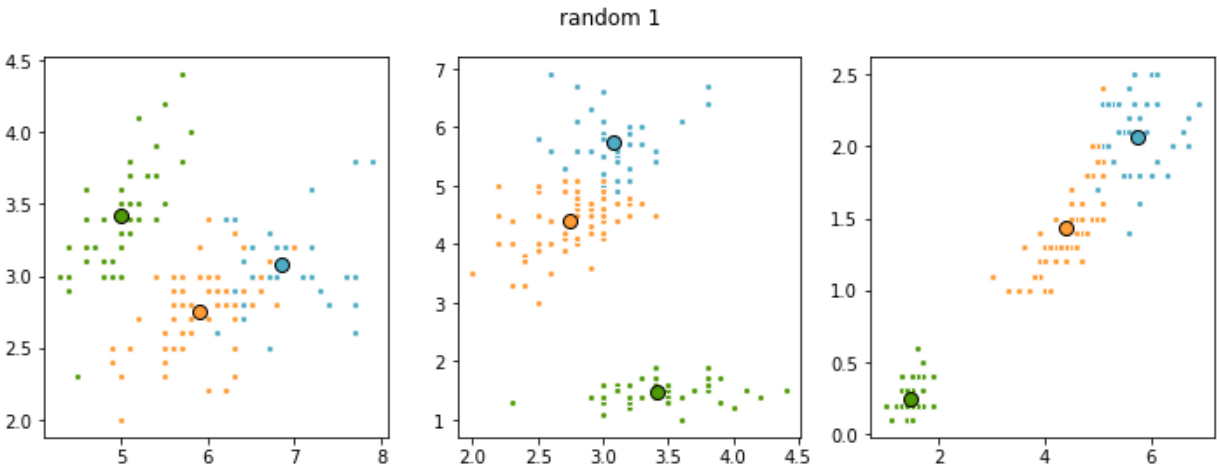

Рисунок 4 - Первое выполнение со случайным выбором центров

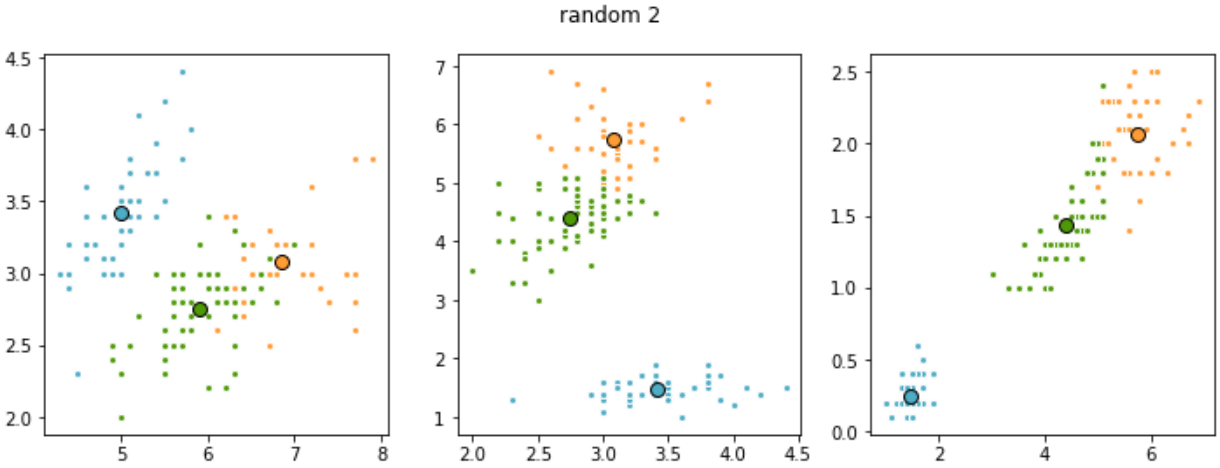

Рисунок 5 - Второе выполнение со случайным выбором центров

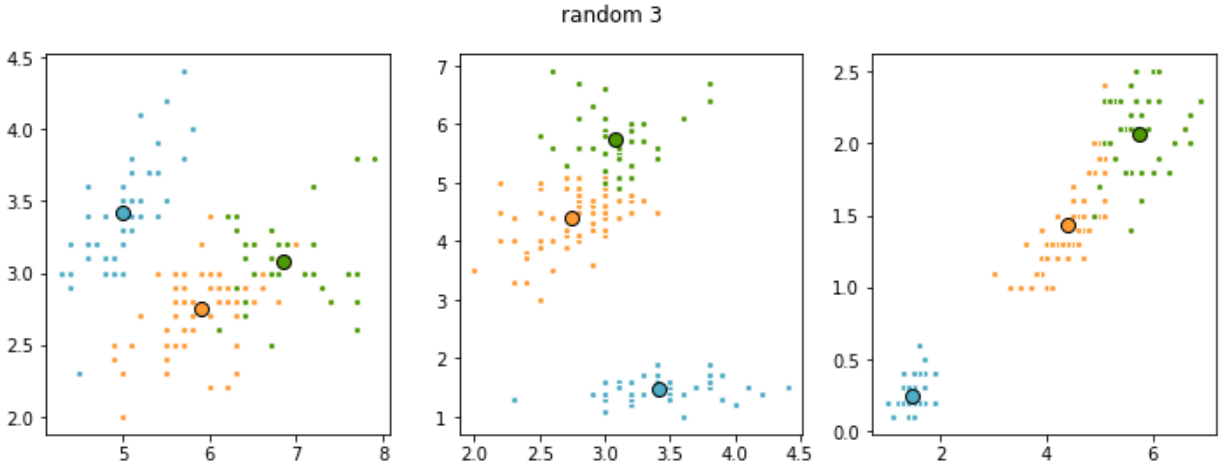

Рисунок 6 - Третье выполнение со случайным выбором центров

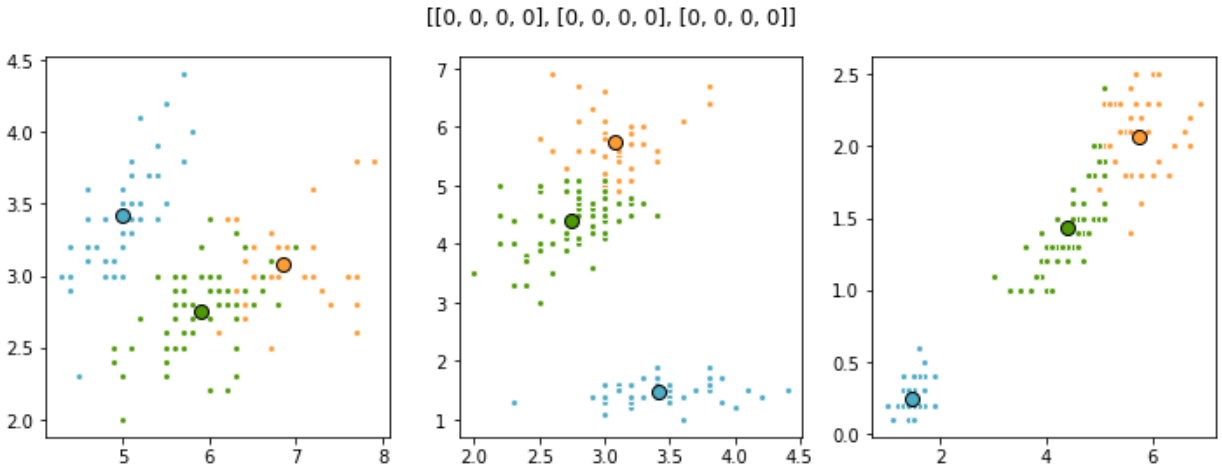

Рисунок 7 - Выполнение сзаранее выборанными центрами

6) Было найдено наилучшее количество кластеров методом локтя. Самое сильное улучшение результата происходит в районе 2-3 кластеров, следовательно, подходящее число кластеров 2 или 3.

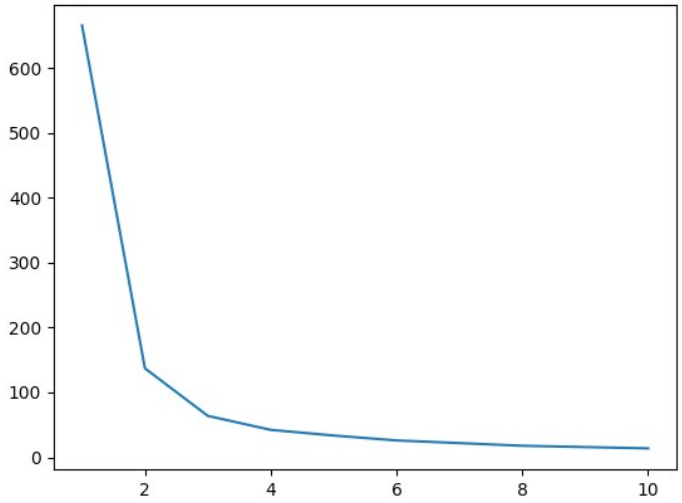

Рисунок 8 - Метод локтя

7) Была проведена кластеризация с использованием пакетной кластеризаций k-средних. Данный метод осуществляет разделение путем выбора только части значений и вычисления кластеров с их помощью. Это повышает быстродействие алгоритма. Была построена диаграмма рассеяния, на которой синим цветом выделены точки, которые для разных методов попали в разные кластеры.

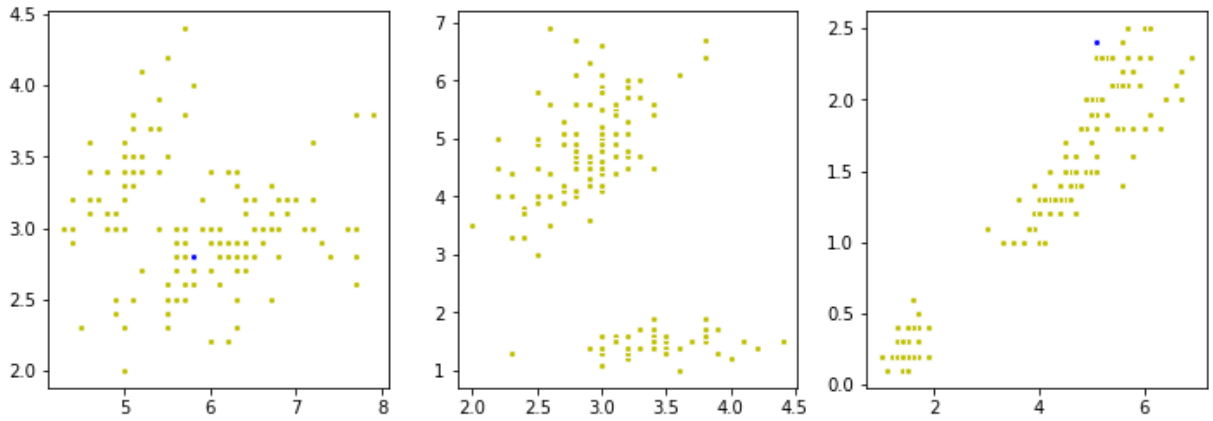

Рисунок 9 - Сравнение MiniBatch K-means и K-means

3. Иерархическая кластеризация.

1) Была проведена иерархическая кластеризация тех же данных.
2) В отличие от k-средних иерархическая кластеризация постипенно объединяет минимальные кластеры, минимизируя определенную характеристику.

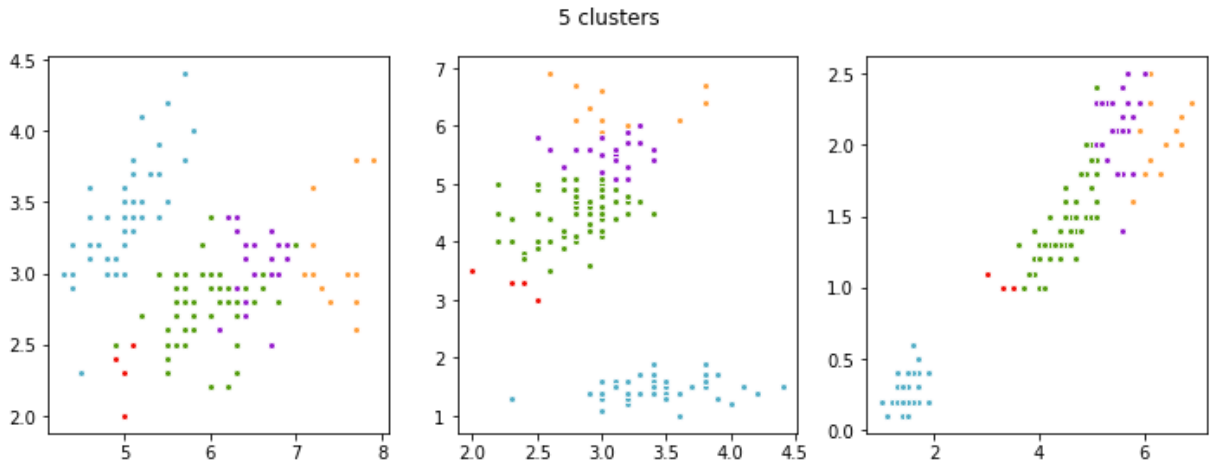

Рисунок 10 – Иерархическая кластеризация для 5 кластеров

3) Было проведено исследование для различного размера кластеров (2-4).

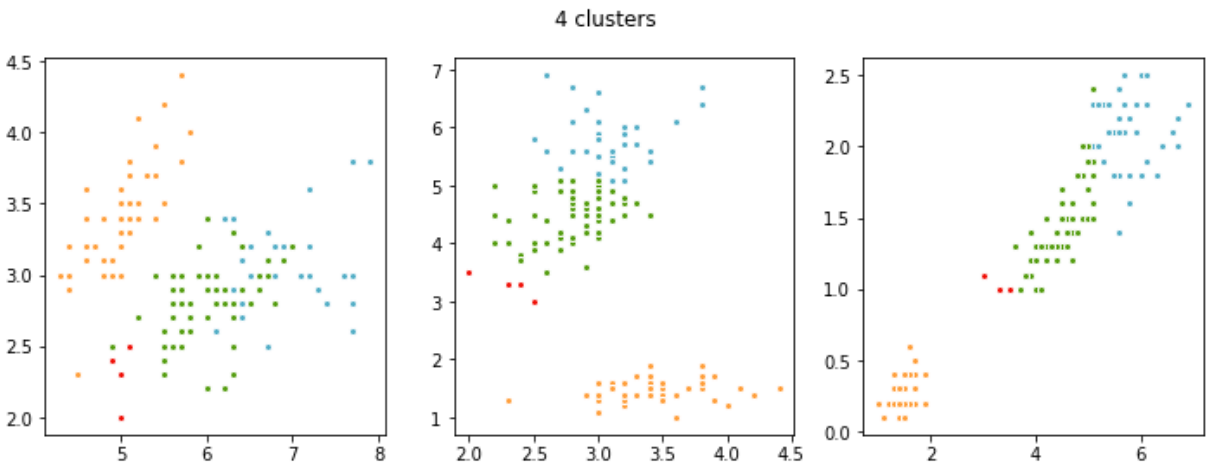

Рисунок 11 - Иерархическая кластеризация для 4 кластеров

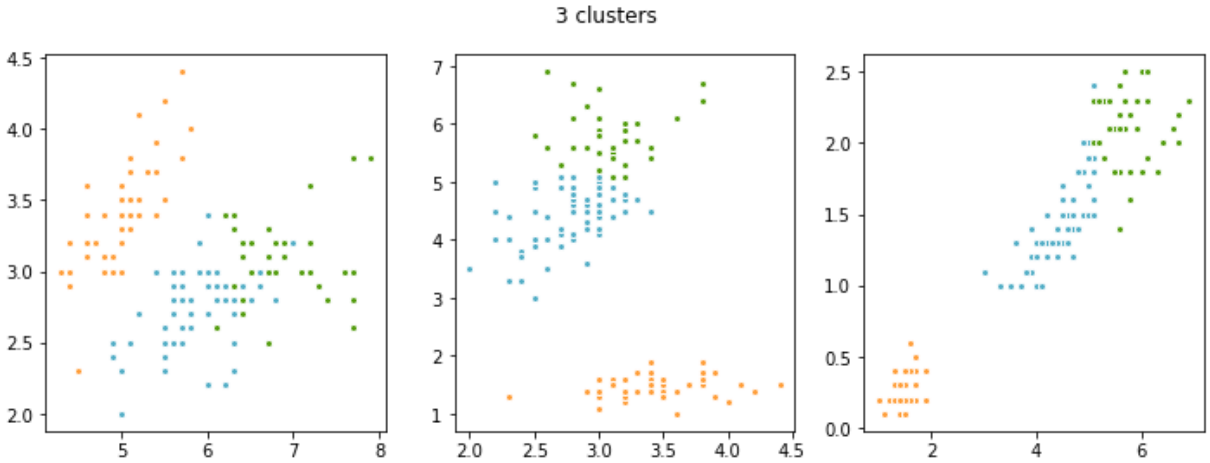

Рисунок 12 - Иерархическая кластеризация сдля 3 кластеров

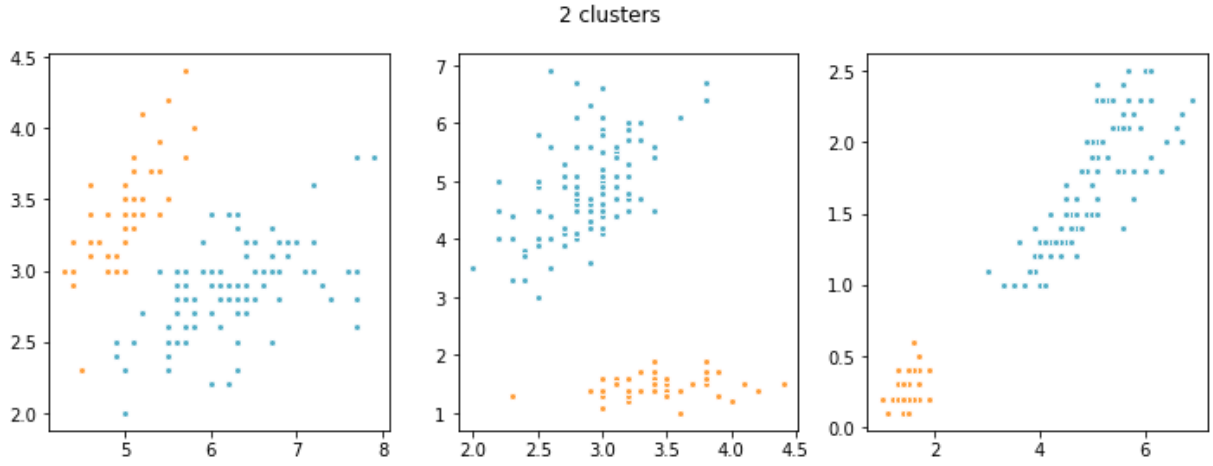

Рисунок 13 - Иерархическая кластеризация для 2 кластеров

4) Была нарисована дендограмма до уровня 6.

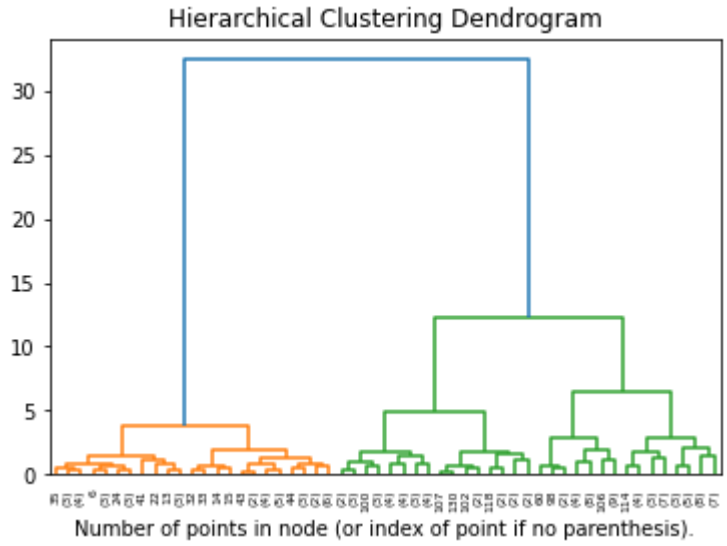

Рисунок 14 - Дендограмма

5) Были сгенерированы данные в виде двух колец.
6) Была проведена иерархическая кластеризация сгенерированных данных.

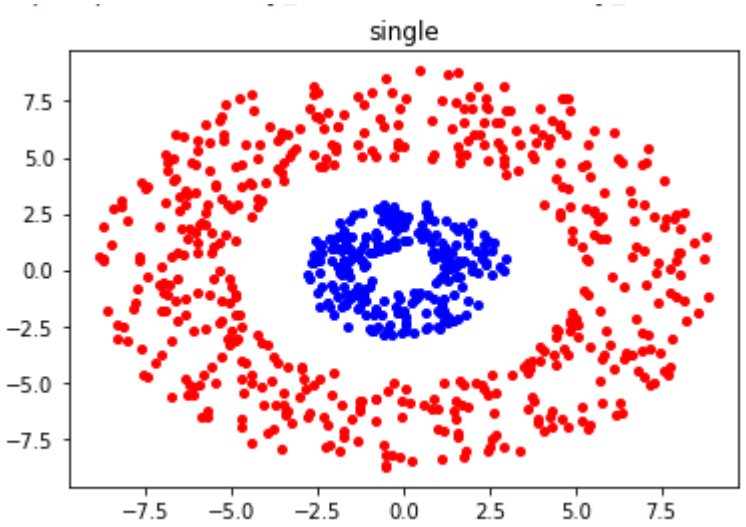

Рисунок 15 - Кластеризация с параметром single

7) Была исследована кластеризация при всех параметрах linkage.
   - ward минимизирует дисперсию;
   - average минимизирует среднее расстояние;
   - complete or maximum минимизирует максимальное расстояние;

- single минимизирует минимальное расстояние.

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

Ward наиболее надежен, однако не может менять настройки расчета расстояния.

Complete и *average* хорошая замена *ward* когда необходимо рассчитать неевклидово расстояние.

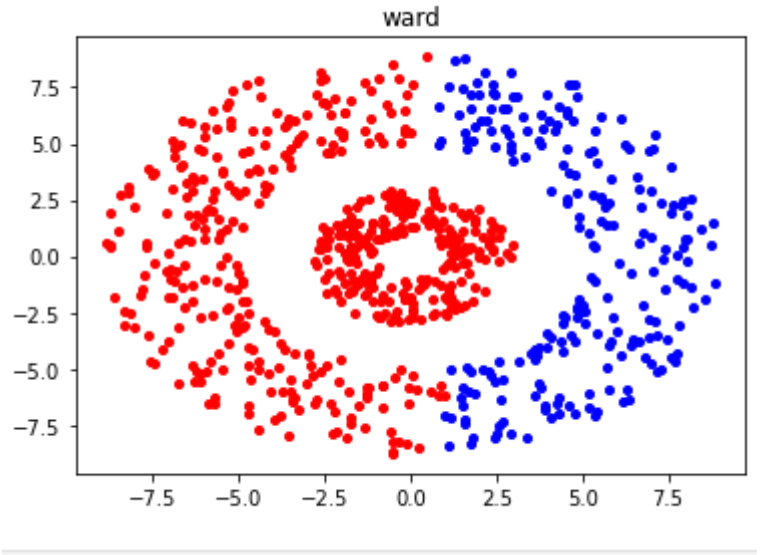

Рисунок 16 - Кластеризация с параметром ward

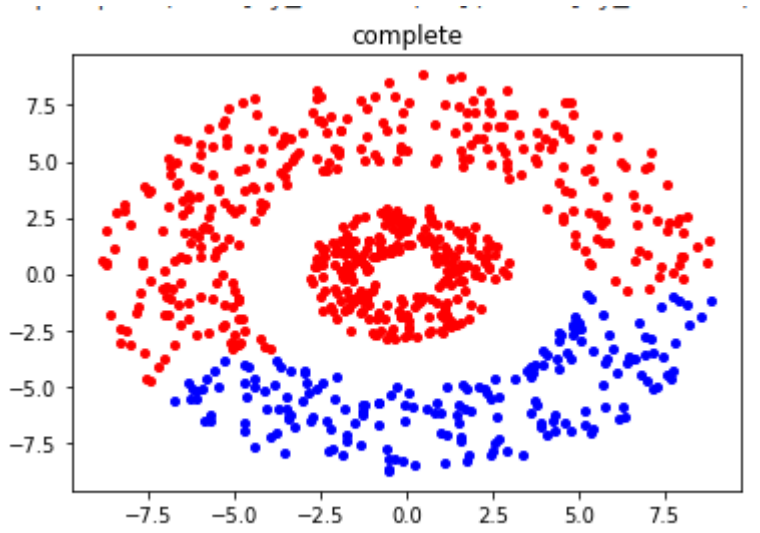

Рисунок 17 - Кластеризация с параметром complete

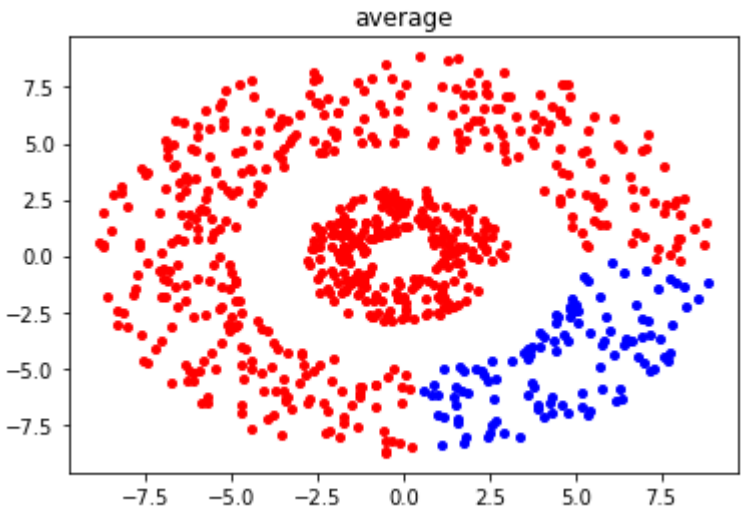

Рисунок 18 - Кластеризация с параметром average

Выводы

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