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 _____ Ястребков А. С.

Преподаватель _____ Жангиров Т. Р.

Санкт-Петербург

2020

Цель работы:

Ознакомиться с методами кластеризации модуля Sklearn.

Ход работы

Загрузка данных.

Был загружен датасет iris.data с информацией о разновидностях ирисов (фрагмент исходного датасета показан на рис. 1).

	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

Рис. 1. Фрагмент исходного датасета.

1. Для датасета была выполнена классификация методом k-средних. Результаты попарно для 4 признаков показаны на рис. 2. По ним видно, что наилучшее разделение на кластеры произошло по признакам 3 и 4. Изменение параметра n_init на результаты видимого воздействия не оказало.

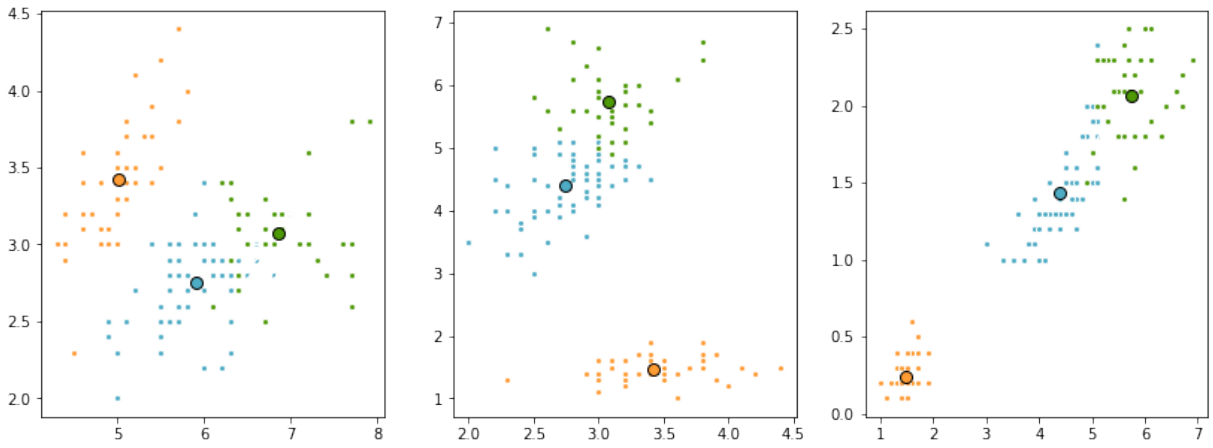

Рис. 2. Результаты кластеризации методом k-средних.

2. При помощи метода главных компонент (PCA) было выполнено понижение размерности пространства до 2 и построена карта области значений, где каждый кластер занимает определённую область. Результат показан на рис. 3.

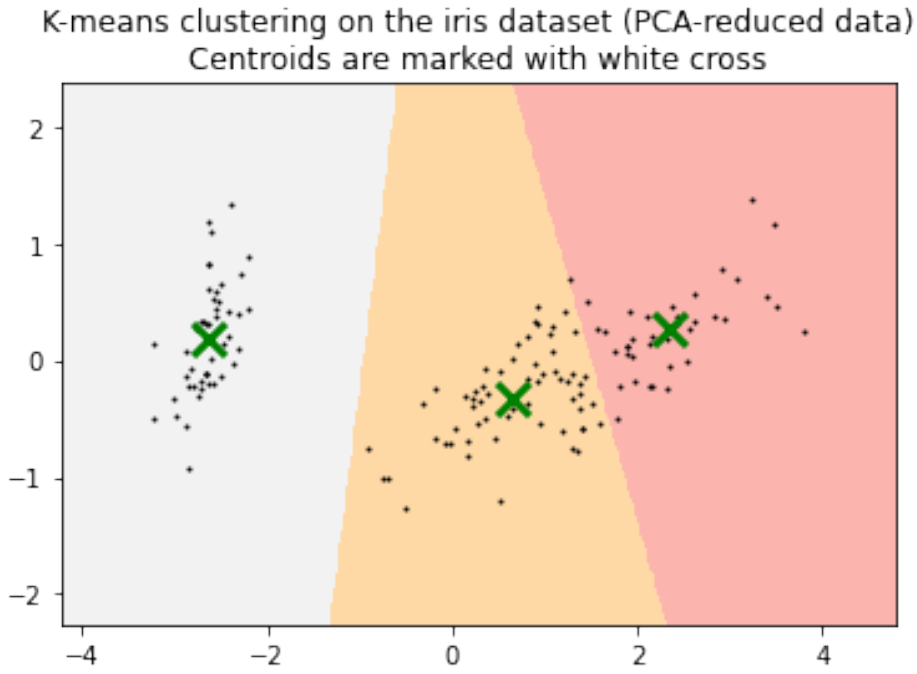

Рис. 3. Карта области значений, полученная после уменьшения размерности данных.

3. На рис. 4-7 показаны результаты для кластеризации методом k-средних при различных значениях параметра init (случайных и заданных вручную) и

разном числе итераций. Заметно отличается только результат на рис. 5, что можно объяснить меньшим числом итераций алгоритма.

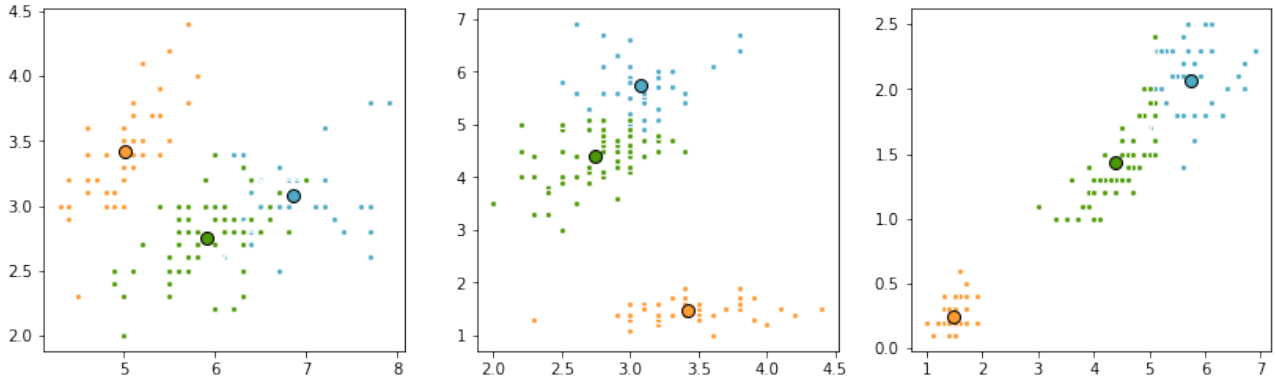

Рис. 4. Кластеризация при init='random', n_init=15, max_iter=300.

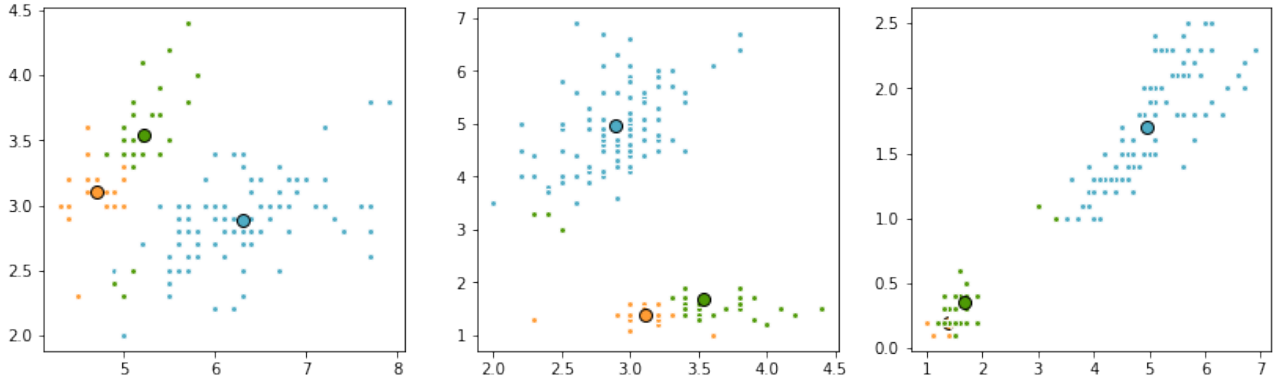

Рис. 5. Кластеризация при init='random', n_init=1, max_iter=30.

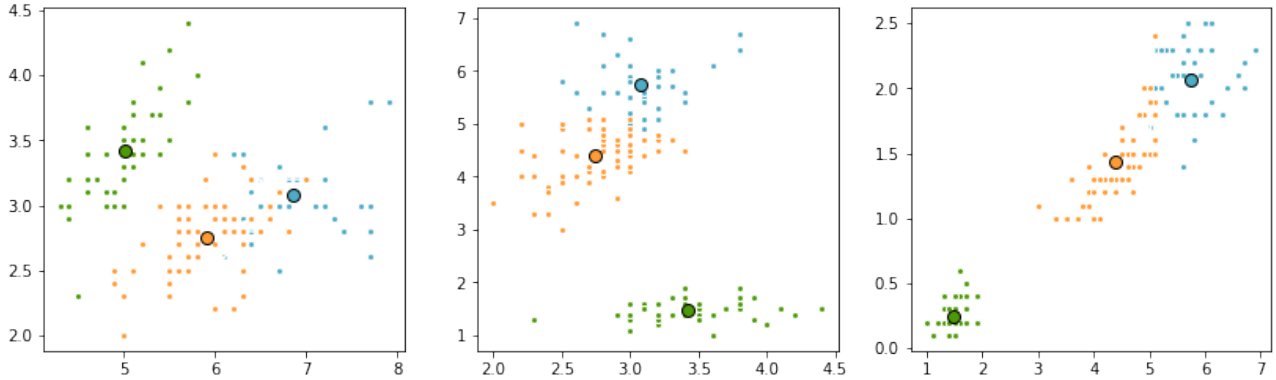

Рис. 6. Кластеризация при init=[[0,0,0,0],[1,1,1,1],[2,2,2,2]].

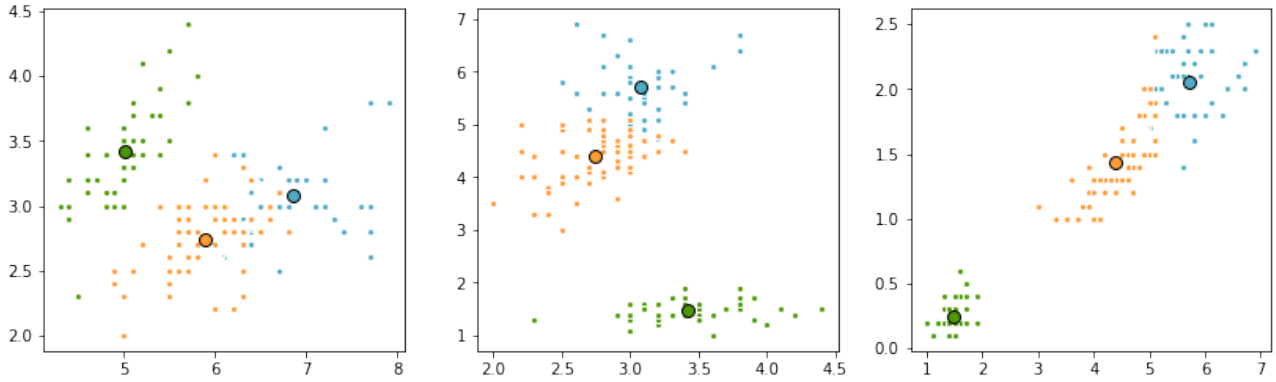

Рис. 7. Кластеризация при init=[[10,10,10,10],[10,10,10,10],[20,20,20,20]].

4. Методом локтя наилучшее количество кластеров было определено как 2, график приведён на рис. 8.

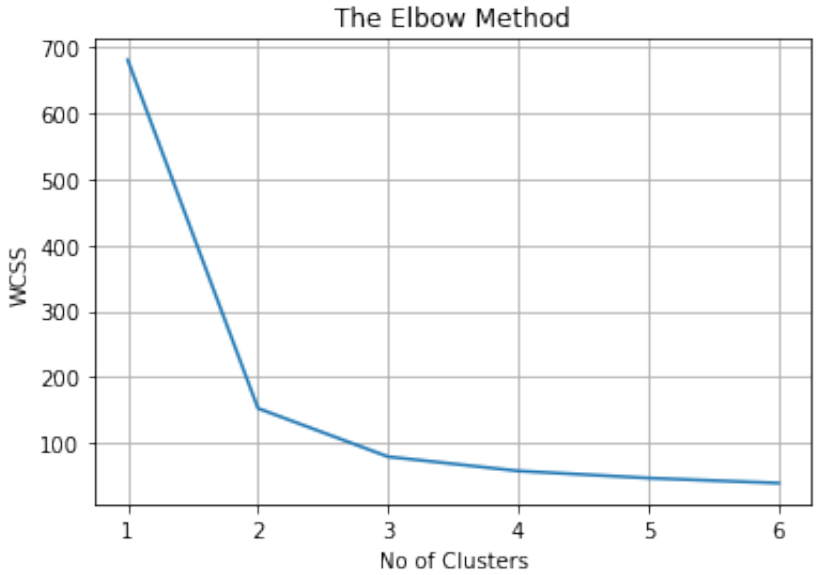

Рис. 8. Выбор количества кластеров методом локтя.

5. Была проведена кластеризация алгоритмом MiniBatchKMeans (результат показан на рис. 9). В отличие от KMeans, этот алгоритм работает с пакетами данных, а не с полным набором данных, что повышает скорость, но может сказаться на точности. В случае рассматриваемого датасета объём данных невелик, и два алгоритма дают практически идентичный результат, что показано на рис. 10.

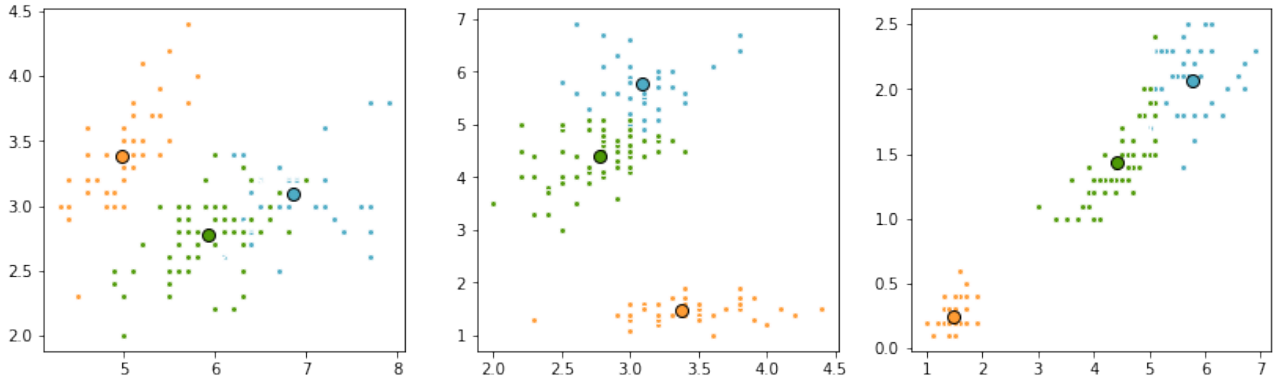

Рис. 9. Результат кластеризации алгоритмом MiniBatchKMeans.

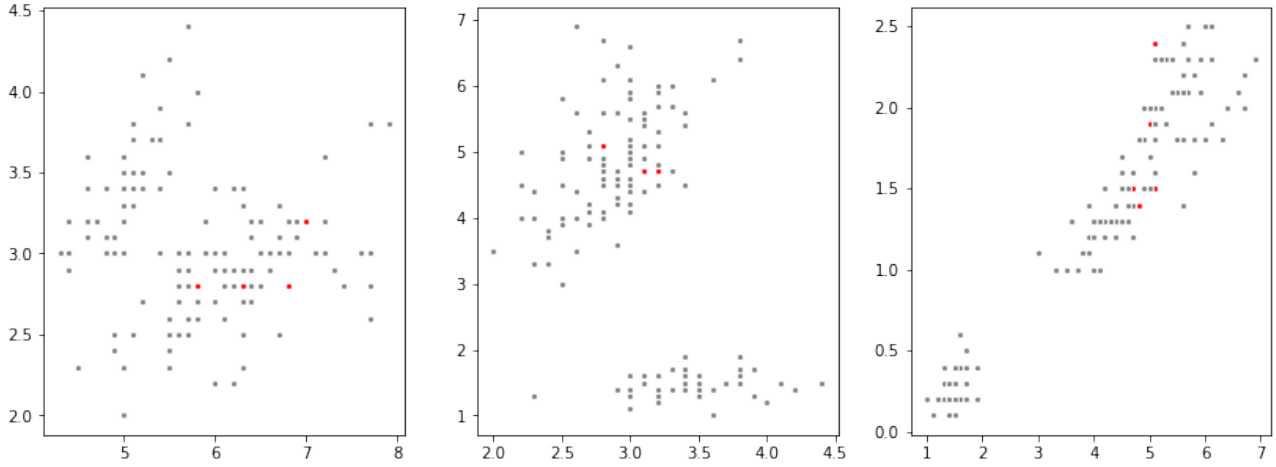

Рис. 10. Сравнение алгоритмов KMeans и MiniBatchKMeans.

6. Была проведена кластеризация данных алгоритмом иерархической сортировки AgglomerativeClustering, результат показан на рис. 11. Этот алгоритм начинает с предположения, что каждая точка является тривиальным кластером из самой себя. Далее кластеры объединяются по какой-либо метрике.

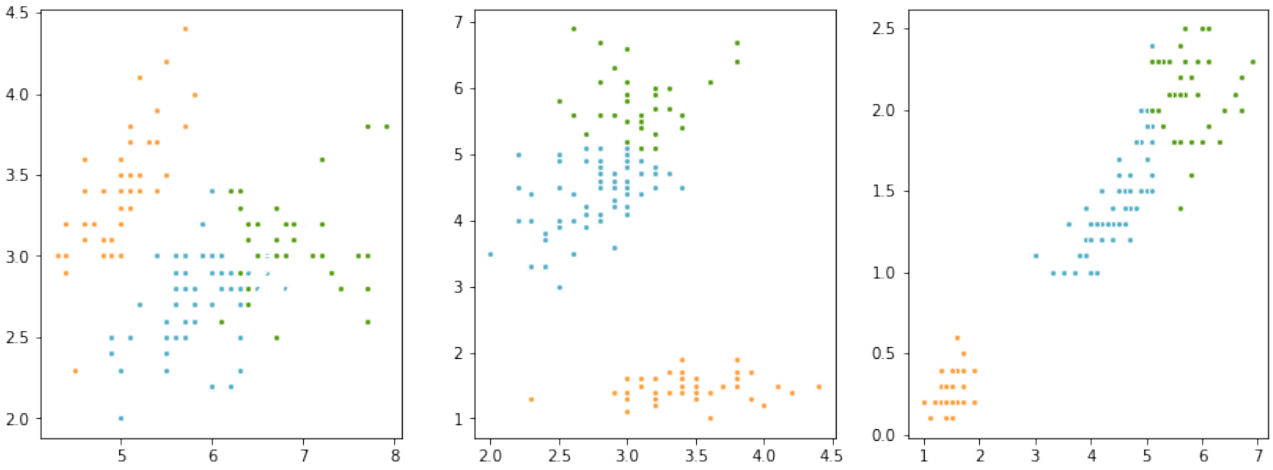

Рис. 11. Результат иерархической кластеризации.

На рис. 12-14 показаны результаты иерархической кластеризации для 2, 4 и 5 кластеров.

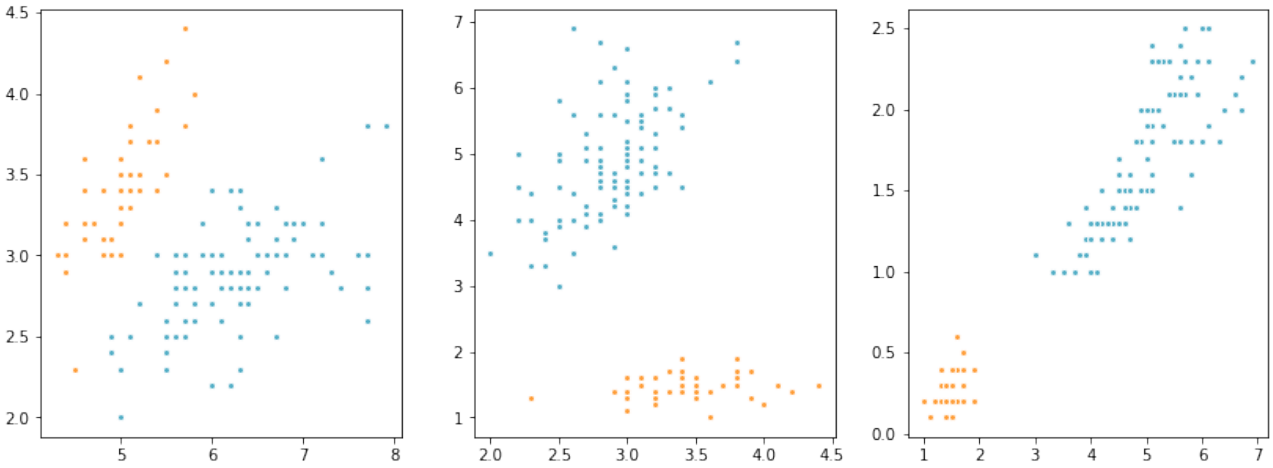

Рис. 12. Результаты иерархической кластеризации на 2 кластера.

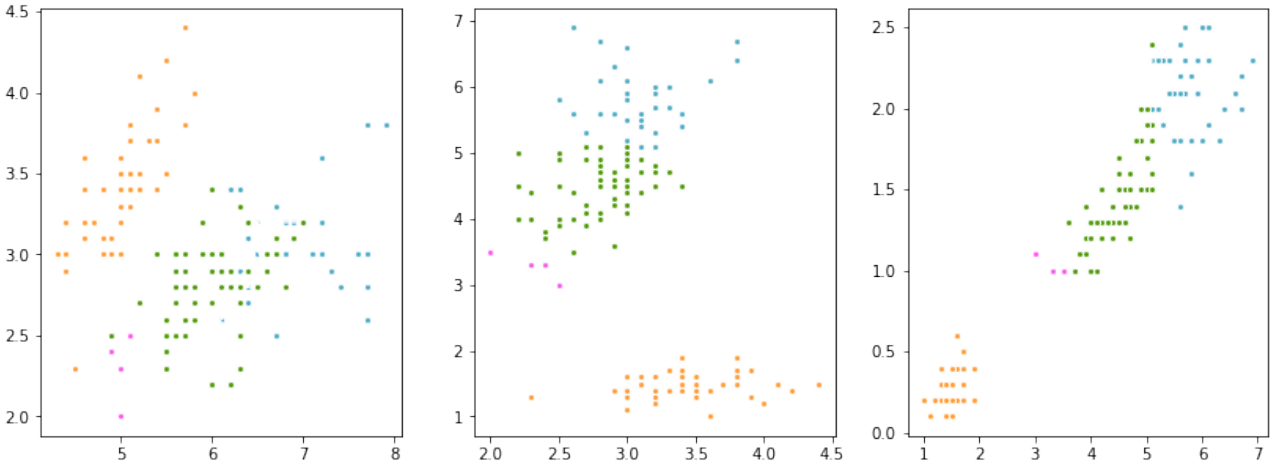

Рис. 13. Результаты иерархической кластеризации на 4 кластера.

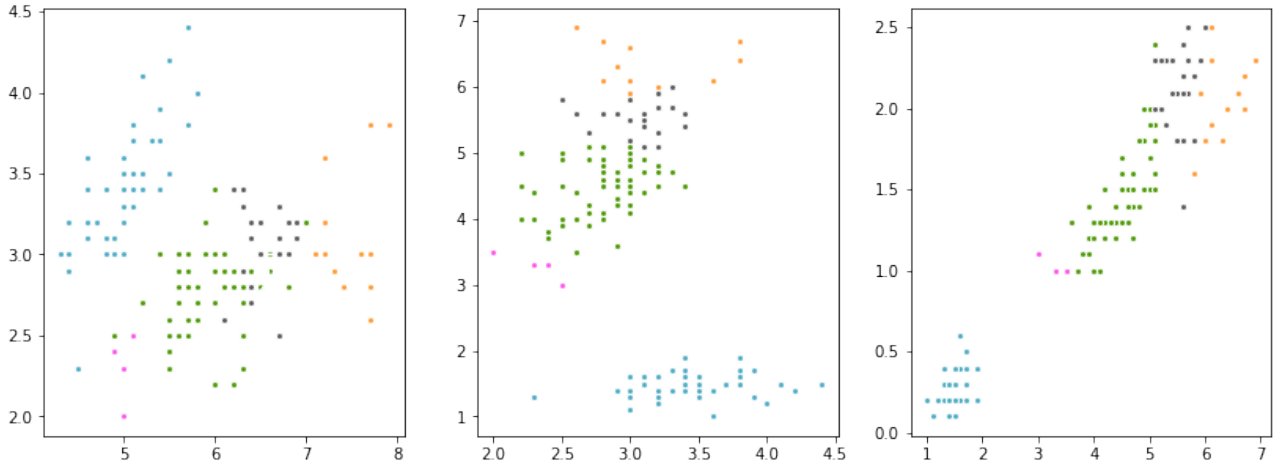

Рис. 14. Результаты иерархической кластеризации на 5 кластеров.

7. Была построена дендрограмма до уровня 6, результат показан на рис. 15.

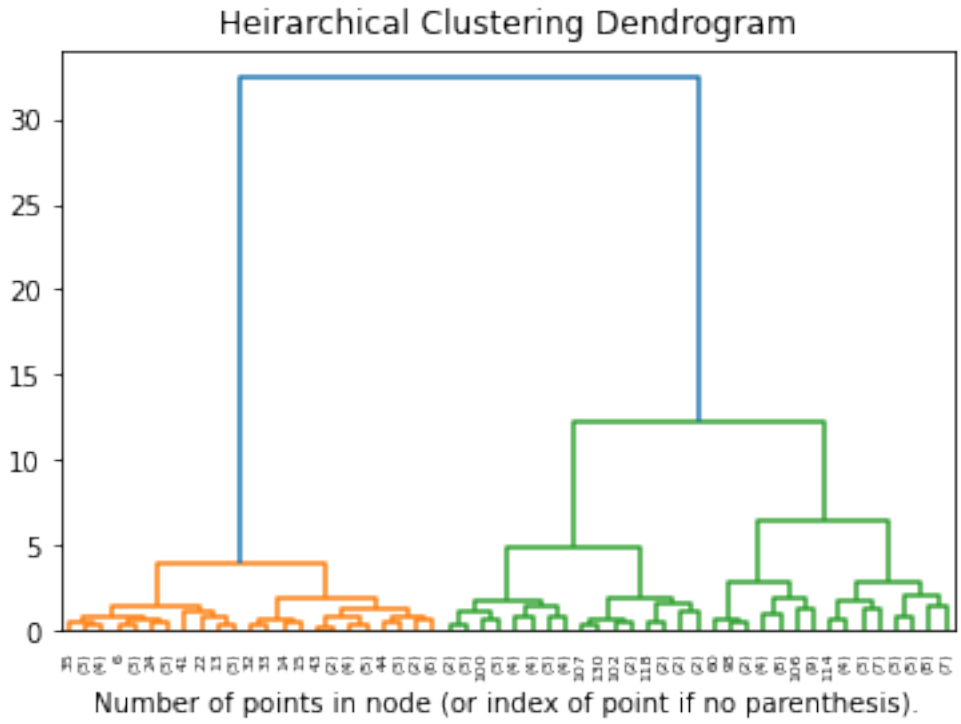

Рис. 15. Дендрограмма до уровня 6.

8. Было проведено исследование того, как параметр linkage влияет на иерархическую кластеризацию для случая двух сложных кластеров (два случайно сгенерированных концентрических кольца). Результат показан на рис. 16. Видно, что лишь значение linkage = „single“ дало ожидаемую

кластеризацию, значения Complete Link, Ward и Group Average дали расстояние между точками разных колец меньше, чем между точками одного кольца.

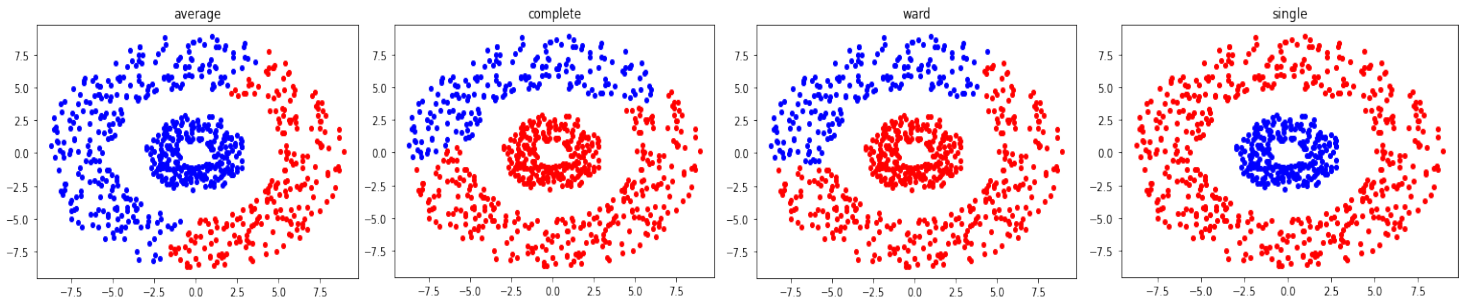

Рис. 16. Сравнение иерархической кластеризации для различных значений параметра linkage на примере кластеров из двух колец.

Вывод:

В результате выполнения лабораторной работы были изучены методы кластеризации k-средних и иерархическая.

Для кластеризации методом k-средних было изучено влияние параметров на вид получаемых кластеров. Размер использованного датасета не позволил в полной мере проследить различия между кластеризацией при разных параметрах, однако показано, что решающее влияние оказывает количество итераций алгоритма. На использованном датасете различия между классическим и пакетным методом k-средних были незначительны.

Метод иерархической кластеризации при изменении параметров оказался корректен для кластеризации данных с нелинейной зависимостью.