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. 8303 _____ Гришин К. И.

Преподаватель _____ Жангиров Т.Р.

Санкт-Петербург

2021

Цель работы

Ознакомиться с методами кластеризации из библиотеки *Sklearn*.

Ход выполнения работы

Загрузка данных

1. Скачать датасет по ссылке: https://archive.ics.uci.edu/ml/datasets/iris.
2. Загрузить данные в датафрейм (табл. 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

Таблица 1. Загруженные данные.

Данные отображены на диаграммах рассеяния (рис. 1) в соответствие с цветами на рисунке 2.

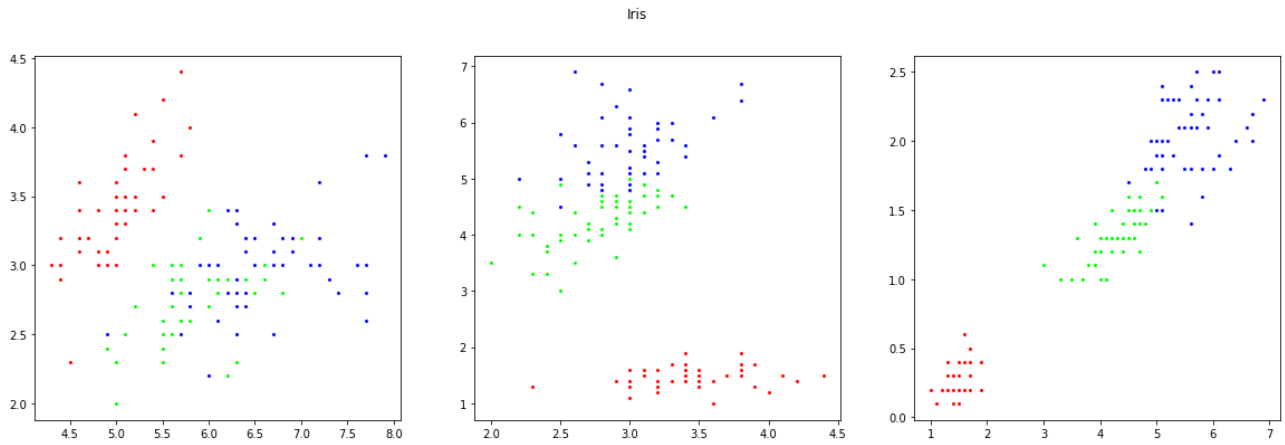

Рисунок 1. Входные данные.

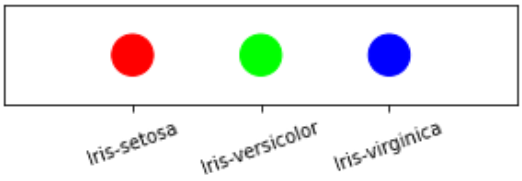

Рисунок 2. Название класса и цвет.

K-Means

1. Провести кластеризацию методов k-средних.
2. Получить центры кластеров и определить какие наблюдения в какой кластер попали.
3. Результаты классификации (рис. 3).

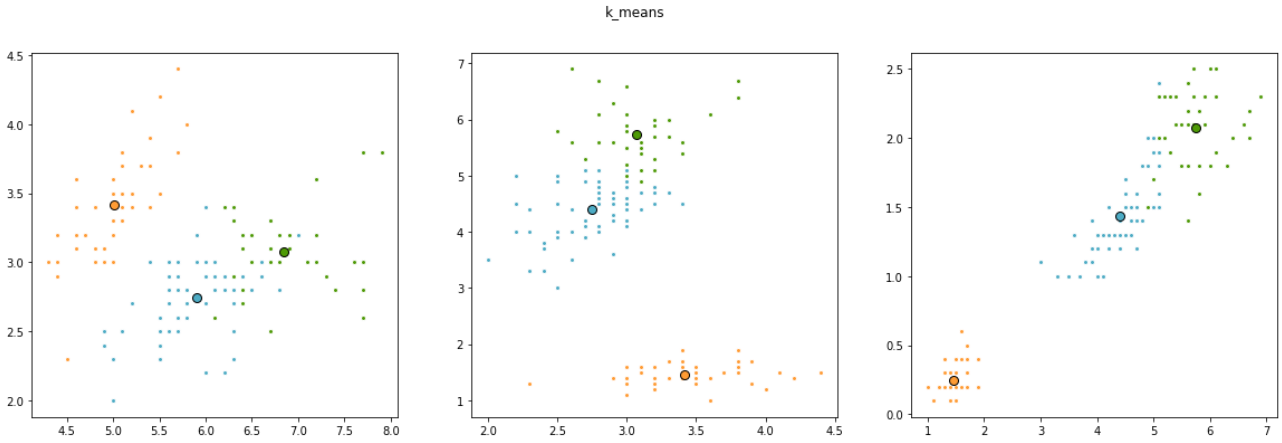

Рисунок 3. Классификация методом K-Means.

4. Уменьшить данные до размерности 2 и отобразить области значений (рис. 4).

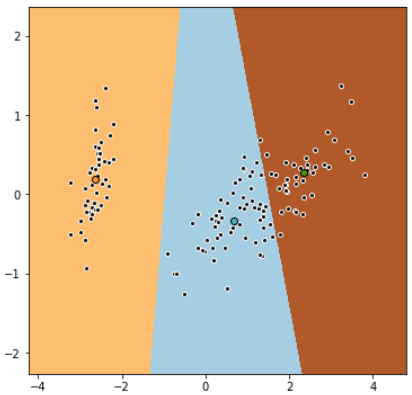

Рисунок 4. Области значений при размерности 2.

5. Исследовать алгоритм при различных параметрах *init*.

   Трижды выполнена классификация с *init='random'* (рис. 5, 6, 7).

   Дважды выполнена классификация с установкой начального значения каждого класса в одной точке. Для точки *(0, 0, 0, 0)* (рис. 8) и центральной точки *(6, 3.25, 4, 1.25)* (рис. 9).

   Беспорядочно выбраны точки для каждого класса вручную *(1, 2, 3, 4)*, *(4, 3, 2, 1)*, *(2, 1, 1, 2)* (рис. 10).

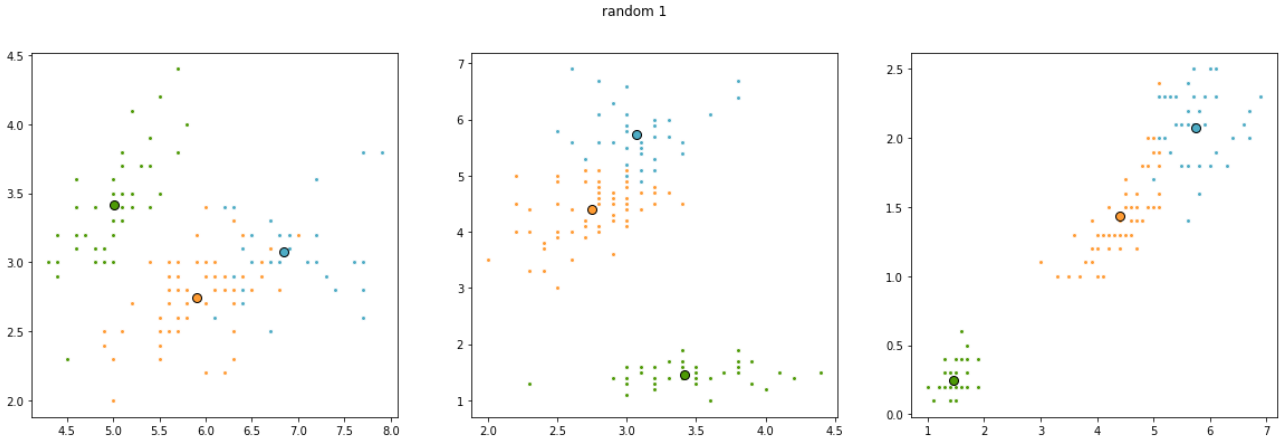

Рисунок 5. Случайный выбор начальных точек из существующих.

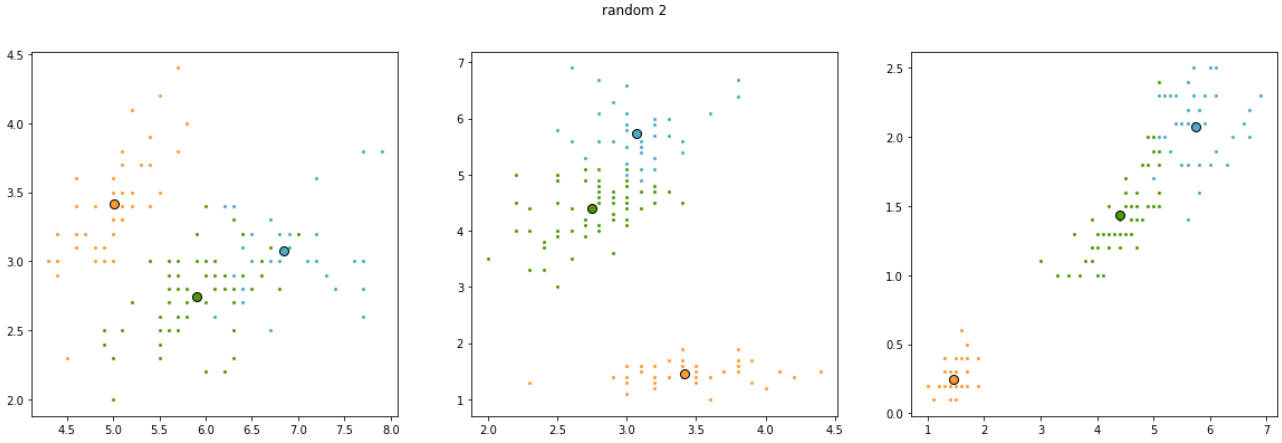

Рисунок 6. Случайный выбор начальных точек из существующих.

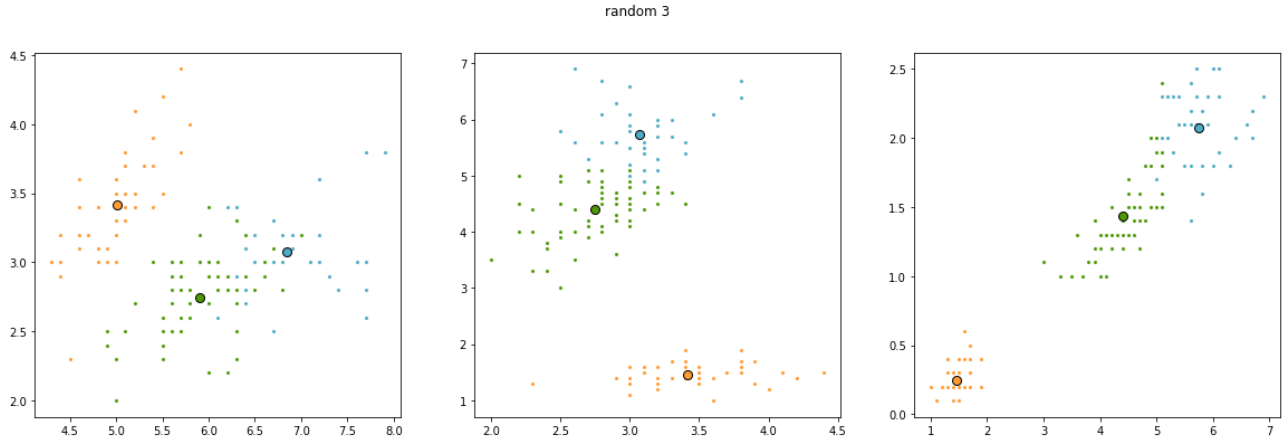

Рисунок 7. Случайный выбор начальных точек из существующих.

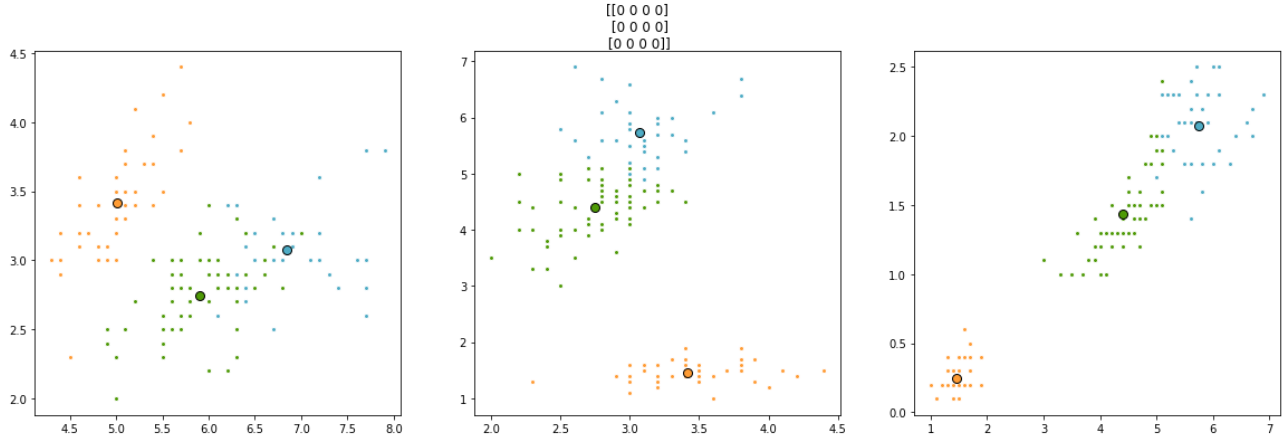

Рисунок 8. Начальная позиция каждого класса в нуле.

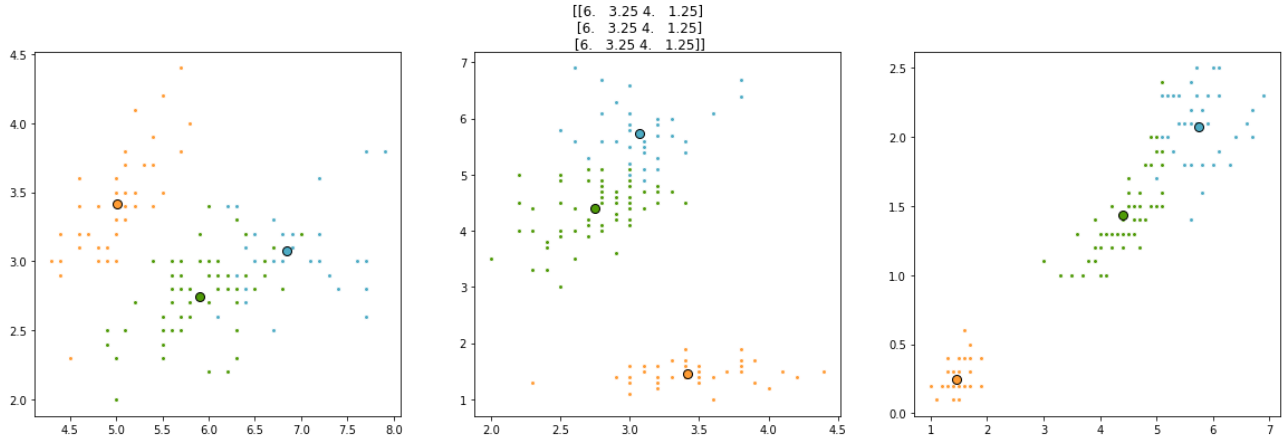

Рисунок 9. Начальная позиция каждого класса в центре облака данных.

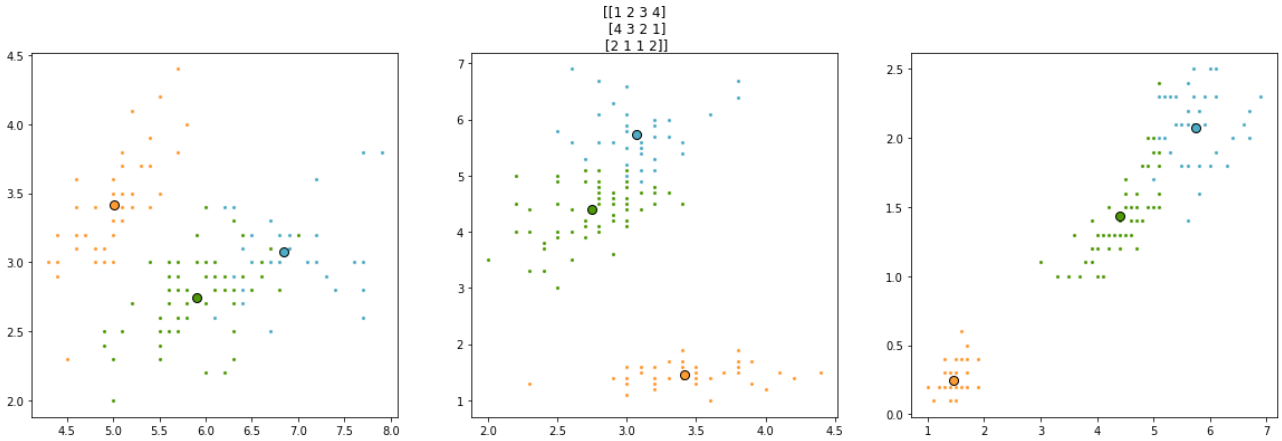

Рисунок 10. Начальная позиция каждого класса указана вручную.

6. Определить наилучшее количество классов методом локтя (рис. 11).

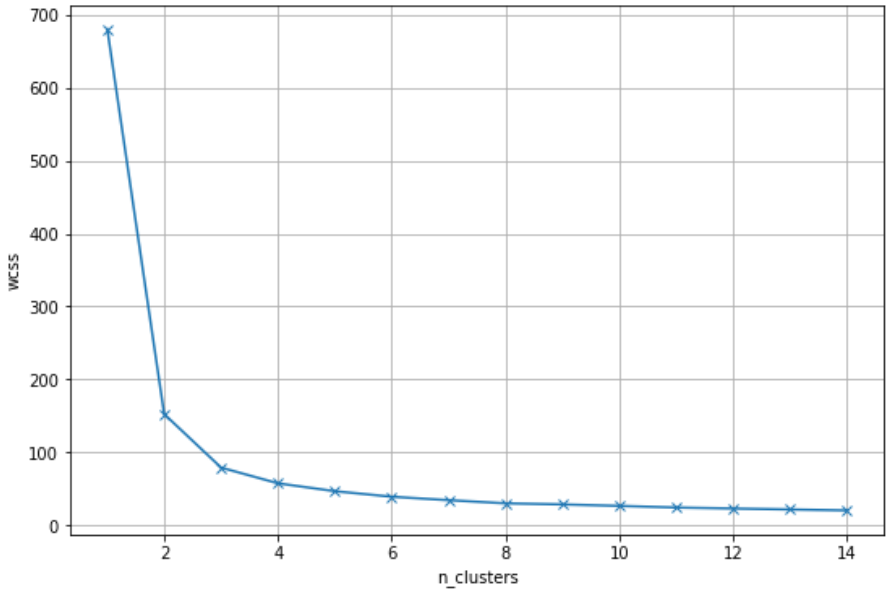

Рисунок 11. Метод локтя.

После трех кластеров, квадраты расстояний между элементами кластеров резко уменьшается, что говорит о том, что оптимальным выбором будет 3 кластера.

7. Пакетная кластеризация k-средних (рис. 12).

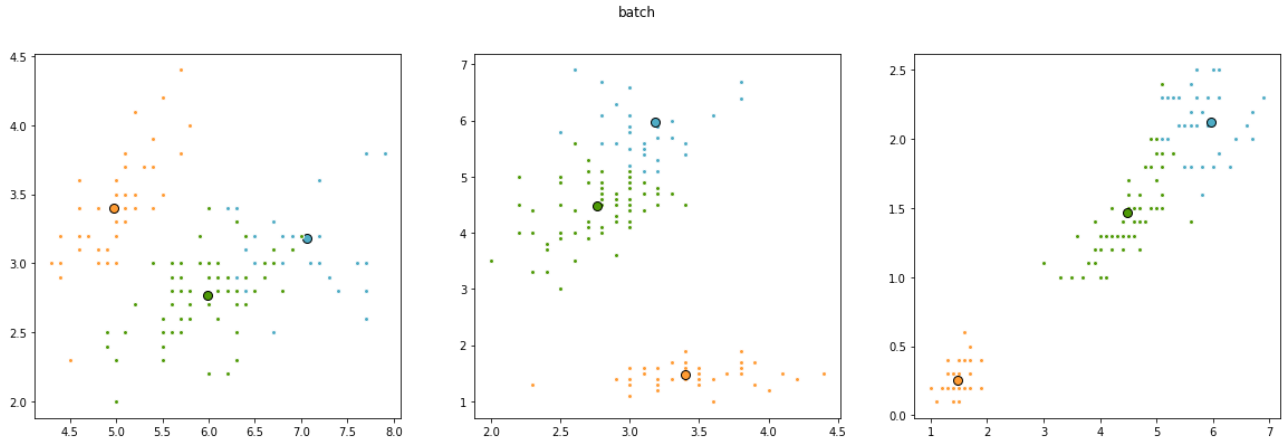

Рисунок 12. Пакетная кластеризация K-Means.

Перерасчет центроидов происходит только за счет части случайно выбранных точек, в отличие от k-means, что приводит к увеличению быстродействия, однако в таком случае страдает точность алгоритма.

Различие в определении кластеров показаны на рисунке 13.

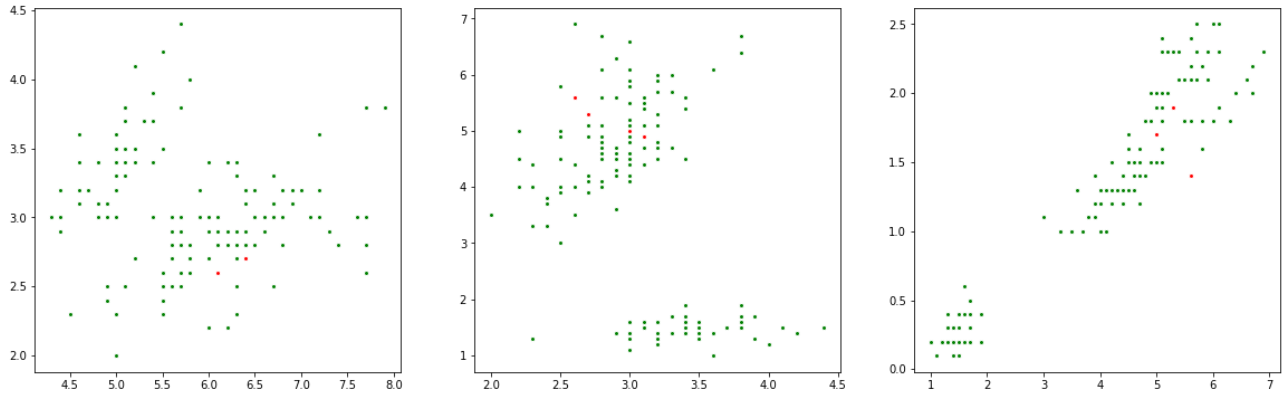

Рисунок 13. Совпадение классов различных наблюдений.

Можно заметить, что точки, попавшие в различные кластеры, находятся на стыке, однако таких точек достаточно мало и на итоговый результат они влияют не сильно.

Иерархическая кластеризация

1. Провести иерархическую кластеризацию.

   Проведена кластеризация на три кластера с функцией поиска расстояния между кластерами *«average»*

2. Отобразить результаты кластеризации (рис. 14)

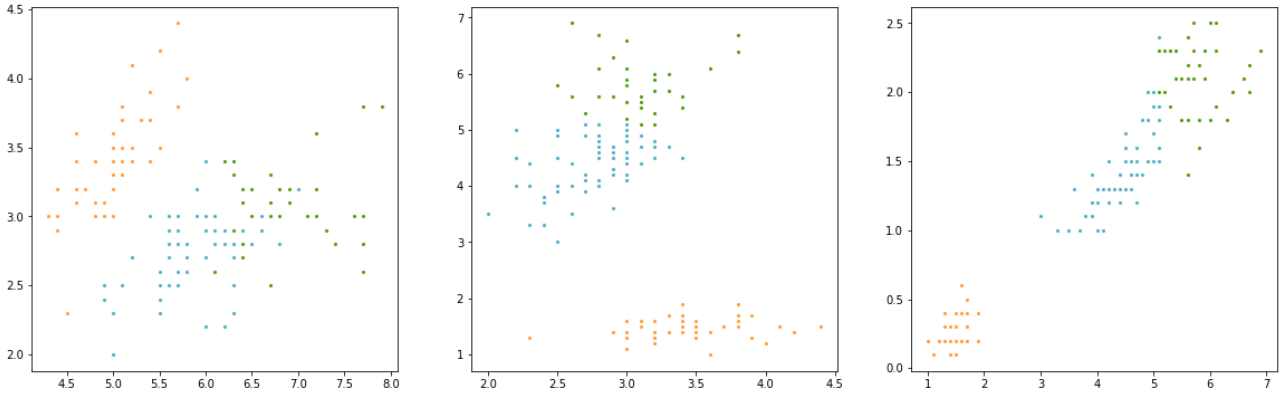

Рисунок 14. Разбиение на три кластера с помощью *AgglomerativeClustering*.

3. Исследовать алгоритм для различного размера кластеров.

   Проведена кластеризация на 2, 4, 5 кластеров (рис. 15, 16, 17 соотв.)

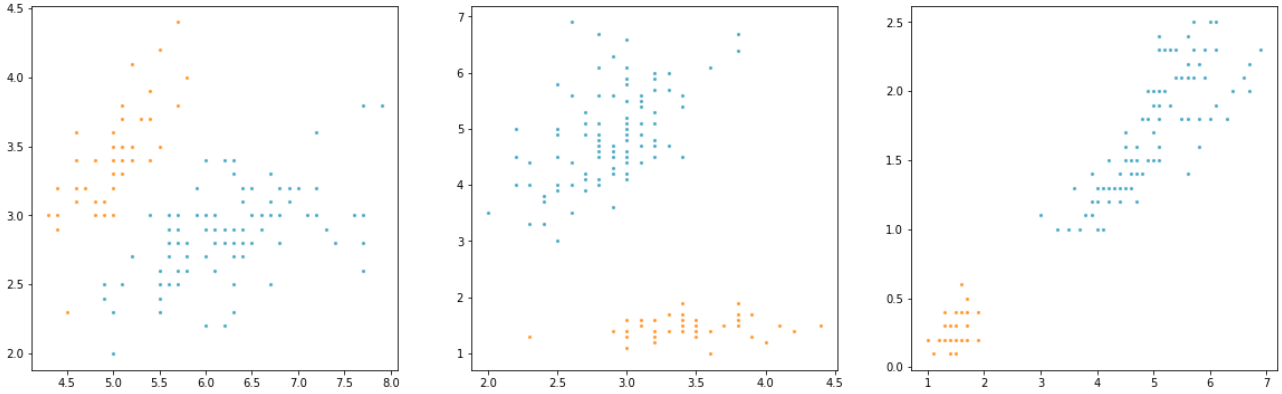

Рисунок 15. Разбиение на два кластера с помощью *AgglomerativeClustering*.

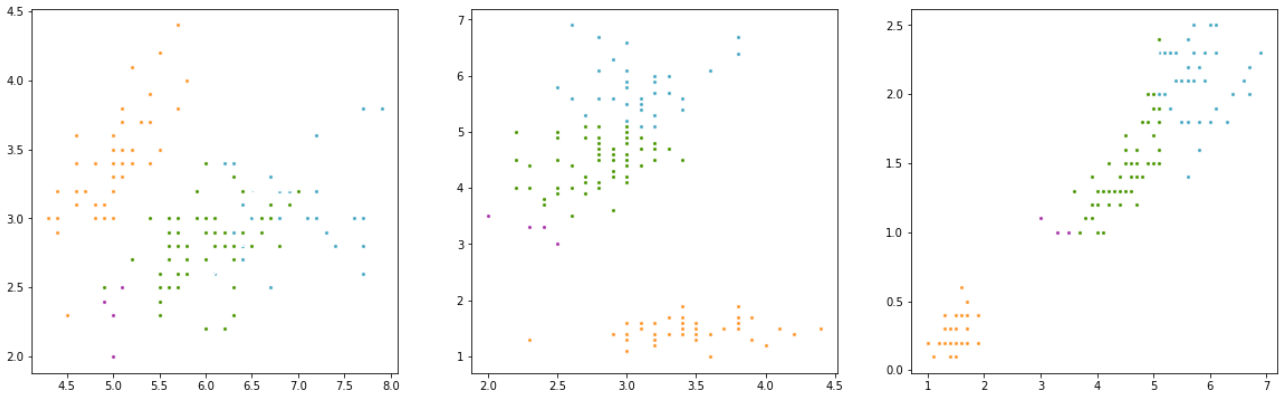

Рисунок 16. Разбиение на четыре кластера с помощью *AgglomerativeClustering*.

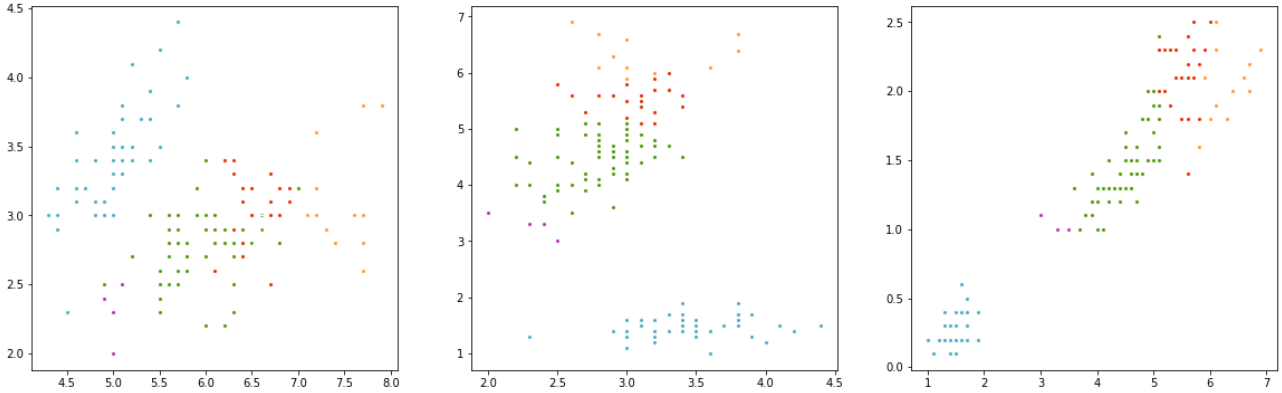

Рисунок 17. Разбиение на пять кластеров с помощью *AgglomerativeClustering*.

4. Дендограмма до шестого уровня (рис. 18)

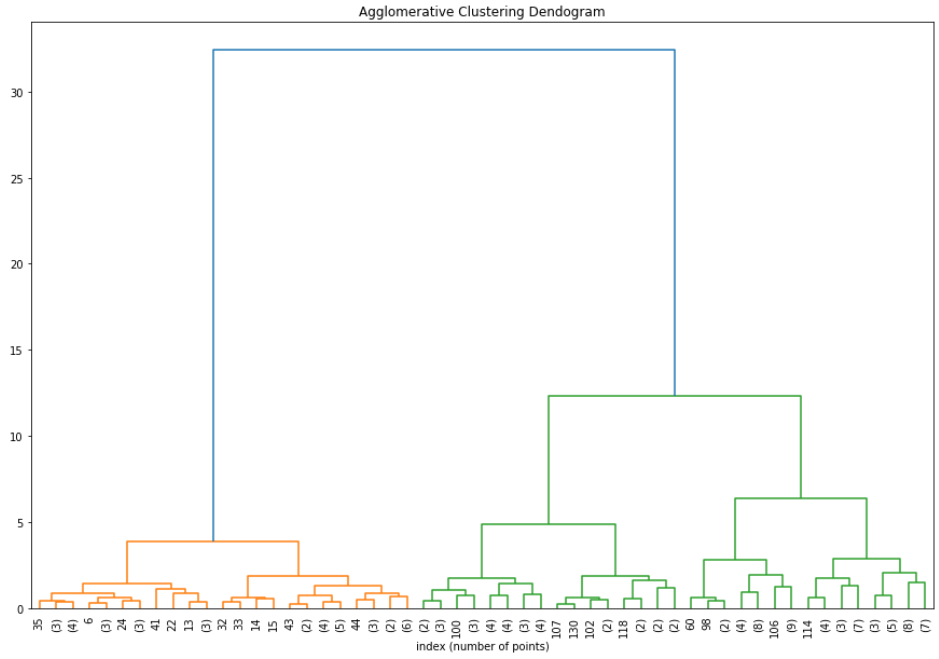

Рисунок 18. Дендограмма данных до шестого уровня.

По дендограмме с рисунка 18 можно заметить резкий рост расстояния между кластерами, когда их стало 3, соответственно кластеризация на три кластера является лучшим вариантом.

5. Сгенерированы случайные данные в виде двух колец (рис. 19).

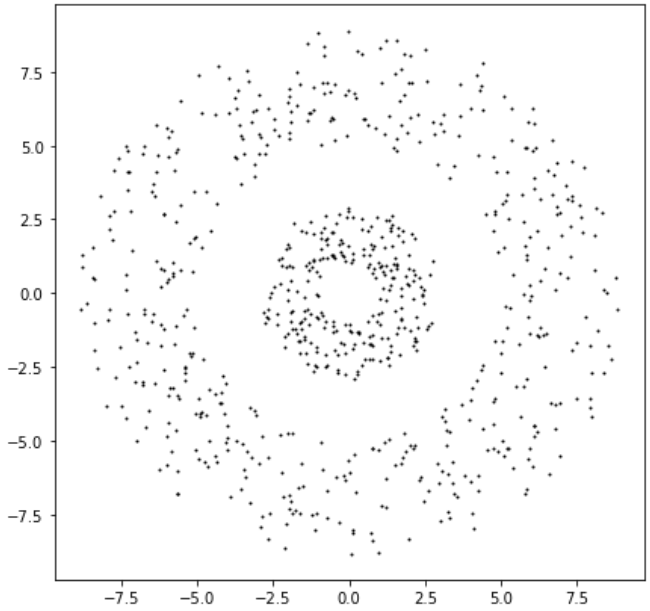

Рисунок 19. Сгенерированные данные в виде двух колец.

6. Проведена иерархическая кластеризация.
7. Результаты кластеризации (рис. 20).

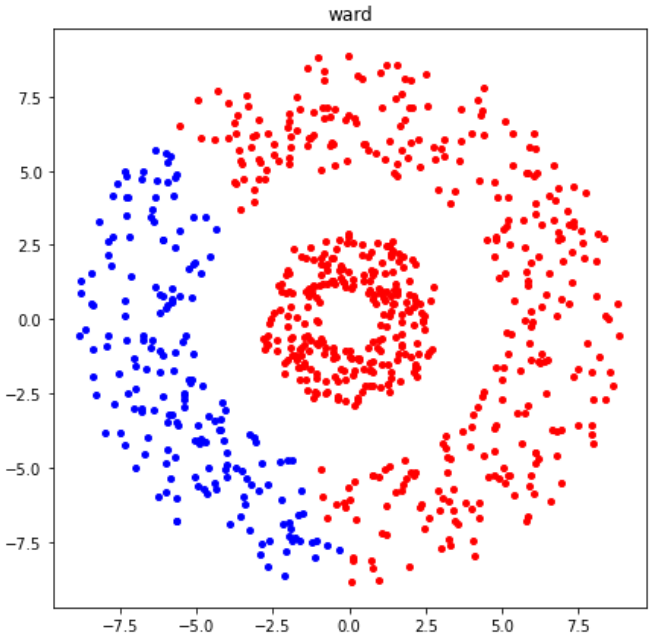

Рисунок 20. Результат кластеризации сгенерированных данных.

8. Исследование параметра *linkage* (рис. 21).

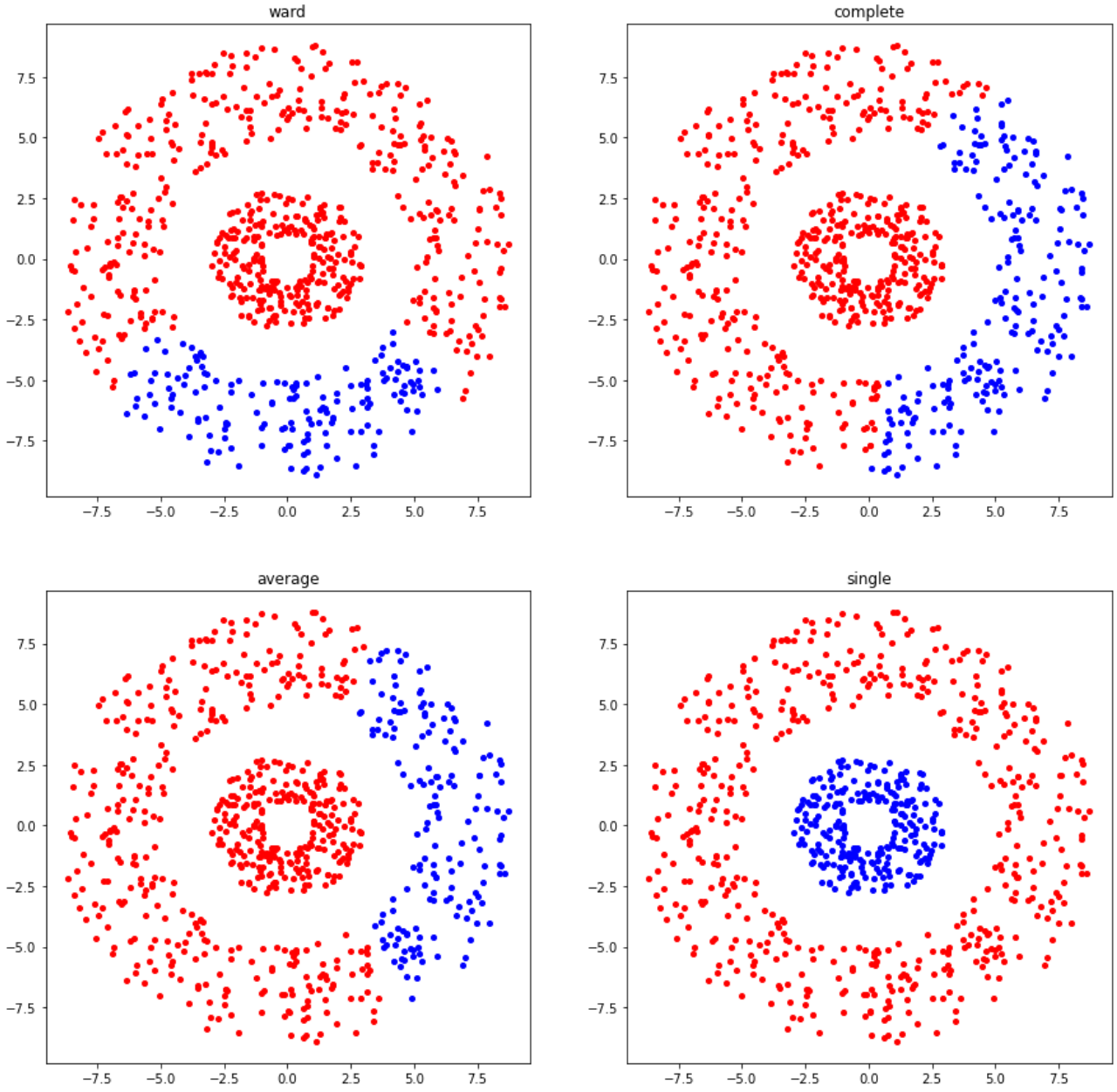

Рисунок 21. Кластеризация с различным параметром *linkage*.

Для сгенерированных данных лучший семантический результат показал метод «*single*».

Ward – опирается на дисперсию, а значит данные в кластере будут равномерные

Single – подходит для неглобулярных данных, однако очень неустойчив к зашумленным данным

Average – позволяет разбить на кластеры данные, которые имеют не-евклидовые показатели.

Complete – объединяет кластеры, крайние точки которых наименее удалены, это позволяет найти классы с четкими и ровными границами, но не обязательно компактными внутри.

Вывод

В ходе лабораторной работы были изучены методы кластеризации *K-Means* и *AgglomerativeClustering* (иерархическая кластеризация).

K-Means ищет классы опираясь на центроиды, что позволяет линейно разбить данные. Данный метод подходит для сконцентрированных вокруг определенных точек данных, в ином случае результат кластеризации может сильно зависеть от выбранных начальных точек.

AgglomerativeClustering производит наращивание кластеров путем их совмещения опираясь на различные методы подсчета расстояния. Выбор метода подсчета расстояния между кластерами напрямую зависит от формы поступающих данных.

Изучены способы определения оптимального количества кластеров: *Elbow-Method* для K-Means и *дендограмма* для иерархической кластеризации.