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

ОТЧЁТ

по лабораторной работе №5
по дисциплине «Машинное обучение»
Тема: Кластеризация (k-средних, иерархическая)

Студент гр. 6304 _____ Корытов П.В.
Преподаватель _____ Жангиров Т.Р.

Санкт-Петербург
2020

1. Цель работы

Ознакомиться с методами кластеризации модуля Sklearn.

2. Выполнение

2.1. K-means

1. Проведена загрузка данных. Часть данных представлена на листинге 1.

Листинг 1. Набор данных

```
      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

[150 rows x 5 columns]
```

2. Проведена кластеризация методом k-средних. Результаты представлены попарно для 4-х признаков на рис. 1.

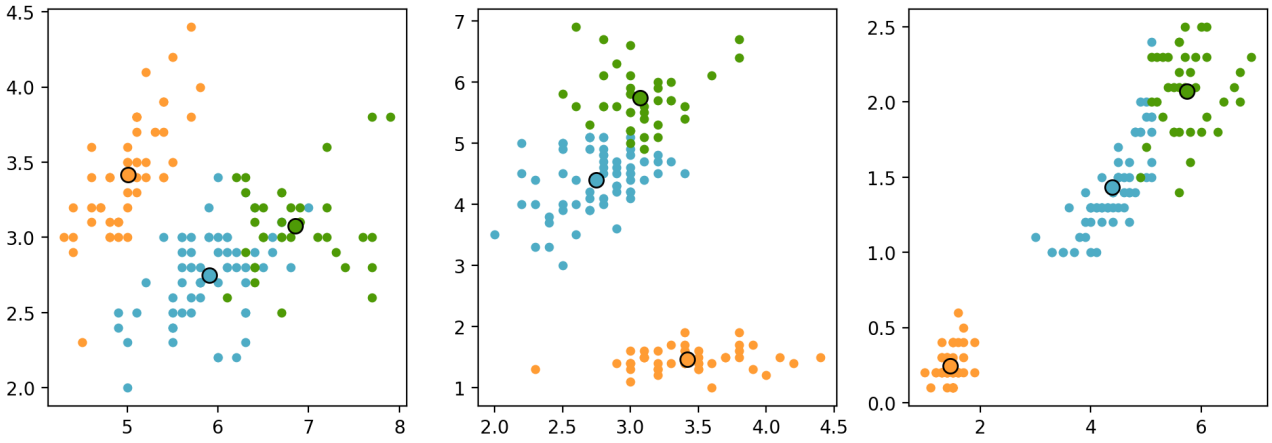

Рисунок 1 – Попарные результеты

Исходя из рисунка, наилучшее разделение прошло по признакам 3 и 4. Параметр `n_init` в данном случае не оказал видимого влияния на результаты.

3. Произведено уменьшение размерности данных до 2 через PCA и составлена карта области значений, на которой каждый кластер занимает определенную область со своим цветом. Результаты на рис. 2.

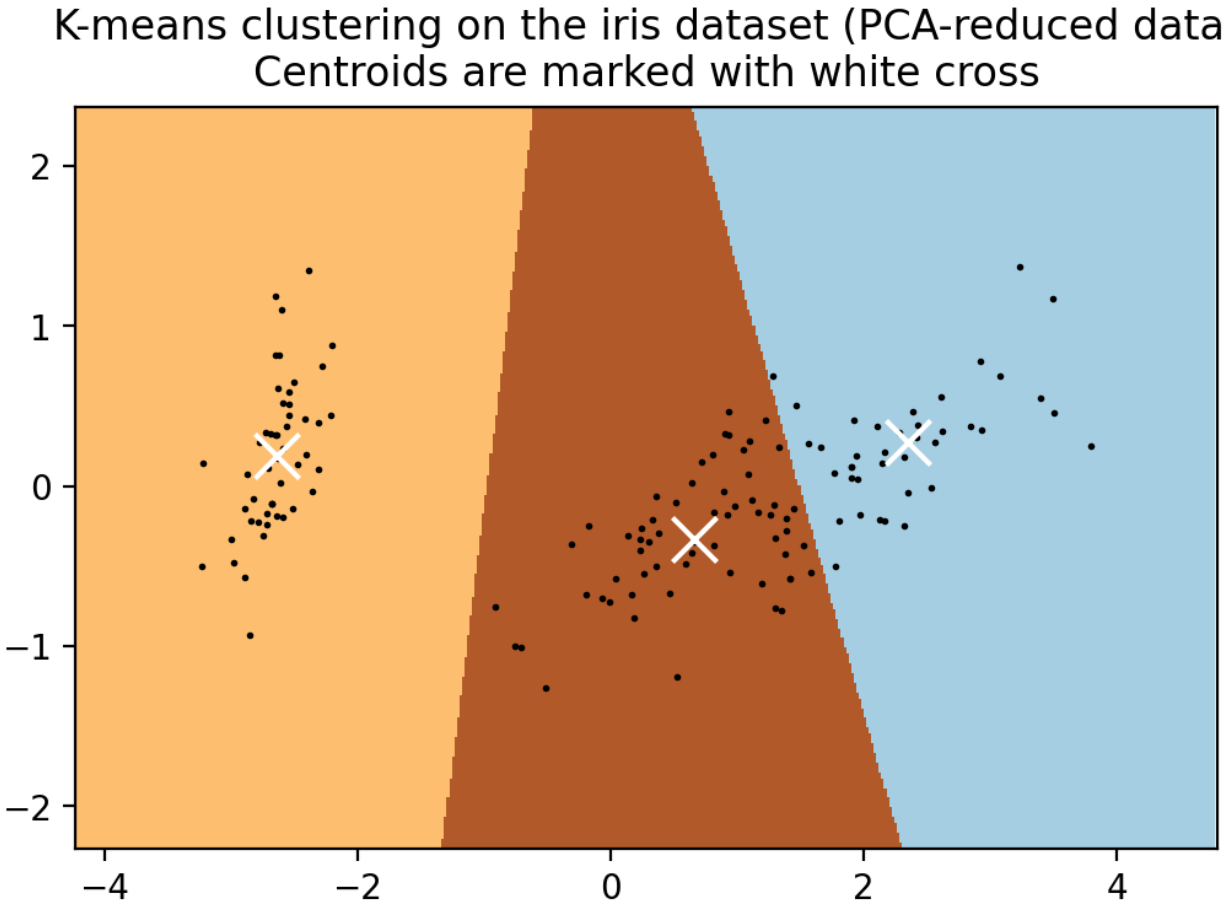

Рисунок 2 – Карта области значений с уменьшением размерности

4. Исследована работа алгоритма для различных параметров `init`. Результаты представлены на рис. 3.

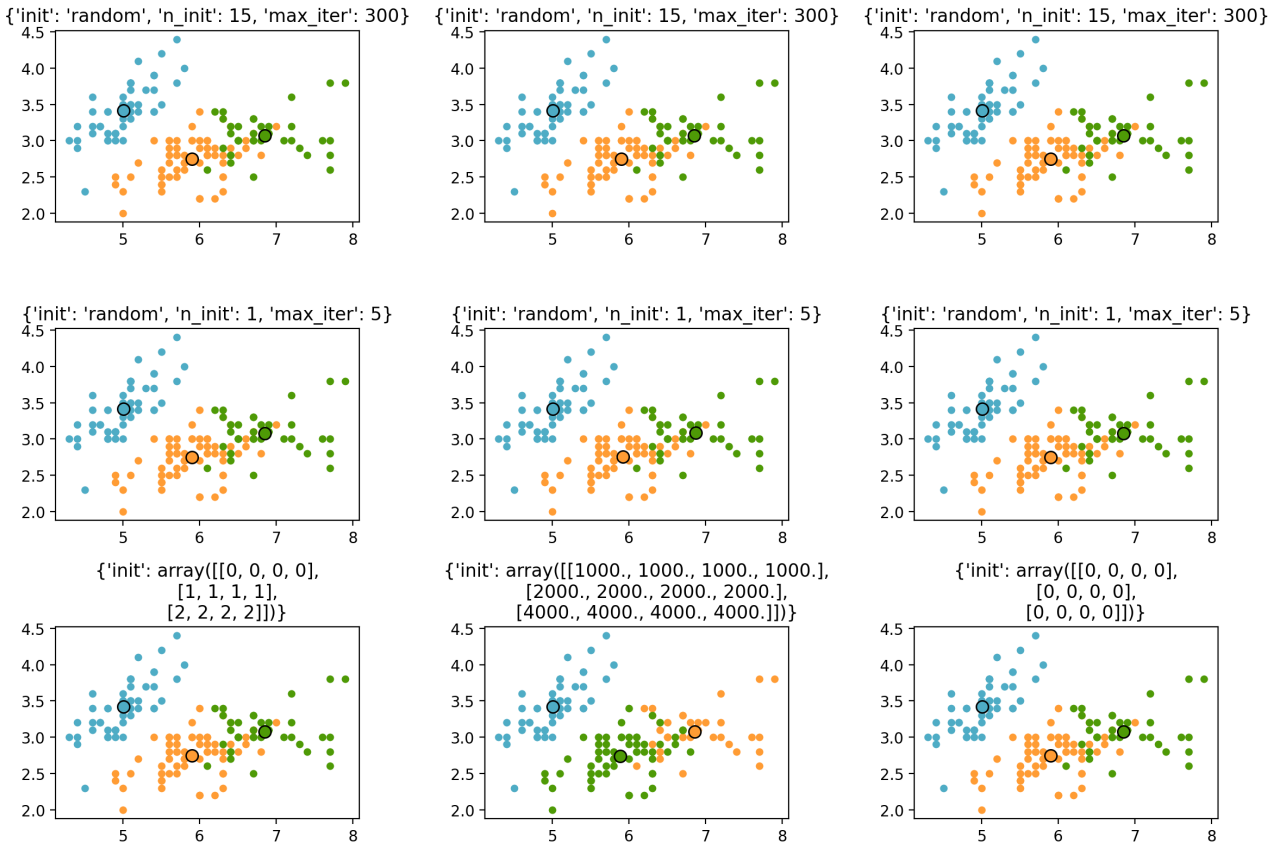

Рисунок 3 – Исследование init

Как можно заметить, `n_init=1` привел к смещению центров кластеров в случае `max_iter=5`. Ручное задание точек к видимым измененим не привело, за исключением изменения порядка следования меток.

5. Определено наилучшее количество кластеров методом локтя. Результаты представлены на рис. 4.

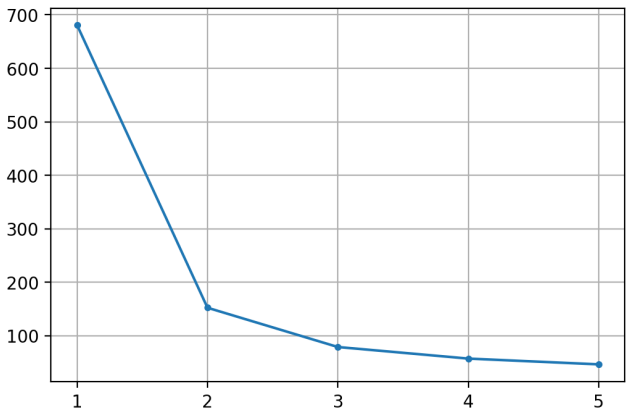

Рисунок 4 – Метод локтя

Как можно заметить, по этому методу, наилучшее количество кластеров — 2.

6. Проведена кластеризация с использованием пакетной кластеризации методом k-средних. Отличия в результатах относительно обычного метода представлены на рис. 5.

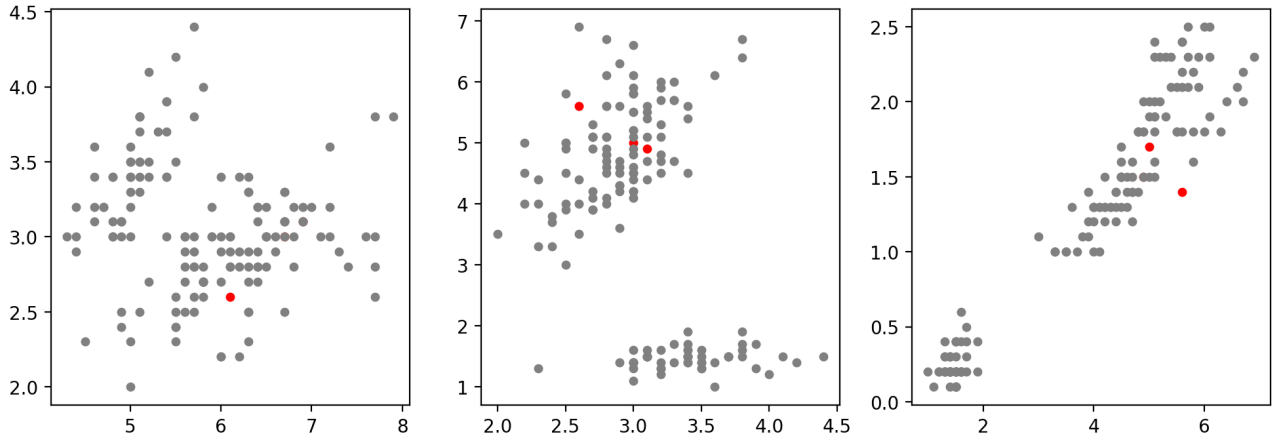

Рисунок 5 – Отличия в результатах между `KMeans` и `MiniBatchKMeans` (красные точки)

Отличие между двумя методами в том, что во втором случае алгоритму на вход подаются пакеты данных, а не полный набор. Это приводит к увеличению скорости работы, но и к некоторому снижению точности.

2.2. Иерархическая кластеризация

1. Проведена иерархическая кластеризация на тех же данных. Результаты представлены на рис. 6.

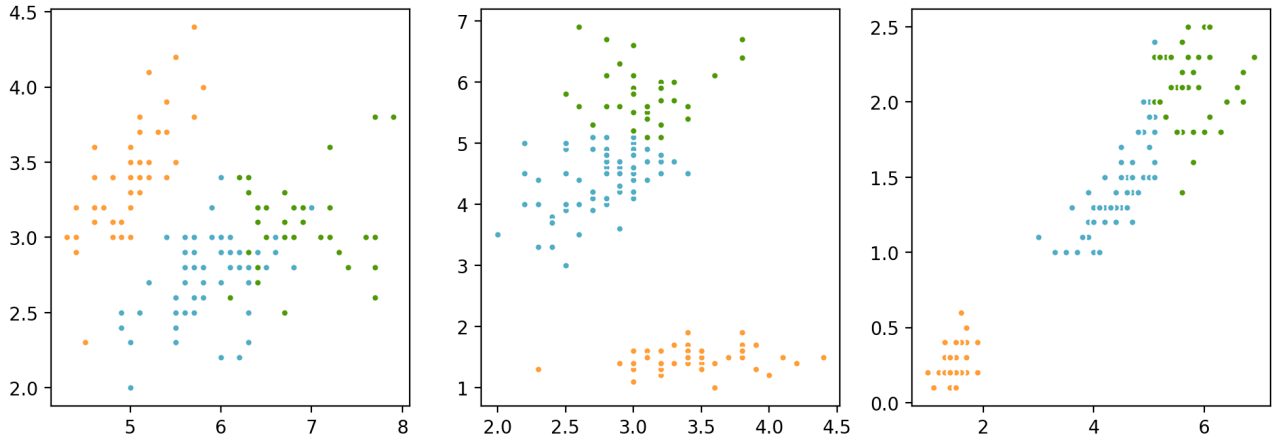

Рисунок 6 – Результаты иерархической кластеризации

Отличие `AgglomerativeClustering` от `KMeans` в алгоритме. Первый начинает с состояния, где все точки принадлежат своему кластеру из одной

этой точки, и объединяет ближайшие кластеры на основе какой-либо метрики. Таким образом, начальный и конечный результат (кластер из всех точек) — тривиальны, но посередине могут оказаться интересные варианты.

2. Проведено исследования для различного размера кластеров. Результаты на рис. 7.

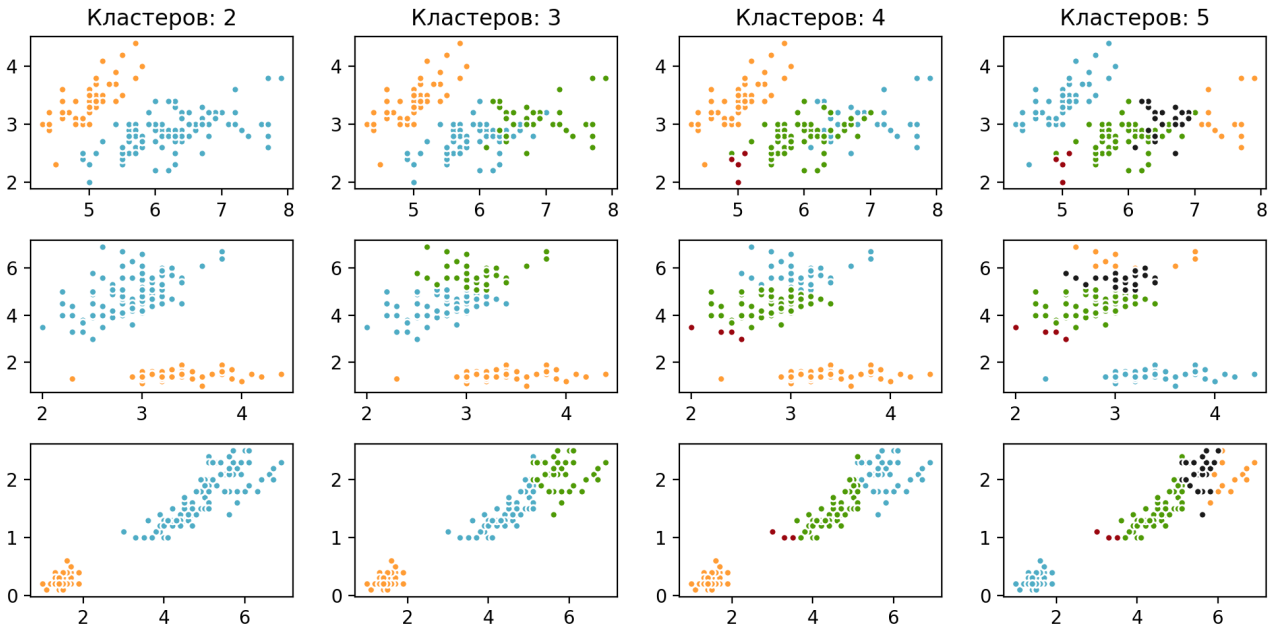

Рисунок 7 – Результаты для различного размера кластеров

3. Нарисована дендрограмма. Результаты на рис. 8.

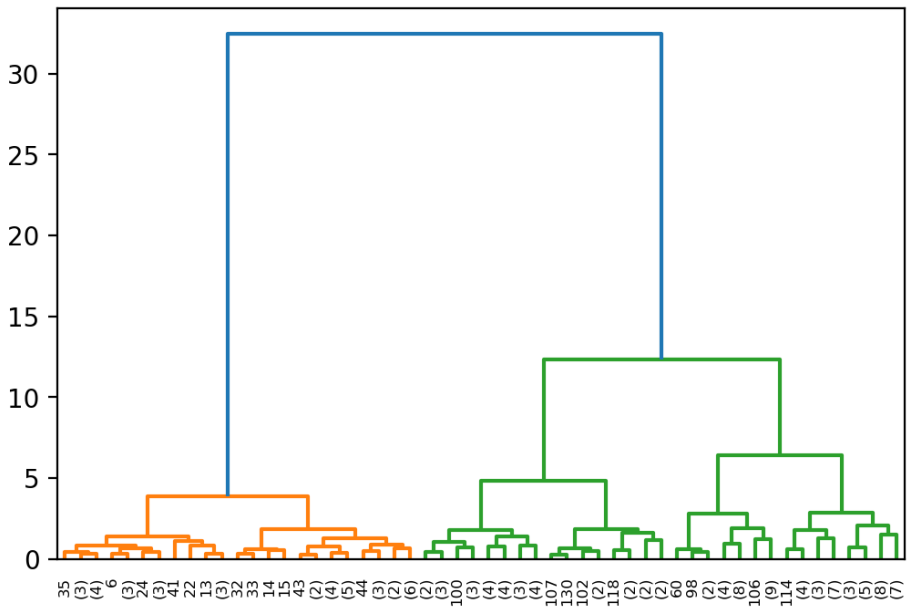

Рисунок 8 – Дендрограмма

4. Сгенерированы случайные данные в виде двух колец. Проведена иерархическая кластеризация с методом Уорда для определения расстояния. Результаты на рис. 9.

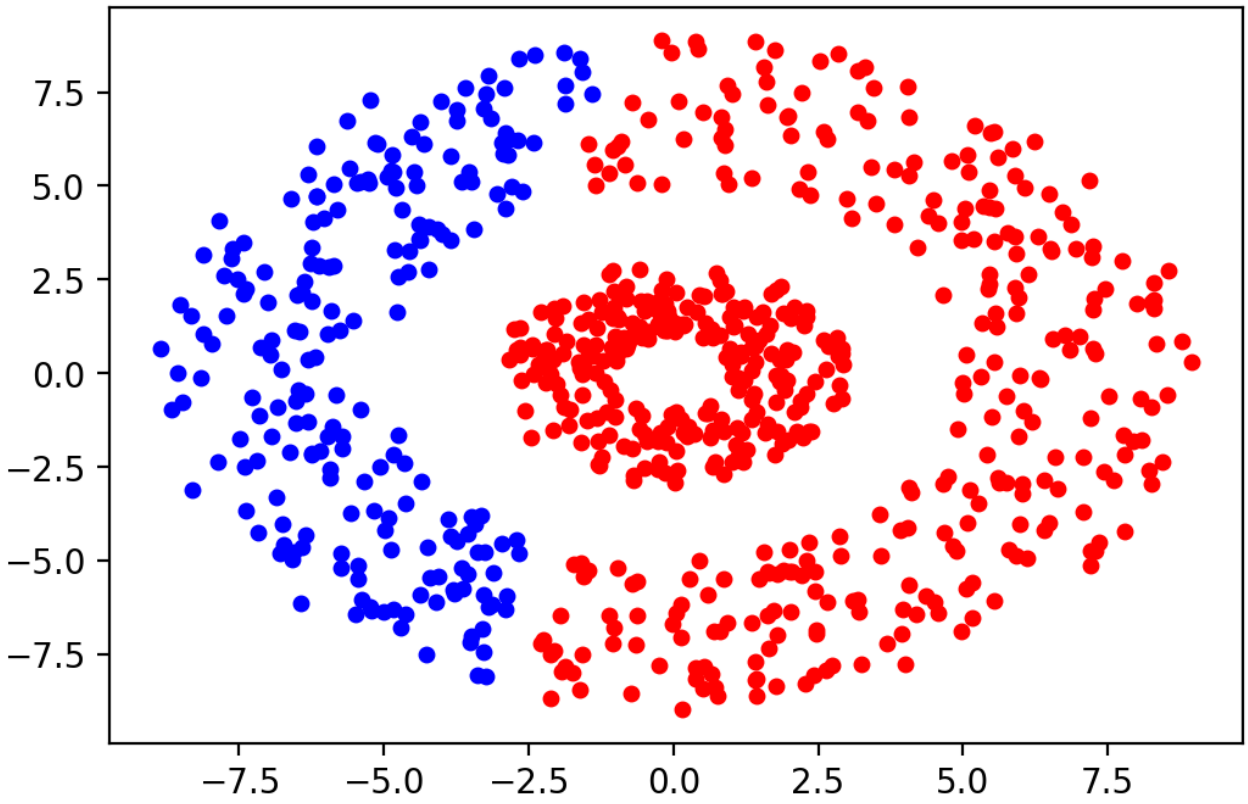

Рисунок 9 – Кольца

5. Проведено исследования для различных параметров `linkage`

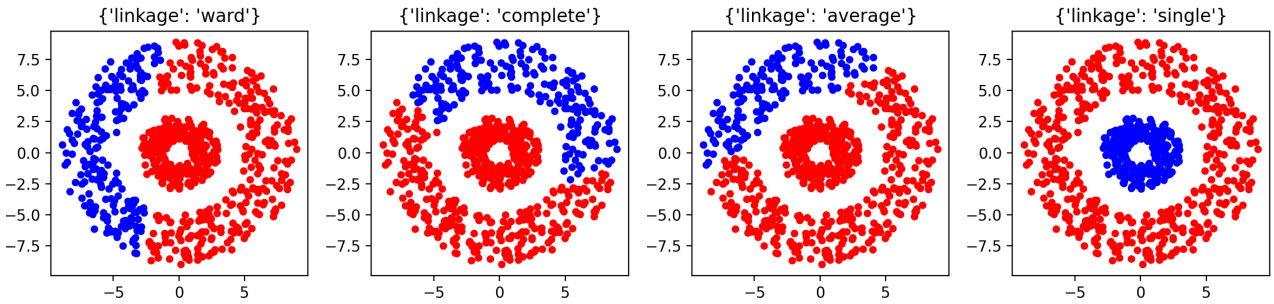

Рисунок 10 – Иерархическая кластеризация при различных параметрах linkage

Во всех случаях использовалась Евклидова дистанция.

Как видно из рисунка, разделение колец произошло только по Single Link, т.е. с расстоянием между кластерами как расстоянием между ближайшими точками.

В случае Complete Link, Group Average или метода Уорда расстояние между кластерами, лежащими на разных кольцах, оказывается меньше, чем между кластерами на одном кольце.

3. Выводы

Произведено знакомство с методами кластеризацией методом k-средних и иерархической кластеризацией в модуле Sklearn.

Использование пакетного метода k-средних приводит к небольшим изменениям результата в сравнении с полным k-средних.

Метод иерархической кластеризации при правильной настройке смог определить нелинейную зависимость между синтетическими данными.