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 _____ Иванов Д.В.

Преподаватель _____ Жангиров Т. Р.

Санкт-Петербург

2020

Цель работы

Ознакомиться с методами кластеризации модуля Sklearn.

Загрузка данных

1. Датасет скачан и загружен в датафрейм.

```
data = pd.read_csv('iris.data',header=None)
```

K-means

1. Проведена кластеризация методом k-средних.

```
k_means = KMeans(init='k-means++', n_clusters=3, n_init=15)
k_means.fit(no_labeled_data)
```

2. Получены центры кластеров и определены какие наблюдения в какой кластер попали.

```
k_means_cluster_centers = k_means.cluster_centers_
k_means_labels = pairwise_distances_argmin(no_labeled_data,
k_means_cluster_centers)
```

3. Построены результаты классификации для признаков попарно (1 и 2, 2 и 3, 3 и 4)

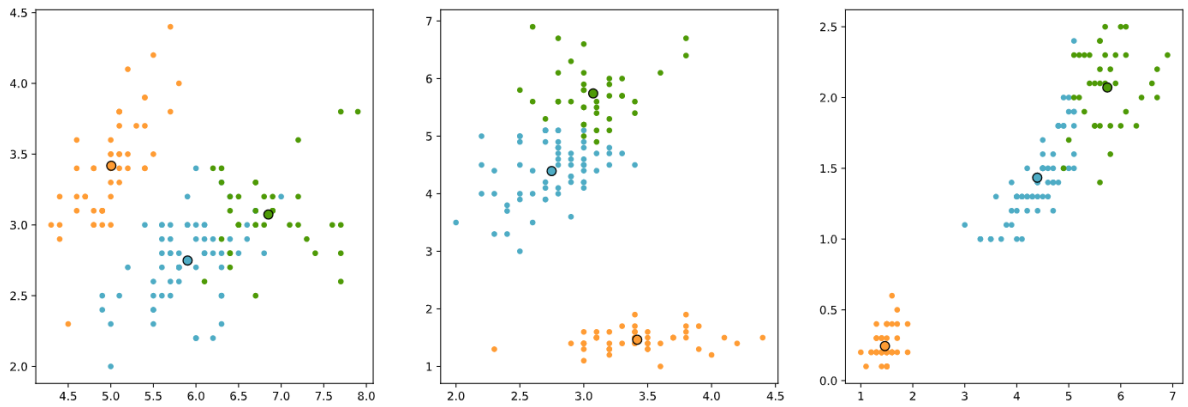

Рис. 1 — Результаты классификации признаков

Можно сделать вывод, что наилучшее разделение произошло по признакам 3 и 4. Различные значения параметра *n_init* не имеют видимых результатов.

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

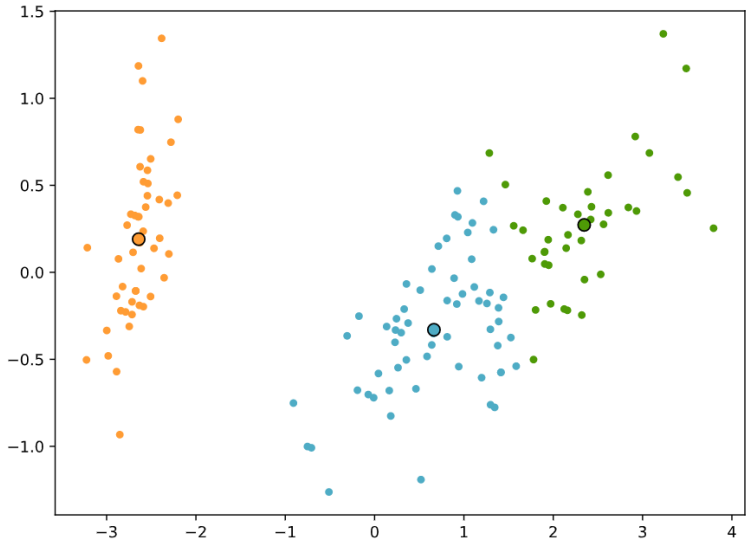

Рис. 2 — Результаты классификации с уменьшенной размерностью данных

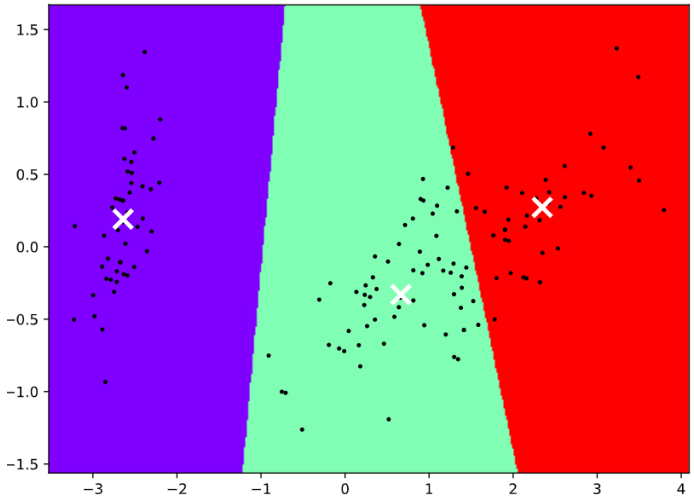

Рис. 3 — Карта области значения

5. Исследована работа алгоритма k-средних при различных параметрах init: сначала алгоритм запущен несколько раз с параметром ‘random’, затем для вручную выбранных точек.

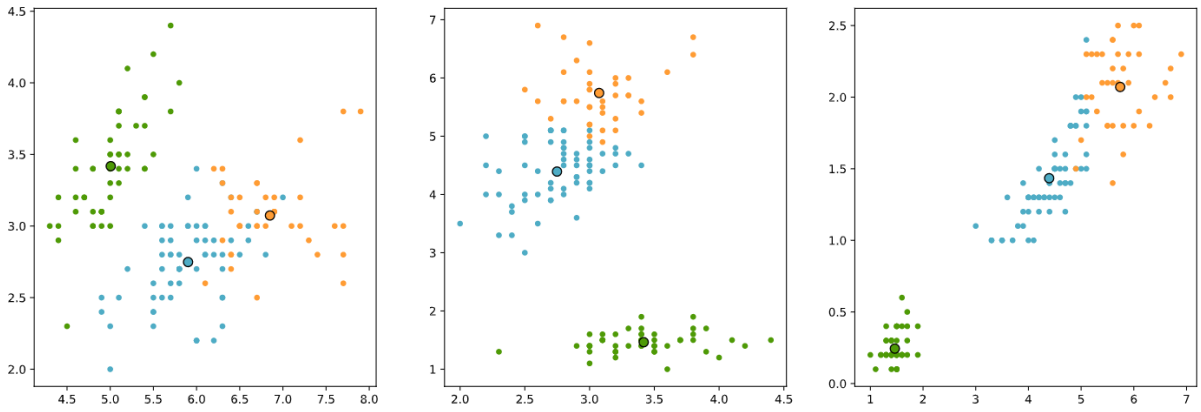

Рис. 4 — init = ‘random’, max_iter = 1

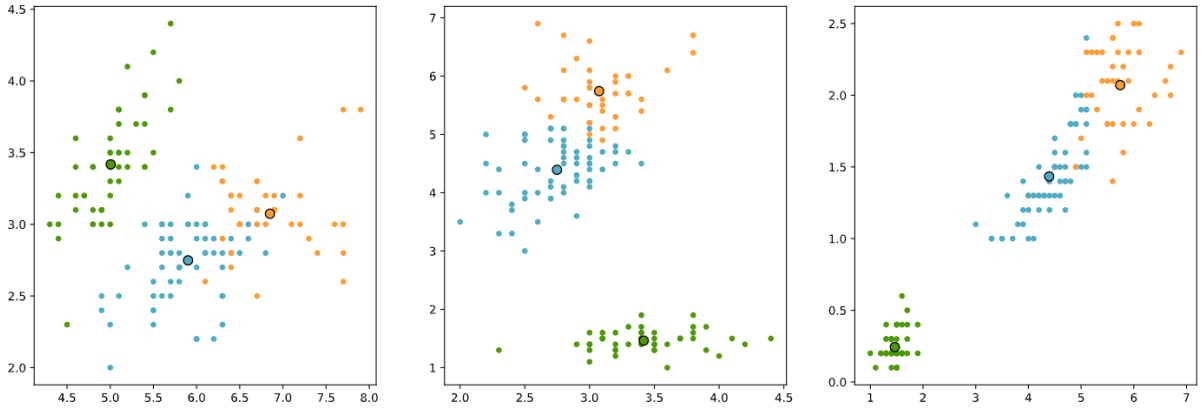

Рис. 5 — init = ‘random’, max_iter = 5

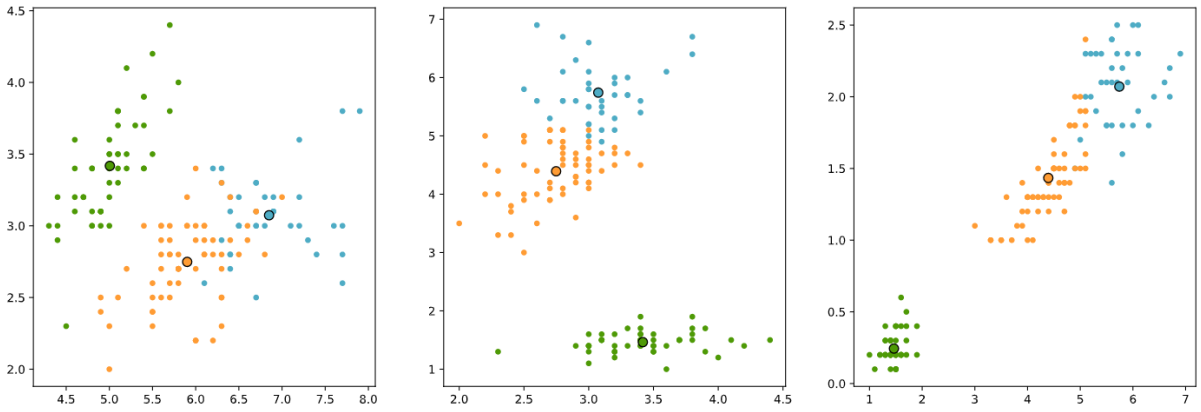

Рис. 6 — init = ‘random’, max_iter = 100

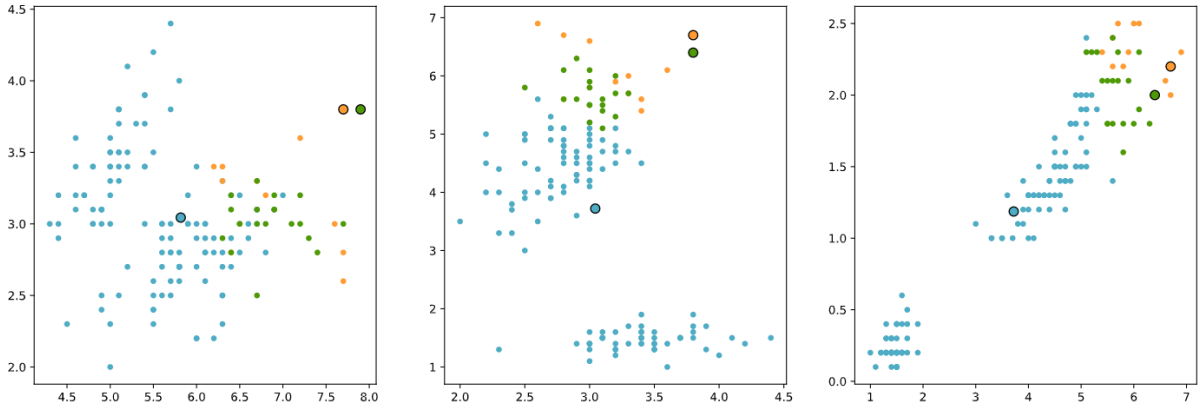

Рис. 7 — init=np.array([[0,0,0,0],[0,0,0,0],[0,0,0,0]]), max_iter = 1

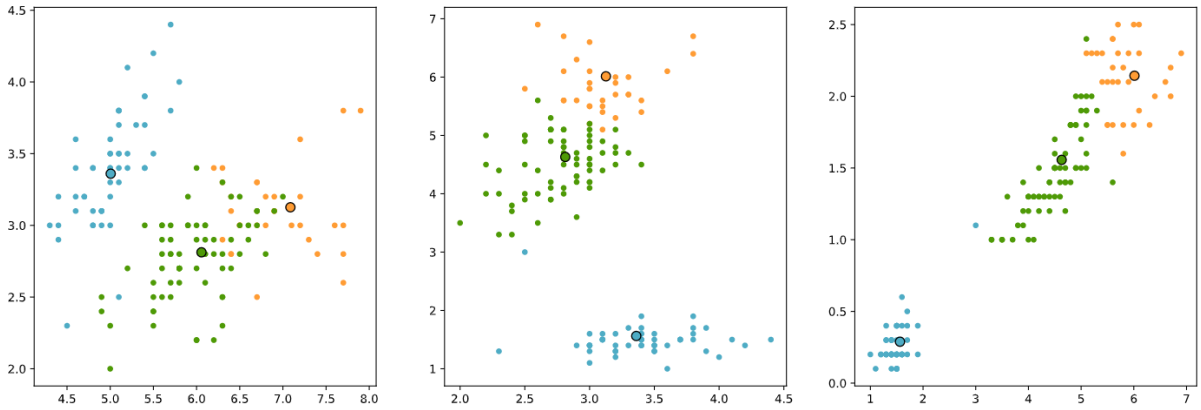

Рис. 8 — init=np.array([[0,0,0,0],[0,0,0,0],[0,0,0,0]]), max_iter = 5

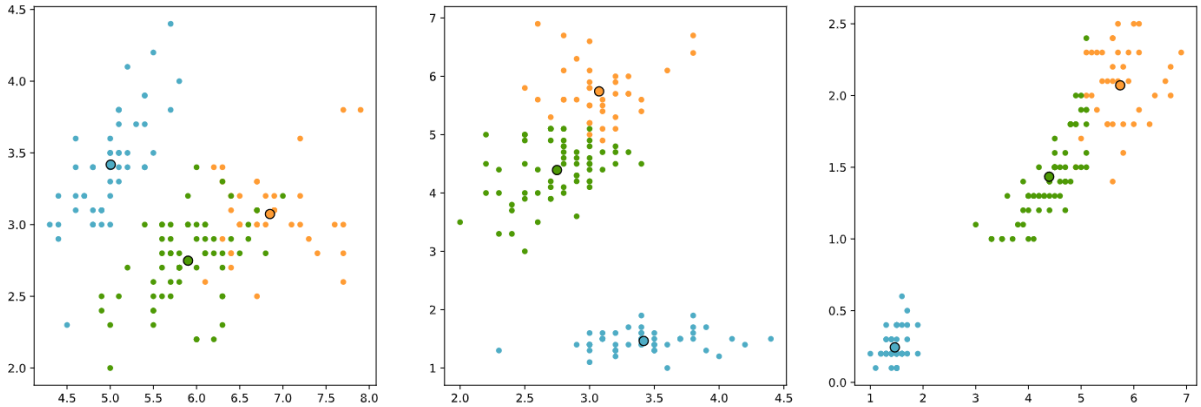

Рис. 9 — init=np.array([[0,0,0,0],[0,0,0,0],[0,0,0,0]]), max_iter = 100

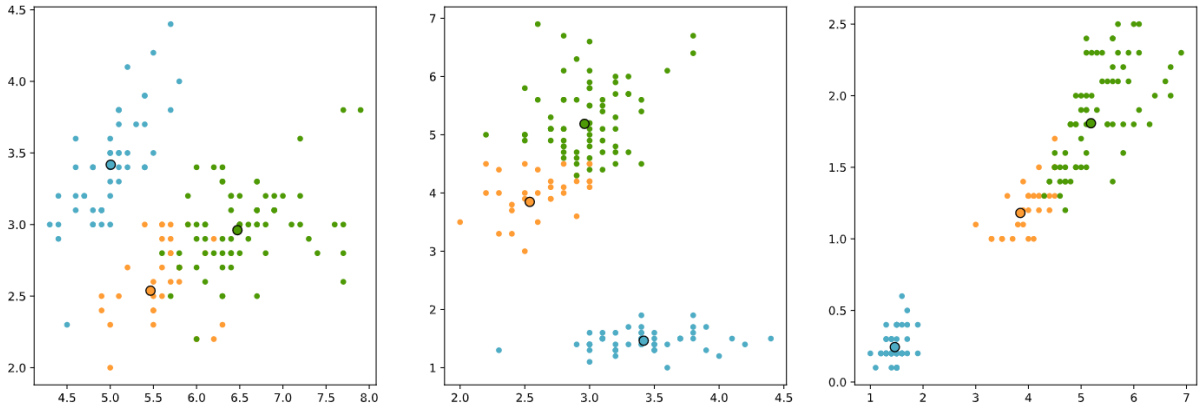

Рис. 10 — init=np.array([[5,3,1,0],[5,2,4,1],[6,3,5,2]]), max_iter = 1

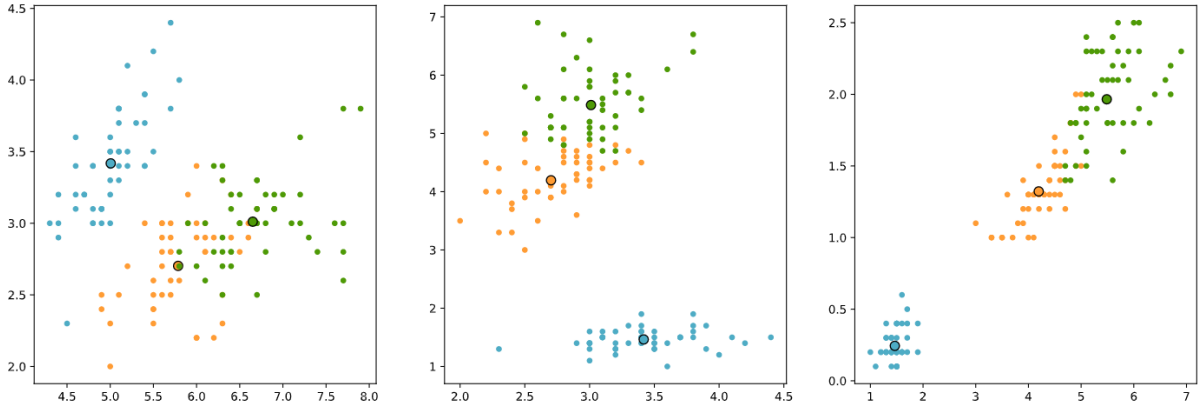

Рис. 11 — init=np.array([[5,3,1,0],[5,2,4,1],[6,3,5,2]]) , max_iter = 5

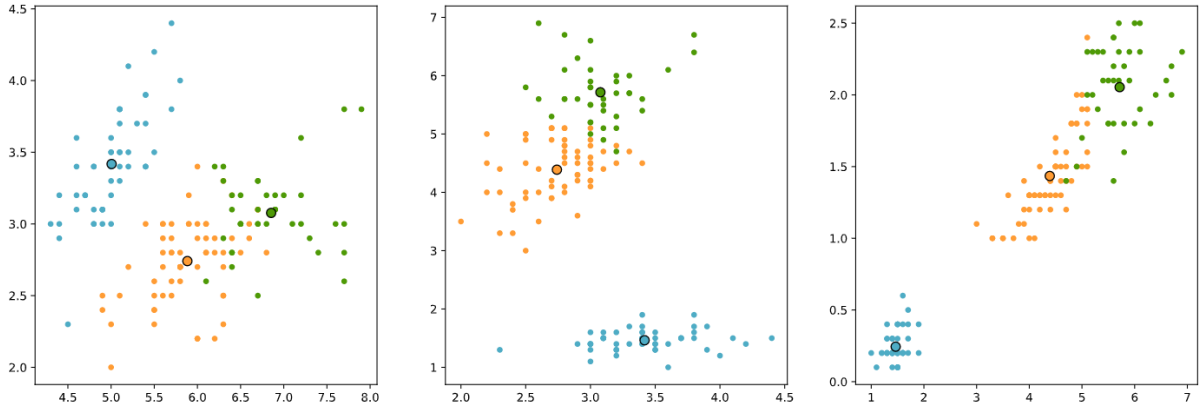

Рис. 12 — init=np.array([[5,3,1,0],[5,2,4,1],[6,3,5,2]]), max_iter = 100

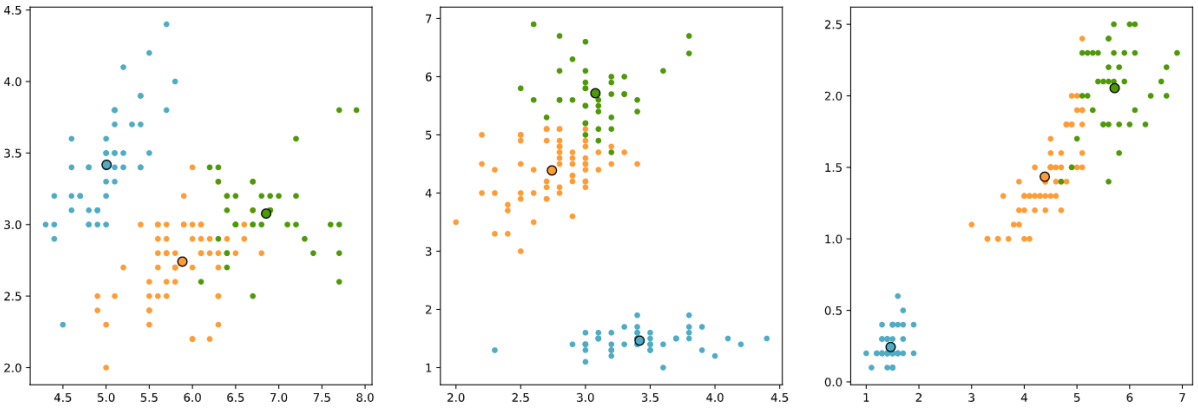

Рис. 13 — init=np.array([[5.0,3.4,1.5,0.2],[5.8,2.2,4.4,1.5],[6.8,3.1,5.9,2.2]]), max_iter = (1 or 5)

6. Методом локтя определено наилучшее количество кластеров.

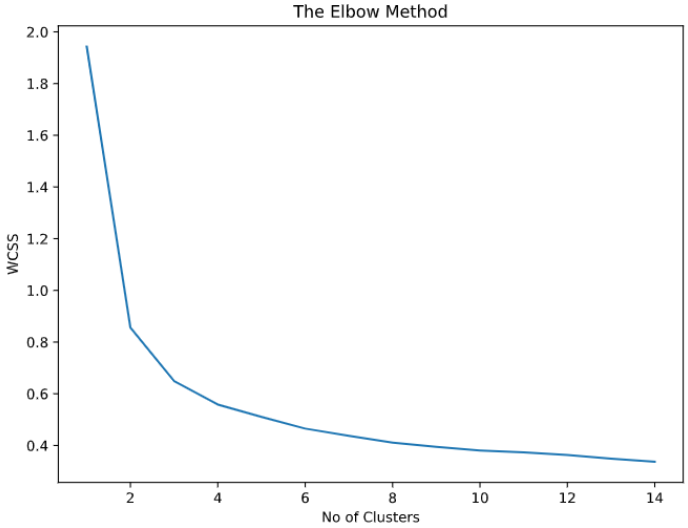

Рис. 14 — Метод локтя

7. Проведена кластеризация с использованием пакетной кластеризации k-средних. Построена диаграмма рассеяния, на которой выделены точки, которые для разных методов попали в разные кластеры. Методы различаются тем, что MiniBatchKMeans на вход подаются пакеты данных, а не полный набор: это увеличивает скорость работы, но снижает точность.

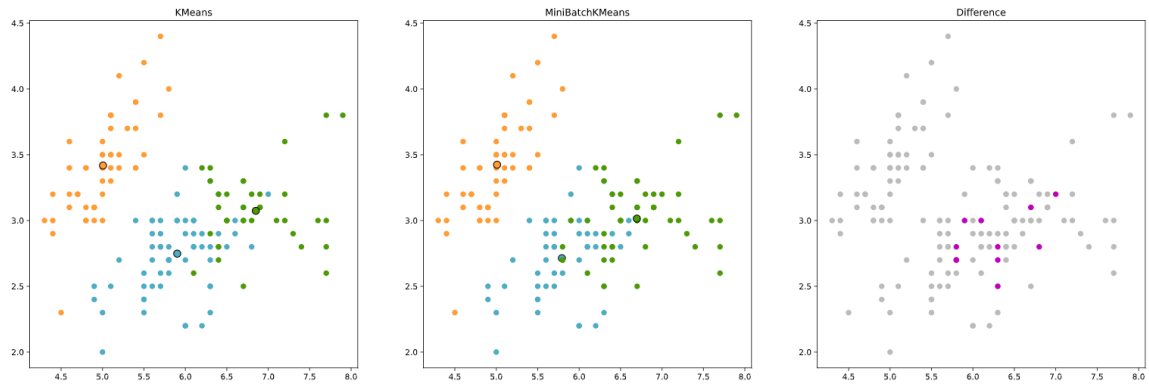

Рис. 14 — Отличия результатов KMeans и MiniBatchKMeans (фиолетовый цвет)

Иерархическая кластеризация

1. На тех же данных проведена иерархическая кластеризация (рис. 15). AgglomerativeClustering отличается от KMeans алгоритмом: изначально все точки принадлежат собственному кластеру, состоящему из одной точки, алгоритм объединяет ближайшие кластеры на основе выбранной метрики.

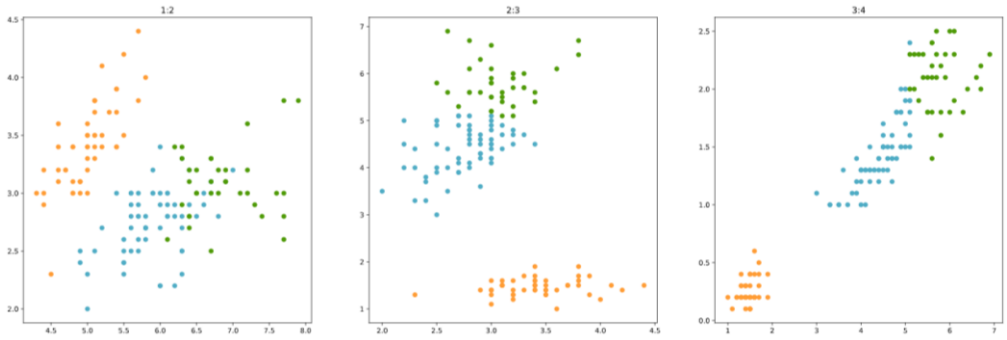

Рис. 15 — Результат иерархической кластеризации

2. Проведено исследование для различного количества кластеров.

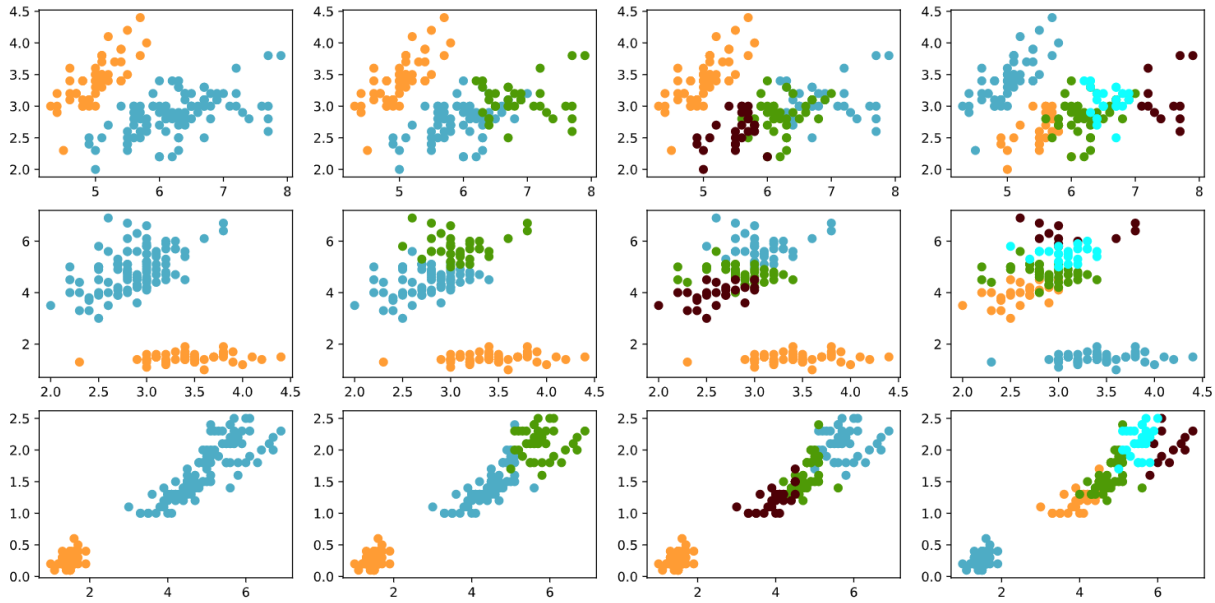

Рис. 16 — Результаты для различного количества кластеров

3. Нарисована дендрограмма до уровня 6.

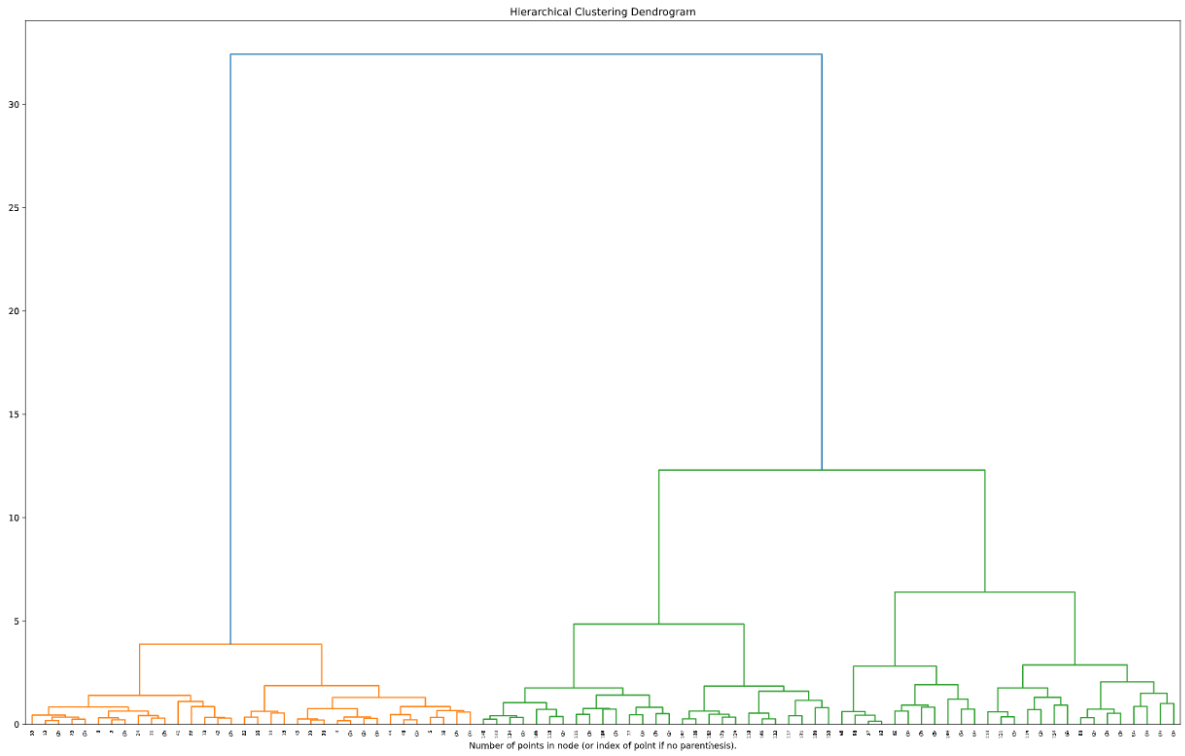

Рис. 17 — Дендрограмма

4. Сгенерируйте случайные данные в виде двух колец.

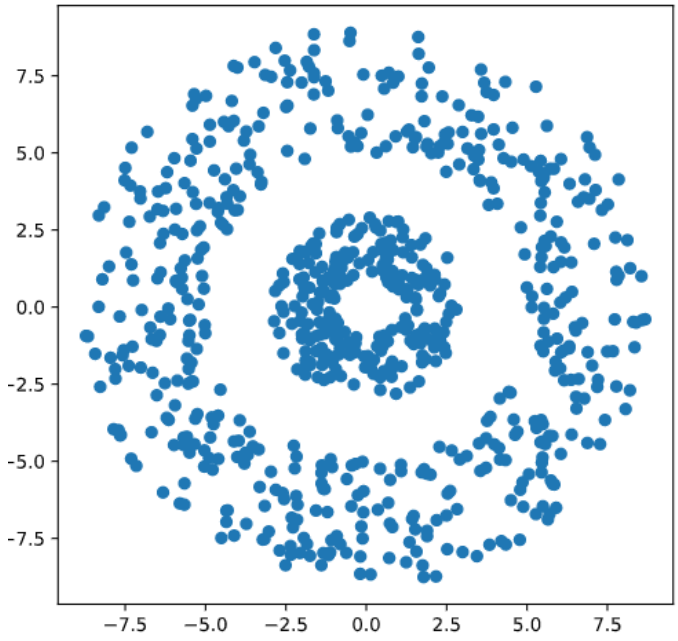

Рис. 18 — Сгенерированные данные

5. Проведена иерархическая кластеризация при использовании метрики Уорда.

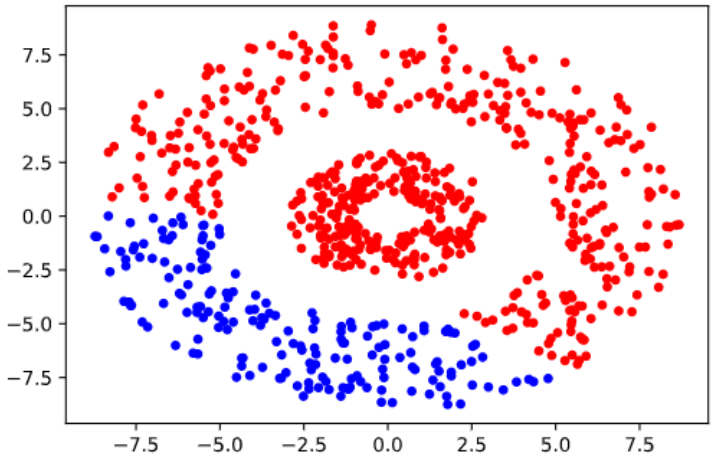

Рис. 19 — Результат иерархической кластеризации

6. Исследована кластеризация при всех параметрах linkage.

- Ward – минимизация суммы квадратов разностей
- Complete – минимизация максимального расстояния
- Average – минимизация среднего расстояния
- Single – минимизация расстояния

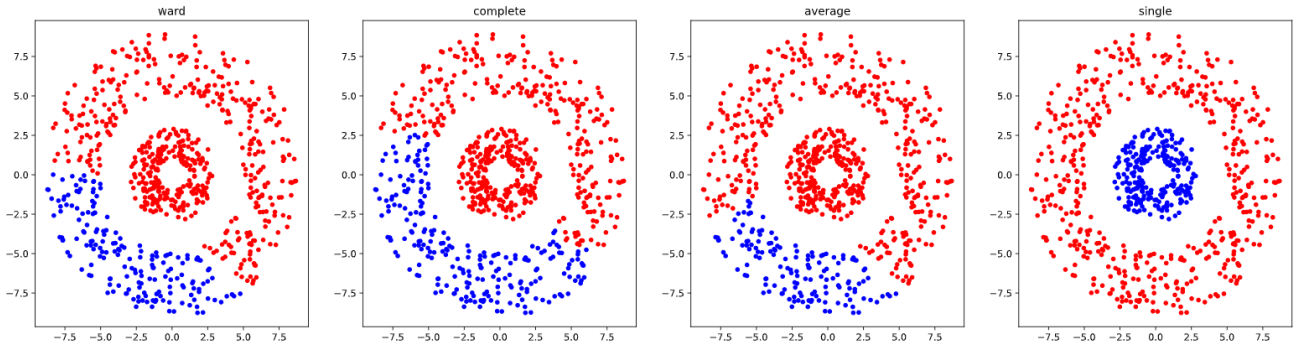

Рис. 20 — Иерархическая кластеризация при различных параметрах linkage

По результатам видно, что разделение колец произошло только при использовании метрики Single.

Вывод

Произведено знакомство с кластеризацией методом k-средних и иерархической кластеризацией в модуле Sklearn. Пакетный метод k-средних имеет несколько иной результат, отличающийся от результата стандартного. При правильном выборе меры иерархическая кластеризация определяет нелинейную зависимость между синтезированными данными.