**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

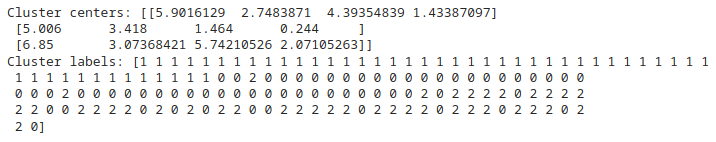
**Цель работы**

Ознакомиться с методами кластеризации модуля Sklearn.

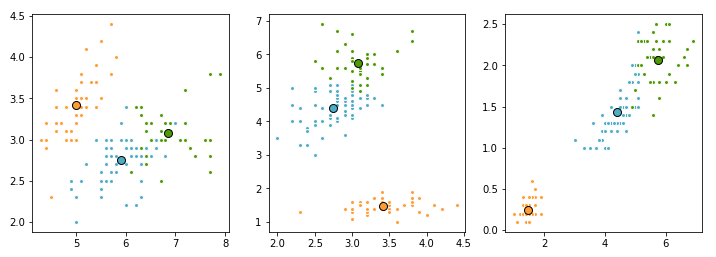
**Ход работы**

* 1. **Загрузка данных.**

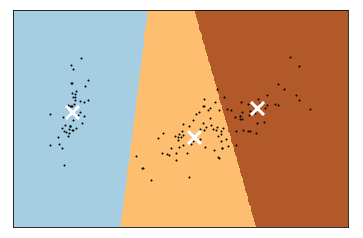
1. Был загружен по ссылке требуемый датасет.
2. **K-means.**
3. Была проведена кластеризация методом k-средних.
4. Были получены центры кластеров и определено, какие наблюдения попали в какой кластер.

Рисунок 1 - Центры и значения кластеров

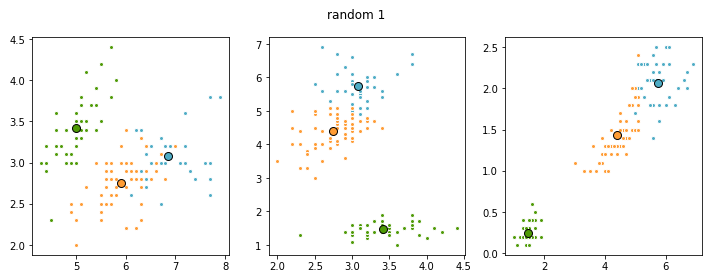
1. Были построены результаты классификации для признаков попарно (1 и 2, 2 и 3, 3 и 4). Кластеризация для каждого случая произведена одинаково эффективно. Параметр n\_init влияет на количество итераций алгоритма.

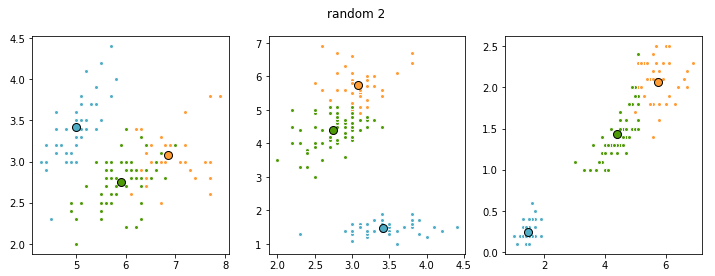
Рисунок 2 - Обработка данных методом k-средних

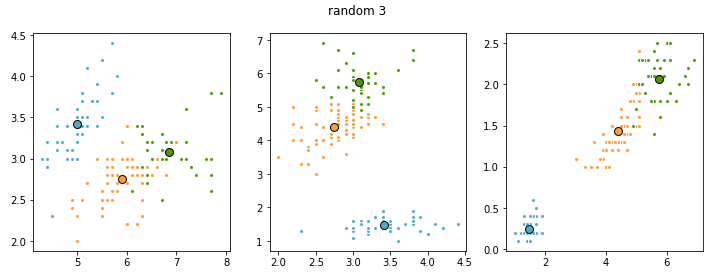
1. Размерность данных была уменьшена до 2 с использованием метода главных компонент, была нарисована карта для всей области значений, на которой каждый кластер занимает определенную область со своим цветом.

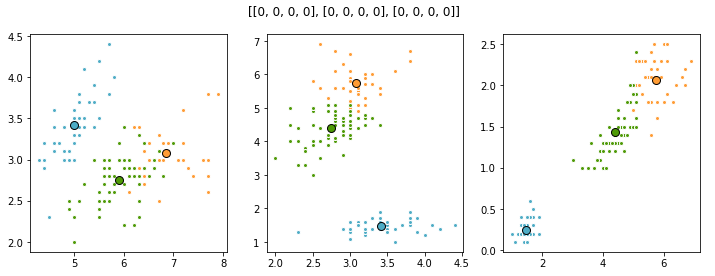
Рисунок 3 - Полученные кластеры

1. 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

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

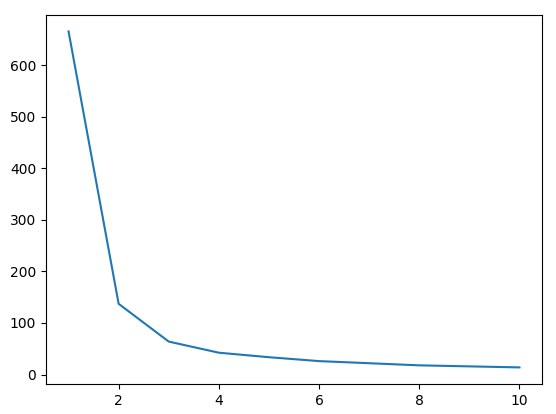
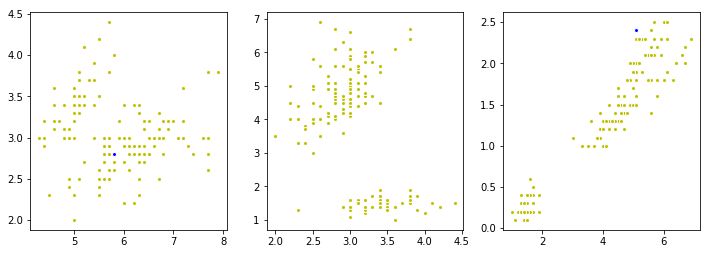
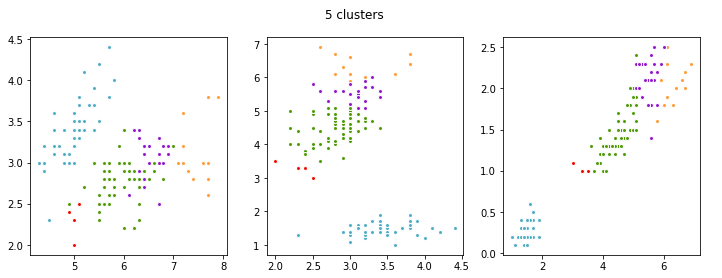


Рисунок 8 - Метод локтя

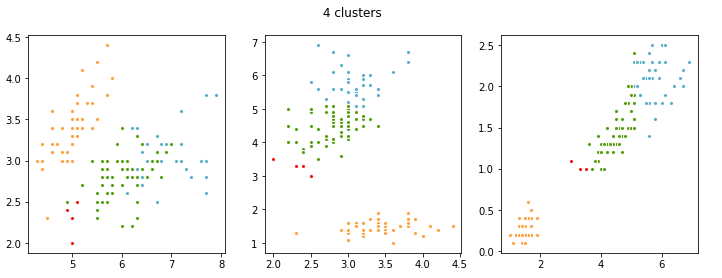
1. Была проведена кластеризация с использованием пакетной кластеризаций k-средних. Данный метод осуществляет разделение путем выбора только части значений и вычисления кластеров с их помощью. Это повышает быстродействие алгоритма. Была построена диаграмма рассеяния, на которой синим цветом выделены точки,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

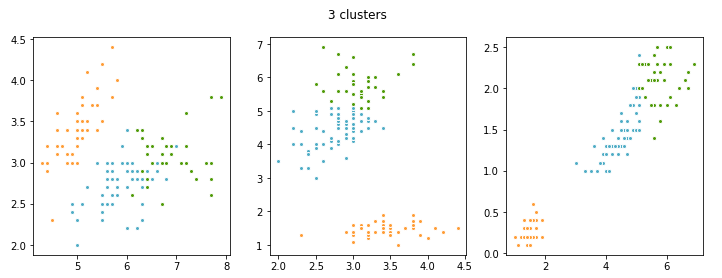
Рисунок 9 - Сравнение MiniBatch K-means и K-means

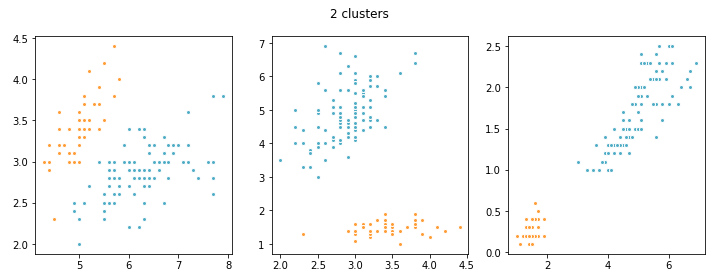
1. **Иерархическая кластеризация.**
2. Была проведена иерархическая кластеризация тех же данных.
3. В отличие от k-средних иерархическая кластеризация постипенно объединяет минимальные кластеры, минимизируя определенную характеристику.

Рисунок 10 – Иерархическая кластеризация для 5 кластеров

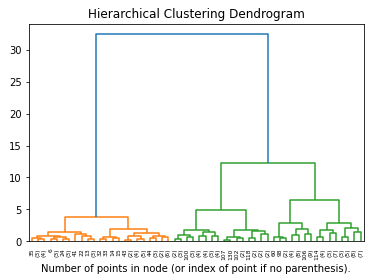
1. Было проведено исследование для различного размера кластеров (2-4).

Рисунок 11 - Иерархическая кластеризация для 4 кластеров

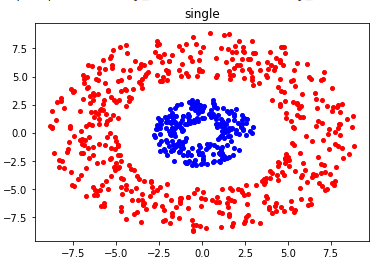
Рисунок 12 - Иерархическая кластеризация сдля 3 кластеров

Рисунок 13 - Иерархическая кластеризация для 2 кластеров

1. Была нарисована дендограмма до уровня 6.

Рисунок 14 - Дендограмма

1. Были сгенерированы данные в виде двух колец.
2. Была проведена иерархическая кластеризация сгенерированных данных.

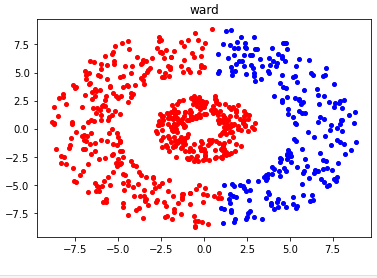
Рисунок 15 - Кластеризация с параметром single

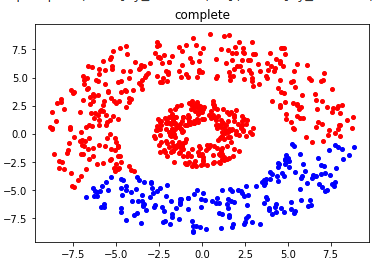
1. Была исследована кластеризация при всех параметрах linkage.
   * ward минимизирует дисперсию;
   * average минимизирует среднее расстояние;
   * complete or maximum минимизирует максимальное расстояние;
   * single минимизирует минимальное расстояние.

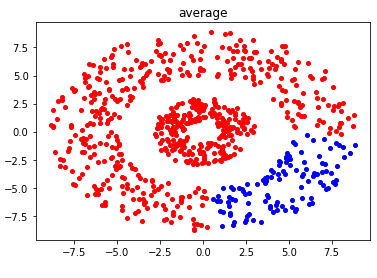
*Single* может применяться на больших наборах данных и на данных, однако неустойчив к выбросам.

*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.