

Кластеризация

Беляков Дмитрий

December 27, 2021

1 Описание задачи

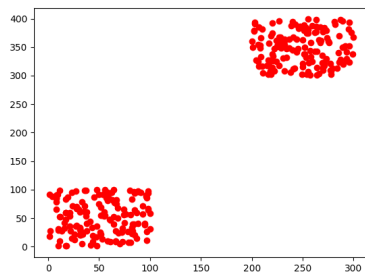
Задача: Провести кластерный анализ точек на плоскости, использовать иерархическую кластеризацию на малом наборе данных, проверить возможно ли использовать k-means. В случае успеха использовать k-means, с центрами из анализа на малом наборе данных

2 Описание метода

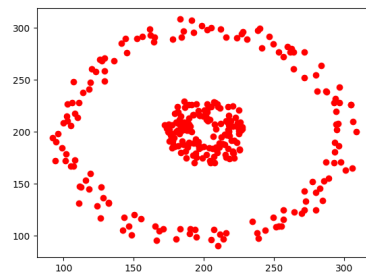
Используем методы Kmeans и AgglomerativeClustering библиотеки sklearn.cluster для алгоритмов K-means и иерархической кластеризации.

Сначала возьмем выборку из небольшого числа точек и разобьем на два кластера с помощью иерархической кластеризации. Если расстояние между центрами полученных кластеров достаточно большое, а также в каждом кластере достаточно много точек, проведем кластеризацию с помощью k-means. Иначе проведем анализ всех данных иерархической кластеризацией.

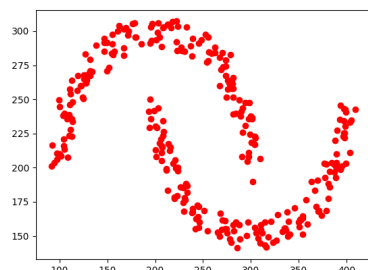
3 Описание данных



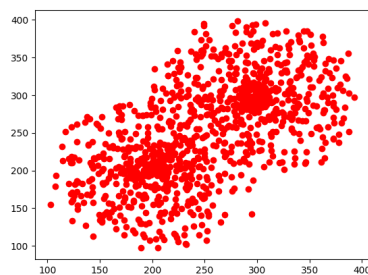
Тест 1



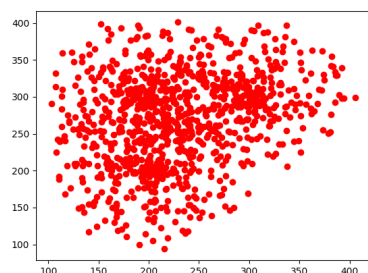
Тест 2



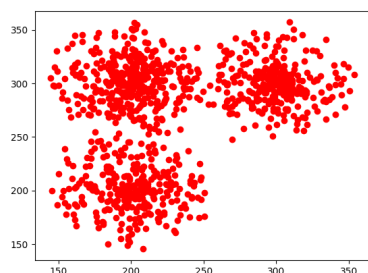
Тест 3



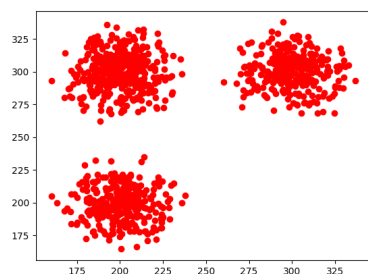
Тест 4



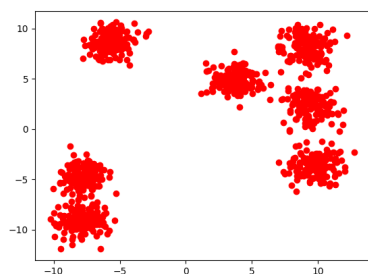
Тест 5



Тест 6



Тест 7



Тест 8

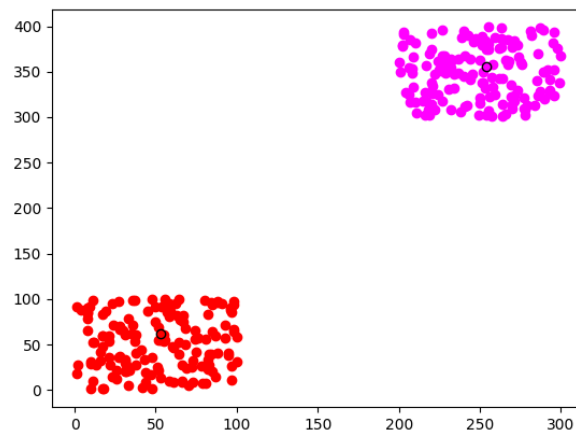
Данные генерируем самостоятельно, рассмотрим 8 случаев:

- Два кластера на большом расстоянии друг от друга
- Два кластера, один внутри другого
- Два кластера, огибающие друг друга
- Два кластера, близкие друг к другу

- Три почти свдвигающих кластера
- Три кластера неподалеку друг от друга
- Три кластера далеко друг от друга
- Семь случайно разбросанных кластеров

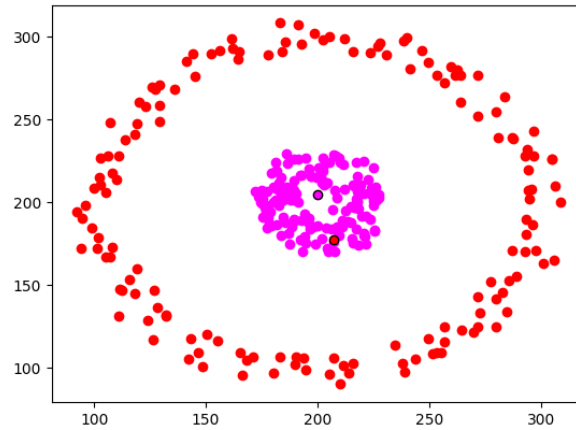
4 Результаты

4.1 Тест 1



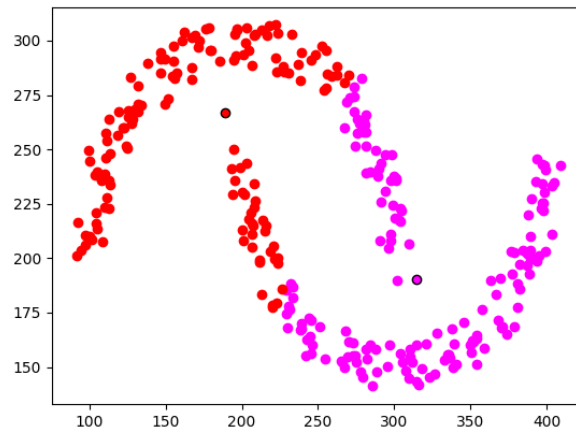
В тесте 1 программа без проблем выявила два кластера и привела кластерный анализ с помощью K-means

4.2 Тест 2



В тесте 2 при анализе малой выборки программа поняла, что расстояние между центрами кластеров невелико и k-means невозможен, поэтому использовала иерархическую кластеризацию

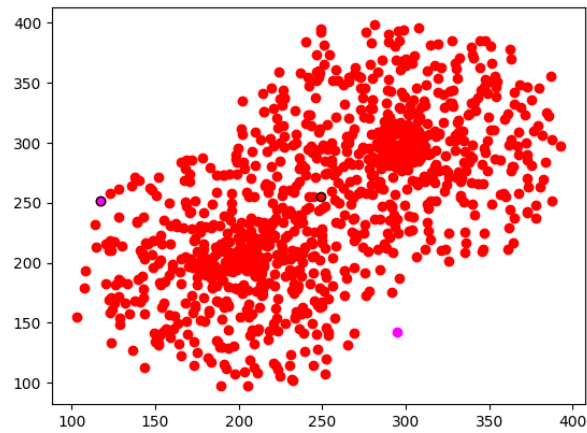
4.3 Тест 3



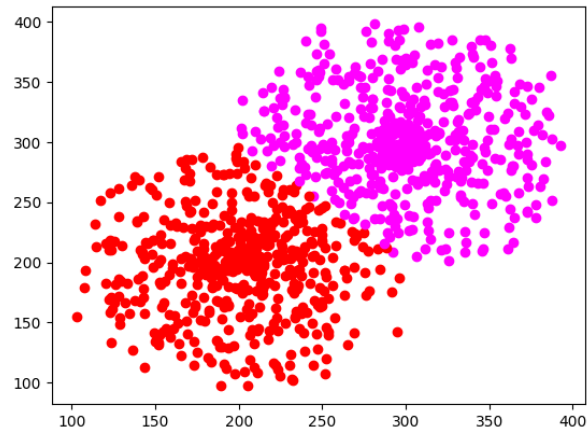
В тесте 3 при анализе малой выборки программа поняла, что расстояние между центрами кластеров достаточно велико и k-means возможен. Однако

в силу особенностей алгоритма программа поделила множество точек на два кластера не соответствующих оптимальному разбиению

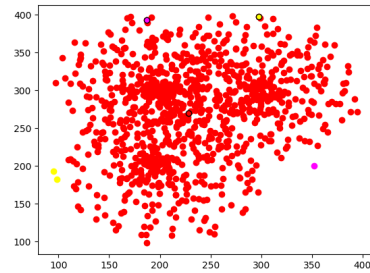
4.4 Тест 4



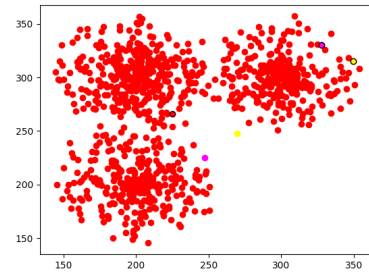
В тесте 4 при анализе малой выборки программа поняла, что в одном кластере много больше точек чем в другом поэтому k-means невозможен. В итоге программа выделила два кластера, один из которых во много больше второго. Заметим что если бы программа провела k-means без анализа малой выборки то разбиение было бы близким к оптимальному



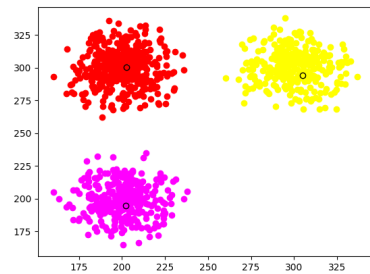
4.5 Тесты 5-7



Тест 5



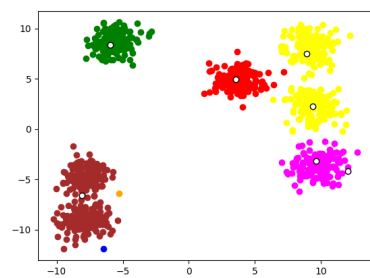
Тест 6



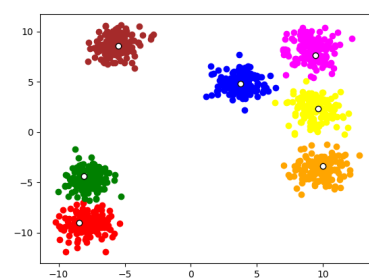
Тест 7

В тестах 5-7 видно, что на достаточно близких кластерах алгоритм не работает, а на далеких - работает

4.6 Тест 8



Иерархическая кластеризация



K-means

В 8 примере четко видно, что при наличии рядом стоящих кластеров программа объединяет их в один. Однако если убрать проверку на количество точек

в кластерах, K-means без проблем разделит точки

5 Вывод

Метод может сэкономить время для поиска центров в k-means (тест 1). А также в некоторых случаях определить, что k-means не оптимален (тест 2). Однако в части тестов программа не смогла определить, что k-means не оптимален (тест 3), а иногда вывод о неоптимальности k-means казывается ложным (Тест 4)