Кластеризация

Беляков Дмитрий

December 22, 2021

1 Описание задачи

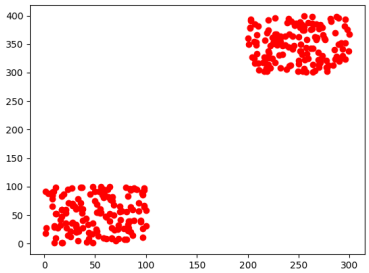
Задача: Провести кластерный анализ точек на плоскости, испльзовать иерархическую кластеризацию на малом наборе данных, проверить возможно ли использовать k-means. В случае успеха использовать k-means, с центрами из анализа на малом набре данных

2 Описание метода

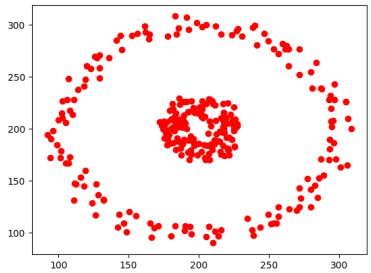
Используем методы Kmeans и AgglomeraiveClustering библиотеки sklearn.clustering для алгоритмов K-means и иерархической кластеризации.

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

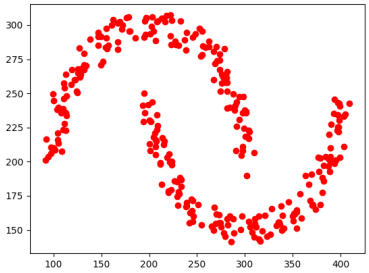


Тест 1

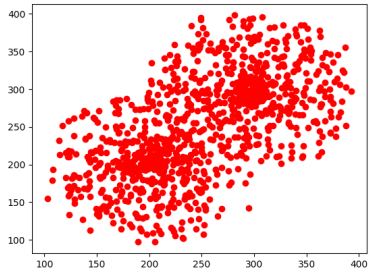


Тест 2

Данные генерируем самостоятельно, рассмотрим 4 случая:



Тест 3

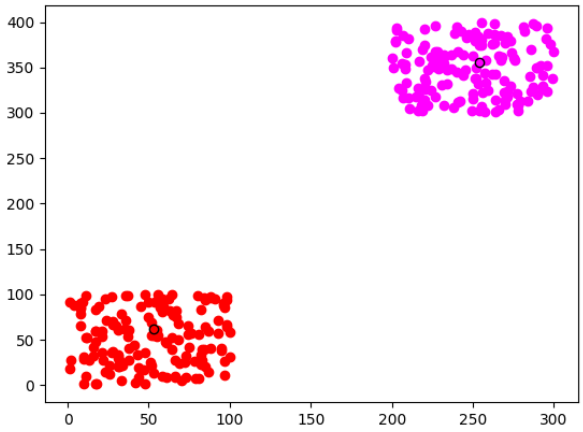


Тест 4

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

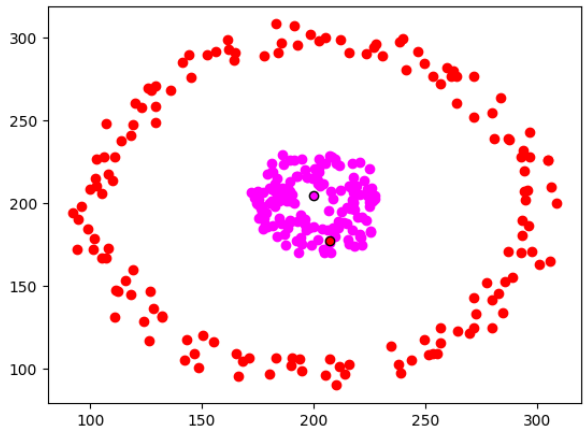
4 Результаты

4.1 Тест 1



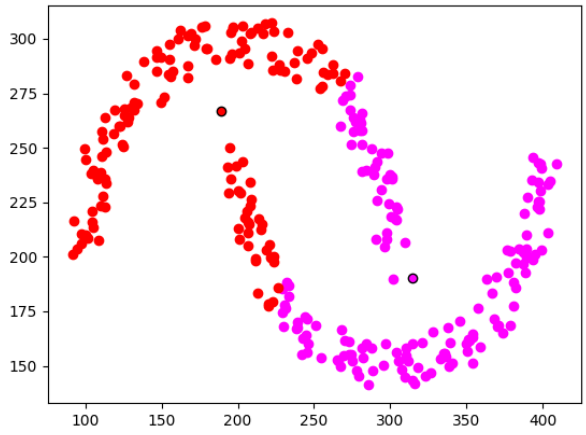
В тесте 1 программа без проблем выявила два кластера и првела кластерный анализ с помощью K-means

4.2 Тест 2



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

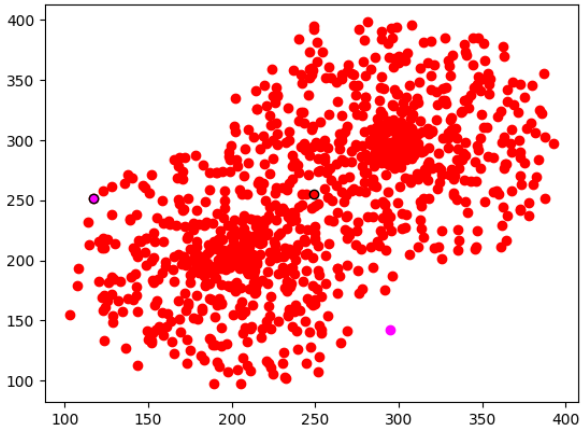
4.3 Тест 3



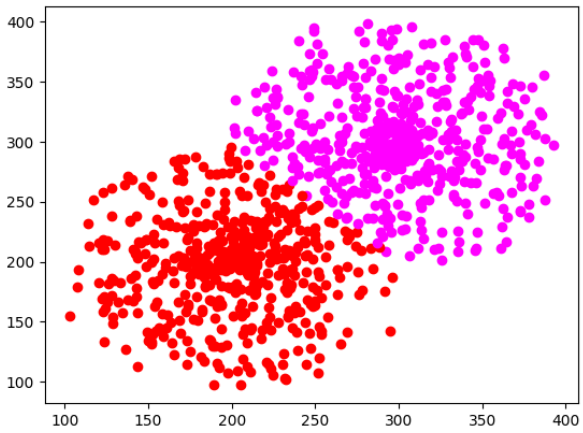
В тесте 3 при анализе малой выборки программа поняла, что расстояние между центрами кластеров достаточно велико и k-means возможен. Однако

в силу особенностей алгоритма программа поделила множество точек на два кластера не сотвествующих 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 неозможен. 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



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 смгла определить, что k-means не оптимален (тест 3), а иногда вывод о неоптимальности k-means казывался ложным (Тест 4)