# Алгоритм K-mean

Филютович Дмитрий

28 декабря 2021 г.

## 1 Описание задачи

Исследовать алгоритм кластеризации K-mean на тестовых данных.

## 2 О тестовых данных

Я использую несколько типов данных, созданных с помощью библиотеки sklearn.datasets: кляксы, луны и круги. Все они на 300 точек. Также я использую мною найденный датасет на 900 точек.

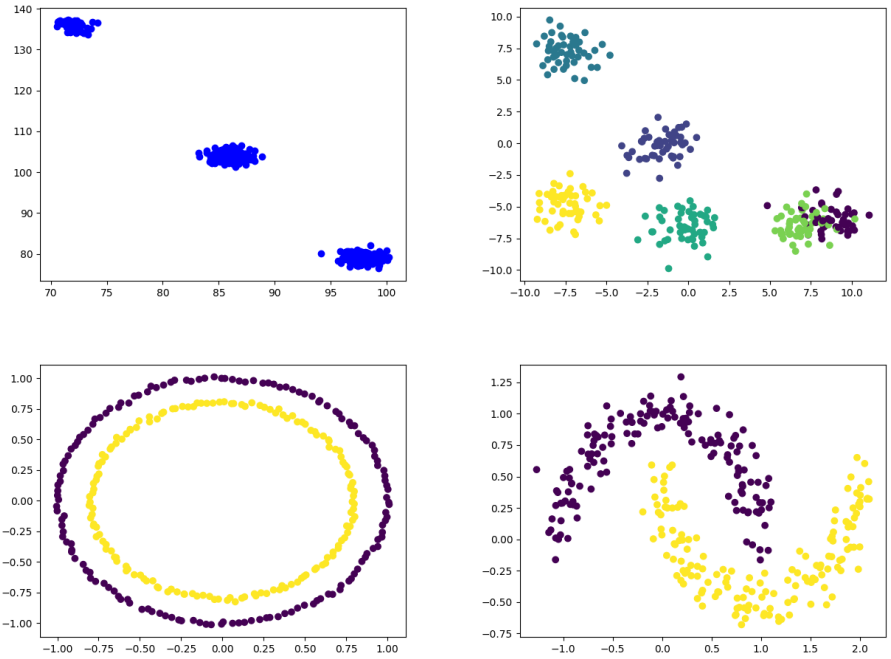

Рис. 1: Используемые тестовые данные.

## 3 Немного о поиске количества кластеров

Для поиска оптимального количества кластеров я использовал иерархическую сортировку. А точнее иерархическая кластеризация прекращается, когда расстояние между новыми кластерами меньше определенного. И из иерархичсекой сортировки я беру количество образовавшихся кластеров и использую его в K-mean.

# 4 Работа с данными

## 4.1 Кляксы

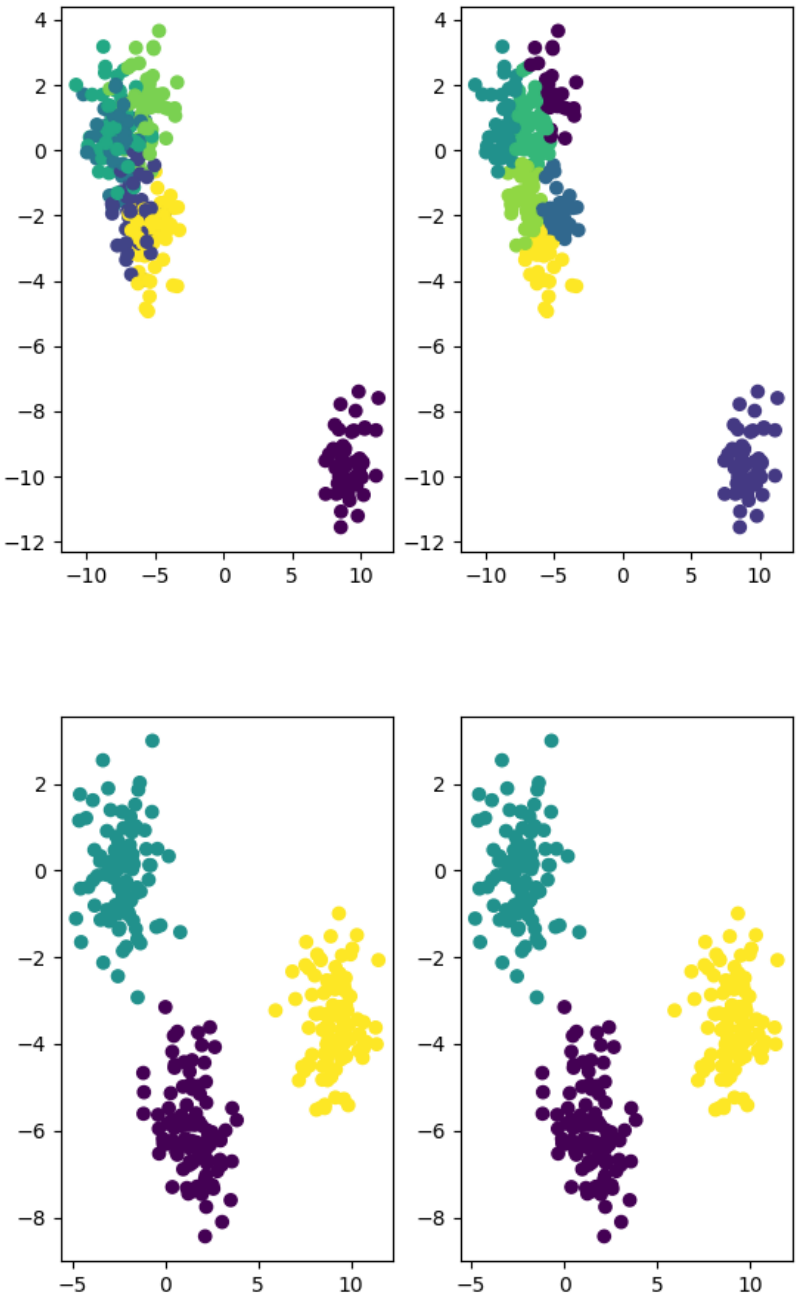


Рис. 2: Работа на кляксах

Как видно на рисунках K-mean определяет кластеры достаточно точно, но зависит от заданного расстояния в иерахической сортировке. На 300 точках и 3 кластерах алгоритм работает сравнительно недолго 100ms. На 7 кластерах алгоритм работает дольше. Уже 150ms.

### 4.2 Круги

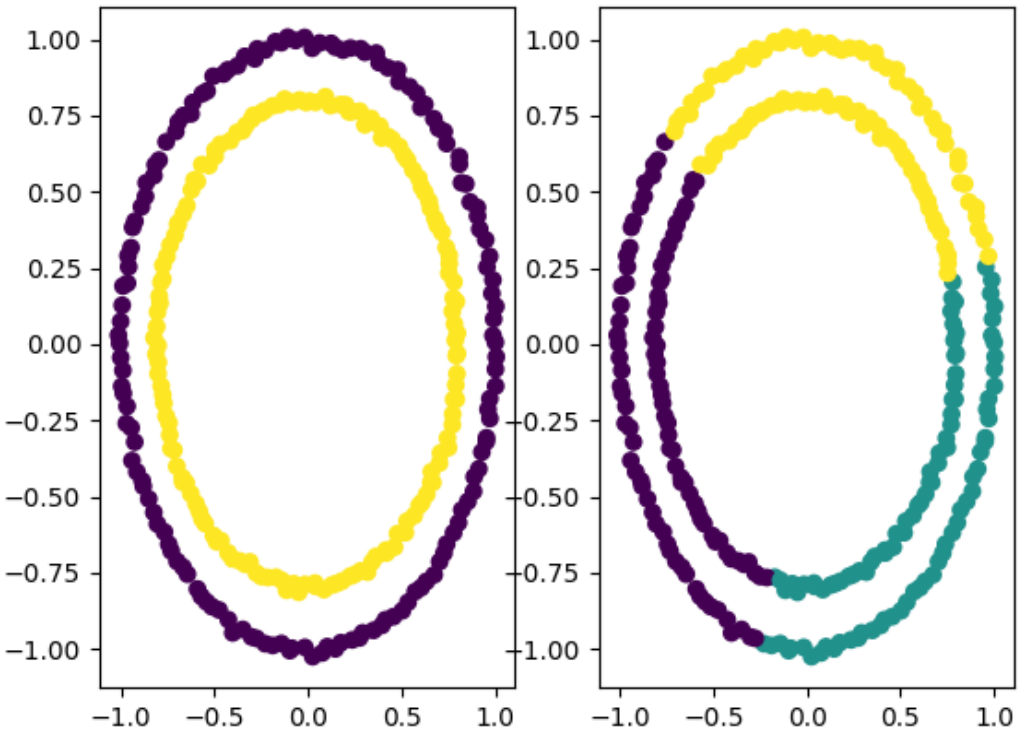


Рис. 3: Работа на кругах

На кругах алгоритм работает примерно 100ms, но как видно кластеризует алгоритм плохо.

### 4.3 Луны

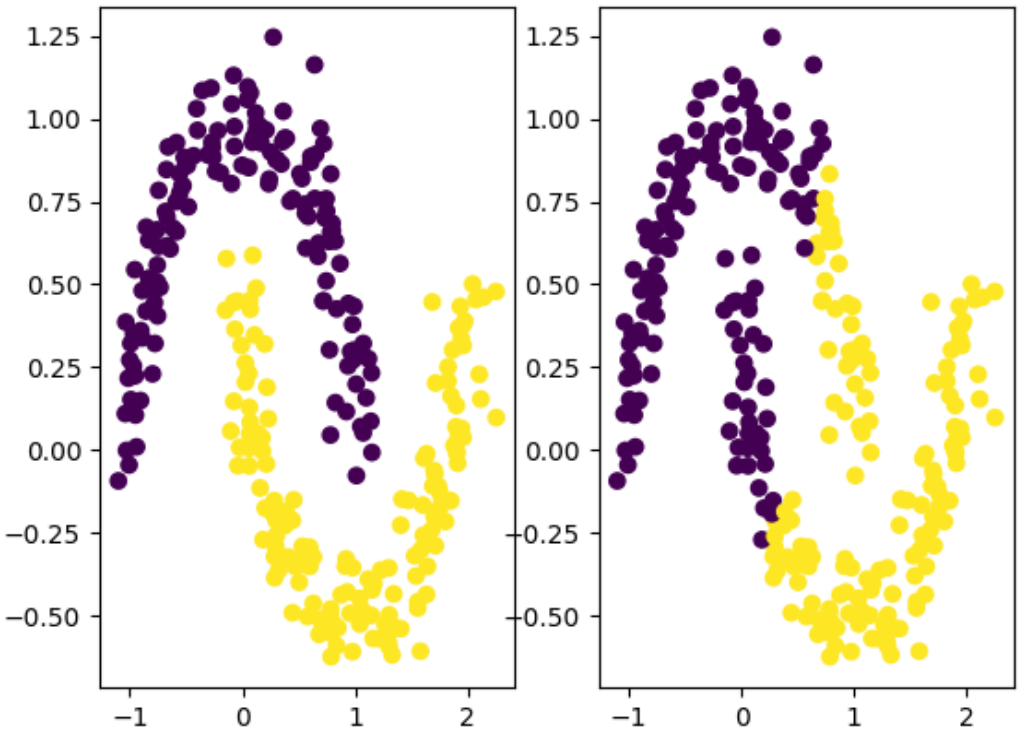


Рис. 4: Работа на лунах

Иерархическая кластеризация верно определяет количество кластеров, но K-mean кластеризует также плохо, как и круги. По времени работы получается также примерно 100ms

### 4.4 Даатсет

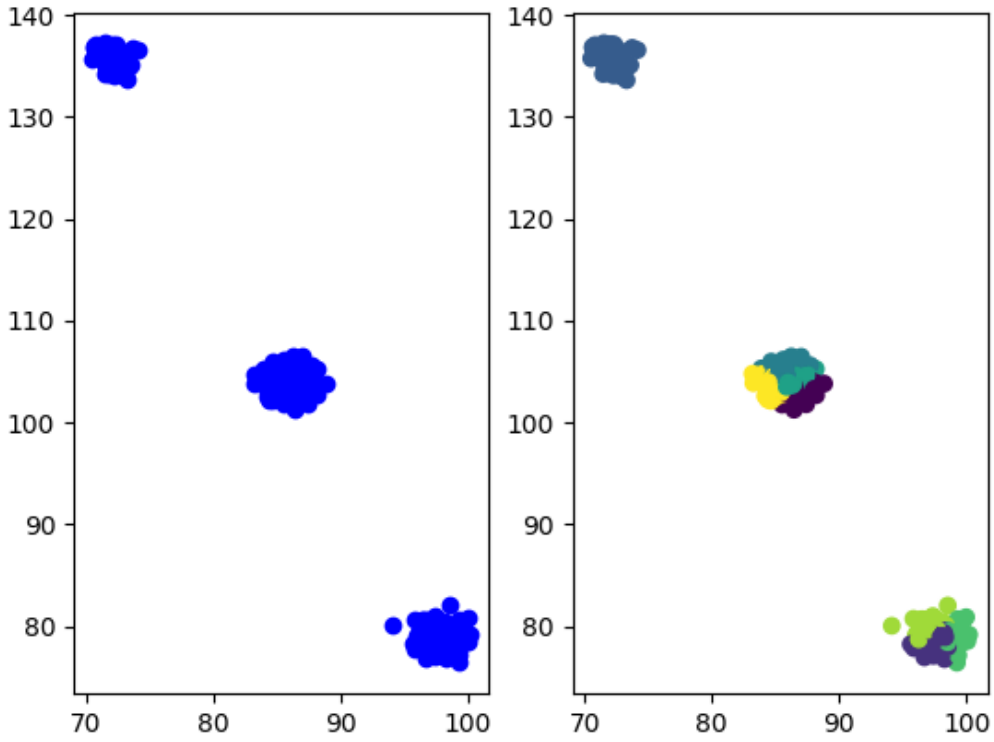


Рис. 5: Работа на датасете

На датасете K-mean кластеризует относительно точно, но на данном датасете это не очень хорошо видно. На 900 точках алгоритм работает 150ms.

## 5 Вывод

K-mean хорошо кластеризует данные, которые хорошо делятся на части плоскости. На данных вида кругов или луны алгоритм работает не точно. Время работы зависит в основном от количества точек, а также от обнаруженного количеств кластеров.