Алгоритмы кластеризации

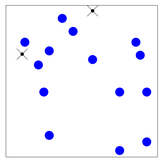
Корышева Юлия; Переверзев Виктор; Ляшев Глеб

Факультет экономических наук НИУ ВШЭ

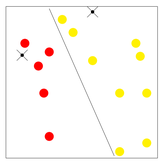
k-means

Алгоритм разбивает объекты на заранее известное число кластеров, при условии минимизации среднего расстояния между каждым объектом кластера и его центроидом.

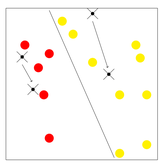
1. На первом шаге фиксируются стартовые центроиды.



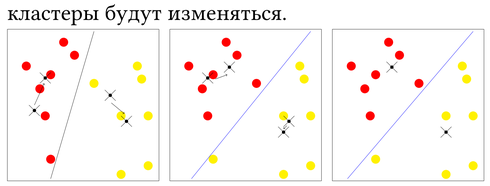
2. На втором шаге объекты разбиваются на кластеры, при условии минимизации расстояния от объектов до центроида.



3. Алгоритм пересчитывает значения центроидов.



4. Алгоритм выполняет шаги 2-3 до тех пор, пока кластеры будут изменяться.



Начальные центроиды

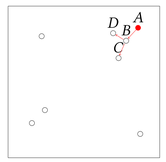
Выбор начальных центроидов может осуществляться следующим образом:

- Случайный выбор k-объектов.
- Выбор первых k-объектов.

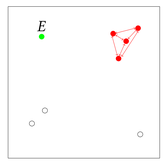
DBSCAN

Суть работы алгоритма заключается в захвате кластеров через соседние точки. При заданном минимальном расстоянии для захвата и минимальном количестве точек для создания кластеров захваченным объектам присваиваются метки кластера или иначе же метки выбросов.

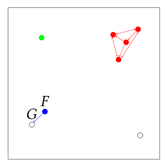
1. DBSCAN нашел случайную точку A в пространстве и «заразил» ее. Далее происходит последовательный захват соседей (B, C, D), находящихся друг от друга в пределах требуемого расстояния.



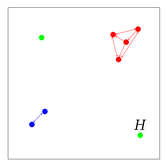
2. Когда все цели поражены, на захваченных территориях DBSCAN принимает решение создать тоталитарное государство-кластер. Но и этого ему мало. Следующим что на пути безжалостной машины попался объект E.



3. Найдя точку G рядом с соседом F, алгоритм захватывает эти два объекта и присваивает им метку кластера. Снова запускается поиск.



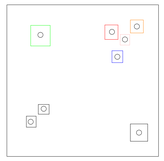
4. На последнем найденном объекте H алгоритм заканчивает работу.



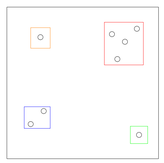
Иерархическая кластеризация

Иерархическая кластеризация выводит иерархию, структуру, которая в целом информативнее, чем набор предсказаний от других видов кластеризации. Иерархическая кластеризация не требует заранее определять количество кластеров, и большинство популярных иерархических алгоритмов является неслучайными.

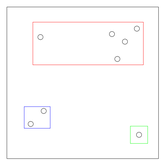
1. На первом шаге алгоритм присваивает каждому объекту отдельный кластер.



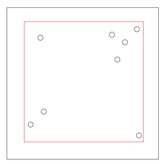
2. На следующем шаге алгоритм рассчитывает расстояние с соседними объектами. Объекты, которые находятся близко друг к другу, объединяются в один кластер.



3. На следующем этапе алгоритм объединяет наиболее близкие кластеры в один.



4. Алгоритм повторяет шаг 3, пока все объекты не принадлежат к одому кластеру.



Понятие кластеризации

Кластеризация (или кластерный анализ) — это задача разбиения множества объектов на группы, называемые кластерами. Внутри каждой группы должны оказаться «похожие» объекты, а объекты разных группы должны быть как можно более отличны. Главное отличие кластеризации от классификации состоит в том, что перечень групп четко не задан и определяется в процессе работы алгоритма.

Формальная постановка задачи

Пусть дано множество объектов, требуемое количество кластеров и функция для определения расстояния. Задачей является построение алгоритма, определяющего для каждого объекта номер кластера, к которому он принадлежит, при условии, что расстояние между объектами минимально.

Расстояния

- Евклидово расстояние:

$$d(a,b) = \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2 + \ldots + (a_n - b_n)^2}$$

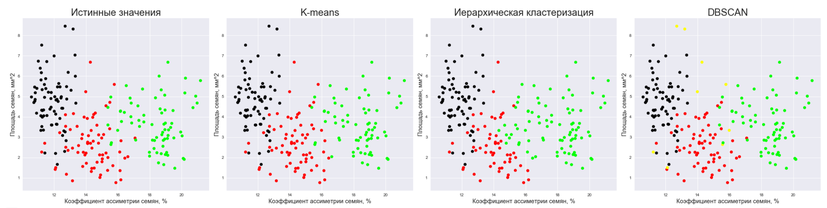
- Манхэттенское расстояние:

$$d(a,b) = |a_1 - b_1| + |a_2 - b_2| + \ldots + |a_n - b_n|$$

- Расстояние Чебышева:

$$d(a,b) = \max |a_1 - b_1| + |a_2 - b_2| + \ldots + |a_n - b_n|$$

Пример работы алгоритмов



Боевая задача: 210 элементов в датасете, 3 сорта семян пшеницы: Kama, Rosa и Canadian. Цель алгоритмов – разделить семена на виды по признакам после того, как мы отрезали столбик с верными ответами.

Все алгоритмы практически верно разделили объекты на кластеры, не зная правильного результата. Один лишь DBSCAN отстал от своих собратьев и впопыхах выделил несколько объектов в класс выбросов (отмечены жёлтым).

Код работы Вы можете посмотреть код работы и оценить прочие вкусняшки на github.com/PereverzevVV