Алгоритмы кластеризации

Всякие людишки

Факультет экономических наук НИУ ВШЭ

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

- $\Omega = \{\omega_1, \omega_2, \ldots, \omega_k\}$ – множество кластеров.
- $X = \{x_1, x_2, \ldots, x_n\}$ – множество объектов.
- $\mu(\omega) = \frac{1}{|\omega|} \sum_{x \in \omega} x$ – центроид кластера, где $|\omega|$ – размер кластера.

Меры расстояния

- Евклидово расстояние:

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

- Квадрат еквлидова расстояния:

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

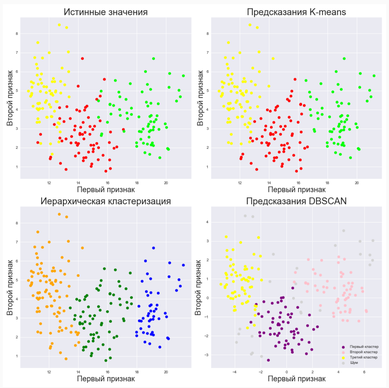
- Манхэттенское расстояние:

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

- Расстояние Чебышева:

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

Предсказание методов

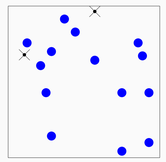


Прогноз рассматриваемых алгоритмов

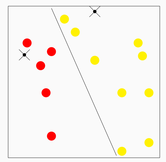
K-means

Алгоритм разбивает множество элементов на известное число кластеров. Основной идеей алгоритма является минимизировать среднее расстояние между каждым из объектов в кластере и его центроидом.

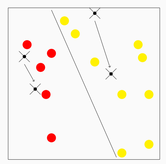
1. На первом шаге фиксируются стартовые центроиды.



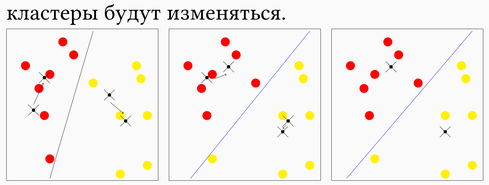
2. На втором шаге объекты разбиваются на кластеры, при условии минимизации расстояния от объектов до центроида.



3. Пересчитываются значения центроидов.



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



Начальное приближение

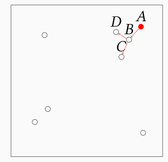
Существует несколько способов задать начальные центроиды:

- Использование в качестве начальных центроидов результат работы другого алгоритма кластеризации.
- Запуск алгоритма несколько раз из различных начальных положений и последующий выбор наилучшего.

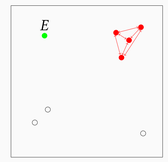
DBSCAN

Суть работы алгоритма заключается в захвате кластеров через соседние точки. Если у случайно найденной точки есть соседи, то он их «заражает», и уже от них производит поиск новых объектов. Если количество подряд зараженных объектов соответствует требованию, то они превращаются в кластер, иначе же в выброс.

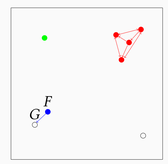
1. DBSCAN нашел случайную точку A в пространстве и «заразил» ее. Далее происходит последовательный захват соседей (B, C, D), находящихся друг от друга в пределах требуемого расстояния.



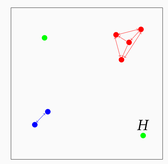
2. Когда все цели поражены, на захваченных территориях DBSCAN принимает решение создать тоталитарное государство-кластер. Но и этого ему мало. Следующим что на пути безжалостной машины попался объект E.



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



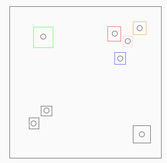
4. На последнем найденном объекте H алгоритм заканчивает работу.



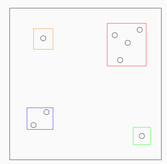
Иерархическая кластеризация

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.

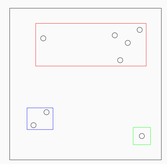
1. На первом шаге алгоритм присваивает каждому объекту отдельный кластер.



2. Затем он подсчитывает схожесть (через расстояние) с соседними объектами, и в случае положительного результата объединяется с ближайшими объектами. На этом уровне кластеризации уже можно сделать разрез.



3. Алгоритм повторяет этот шаг, пока не останется других кластеров. Ближайший объект присоединяет к себе.



4. Наконец, алгоритм захватывает все точки на плоскости, объединяя их в один единый кластер.