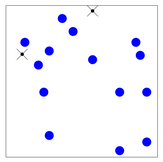
Алгоритмы кластеризации

Факультет экономических наук НИУ ВШЭ

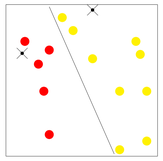
k-means

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

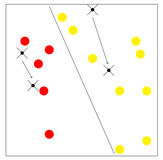
1. На первом шаге фиксируются стартовые центроиды.



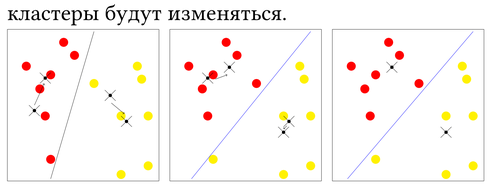
2. На втором шаге объекты разбиваются на кластеры, при условии минимизации расстояния от объектов до центроида.



3. Алгоритм пересчитывает значения центроидов.



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



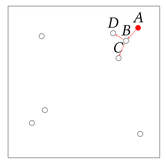
Выбор начальных центроидов

- Использование результата работы другого алгоритма кластеризации.
- Запуск алгоритма несколько раз из различных начальных положений и последующий выбор наилучшего.
- Использование метода k-means++

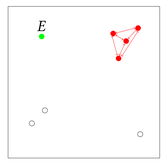
DBSCAN

Алгоритм образует кластеры из тех точек, которые находятся друг от друга на заданном расстоянии. Если количество точек в кластере соответствует начальному требованию, то они превращаются в кластер, иначе же в выброс.

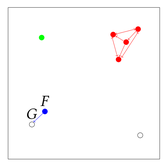
1. На первом шаге DBSCAN находит точку A. Далее алгоритм образует кластер из этой точки и точек B, C, D, которые находятся в пределах требуемого расстояния от неё.



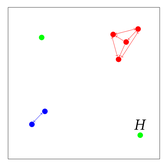
2. На следующем шаге алгоритм находит точку E. Так как в пределах требуемого расстояния нет других точек, точка E определяется как выброс.



3. Найдя точку G рядом с соседом F, алгоритм присваивает этим точкам отдельный кластер.



4. На последнем найденном объекте H алгоритм заканчивает работу.



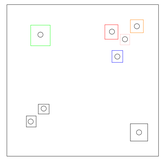
GitHub

https://github.com/YuliaKorysheva

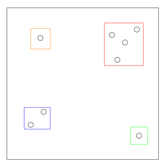
Иерархическая кластеризация

Существует два вида алгоритмов иерархической кластеризации – агломеративные и дивизионные. Агломеративные алгоритмы последовательно объединяют мелкие кластеры в более крупные. Дивизионные алгоритмы работают наоборот – разделяют крупные кластеры на более мелкие. Иерархическая кластеризация визуализируется с помощью дендрограммы. Она показывает степень схожести отдельных объектов и кластеров, а также иллюстрирует процесс объединения или разделения кластеров.

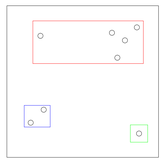
1. На первом шаге агломеративный алгоритм присваивает каждому объекту отдельный кластер.



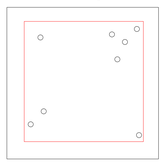
2. На следующем шаге алгоритм рассчитывает расстояние с соседними объектами. Объекты, которые находятся близко друг к другу объединяются в один кластер.



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 – это задача разбиения объектов на кластеры. Внутри каждого кластера должны оказаться «похожие» объекты, а объекты разных кластеров должны быть как можно более отличны. Главное отличие задачи кластеризации от классификации состоит в том, что изначально данные не размечены.

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 в одном кластере между объектами минимально, а между объектами из разных кластеров максимально.

Расстояние между кластерами

С помощью формулы Ланса-Уильямса можно вычислить расстояние между кластером, который получается после слияния кластеров U и V, и кластером S. Если известны расстояния между более мелкими кластерами, эта формула позволяет рекурсивно получать расстояния между большими кластерами.

$$R(U \cup V, S) = \alpha_U R(U, S) + \alpha_V R(V, S) + \beta R(U, V) + \gamma |R(U, S) - R(V, S)|$$

Разные коэффициенты позволяют вычислять расстояние между кластерами разными способами:

- Расстояние между ближайшими соседями (при $\alpha_U = \alpha_V = \frac{1}{2}, \beta = 0, \gamma = -\frac{1}{2}$):
$$R(W, S) = \min_{w \in W, s \in S} \rho(w, s)$$
- Расстояние между дальними соседями (при $\alpha_U = \alpha_V = \frac{1}{2}, \beta = 0, \gamma = \frac{1}{2}$):
$$R(W, S) = \max_{w \in W, s \in S} \rho(w, s)$$
- Среднее расстояние между объектами разных кластеров(при $\alpha_U = \frac{|U|}{|W|}, \alpha_V = \frac{|V|}{|W|}, \beta = \gamma = 0$):
$$R(W, S) = \frac{1}{|W||S|} \sum_{w \in W} \sum_{s \in S} \rho(w, s)$$

Пример работы алгоритмов