Метод k-средних (K-means)

Разделы: Алгоритмы
Loginom: Кластеризация (обработчик)

Метод k-средних используется для кластеризации данных на основе алгоритма разбиения векторного пространства на заранее определенное число кластеров k. Алгоритм представляет собой итерационную процедуру, в которой выполняются следующие шаги:

1. Выбирается число кластеров k.
2. Из исходного множества данных случайным образом выбираются k наблюдений, которые будут служить начальными центрами кластеров.
3. Для каждого наблюдения исходного множества определяется ближайший к нему центр кластера (расстояния измеряются в метрике Евклида). При этом записи, «притянутые» определенным центром, образуют начальные кластеры.
4. Вычисляются **центроиды** — центры тяжести кластеров. Каждый центроид — это вектор, элементы которого представляют собой средние значения соответствующих признаков, вычисленные по всем записям кластера.
5. Центр кластера смещается в его центроид, после чего центроид становится центром нового кластера.
6. 3-й и 4-й шаги итеративно повторяются. Очевидно, что на каждой итерации происходит изменение границ кластеров и смещение их центров. В результате минимизируется расстояние между элементами внутри кластеров и увеличиваются междукластерные расстояния.

Остановка алгоритма производится тогда, когда границы кластеров и расположения центроидов не перестанут изменяться от итерации к итерации, т.е. на каждой итерации в каждом кластере будет оставаться один и тот же набор наблюдений. На практике алгоритм обычно находит набор стабильных кластеров за несколько десятков итераций.

Преимуществом алгоритма являются скорость и простота реализации. К **недостаткам** можно отнести неопределенность выбора начальных центров кластеров, а также то, что число кластеров должно быть задано изначально, что может потребовать некоторой априорной информации об исходных данных.

Существуют методы кластеризации, которые можно рассматривать как происходящие от k-средних. Например, в методе k-медиан (k-medians) для вычисления центроидов используется не среднее, а медиана, что делает алгоритм более устойчивым к аномальным значениям в данных.

Алгоритм g-средних (от gaussian) строит кластеры, распределение данных в которых стремится к нормальному (гауссовскому) и снимает неопределенность выбора начальных кластеров. Алгоритм C-средних использует элементы нечеткой логики, учитывая при вычислении центроидов не только расстояния, но и степень принадлежности наблюдения к множеству объектов в кластере. Также известен алгоритм Ллойда, который в качестве начального разбиения использует не множества векторов, а области векторного пространства.

Идея метода k-средних была одновременно сформулирована Гуго Штейнгаузом и Стюартом Ллойдом в 1957 г. Сам термин «k-средних» был впервые введен Дж. Маккуинном в 1967 г.