Метод локтя (Elbow method)

Разделы: Алгоритмы

Метод локтя (Elbow method) — инструмент анализа данных, направленный на оптимизацию числа кластеров в алгоритмах кластеризации. Впервые был предложен Робертом Л. Торндайком в 1953 году.

Правильно подобранное количество кластеров в алгоритмах позволяет найти баланс между погрешностью вычисляемой дисперсии и сложностью модели. Использование метода позволяет избежать недообучения или переобучения алгоритма кластеризации.

Метод применим к алгоритму k-средних и заключается в неоднократном повторении сценария. При использовании метода для каждого натурального числа k из некоторого диапазона строится значение целевой функции, равной сумме внутрикластерных расстояний. Количество кластеров — гиперпараметр, т.е. он будет определен перед запуском модели.

Использование метода локтя подразумевает прохождение трех этапов.

На **первом этапе** для различных значений числа кластеров k вычисляется сумма квадратов расстояний каждой точки данных до их центроида (центра тяжести) ($WCSS$) по формуле:

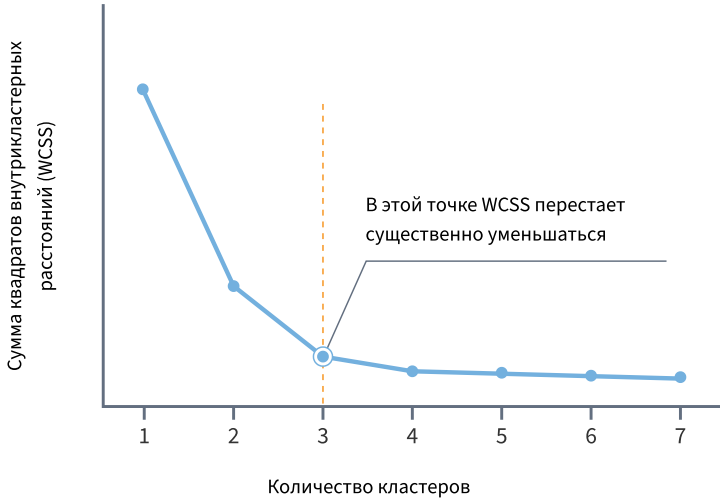
$$WCSS = \sum_{j=1}^{k} \sum_{i=1}^{n} \min(\| x_i^{(j)} - c_j \|)^2,$$

где k — число кластеров, n — количество наблюдений, $x_i^{(j)}$ — i-ое наблюдение в j-том кластере, c_j — центроид j-того кластера.

Второй этап содержит построение графика зависимости $WCSS$ от количества кластеров, где по оси X откладывается число кластеров k, а по оси Y — соответствующая сумма квадратов расстояний.

Третий этап заключается в поиске точки излома («локтя») на графике, которая указывает на оптимальное число кластеров. Оптимальным k будет то, при котором ошибка перестает существенно уменьшаться, т.е. начинает сглаживаться.

График может иметь следующий вид:



На основание данного графика можно определить, что оптимальным будет использование трех кластеров.

В методе локтя основной акцент делается на визуальный анализ. Если линейный график выглядит как рука, то «локоть» (точка перегиба на кривой) является наилучшим значением k. При том «рука» может быть направлена как вверх, так и вниз.

Основными недостатками локтевого метода считаются субъективность и ненадежность. На практике выбор «локтя» весьма неоднозначен поскольку на графике не всегда можно проследить точку перегиба, которая определяет оптимальное число кластеров. Это справедливо даже в тех случаях, когда все другие методы определения количества кластеров в наборе данных дают такой же результат.