K-means. Метод кластеризации

Дисциплина: Основы машинного обучения.

гр. 5030102/20201
Смирнова А. П.
Грушин А. Д.

Введение в кластеризацию

Кластеризация

— это задача разделения набора объектов на группы (кластеры) таким образом, чтобы объекты внутри одной группы были похожи между собой, а объекты из разных групп — как можно более различны.

Кластеризацию полезно использовать для выявления скрытых закономерностей в данных, сегментации рынка, обработки изображений и в других областях.

Введение в k-means

Метод k-means — это алгоритм кластеризации, направленный на разбиение набора данных на k кластеров, таких, что каждый объект принадлежит кластеру с ближайшим к нему центром.

Цель алгоритма — минимизировать сумму квадратов расстояний точек кластеров от их центров.

Алгоритм k-means

Шаг 1

Выбираются k случайных центров кластеров для набора данных. Эти центры могут быть выбраны случайным образом или используя специальные методы инициализации.

Шаг 2

Каждая точка данных присваивается к ближайшему центру кластера. Это делается путем расчета расстояния (обычно евклидово) между каждой точкой данных и каждым центром.

Алгоритм k-means

Шаг 3

Центры кластеров обновляются как среднее значение всех точек, принадлежащих каждому кластеру.

Шаг 4

Шаги 2 и 3 повторяются до тех пор, пока центры не перестанут меняться или до достижения заданного критерия сходимости.

Количество кластеров

Алгоритм k-means требует заранее заданное значение k — количество кластеров. Важно выбрать правильное значение, ведь это влияет на качество кластеризации:

1. Если k слишком маленькое, то данные будут сильно обобщены, и точная структура данных будет утрачена.

2. Если k слишком большое, кластеры будут слишком раздроблены, и значимые закономерности могут быть скрыты.

Методы выбора числа кластеров

Метод локтя (Elbow method)

Строится график зависимости суммы квадратов расстояний от числа кластеров k.

По мере увеличения k, ошибка уменьшается, но после некоторого момента улучшения становятся незначительными.

Точку «излома» (локоть на графике) и считают оптимальным числом кластеров.

Метод силуэта (Silhouette method)

Оценивается качество кластеризации для каждой точки:

$$s = \frac{b - a}{\max(a, b)}$$

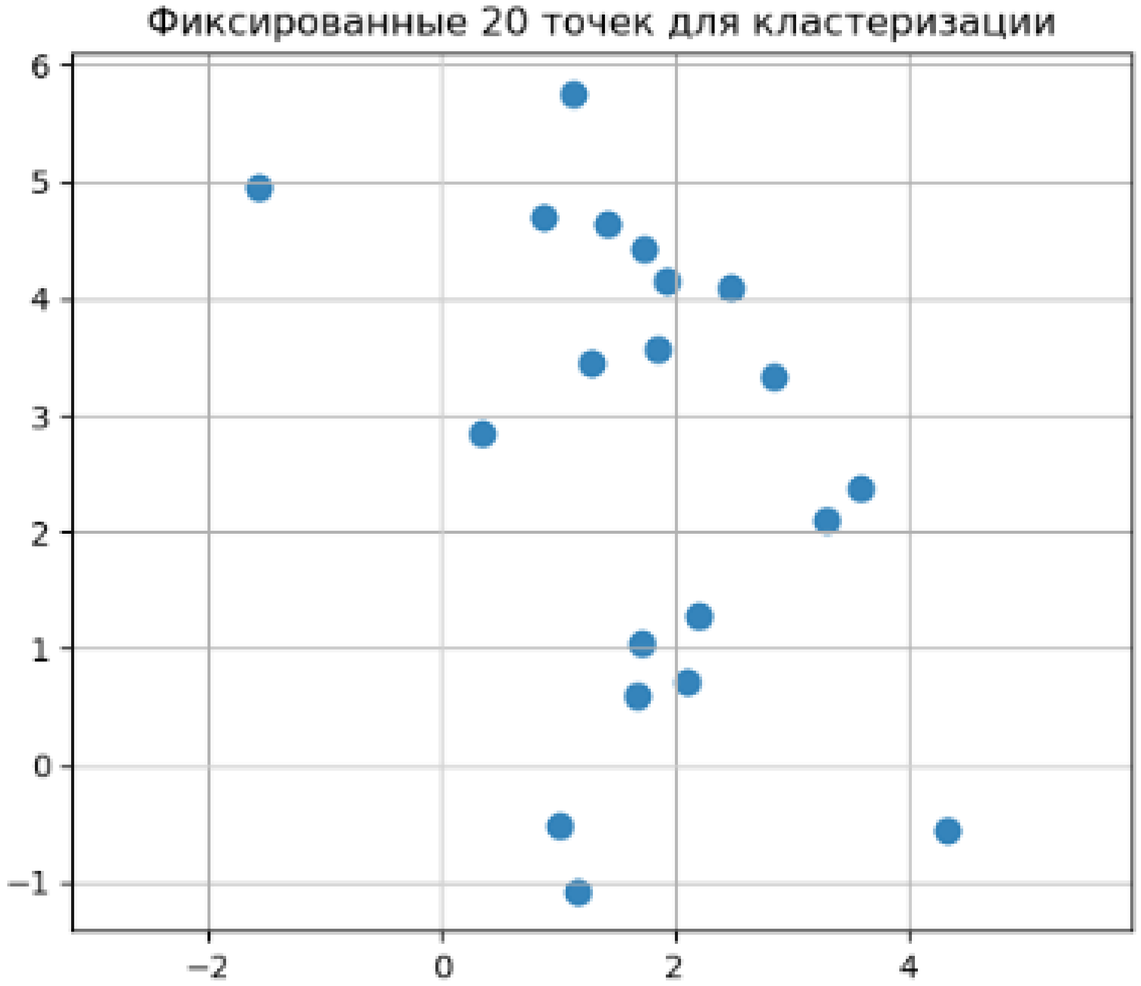
где a — среднее внутрикластерное расстояние, b — среднее расстояние до ближайшего кластера

Вычисляется значение силуэта как среднее значение коэффициента для всех объектов.

Выбирается k, при котором значение силуэта максимально.

Пример

Рассмотрим набор из 20 зафиксированных точек и разделим их на 2 кластера с помощью метода k-means.



Ограничения метода

Локальные минимумы

Алгоритм K-means может застрять в локальном минимуме, зависящем от начальной инициализации центров.

Количество кластеров

Необходимо заранее определить количество кластеров k, которое нужно найти в данных.

Начальная инициализация

Разные начальные центры могут привести к разным кластерам.

Преимущества метода

Простота

Алгоритм K-means относительно прост в реализации и понимании.

Скорость

K-means может работать с огромными наборами данных, он может быть использован для решения многих сложных задач машинного обучения.

Широкая поддержка

Алгоритм K-means реализован в различных фреймворках машинного обучения, что облегчает его использование

Заключение

Метод k-means — это простой и эффективный способ кластеризации данных. Он активно используется в различных областях.

Основная его проблема — правильный выбор числа кластеров k и чувствительность к начальным условиям, но его простота и скорость делают его очень популярным в практике машинного обучения.