Что такое кластеризация?

Кластеризация — это процесс разделения всех данных на группы (также известные как кластеры) на основе закономерностей в данных. Кластеризация — это задача обучения без учителя!

Алгоритм кластеризации K-средних вычисляет центроиды и повторяется до тех пор, пока не будет найден оптимальный центроид. Предположительно известно, сколько кластеров имеется. Он также известен как алгоритм плоской кластеризации. Количество кластеров, найденных из данных методом, обозначается буквой «K» в K-средних.

В этом методе точки данных назначаются кластерам таким образом, чтобы сумма квадратов расстояний между точками данных и центроидом была как можно меньше. Важно отметить, что уменьшение разнообразия внутри кластеров приводит к большему количеству идентичных точек данных в одном и том же кластере.

Работа алгоритма K-средних

Следующие этапы помогут нам понять, как работает метод кластеризации K-средних:

Шаг 1: Во-первых, нам нужно указать количество кластеров K, которые необходимо сгенерировать с помощью этого алгоритма.

Шаг 2: Затем выберите случайным образом K точек данных и назначьте каждую кластеру. Вкратце, классифицируйте данные на основе количества точек данных.

Шаг 3: Теперь будут вычислены центроиды кластера.

Шаг 4: Повторяйте описанные ниже шаги, пока не найдем идеальный центроид, то есть присваивание точек данных кластерам, которые не меняются.

4.1 Сначала будет рассчитана сумма квадратов расстояний между точками данных и центроидами.

4.2 На этом этапе нам нужно распределить каждую точку данных по кластеру, ближайшему к другим (центроиду).

4.3 Наконец, вычислите центроиды для кластеров, усредняя все точки данных кластера.

K-means реализует стратегию максимизации ожиданий для решения проблемы. Шаг ожидания используется для назначения точек данных ближайшему кластеру, а шаг максимизации используется для вычисления центроида каждого кластера.

При использовании алгоритма K-средних мы должны помнить о следующих моментах:

Предлагается нормализовать данные при работе с алгоритмами кластеризации, такими как K-Means, поскольку такие алгоритмы используют измерение на основе расстояния для определения сходства между точками данных.

Из-за итеративного характера K-средних и случайной инициализации центроидов K-средние могут застрять в локальном

минимуме и не сойтись к глобальному минимуму. В результате рекомендуется использовать различные инициализации центроидов.

Реализация графической формы кластеризации K-средних

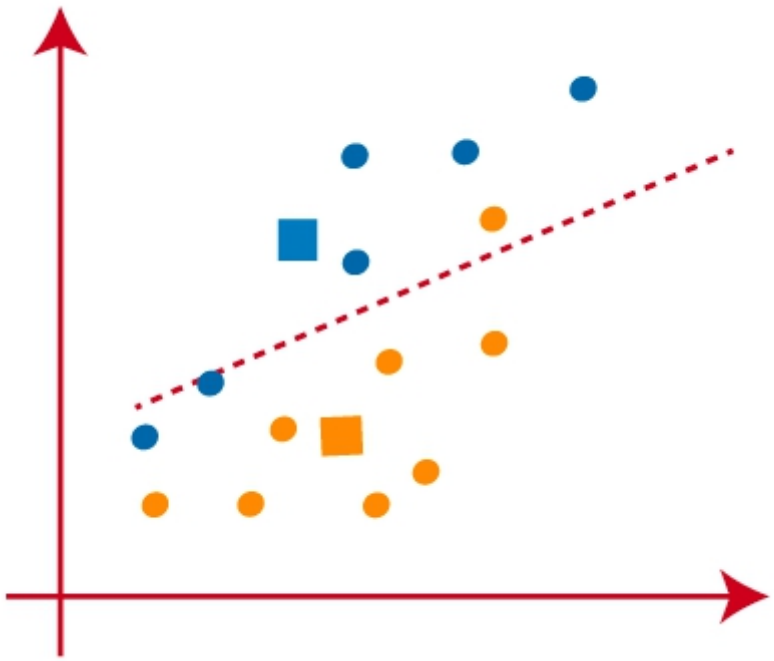
ШАГ 1: Давайте выберем k кластеров, т. Е. K = 2, чтобы разделить набор данных и назначить его соответствующим кластерам. Мы выберем два случайных места, которые будут функционировать как центр тяжести кластера.

ШАГ 2: Теперь каждой точке данных будет присвоена диаграмма рассеяния в зависимости от ее расстояния от ближайшей K-точки или центроида. Это будет достигнуто путем установления медианы между обоими центроидами. Рассмотрим следующую иллюстрацию:

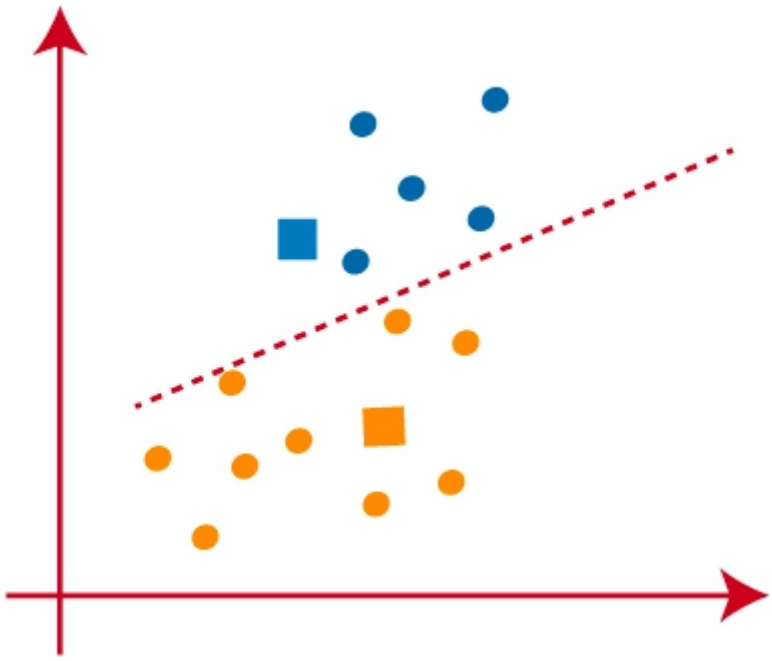
ШАГ 3: Точки на левой стороне линии близки к синему центроиду, а точки на правой стороне линии близки к желтому центроиду. Левый кластер формы имеет синий центроид, тогда как правый кластер формы имеет желтый центроид.

ШАГ 4: Повторите процедуру, на этот раз выбрав другой центроид. Чтобы выбрать новые центроиды, мы определим их новый центр тяжести, который представлен ниже:

ШАГ 5: После этого мы повторно назначим каждую точку данных ее новому центроиду. Мы повторим описанную выше процедуру (используя срединную линию). Синий кластер будет содержать желтую точку данных на синей стороне срединной линии.

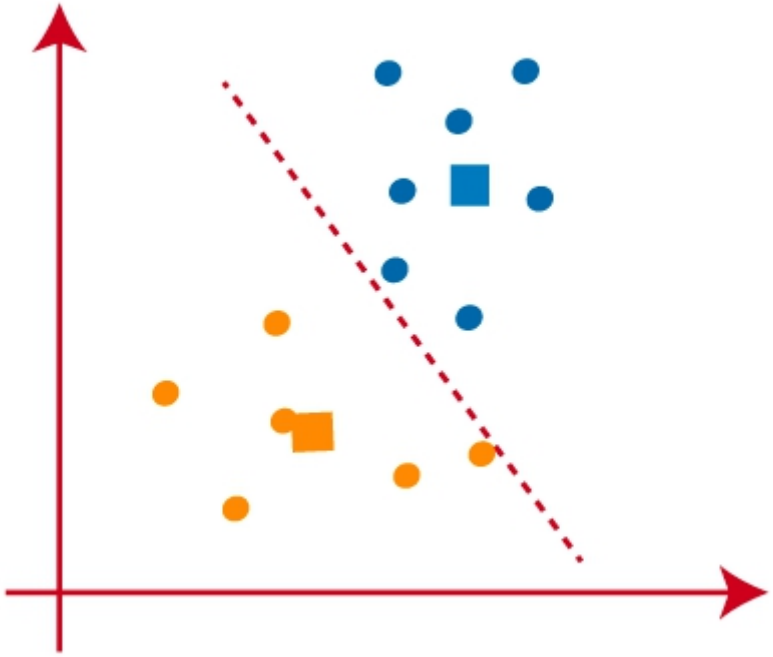


ШАГ 6: Теперь, когда произошло переназначение, мы повторим предыдущий шаг поиска новых центроидов.

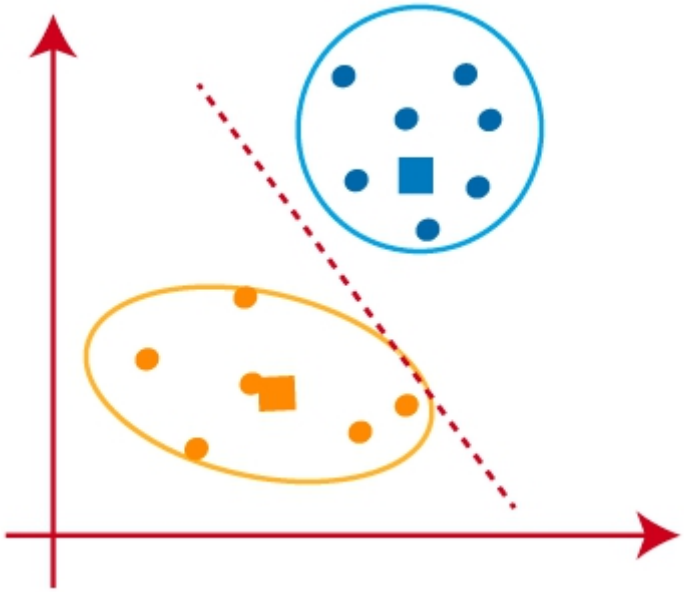


ШАГ 7: Мы повторим описанную выше процедуру для определения центра тяжести центроидов, как показано ниже.

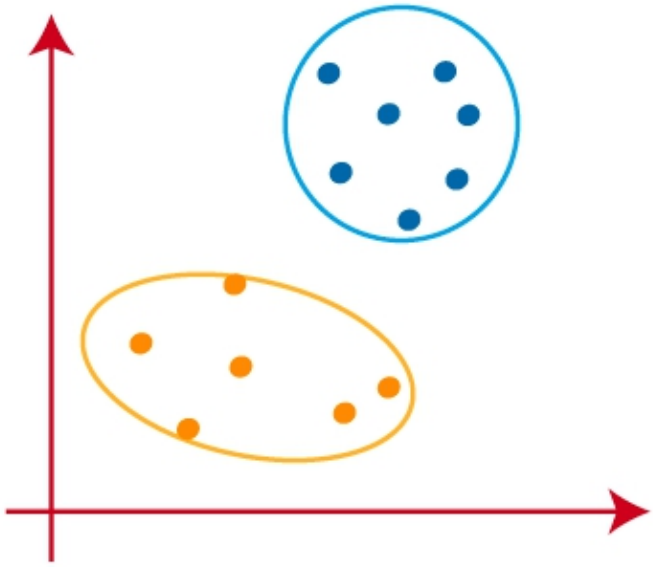
ШАГ 8: Как и на предыдущих этапах, мы проведем срединную линию и переназначим точки данных после определения новых центроидов.



ШАГ 9: Наконец, мы сгруппируем точки в зависимости от их расстояния от срединной линии, убедившись, что установлены две отдельные группы и что в одну группу не включены разнородные точки.



Окончательный кластер выглядит следующим образом:



Преимущества и недостатки
Преимущества

Ниже приведены некоторые особенности алгоритмов кластеризации K-средних:

Это просто понять и применить на практике.
K-средние были бы быстрее, чем иерархическая кластеризация, если бы у нас было большое количество переменных.
Кластер экземпляра можно изменить при повторном вычислении центроидов.
По сравнению с иерархической кластеризацией K-средние создают более плотные кластеры.

Недостатки

Некоторые из недостатков методов кластеризации K-средних заключаются в следующем:

Количество кластеров, т. е. значение k, трудно оценить.
Основное влияние на выпуск оказывают начальные входные данные, такие как количество кластеров в сети (значение k).
Последовательность, в которой вводятся данные, оказывает значительное влияние на конечный результат.
Он довольно чувствителен к масштабированию. Если мы масштабируем наши данные, используя нормализацию или стандарты, результат будет совершенно другим. конечный результат
Нецелесообразно выполнять задачи кластеризации, если кластеры имеют сложную геометрическую форму.