Интерпретация. Метод k-средних

Андрей Мещеряков
Senior Data Scientist в EPAM Systems

Skillbox

Андрей Мещеряков

5+ лет опыта

в обучении и развитии персонала

Создавал

решения на основе машинного обучения для ритейла и маркетинга

Цель модуля

Освоить аналитические методы подбора гиперпараметров и познакомиться с методами бизнес-интерпретации результатов кластеризации.

Метод k-средних. Выбор числа кластеров

Цель видео

Изучить аналитические методы подбора гиперпараметров метода k-средних.

Выбор оптимального числа кластеров

Ранее вы узнали, как использовать метод локтя и метрику инерции для выбора оптимального числа кластеров.

У этого способа есть ряд недостатков:

- точку перелома на графике зависимости инерции от количества кластеров необходимо искать глазами
- метрика инерции учитывает только внутрикластерные расстояния, то есть расстояния от точек до ближайших центроид

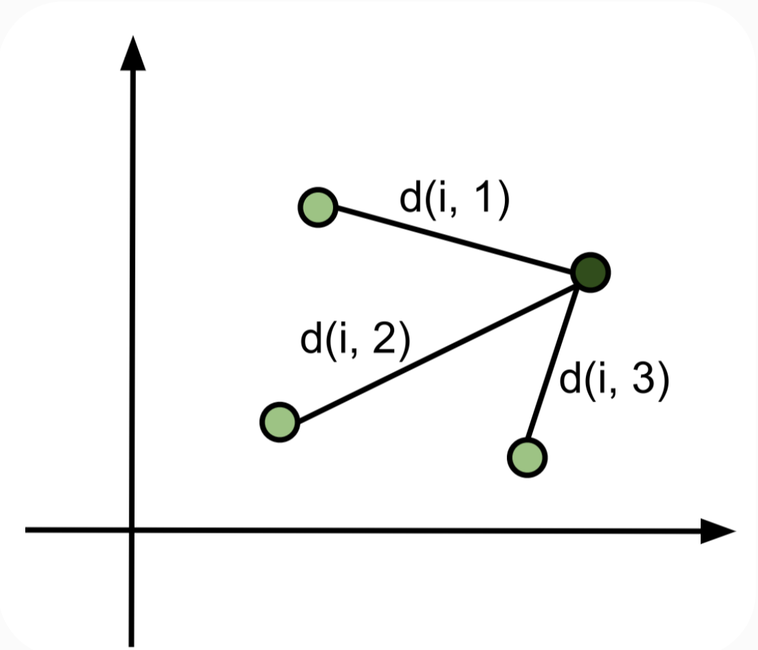
Коэффициент силуэта

Коэффициент силуэта совмещает в себе как внутрикластерные, так и межкластерные расстояния.

Для начала нужно дать определения этим расстояниям.

Внутрикластерное расстояние

Среднее внутрикластерное расстояние для какой-либо точки данных — это среднее расстояние от этой точки до всех остальных точек данных, принадлежащих тому же кластеру.



Изображение: работа спикера Мещерякова Андрея

Внутрикластерное расстояние

Среднее внутрикластерное расстояние для какой-либо точки данных — это среднее расстояние от этой точки до всех остальных точек данных, принадлежащих тому же кластеру.

$$a_i = \frac{1}{|C_i| - 1} \sum_{j \in C_i, i \neq j} d(i, j)$$

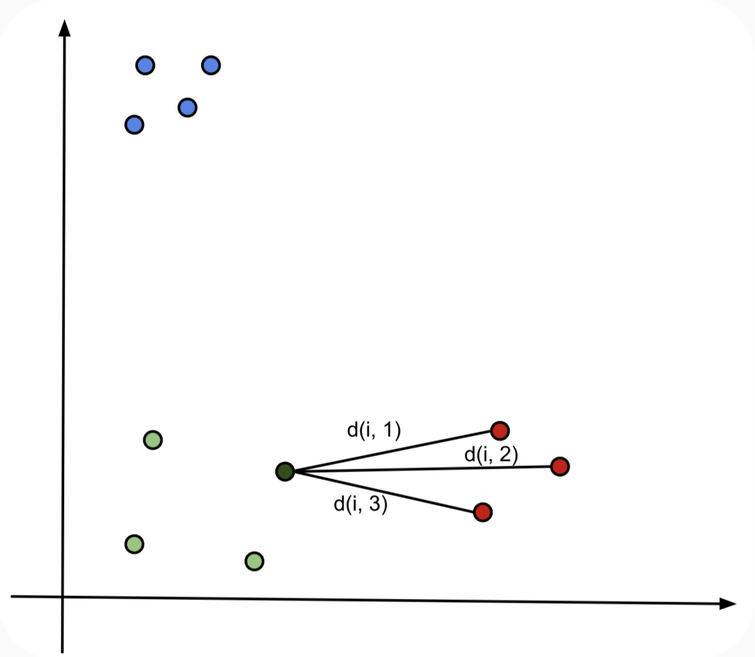
Количество точек в кластере

Расстояние от точки i до точки j

Изображение: работа спикера Мещерякова Андрея

Межкластерное расстояние

Среднее межкластерное расстояние для какой-либо точки данных — это среднее расстояние от этой точки данных до всех точек данных ближайшего к ней другого кластера.



Изображение: работа спикера Мещерякова Андрея

Межкластерное расстояние

Среднее межкластерное расстояние для какой-либо точки данных — это среднее расстояние от этой точки данных до всех точек данных ближайшего к ней другого кластера.

$$b_i = \min_{k \neq i} \frac{1}{|C_k|} \sum_{j \in C_k} d(i, j)$$

Чтобы найти ближайший кластер, нужно посчитать расстояние до всех и выбрать минимальное

Количество точек в кластере

Расстояние от точки i до точки j

Изображение: работа спикера Мещерякова Андрея

Коэффициент силуэта

Коэффициент силуэта для точки i определяется по следующей формуле:

$$s_i = \frac{b_i - a_i}{max(b_i, a_i)}$$

Если среднее межкластерное расстояние точки больше внутрикластерного, то коэффициент силуэта будет близок к 1.

Если же внутрикластерное расстояние больше межкластерного, то коэффициент силуэта будет близок к –1. Это говорит вам о том, что эта точка была некорректно приписана к кластеру.

Silhouette Score

Sulhouette Score для всей модели кластеризации определяется как средний коэффициент силуэта для всех точек данных, участвующих в кластеризации.

Силуэт скор также может принимать значения от –1 до 1.

Выводы

Вы изучили аналитический способ интерпретации результатов алгоритма кластеризации с помощью silhouette score

Преимущества этого способа перед методом локтя в том, что silhouette score учитывает как внутрикластерные, так и межкластерные расстояния, что позволяет оценить, насколько хорошо кластеры отделены друг от друга.