Кластерный анализ
метод к – средних

k-means

Изобретатели

Steinhaus	1956
Lloyd	1957
MacQueen	1967

Когда-то считался кластеризацией на скорую руку

В пакете SPSS Quick Cluster.

В пакете SAS – процедура FASTCLUS.

Быстрый не значит небрежный.

Алгоритм метода к средних-1

Заранее определяется k - число кластеров.

Выбирается k точек — центры кластеров.

Процедуру для определения числа кластеров обсудим позднее.

Алгоритм метода к средних-2

Итеративно.

Правило 1

Объект приписывается к тому кластеру, чей центр ближайший.

Правило 2

Центр кластера переносим в центр тяжести кластера.

Рассмотрим работу метода на примере.

Скрипт k_means_ex_pictures_2.r

Схожесть объектов

Используется только евклидово расстояние.

Недостаток исправляется в других вариантах метода к-средних.

Например k-медоиды

Реализован в пакете flexclust (R)

Результат кластеризации зависит от расположения начальных центров кластеров

Начальное расположение центров кластеров.

Наиболее популярны три метода.

1 **Forgy** (фамилия).

Случайным образом выбираются k наблюдений. Они и будут начальными центрами кластеров.

2. **Случайное разбиение (Random Partition)**.

Каждое наблюдение случайным образом приписывается к одному из кластеров. Находятся центры тяжести кластеров. Они и будут начальными центрами.

Начальное расположение центров кластеров.

- 3) **K-means++**
- Фанатики метода считают подход самостоятельным способом кластеризации
- Отличается от K-means только процедурой создания начального расположения центров
- Популярен среди пользователей Питона

Алгоритм K-means++

1) Случайное наблюдение равновозможно выберите в качестве первого центроида.
2) Вычислите расстояние от каждого наблюдения до ближайшего из ранее выбранных центроидов.
3) Выберите очередной центроид среди наблюдений случайным образом. Вероятность выбора наблюдения прямо пропорциональна ее расстоянию до ближайшего из уже выбранных центроидов. В результате максимальные шансы быть выбранной очередным центроидом у той точки, чье расстояние до ближайшего центроида максимально.
4) Повторите шаги 2 и 3, пока не будут выбраны *k* центроидов.

Определение числа кластеров

Задаем разное число кластеров

k = 2, 3, …, 100

k = 2, 4, 8, …, 512

Строим график каменистая осыпь

Выбираем лучшую кластеризацию.

Объем вычислений возрастает в 100 раз...

Математическая модель

$$W_S = \sum_{i=1}^{k} \sum_{x \in S_i} \|x - \bar{x}_i\|^2$$

$$S_{optim} = \underset{S}{argmin} \sum_{i=1}^{k} \sum_{x \in S_i} \|x - \bar{x}_i\|^2$$

Процедура Python

- sklearn.cluster.Kmeans
- Используются только квадраты евклидова расстояния¶

Отступление

- Расстояние Варда в иерархическом кластерном анализе

  https://en.wikipedia.org/wiki/Nearest-neighbor_chain_algorithm#Complete_linkage_and_average_distance

Недостатки k-means

Только евклидово расстояние.

Решение зависит от начальных центров.

Надо определять число кластеров

Слишком много вычислений расстояний.

На поздних итерациях мало точек меняют кластер, вычисления для "определившихся" точек можно исключить. Только как?

-
- Что делать, если вместо матрицы данных имеем матрице попарных расстояний?

- Метод k - medoids

Родственники k-means

- X-means
- C-means (Нечеткий алгоритм кластеризации)
- Форель (FOREL) Новосибирск, Загоруйко
- Mini Batch K-Means (Питон)

Mini Batch K-Means

- Mini-batches — подмножества набора данных, случайно выбираются на каждой итерации
- При обновлении центров кластеров на каждой итерации исользуется только свой Mini-batch
- In practice this difference in quality can be quite small (https://scikit-learn.org/stable/modules/clustering.html)

Метод к-средних и уменьшение дисперсии

- После кластеризации выборочное среднее заменяется на выборочные средние для каждого кластеров

- Цена вопроса: вместо одного типичного значения появляется несколько, но нас это устраивает. (Уже обсуждали, когда на гистограмме видели мультимодальность распределения)

- Выигрыш: уменьшение дисперсии

- Риски: логнормальное распределение и ленточные кластеры