Кластерный анализ
метод к – средних

K-means

Версия 07.22

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

Результат кластеризации зависит от расположения начальных центров кластеров

Схожесть объектов

Используется только евклидово расстояние.

Недостаток исправляется в модификациях метода к-средних.

Например k-медоиды

Реализован в пакете flexclust (R)

Математическая модель

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

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

- Lloyd

- Forgy
- MacQueen
- Hartigan–Wong

В следующей части

- определять число кластеров,
- как считается расстояние варда
  в иерархическом кластерном анализе
- обсудим недостатки метода к-средних
  и методы их лечения
- метод к-медоидов
- обобщения метода к-средних

Начальное расположение центров кластеров.

Наиболее популярны три метода.

1 **Forgy** (фамилия).

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

В процедуре sklearn.cluster.Kmeans

задается значением параметра

init= „random“

Начальное расположение центров кластеров.

- 2. **Случайное разбиение (Random Partition)**.
- Каждое наблюдение случайным образом приписывается к одному из кластеров. Находятся центры тяжести кластеров. Они и будут начальными центрами.

В процедуре sklearn.cluster.Kmeans отсутствует.

Начальное расположение центров кластеров.

- 3) **K-means++**
- Используется в процедуре sklearn.cluster.Kmeans по умолчанию

Алгоритм K-means++

1) Первым центроидом назначьте случайное наблюдение (равновозможно).
2) Вычислите расстояние от каждого наблюдения до ближайшего из ранее выбранных центроидов.
3) Выберите случайным образом наблюдение в качестве очередного центроида. Вероятность выбора наблюдения прямо пропорциональна его 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…

График «каменистая осыпь»

Процедура Python

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

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 нет матрицы данных,

но есть матрица попарных расстояний?

Поиск наилучшей проекции (projection pursuit)

- многомерное шкалирование

  (multidimensional scaling)
- множественный анализ соответствий

  (multiple correspondence analysis)

Метод k - medoids

Медоид - точка кластера,
для которой минимальна сумма расстояний
до других точек кластера

Медоид вместо центроида

Родственники k-means

- modified k means clustering algorithm
- Google: 49 миллионов страниц

Родственники k-means

- X-means
- C-means (Нечеткий алгоритм кластеризации)
- Форель (FOREL) Новосибирск, Загоруйко

  Формальный Элемент

  правильно произносить ФорЭл
- 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 используется только свой Mini-batch
- На практике различия в качестве кластеризаций могут быть малыми

  (In practice this difference in quality can be quite small)
  (https://scikit-learn.org/stable/modules/clustering.html)

Mini Batch K-Means

исходная статья

Sculley

Web-scale k-means clustering.

In Proceedings of the 19th international conference on World wide web,

pages 1177–1178. ACM, 2010.

Алгоритм Mini-batch k-Means обозначения

X - таблица данных

k - число кластеров,

b - число наблюдений в mini-batch-е

T - число итераций алгоритма,

C - вектор кодов кластера

$c \in C$ - код кластера

x - наблюдение (случайное) из таблицы X

v - вектор, число наблюдений в кластере

Алгоритм Mini-batch k-Means

```
v ← 0                  //  обнуляем число наблюдений в кластерах
for i = 1 to T do
    M ← из X случайно выбрано b наблюдений
    for x ∈ M do
        d[x] ← f(C,x)    // код ближайшего к  x центра кластера
    end for
    for x ∈ M do
        c ← d[x]              //  код ближайшего к  x центра кластера
        v[c] ← v[c] + 1   // увеличиваем объем кластера
        η ← 1 / v[c]          // вычисляем вес наблюдения
        c ← (1 − η)c + ηx       // Пересчитываем центр кластера
    end for
end for
```

Mini Batch K-Means и K-Means совпадают или нет?

Метод градиентного спуска

Bottou, Bengio

Convergence Properties of the K-Means Algorithms

Advances in Neural Information Processing Systems 7 (NIPS 1994)

Авторы

Steinhaus	1956
Lloyd	1957
MacQueen	1967

Отступление

- Расстояние Варда в иерархическом кластерном анализе

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

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