Кластерный анализ
метод к – средних

предложен MacQueen в 1967 году

(цит. по Kopparapu, Desai Bayesian Approach to Image Interpretation стр 99)

В пакете SPSS Quick Cluster.

В пакете SAS – процедура FASTCLUS.

Быстрый не значит небрежный.

Идея метода

Заранее определяется k - число кластеров.
Это непросто. Хотя ниже обсуждается процедура для определения числа кластеров.

Выбирается k точек — центры кластеров.

Далее в цикле применяем правила.

Правило 1
Объект приписывается к тому кластеру, чей центр ближайший.

Правило 2
Центр кластера — центр тяжести объектов кластера.

Используется только евклидово расстояние.

Недостаток исправляется в других вариантах метода к-средних.
Например k-медоиды
Реализован в пакете flexclust

Рассмотрим работу метода на примере.

Скрипт k_means_ex_pictures_2.r

Результат зависит от начальных центров кластеров

Начальное расположение центров кластеров.

Наиболее популярны два 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.

Определение числа кластеров
То, что надо задать число кластеров, не обременительно, ведь можно прогнать процедуру, задав разное число кластеров.

И выбрать наилучшую кластеризацию.

- Математическая модель

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?