PS5841

Data Science in Finance & Insurance

K-Means Clustering

Yubo Wang

Autumn 2021

K-Means Clustering

- Pre-specify K clusters
- Each observation belongs to at least one of the K clusters
- No observation belongs to more than one cluster
- Clustering driven by minimizing within-cluster variations, e.g. squared Euclidean distance

KMC objective

Global minimum

$$\min_{C_1,\dots C_K} \left\{ \sum_{k=1}^{K} W(C_k) \right\}$$

$$= \min_{C_1,\dots C_K} \left\{ \sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{i,j \in C_k} \sum_{l=1}^{p} \left(x_{il} - x_{jl}\right)^2 \right\}$$

$$= \min_{C_1,\dots C_K} \left\{ \sum_{k=1}^{K} 2 \sum_{i \in C_k} \sum_{l=1}^{p} (x_{il} - \bar{x}_{kl})^2 \right\}$$

Mean for feature l in cluster C_k: $\bar{x}_{kl} = \frac{1}{|C_k|} \sum_{i \in C_k} x_{il}$

KMC Algo

- Repeat the following N times (N initial random clustering)
- Select the one that minimizes the objective

$$\sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{i,j \in C_k} \sum_{l=1}^{p} (x_{il} - x_{jl})^2$$

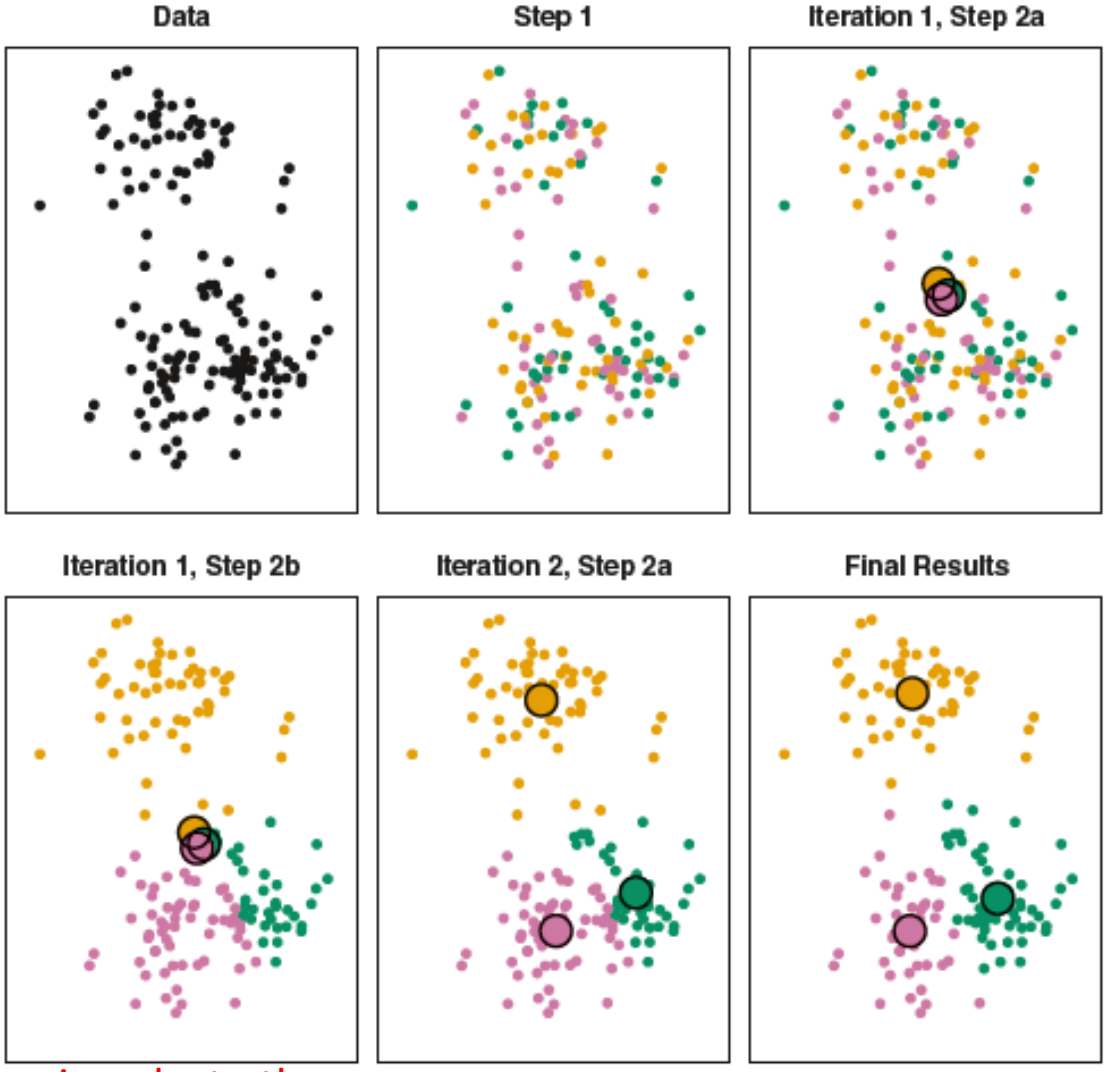
Global minimum

1. Randomly assign a number, from 1 to K, to each of the observations. These serve as initial cluster assignments for the observations.
2. Iterate until the cluster assignments stop changing: Local minimum
   (a) For each of the K clusters, compute the cluster *centroid*. The kth cluster centroid is the vector of the p feature means for the observations in the kth cluster.
   (b) Assign each observation to the cluster whose centroid is closest (where *closest* is defined using Euclidean distance).

K-means Clustering

random assignment

calc cluster centroid



within-cluster variation minimized

assign obs to the nearest centroid

That was