K-means Clustering

Supervised learning: y_i: labels (response)

Unsupervised learning: no y_i response

note: y_i's were used in loss, in testing, ROC/PR

$\{x_i\}_{i=1}^n$ $x_i \in \mathbb{R}^p$ is the data

Task: Learn $z_i \in \{1, \dots, K\}$ (cluster assignments)

eg clusters $C_1 = \{x_1, x_3, x_7\}$ $z_1 = 1, z_3 = 1, z_7 = 1$

$C_2 = \{x_2, x_5, x_6\}$ $z_2 = 2, \dots$

$C_3 = \{x_4, x_8, x_9\}$ $z_4 = 3, \dots$

also learn $m_k \in \mathbb{R}^p$, cluster centers, $k = 1, \dots, K$

Compression interpretation: summarize data w/ only $\{z_i\}_{i=1}^n$, $\{m_k\}_{k=1}^K$

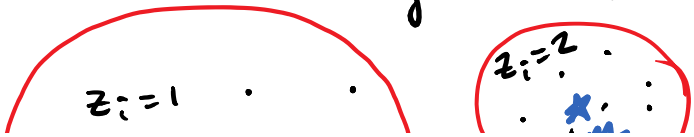
Exploration interpretation: cluster assignment "mean something" about dist^n of data

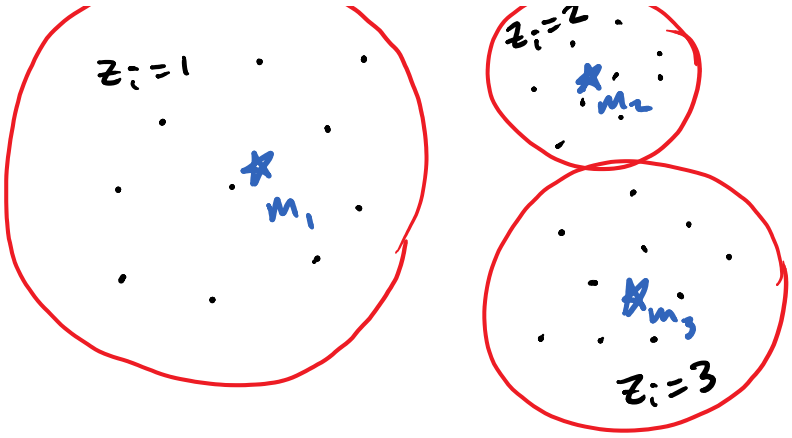
Semi supervised int.: clustering to form features for supervised learning.

K-means alg

Objective: $\min_{z,m} \underbrace{\sum_{i=1}^n \|x_i - m_{z_i}\|^2}_{\text{Distortion } J(z,m)}$

z: cluster assig. $\uparrow$ m: centers





Notice, $J(z,m) = \sum_{k=1}^{K} \sum_{i: z_i = k} \|x_i - m_k\|_2^2$

So fixed z_i, $\min_{m_k} J(z,m) \Rightarrow \hat{m}_k = \frac{1}{|C_k|} \sum_{i \in C_k} x_i$

fixing m_k, $\min_{z_i} J(z,m) \Rightarrow z_i = \operatorname{argmin}_k \|x_i - m_k\|^2$

Lloyd's Algorithm

(1) Init m_k (randomly)

(2) Alternate

(a) Update $z_i \leftarrow \operatorname{argmin}_k \|x_i - m_k\|_2^2$ for all i

(b) Update $m_k \leftarrow \frac{1}{|C_k|} \sum_{i \in C_k} x_i$, $C_k = \{i : z_i = k\}$

- J is non-increasing in iterations
- Finite # of configurations

$\Rightarrow$ Lloyd's will terminate

Hierarchical Clustering

Two main types

Agglomerative : bottom-up

Divisive : top-down

Agglomerative

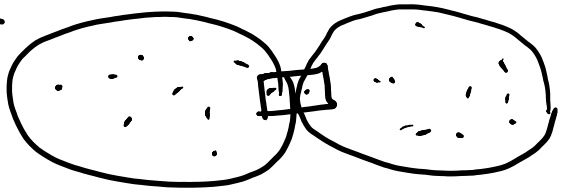
(1) Start w/ $K=n$ and $C_i = \{x_i\}$

(2) Find clusters $C' \neq C$ most similar (*)

(3) Merge clusters $C \cup C'$ repeat (2) ($K \leftarrow K-1$)

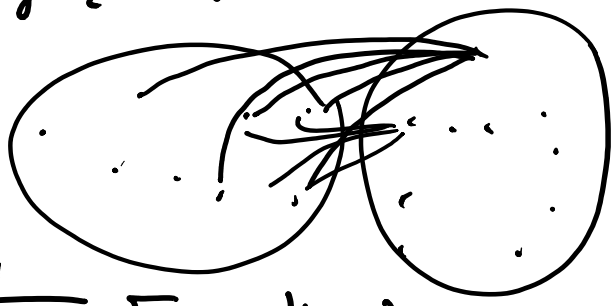
Cluster similarities

Single linkage :



$$d_{sl}(C_1, C_2) = \min_{x \in C_1, y \in C_2} d(x,y)$$

Average linkage



$$d_{al}(C_1, C_2) = \frac{1}{|C_1||C_2|} \sum_{x \in C_1, y \in C_2} d(x,y)$$

Complete linkage :

$$d_{cl}(C_1, C_2) = \max_{x \in C_1, y \in C_2} d(x,y)$$

- Typically, sl tends to produce unbalanced clusters

Dendrogram : visualization tool

Dendrogram : visualization tool