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 $\{z_i\}_{i=1}^n$, $\{m_k\}_{k=1}^K$

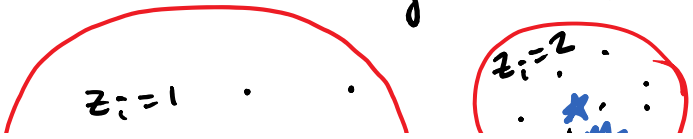
Exploration interpretation: cluster assignment "mean something" about dist^n of data

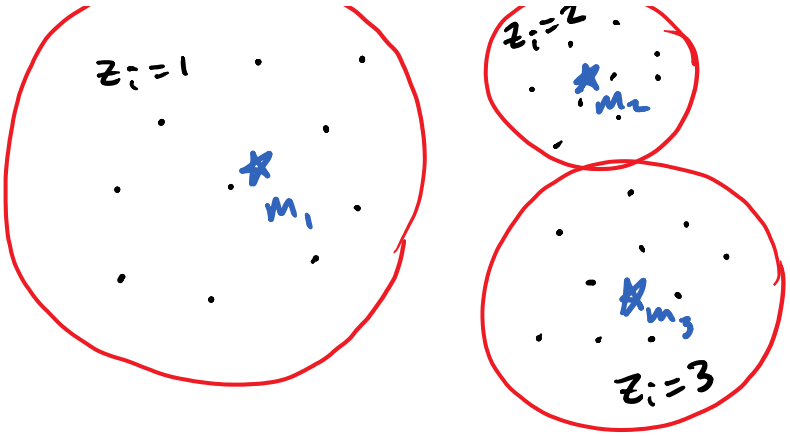
Semi supervised int.: clustering to form features for supervised learning.

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. 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

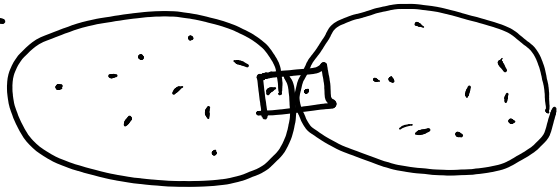
(1) Start w/ $K=n$ and $C_i = \{x_i\}$

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

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

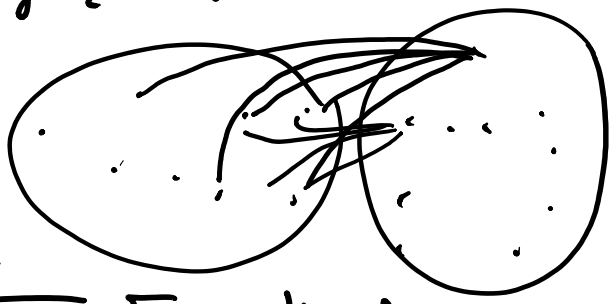
Cluster similarities

Single linkage :



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

Average linkage



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

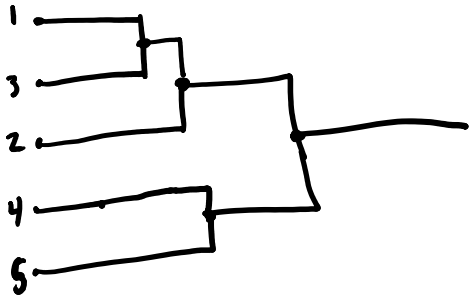
Complete linkage :

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

- Typically, sl tends to produce unbalanced clusters

Dendrogram : visualization tool

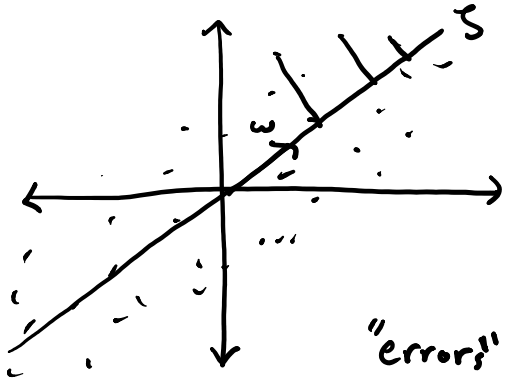
Dendrogram : visualization tool



Principle Component Analysis

Wednesday, May 1, 2019 1:49 PM

$x_i \in \mathbb{R}^p \quad i=1,\dots,n$ (I assume centered $\frac{1}{n}\sum_i x_i = 0$)



w is such that the subspace is $\text{span}\{w\}$, $\|w\|=1$ (currently we are thinking of 1D subspaces)

"errors" $\varepsilon_i^2 = \|x_i - w^T x_i w\|^2$

b.c. $w^T x_i w$ is projection onto S

$$\varepsilon_i^2 = x_i^T x_i - 2x_i^T(w^T x_i w) + (w^T x_i w)^T(w^T x_i w)$$
$$= x_i^T x_i - 2(w^T x_i)(x_i^T w) + (w^T x_i)^2 (w^T w)$$
$$= x_i^T x_i - (w^T x_i)^2$$

$$\min_{\|w\|=1} \sum_i \varepsilon_i^2 \iff \max_{\|w\|=1} \sum_i (w^T x_i)^2 \quad (\dagger)$$

$$\frac{1}{n}\sum_i (w^T x_i)^2 = \frac{1}{n}\sum_i w^T (x_i x_i^T) w = w^T \left(\frac{\sum_i x_i x_i^T}{n}\right) w \quad (\star)$$

Because x_i's are centered $(\star)$ is covariance matrix

$$\hat{\Sigma} = \frac{1}{n}\sum_i x_i x_i^T = \frac{1}{n} X^T X$$

$(\dagger) \iff \max_{\|w\|=1} w^T \hat{\Sigma} w$ (w is eigenvector corresp. to largest eigenvalue of $\hat{\Sigma}$)

Recall Singular Value Decomp. (SVD):

$X = U D V^T \quad U^T U = I, \; V^T V = I, \; D$ is diagonal

U: $n \times n$, D: $n \times p$, V^T: $p \times p$

$$\hat{\Sigma} = \frac{1}{n} X^T X = \frac{1}{n}(UDV^T)^T(UDV^T) = \frac{1}{n}(V D^T \underbrace{U^T U}_{I} D V^T)$$

$= \frac{1}{n} V D^T D V^T$ $\quad d_{ii}$ - singular values of X

s.t. $d_{11}^2 \geq d_{22}^2 \geq \ldots$

$\quad\quad \| \quad\quad\quad \|$

$\quad\quad \lambda_{11} \quad\quad \lambda_{22}$

Eigenvectors of $\hat{\Sigma}$ are right eigenvectors of X

and eigenvalues of $\hat{\Sigma}$ are square of " of X

Let $\Lambda = D^T D$ then we have that λ_{ii} is the i^{th} eigenvalues of $\hat{\Sigma}$.

$$\hat{X}_k = U D_k V^T \quad w/ \quad D_k = \begin{pmatrix} d_{11} & & & & \\ & d_{22} & & 0 & \\ & & \ddots & & \\ & 0 & & d_{kk} & \\ & & & & 0 \\ & & & & & \ddots \\ & & & & & & 0 \end{pmatrix}$$

<u>result</u> $\hat{X}_k$ is the best rank k approximation of X

in the sense that $\sum_i \| X_i - \hat{X}_{ki} \|^2$ is smallest

$$\hat{X}_{ki} = u_i^T D_k V^T = \sum_j u_{ij} (D_k V^T)_j^{\leftarrow \text{column}}$$

$$= \sum_j u_{ij} (D_k)_{jj} V_j = \sum_{j=1}^{k} u_{ij} d_{jj} V_j$$

$k=1$ $\quad w = v_1$ and $X_i^T v_1$ is 1^{st} PC loading

$k=2$ $\quad v_1, v_2$ are basis $\quad X_i^T v_1, X_i^T v_2$ are $1^{st}, 2^{nd}$ PC loadings

$\vdots$

$\downarrow$ row $\quad \swarrow$ column

$(XV)_{ij} = x_i^T v_j \qquad XV = UDV^T V = UD$

<u>alg</u> $(n > p)$

Compute $\hat{\Sigma} = \frac{1}{n} X^T X \quad (O(np^2))$

Compute $\hat{\Sigma} = \frac{1}{n} X^T X$ $(O(np^2))$

Compute $\hat{\Sigma} = V \Lambda V^T$ (spectrum)

Compute loadings $x_i^T v_j$ $i = 1, \ldots, n$
$j = 1, \ldots, K$

PCs have loading and reconstruction

$$\hat{X}_i = \sum_{j=1}^{K} (x_i^T v_j) v_j$$