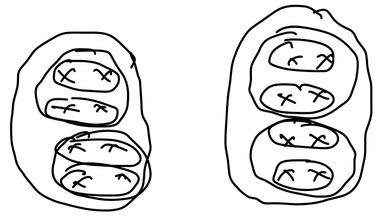
Unsupervised Learning

Supervised learning setting $-$ $(x_i, \boxed{y_i})$, $x_i \in \mathbb{R}^d$

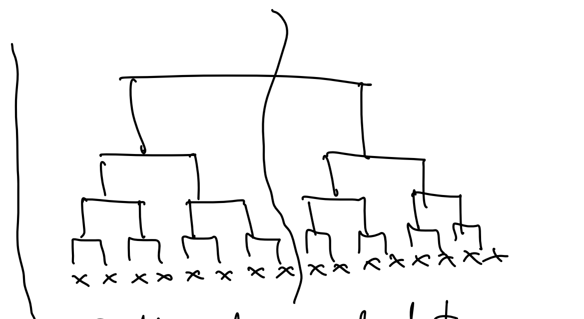
Unsupervised learning setting $-$ $x_i \in \mathbb{R}^d$

Clustering



Hierarchical Agglomerative Clustering

Given : n points, $x_i \in \mathbb{R}^d$



Building hierarchical tree

Let $c = n$, & $e_i = \{x_i\}$, $i = 1, 2, \dots n$

While $c \geq k$

Find the "nearest" pair of distinct clusters, e_i & e_j

Merge e_i & e_j, and call the merged cluster e_i (and delete e_j)

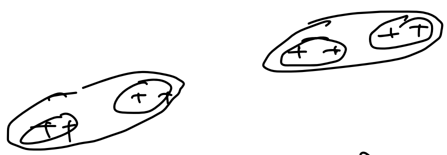
$c \leftarrow c - 1$

End while

Measures of nearness/distance :

$$d_{min}(e_i, e_j) = \min_{\substack{x \in e_i \\ x' \in e_j}} \|x - x'\| \qquad - \text{ Single-link}$$

$$d_{max}(e_i, e_j) = \max_{\substack{x \in e_i \\ x \in e_j}} \|x - x'\| \qquad - \text{ Complete-link}$$



Long-chains (single-link)

$$d_{mean}(e_i, e_j) = \|m_i - m_j\|, \quad m_i \text{ is the mean of cluster } e_i, \quad m_j \text{ " " " " } e_j$$

$$d_{avg}(e_i, e_j) = \frac{1}{n_i n_j} \sum_{\substack{x \in e_i \\ x' \in e_j}} \|x - x'\|, \quad \text{where } n_i = |e_i|, \; n_j = |e_j|.$$

Objective Function for k-clusters

$$\boxed{J = \sum_{i=1}^{k} \underbrace{\left(\sum_{x \in e_i} \|x - m_i\|^2 \right)}_{\text{scatter/width of cluster } i}, \quad m_i \text{ is centroid/mean of } e_i}$$

Goal: Find clustering into k clusters that minimizes J

Agglomerative Clustering - algorithm that greedily optimizes J

$$J_i = \sum_{x \in e_i} \|x - m_i\|_2^2 \quad , \quad J_e = \sum_{x \in e_e} \|x - m_e\|_2^2 \quad , \quad m_i = \frac{1}{n_i}\sum_{x \in e_i} x$$

$$m_j = \frac{1}{n_j}\sum_{x \in e_j} x$$

Suppose we merge e_i & e_e, then the new mean

$$\hat{m}_i = \frac{1}{n_i + n_j} \sum_{\substack{x \in e_i \\ x \in e_j}} x$$

$$\hat{m}_i = \frac{n_i m_i + n_e m_e}{n_i + n_e} = \frac{n_i m_i + n_e m_i + n_e m_e - n_e m_i}{n_i + n_e}$$

$$\hat{m}_i = m_i + \frac{n_e (m_e - m_i)}{n_i + n_e} = m_e + \frac{n_i}{n_i + n_e}(m_i - m_e)$$

Before merging, $J_i + J_e$

New objective after merging

$$\sum_{x \in e_i} \|x - \hat{m}_i\|^2 + \sum_{x \in e_e} \|x - \hat{m}_i\|^2$$

$$\|a - b\|_2^2 = (a-b)^T(a-b) = \|a\|^2 + \|b\|^2 - 2a^T b$$

$$= \sum_{x \in e_i} \left\|(x - m_i) - \frac{n_e}{n_i + n_e}(m_e - m_i)\right\|^2 + \sum_{x \in e_e} \left\|(x - m_e) - \frac{n_i}{n_i + n_e}(m_i - m_e)\right\|^2$$

$$= \sum_{x \in e_i} \left(\|x - m_i\|^2 + \frac{n_e^2}{(n_i + n_e)^2}\|m_e - m_i\|^2 - \cancel{2(x - m_i)^T(m_e - m_i)} \right)$$

$$+ \sum_{x \in e_e} \left(\|x - m_e\|^2 + \frac{n_i^2}{(n_i + n_e)^2}\|m_i - m_e\|^2 - \cancel{2(x - m_e)^T(m_i - m_e)} \right)$$

J_i ; J_e ; $= 0$; $= 0$

$$\boxed{\sum_{x \in e_i}(x - m_i) = 0}$$

$$\left(\sum_{x \in e_i} x\right) - n_i m_i = 0$$

$$\Rightarrow m_i = \frac{1}{n_i}\sum_{x \in e_i} x$$

$$= J_i + J_e + \underbrace{\frac{n_i n_e^2}{(n_i + n_e)^2}\|m_e - m_i\|^2 + \frac{n_e n_i^2}{(n_i + n_e)^2}\|m_i - m_e\|^2}$$

$$\frac{n_i n_e}{(n_i + n_e)^{\cancel{2}}}\|m_e - m_i\|^2 \cancel{(n_e + n_i)}$$

$$= J_i + J_e + \frac{n_i n_e}{n_i + n_e}\|m_i - m_j\|^2$$

$$\boxed{d_{opt} = \sqrt{\frac{n_i n_e}{n_i + n_e}}\,\|m_i - m_j\|}$$

Partitional Algorithms (Top-down clustering)

"k-means" algorithm

"k-means" objective function $J = \sum_{i=1}^{k} \sum_{x \in \mathcal{C}_i} \|x - m_i\|_2^2$, $x_i \in \mathbb{R}^d$

k-means algorithm.

1. Start with some partitioning of the data into k-clusters $\mathcal{C}_1^{(0)}, \mathcal{C}_2^{(0)}, \mathcal{C}_3^{(0)}, \dots, \mathcal{C}_k^{(0)}$, $t=0$.

2. Calculate the mean/centeroid of each cluster $\mathcal{C}_i^{(t)}$:

$$m_i^{(t)} = \frac{1}{n_i^{(t)}} \sum_{x \in \mathcal{C}_i^{(t)}} x \quad , \quad n_i^{(t)} = |\mathcal{C}_i^{(t)}| \quad , \; i = 1,2,..,k$$

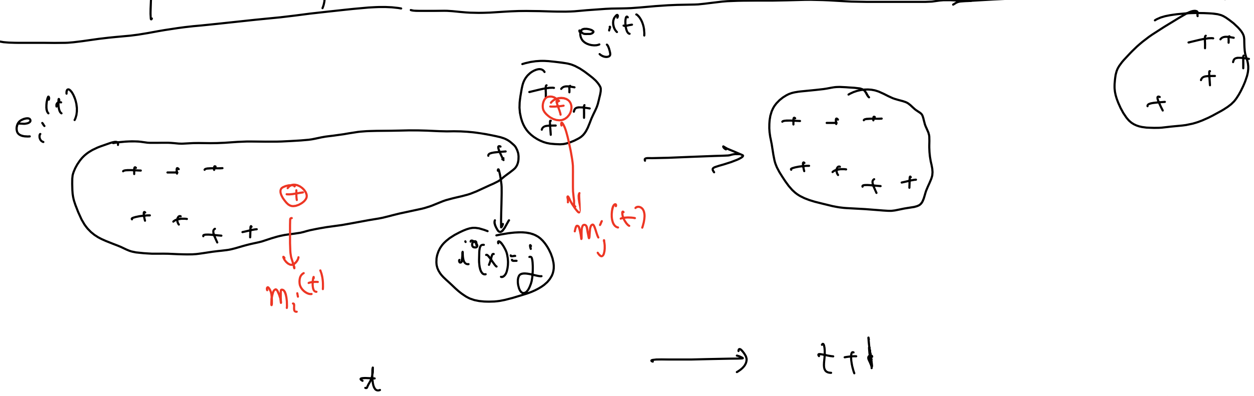
3. For each x, find its cluster index as:

$$i^\circ(x) = \underset{1 \le i \le k}{\operatorname{argmin}} \|x - m_i^{(t)}\|$$

4. Update clusters:

$$\mathcal{C}_i^{(t)} = \{ x \mid i^\circ(x) = i \}$$

5. Repeats steps 2,3 & 4 until "convergence".



Property: kmeans monotonically decreases the objective function J

$$J^{(t)} = \sum_{i=1}^{k} \sum_{x \in \mathcal{C}_i^{(t)}} \|x - m_i^{(t)}\|^2$$

$$J^{(t+1)} \le J^{(t)} \quad ?$$

Proof:

$$J^{(t)} = \sum_{i=1}^{k} \sum_{x \in \mathcal{C}_i^{(t)}} \|x - m_i^{(t)}\|^2$$

$$\ge \sum_{i=1}^{k} \sum_{x \in \mathcal{C}_i^{(t)}} \|x - m_{i^\circ(x)}^{(t)}\|^2 \quad \text{— by step 3 of the k-means algorithm}$$

$$= \sum_{i=1}^{k} \sum_{x \in \mathcal{C}_i^{(t+1)}} \|x - m_{i^\circ(x)}^{(t)}\|^2$$

$$\boxed{y_1, y_2, \dots y_N \qquad \min_z \sum_{i=1}^{N} \|y_i - z\|^2 \Rightarrow z^* = \frac{1}{n}\sum_{i=1}^{N} y_i}$$

$$\geq \sum_{i=1}^{k} \sum_{x \in C_i^{(t+1)}} \|x - m_i^{(t+1)}\| = J^{(t+1)}$$

Hence, the kmeans algorithm monotonically minimizes the kmeans objective function

kmeans separates data by linear surfaces (hyperplanes)

Suppose $k = 2$

Locus of all points equidistant to m_1 & m_2

$\|x - m_1\|^2 = \|x - m_2\|^2$ - equation of a hyperplane