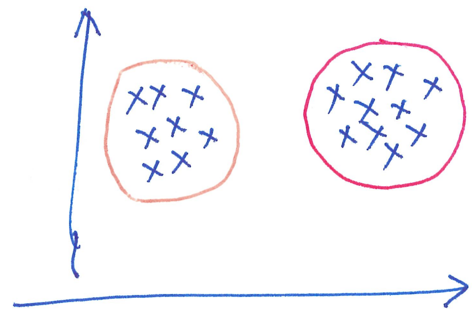
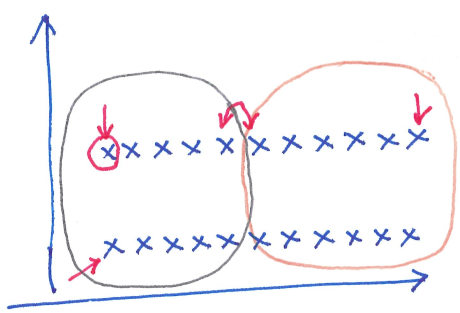
CS 307 Lecture 25

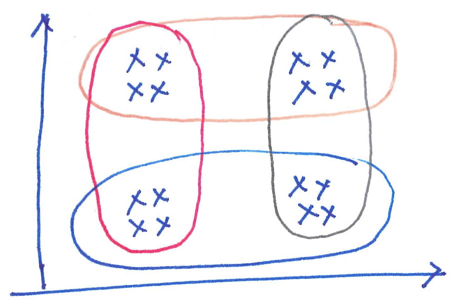
Clustering Unsupervised learning.

Training set $S = \{x^{(i)}\}_{i=1}^{n}$

"Cluster Data": Group examples that are similar in one cluster and keep examples that are dissimilar in different clusters.







~~K-means / Clustering~~:

Clustering setup:

$\rho(x,x) = 0 \quad \forall x \in \mathbb{R}^d$

$\rho(x,y) = \rho(y,x) \quad \forall x,y$

$\rho(x,y) + \rho(y,z) \geq \rho(x,z) \quad \forall x,y,z.$

Input: $S = \{x^{(i)}\}_{i=1}^{n}$ ($x^{(i)} \in \mathbb{R}^d$) and distance metric

$\rho : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, k (# clusters)

Output: $C_1, C_2, C_3 \ldots C_k$ s.t. $C_i \cap C_j = \phi \;\; \forall i,j$.

$C_1 \cup C_2 \cup \cdots \cup C_k = S$.

"$C_1, C_2, C_3 \ldots C_k$" is a partition of S".

k-means Clustering: On input S, ρ, k.

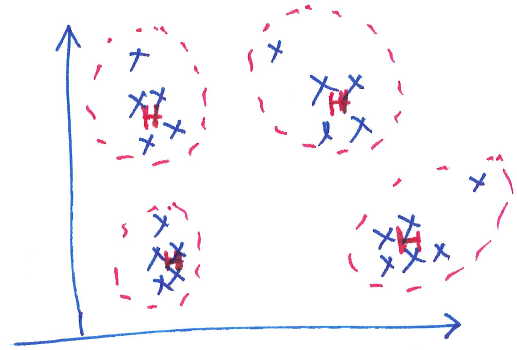
Goal is find $C_1, C_2 \dots C_k$ s.t.

$$\min \sum_{j=1}^{k} \sum_{x^{(i)} \in C_j} \rho(x^{(i)}, \mu_j)$$

$$\mu_j = \frac{\sum_{x^{(i)} \in C_j} x^{(i)}}{|C_j|}$$

NP-hard.

k-means Clustering Algorithm:



- Identify centers $\mu_1 \dots \mu_k$.
- $y^{(i)}$ identifies the cluster to which $x^{(i)}$ belongs.

$\in \{1, 2, 3 \dots k\}$.

$\| \ \|_2^2$

Input: S, ρ, k.

Initially: Let $\mu_1, \mu_2 \dots \mu_k$ be some centers.

y_{old} to something.

while ($y_{old} \neq y$) do

$y_{old} = y$

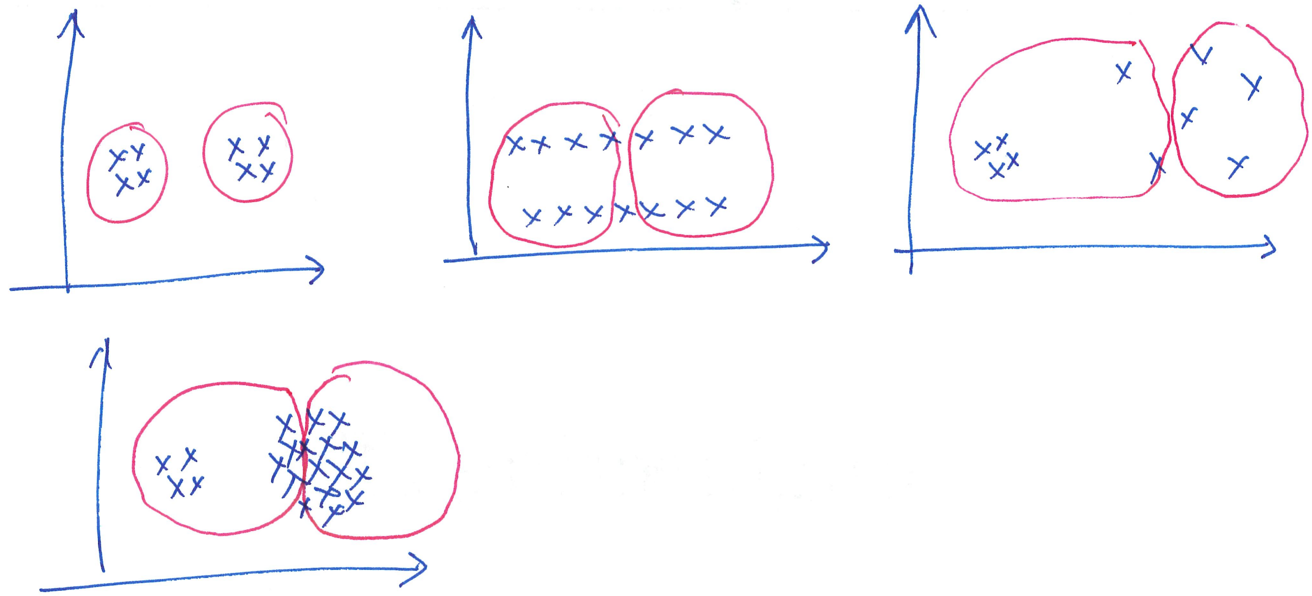
$$\forall i. \quad y^{(i)} = \underset{j \in \{1, \dots k\}}{\operatorname{argmin}} \ \rho(x^{(i)}, \mu_j)$$

$$\forall j \quad \mu_j^* = \frac{\sum_{y^{(i)}=j} x^{(i)}}{|\{ i \mid y^{(i)} = j \}|} = \frac{\sum_{i=1}^{n} 1[y^{(i)} = j]\, x^{(i)}}{\sum_{i=1}^{n} 1[y^{(i)} = j]}$$

$$\underset{C_1, C_2 \dots C_k,\, k}{\operatorname{argmin}} \ \underbrace{\sum_{j=1}^{k} \sum_{x^{(i)} \in C_j} \rho(x^{(i)}, \mu_j)}_{\#\text{clusters} = n = |S|} + \sum_{j=1}^{k} \text{cost}_j$$

(Put each point in its own cluster)

When the data consists of well separated spherical groups of the same size, this algorithm does well.



~~Single Link clustering~~. Linkage based Clustering

Input : S, ρ.

Initialize : $\mathcal{C} = \{ \{ x^{(i)} \} \mid i \in \{1, 2 .. n\} \}$ ← every point in its own cluster.

Repeat

$$C_*, D_* = \operatorname*{argmin}_{C, D \in \mathcal{C}} \text{distance}(C, D)$$

$$\mathcal{C} = (\mathcal{C} - \{C_*, D_*\}) \cup \{C_* \cup D_*\}$$

until ______

$|\mathcal{C}| = k$.

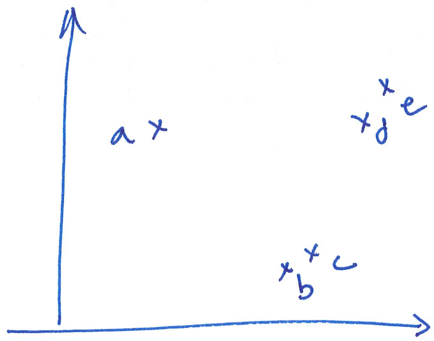
$|\mathcal{C}| = 1$

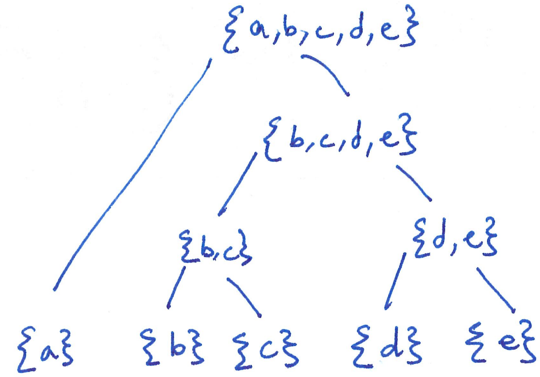
Single ~~link~~ linkage clustering

$$\text{distance}(C, D) = \min_{x \in C, y \in D} \rho(x, y)$$

Average single linkage clustering

$$\text{distance}(C, D) = \frac{1}{|C| \cdot |D|} \sum_{\substack{x \in C \\ y \in D}} \rho(x, y)$$





~~Dendogram~~.

DENDROGRAM.