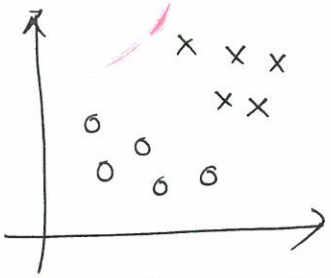
(1)

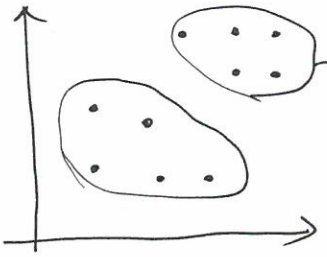
Clustering – Unsupervised learning Algo.

In Supervised learning



Training Data: $\{(x_1, y_1), (x_2, y_2) \ldots (x_N, y_N)\}$

In unsupervised learning



Training Data: $\{(x_1), (x_2), (x_3), \ldots (x_N)\}$

↓

No labels

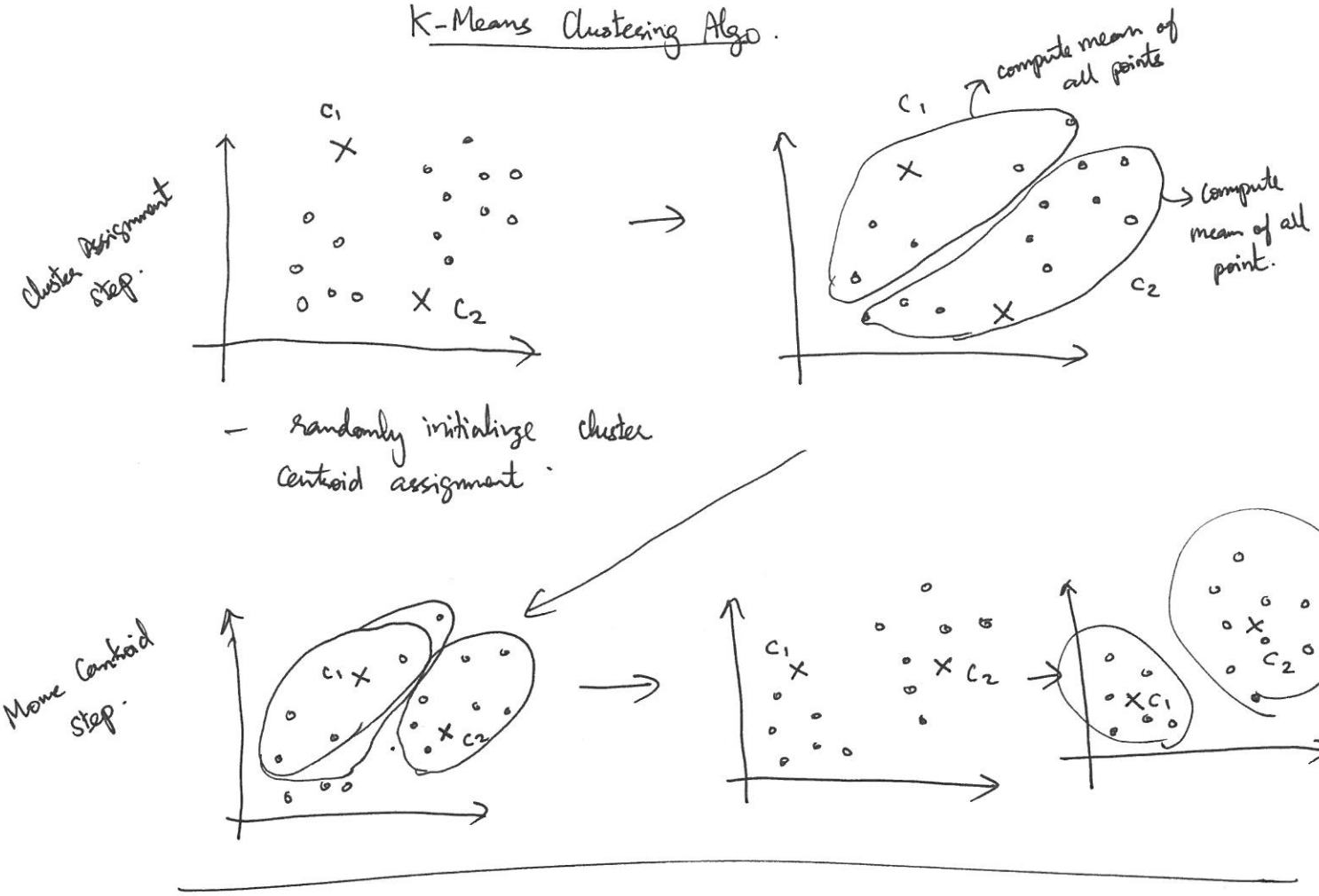
Applications:

- To find structure/patterns in data

1. Market Segmentation
2. Social Network Analysis
3. Organize Computing clusters
4. Space data analysis.

K-Means Clustering.

K-Means Clustering Algo.



- Randomly initialize cluster centroid assignment.

K-Mean Clustering:

Input:

- K (number of clusters)
- Training set $(x_1, x_2 \dots x_N) \Rightarrow$ No $x_0 = 1$ ✗

$$\downarrow$$

$$(x_{11}, x_{12} \dots x_{1d})$$

$$x_i \in \mathbb{R}^d.$$

1. Randomly initialize K-cluster centroids $\mu_1, \mu_2, \dots, \mu_K \in \mathbb{R}^d$
2. Repeat {
3. (Cluster Assignment Step) for $i = 1$ to N

$$c_i := \{\text{index (from 1 to K) of cluster centroid close to } x_i\} \rightarrow \min_k \|x_i - \mu_k\|^2 \rightarrow c_i$$

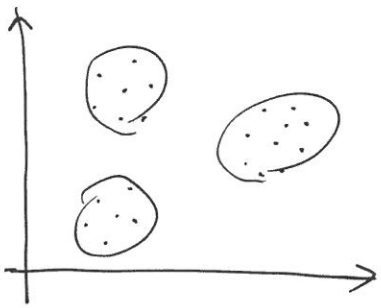
4. (Move centroid step) for $k = 1$ to K

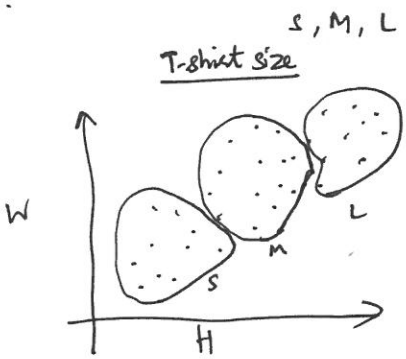
$$\mu_k := \{\text{Average (mean) of points assigned to cluster } k.\}$$

}

Note: if no points assigned to a cluster, then eliminate that cluster.

K-means for Non-seperated Clusters:





K-Means Optimization Objective:

K-means:

- c_i = index of clusters $(1, 2, \cdots K)$ to which x_i is currently assigned.
- μ_k = cluster centroid k. $(\mu_k \in \mathbb{R}^d)$

K ↓ # clusters

$k = \{1, 2, \cdots K\}$ ↓ index

μ_{c_i} = cluster centroid of cluster to which x_i has been assigned.

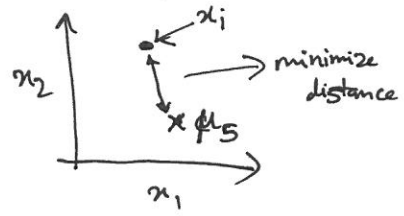
$x_i \to 5 \quad c_i = 5 \Rightarrow \mu_{c_i} = \mu_5$

Optimization Objective:

$$E(c_1, \cdots c_N, \mu_1, \cdots \mu_K) = \frac{1}{N} \sum_{i=1}^{N} \left\| x_i - \mu_{c_i} \right\|^2$$

$$\min_{c_1 \cdots c_N, \mu_1 \cdots \mu_K} E(c_1 \cdots c_N, \mu_1 \cdots \mu_K)$$

↓ distortion cost fn.



Cluster Assignment Step:

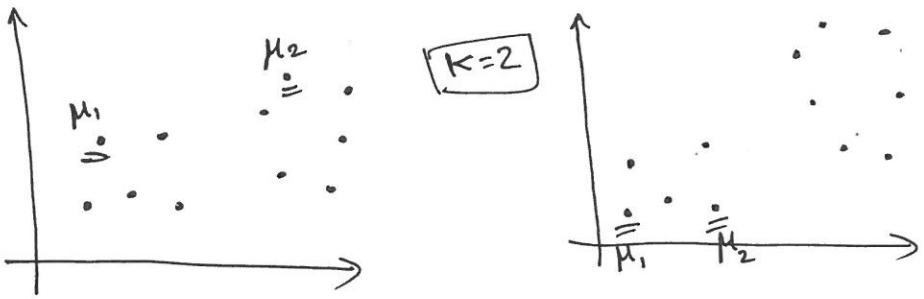
$\min E(\cdots)$ with $(c_1, c_2 \cdots c_N)$ and fixing $(\mu_1, \mu_2, \cdots \mu_K)$

Moving Centroid Step:

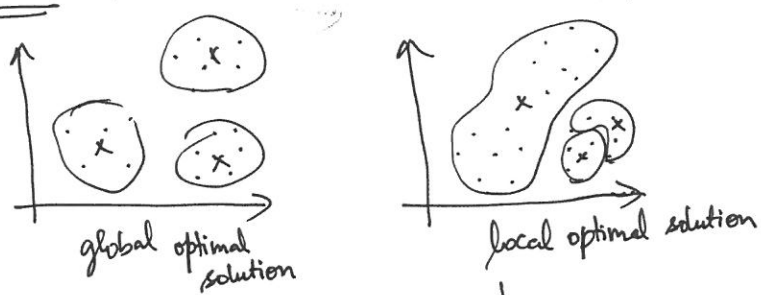
$\min E(..)$ with $(\mu_1, \mu_2, \cdots \mu_K)$ and fixing $(c_1, c_2, \cdots c_N)$

④ Random Initialization:

- Should have $K < N$
- Randomly pick K training samples
- Set $\mu_1 \dots \mu_K$ equal to these K examples.



Local Optima :→ K-mean can stuck at local optima.



↓

To avoid this, initialize and run K-means multiple times

↓

100 -times.

↙

{ Randomly initialize and Run K-means } 100 times

↓

Pick clustering that give lowest cost $E(\dots)$

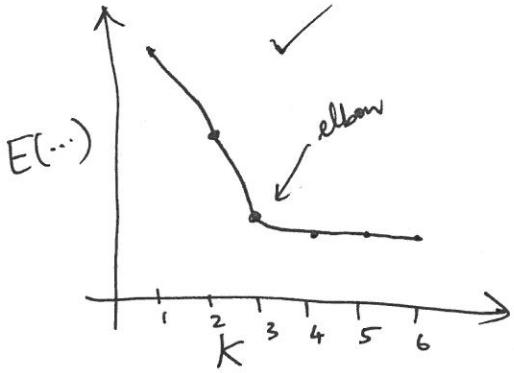
↓

$K = 2 \dots 10$ problem size.

(5)

Choose value of K ? → # clusters

- Choose K manually → depending on application → T-shirt sizes (S, M, L). (XS, .., XL).
- Use Elbow method



(choose K=3 as it is the elbow in the curve)