Unsupervised learning : Find patterns

① Clustering

application: Market segmentation, social network analysis ...

algorithm: k-means

# of clusters.

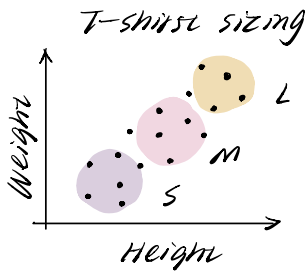
Randomly initialize k cluster centroids.

(i) Cluster assignment (color points based on which is closer)

(ii) Centroid movement (move centroid to the average)

Repeat (i) and (ii) until no change

e.g. k-means for non-separated clusters



K-means optimization objective

→ $c^{(i)}$ = index of cluster (1,2,...,K) to which example $x^{(i)}$ is currently assigned

→ μ_k = cluster centroid k ($\mu_k \in \mathbb{R}^n$) K $k \in \{1,2,\dots,K\}$

$\mu_{c^{(i)}}$ = cluster centroid of cluster to which example $x^{(i)}$ has been assigned $x^{(i)} \to 5$ $c^{(i)} = 5$ $\mu_{c^{(i)}} = \mu_5$

Optimization objective:

→ $J(c^{(1)},\dots,c^{(m)},\mu_1,\dots,\mu_K) = \frac{1}{m}\sum_{i=1}^{m} ||x^{(i)} - \mu_{c^{(i)}}||^2$ Distortion

$$\min_{\substack{c^{(1)},\dots,c^{(m)},\\ \mu_1,\dots,\mu_K}} J(c^{(1)},\dots,c^{(m)},\mu_1,\dots,\mu_K)$$

K-means algorithm

Randomly initialize K cluster centroids $\mu_1, \mu_2, \dots, \mu_K \in \mathbb{R}^n$

Repeat {

Cluster assignment step: Minimize $J(\dots)$ wrt $c^{(1)}, c^{(2)}, \dots, c^{(m)}$ (hold $\mu_1, \dots, \mu_K$ fixed)

for i = 1 to m

$c^{(i)}$:= index (from 1 to K) of cluster centroid closest to $x^{(i)}$

Move centroid: for k = 1 to K

μ_k := average (mean) of points assigned to cluster k

}

Minimize $J(\dots)$ wrt $\mu_1, \dots, \mu_K$

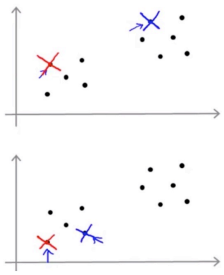
Random initialization

Should have $K < m$

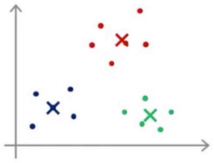
Randomly pick K training examples.

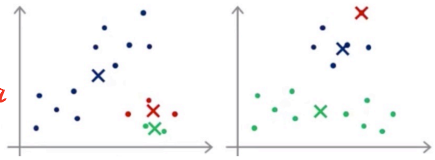
Set $\mu_1, \ldots, \mu_K$ equal to these K examples. $\mu_1 = x^{(i)}$ $\mu_2 = x^{(j)}$

K=2



problem ⇒ local optima





For i = 1 to 100 { 50 - 1000

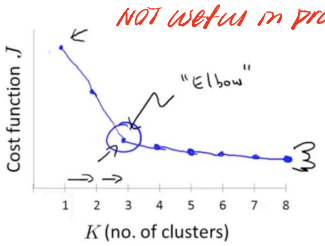
→ Randomly initialize K-means.
Run K-means. Get $c^{(1)}, \ldots, c^{(m)}, \mu_1, \ldots, \mu_K$.
Compute cost function (distortion)
→ $J(c^{(1)}, \ldots, c^{(m)}, \mu_1, \ldots, \mu_K)$
}

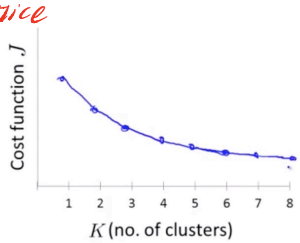
Pick clustering that gave lowest cost $J(c^{(1)}, \ldots, c^{(m)}, \mu_1, \ldots, \mu_K)$

Choosing the value of K

Elbow method: ⇒ problem : No clear elbow

NOT useful in practice





Sometimes, you're running k-means to get clusters to use for some later/downstream purpose. Evaluate k-means based on a metric for how well it performs for later purpose