K-means Clustering Algorithm

Sid

1) Randomly initialize 2 points : cluster centroid.
2) Cluster assignment : assign each data point into 2 groups based on proximity to cluster centroids.
3) Move centroid : compute average for all points inside each centroid group and move centroids to average.
4) Repeat (2 – 3)

Input
- K (# of clusters)
- Training Set $\{ x^{(1)}, \dots, x^{(m)} \}$ $\quad x^{(i)} \in \mathbb{R}^{(n)}$ (drop $x^{(0)}$)

Randomly initialize K cluster centroid : $\mu_1, \dots \mu_K$

Repeat {

Cluster assignment

for $i = 1 \to m$

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

$\underset{k}{\min} \, \| x^{(i)} - \mu_k \|^2$ $\to c^{(i)}$

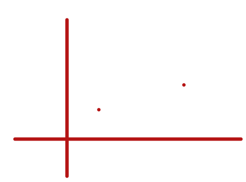
Move cluster

for $k = 1 \to K$

$\mu_k :=$ avg of points assigned to cluster k

$= \frac{1}{n} \left[x^{(k_1)} + x^{(k_2)} + \dots + x^{(k_n)} \right]$

}



Optimization objective

$c^{(i)}$ = index of cluster to which $x^{(i)}$ is assigned to

μ_k = cluster centroid k

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

$$J(c^{(1)}, \ldots, c^{(m)}, \mu_1 \ldots \mu_k) = \frac{1}{m} \sum_{i=1}^{m} ||x^{(i)} - \mu_{c^{(i)}}||^2$$

$\min_{c,\mu} J(c,\mu)$ "distortion" of examples

Random Initialization :

- $k < m$
- Randomly pick k training examples
- Set $\mu_1 \cdots \mu_k$ = to examples

- To avoid local optima:

for $i = 1 \rightarrow 100$

randomly initialize k-means

run k-means → compute 'c' & 'm'

compute cost : $J(c,m)$

pick clustering w/ lowest cost.