

# K Means

## Clustering

- Given:  $N$  unlabeled examples  $\{x_1, \dots, x_N\}$  no. of desired partitions  $K$ .
- Goal: : Group the examples into  $K$  “homogeneous” partitions.

Def.

Given a set of observations  $X = \{x_1, x_2, \dots, x_N\} (x_i \in \mathbb{R}^D)$ , partition the  $N$  observations into  $K$  sets ( $K \leq N$ )  $\{\mathcal{C}_k\}_{k=1, \dots, K}$  such that the sets minimize the within-cluster sum of squares:

$$\arg \min_{\mathcal{C}_k} \sum_{i=1}^k \sum_{x \in \mathcal{C}_i} \|x - \mu_i\|^2$$

where  $\mu$  is the mean of points in set  $\mathcal{C}_i$ .

## K Means Algorithm

Algorithm

- (Re)-Assign each example  $x_i$  to its closest cluster center (based on the smallest Euclidean distance)

$$\mathcal{C}_k = \{x_i \mid \|x_i - \mu_k\|^2 \leq \|x_i - \mu_{k'}\|^2, \text{ for } \forall k' \neq k\}$$

- ( $\mathcal{C}_k$  is the set of examples assigned to cluster  $k$  with center  $\mu_k$ ) -  
Update the cluster means

$$\mu_k = \text{mean}(\mathcal{C}_k) = \frac{1}{|\mathcal{C}_k|} \sum_{x \in \mathcal{C}_k} x$$

- Let  $z_{i,k}$  be an indicator

$$z_{i,k} = \begin{cases} 1, & x_i \in \mathcal{C}_k \\ 0, & \text{otherwise} \end{cases}$$

- and  $z_i = [z_{i,1}, \dots, z_{i,k}]^T$  represents the one-hot encoding of  $x_i$ .
- The loss is  $L(\mu, \mathbf{X}, \mathbf{Z}) = \sum_{i=1}^N \sum_{k=1}^K z_{i,k} \|x_i - \mu_k\|^2 = \|\mathbf{X} - \mathbf{Z}\mu\|^2$
- where  $\mathbf{X} \in \mathbb{R}^{N \times D}$ ,  $\mathbf{Z} \in \mathbb{R}^{N \times K}$ ,  $\mu \in \mathbb{R}^{K \times D}$

## Limitations

- Makes **hard** assignments of points to clusters
- Works well only if the clusters are roughly of equal sizes
- K-means also works well only when the clusters are round-shaped and does badly if the clusters have non-convex shapes

## Kernel K Means

- Basic idea: Replace the Euclidean distance/similarity computations in K-means by the kernelized versions

$$\begin{aligned}d(x_i, \mu_k) &= \|\phi(x_i) - \phi(\mu_k)\| \\ \|\phi(x_i) - \phi(\mu_k)\|^2 &= \|\phi(x_i)\|^2 + \|\phi(\mu_k)\|^2 - 2\phi(x_i)^T \phi(\mu_k) \\ &= k(x_i, x_i) + k(\mu_k, \mu_k) - 2k(x_i, \mu_k)\end{aligned}$$