K-Means Clustering

K-Means Clustering Algorithm

- Dividing the data into K groups or partitions
- Given: Training data, $\mathcal{D} = \{\mathbf{x}_n\}_{n=1}^{N}, \mathbf{x}_n \in \mathbb{R}^d$ and K
- Target: Partition the set $\mathcal{D}$ into K clusters (disjoint subsets), $\{\mathcal{D}_k\}_{k=1}^{K}$
  - Each of the clusters is associated with centers, $\boldsymbol{\mu}_k$, $k=1, 2, \ldots, K$
  - Come up with the centers of clusters
  - Cluster center acts as a cluster representative
- Euclidean distance with center of a cluster can be used as a measure of dissimilarity

2

K-Means Clustering Algorithm

1. Initialize the cluster center, $\boldsymbol{\mu}_k$, k=1, 2, …, K using randomly selected K data points in $\mathcal{D}$
2. Assign each data point $\mathbf{x}_n$ to cluster center k*

$$k^* = \arg\min_k \|\mathbf{x}_n - \boldsymbol{\mu}_k\|^2$$

3. Update $\boldsymbol{\mu}_k$, k=1, 2, …, K: Re-compute $\boldsymbol{\mu}_k$ after assigning all the data points.

$$\widehat{\boldsymbol{\mu}}_k = \frac{\sum_{\mathcal{D}_k} \mathbf{x}_n}{N_k}$$

N_k: Number of examples in cluster k

4. Repeat the steps 2 and 3 until the convergence

3

K-Means Clustering Algorithm

- Convergence criteria:
  - No change in the cluster assignment **OR**
  - The difference between the distortion measure (J) in the successive iteration falls below the threshold
    - Distortion measure (J) : Sum of the squares of the distance of each example to its assigned cluster center

$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk} \|\mathbf{x}_n - \boldsymbol{\mu}_k\|^2$$

z_{nk} is 1 if $\mathbf{x}_n$ belongs to cluster k, otherwise 0

4