Avisek Gupta (avisek003@gmail.com)

Machine Learning

11 – Data Clustering

September 24, 2022

k-Means Clustering

The *k*-Means Clustering Algorithm –

Input: The data *X*, the number of clusters to find *k*
Output: The *k* cluster centers, the cluster memberships of each data instance

1. Initialize the *k* cluster centers by randomly selecting *k* data instances

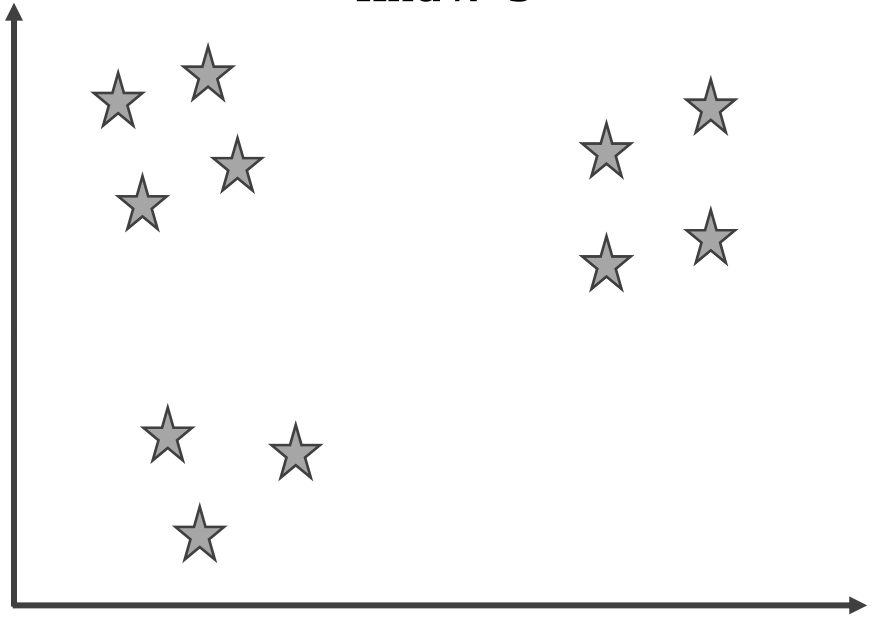
2. Repeat until convergence:

    2(a). Calculate the distance between all *n* data instances and all *k* cluster centers.

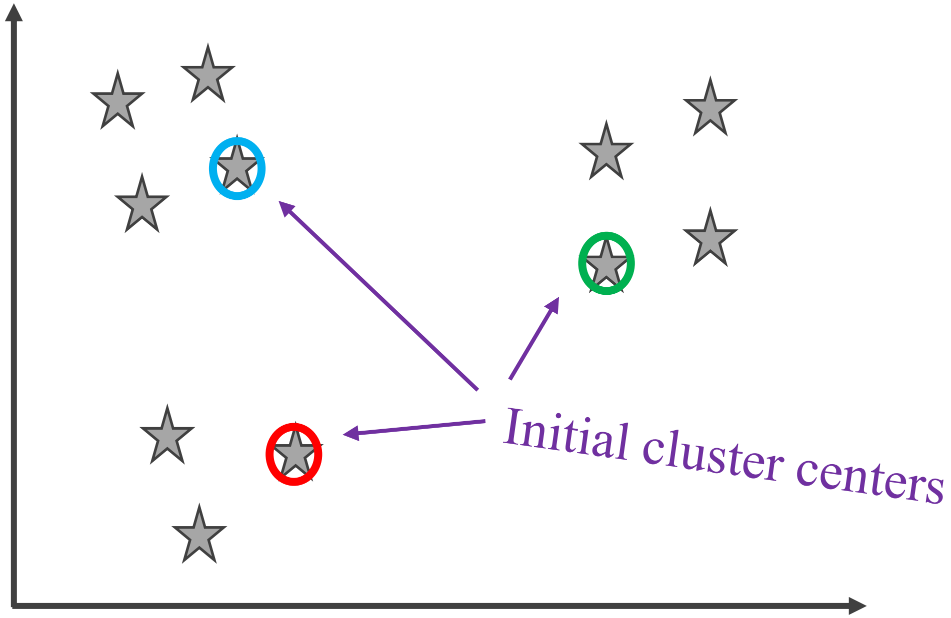
    2(b). Calculate the cluster membership of each data instance, as that cluster whose center lies at the closest distance to the data instance.

    2(c). Update the *k* cluster centers, as the mean of all data instances that have membership to that cluster.

An unlabeled data set, no. of clusters to find *k*=3



Initialize *k*=3 data instances as the initial cluster centers

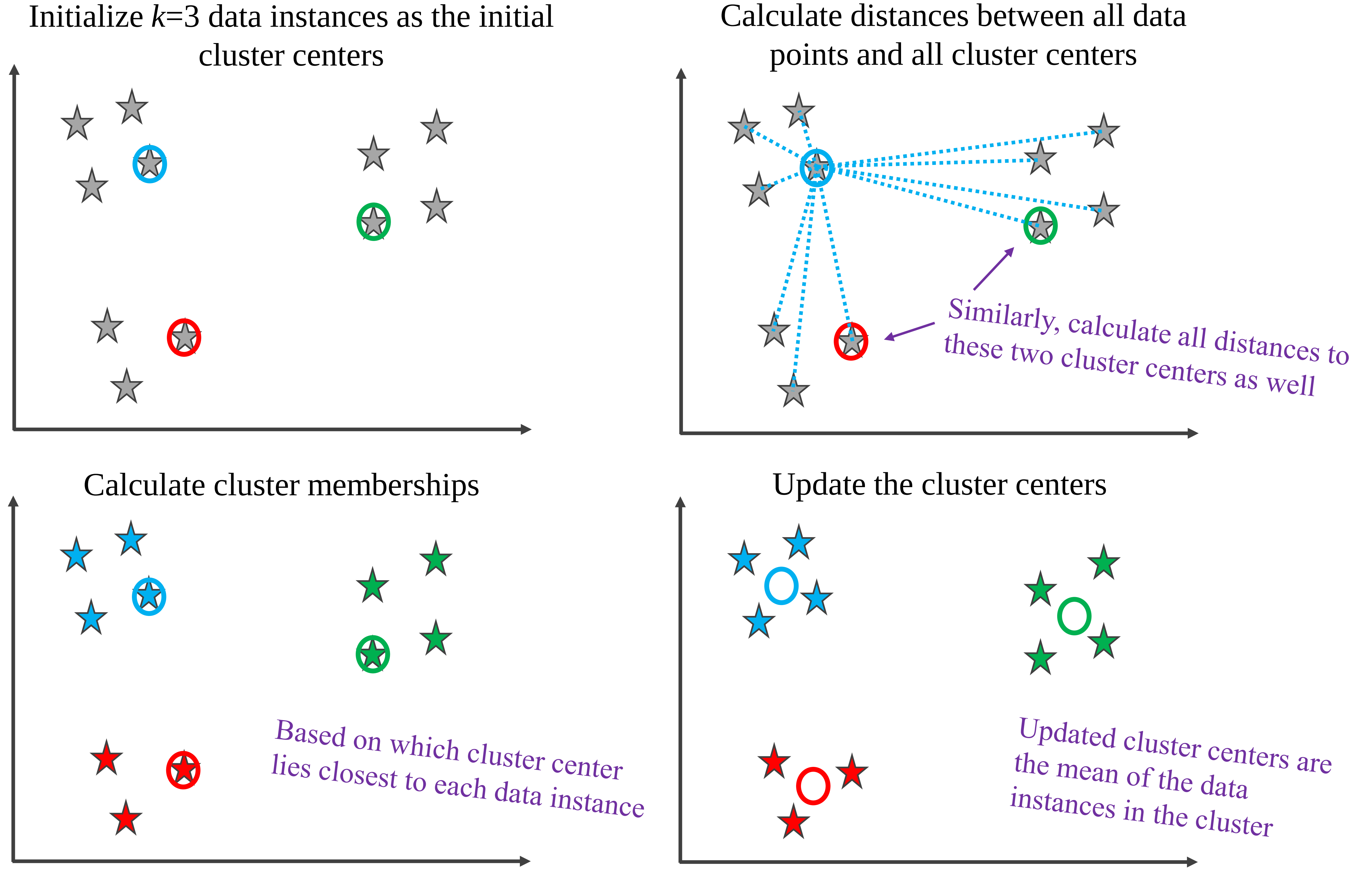


1. Initialize the *k* cluster centers by randomly selecting *k* data instances
2. Repeat until convergence:

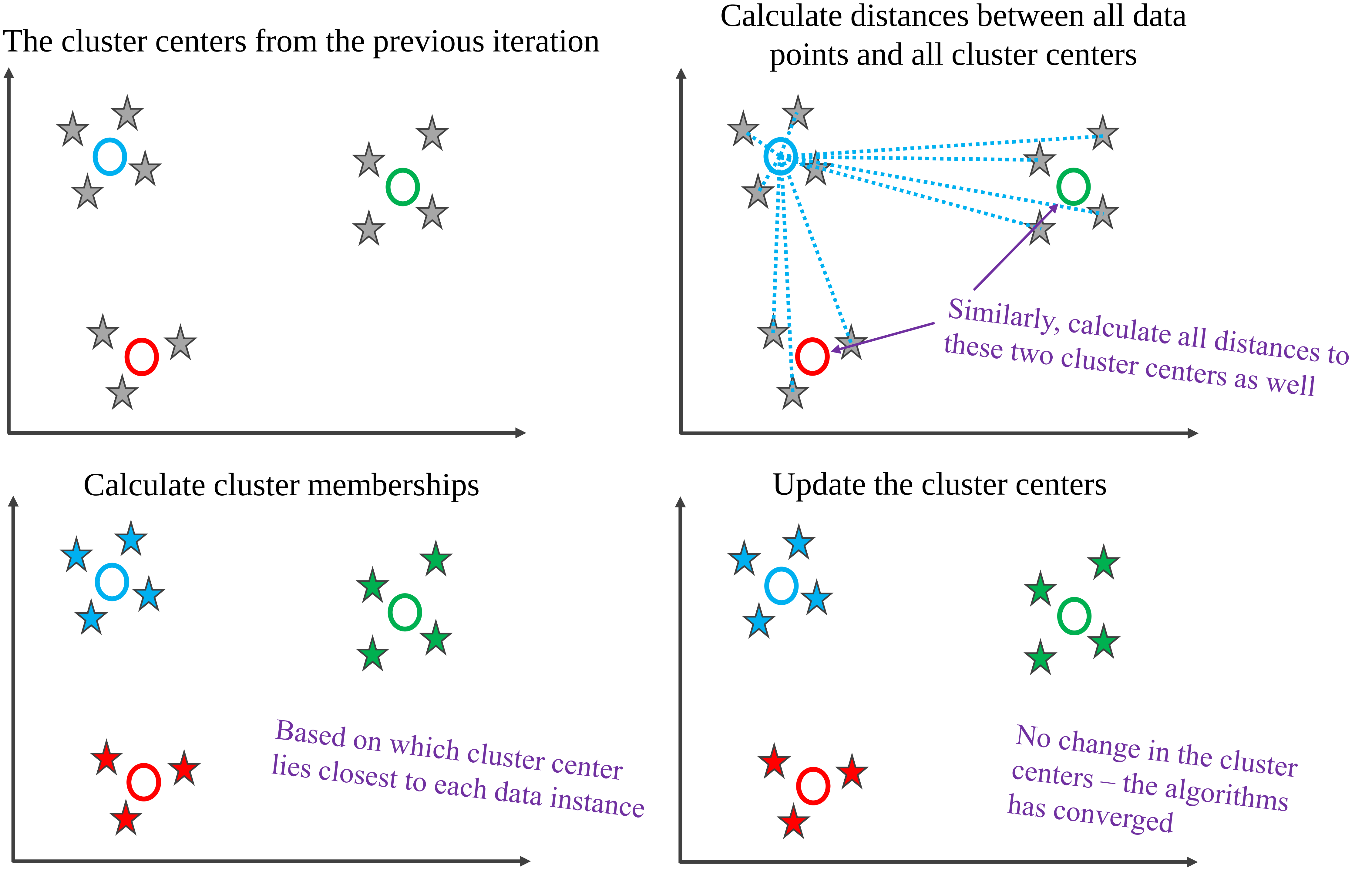
    2(a). Calculate the distance between all *n* data instances and all *k* cluster centers.

    2(b). Calculate the cluster membership of each data instance, as that cluster whose center lies at the closest distance to the data instance.

    2(c). Update the *k* cluster centers, as the mean of all data instances that have membership to that cluster.


Initialize k=3 data instances as the initial cluster centers
Calculate distances between all data points and all cluster centers
Similarly, calculate all distances to these two cluster centers as well
Calculate cluster memberships
Based on which cluster center lies closest to each data instance
Update the cluster centers
Updated cluster centers are the mean of the data instances in the cluster

The next iteration of the algorithm:



k-Means Clustering

The *k*-Means Clustering Algorithm –

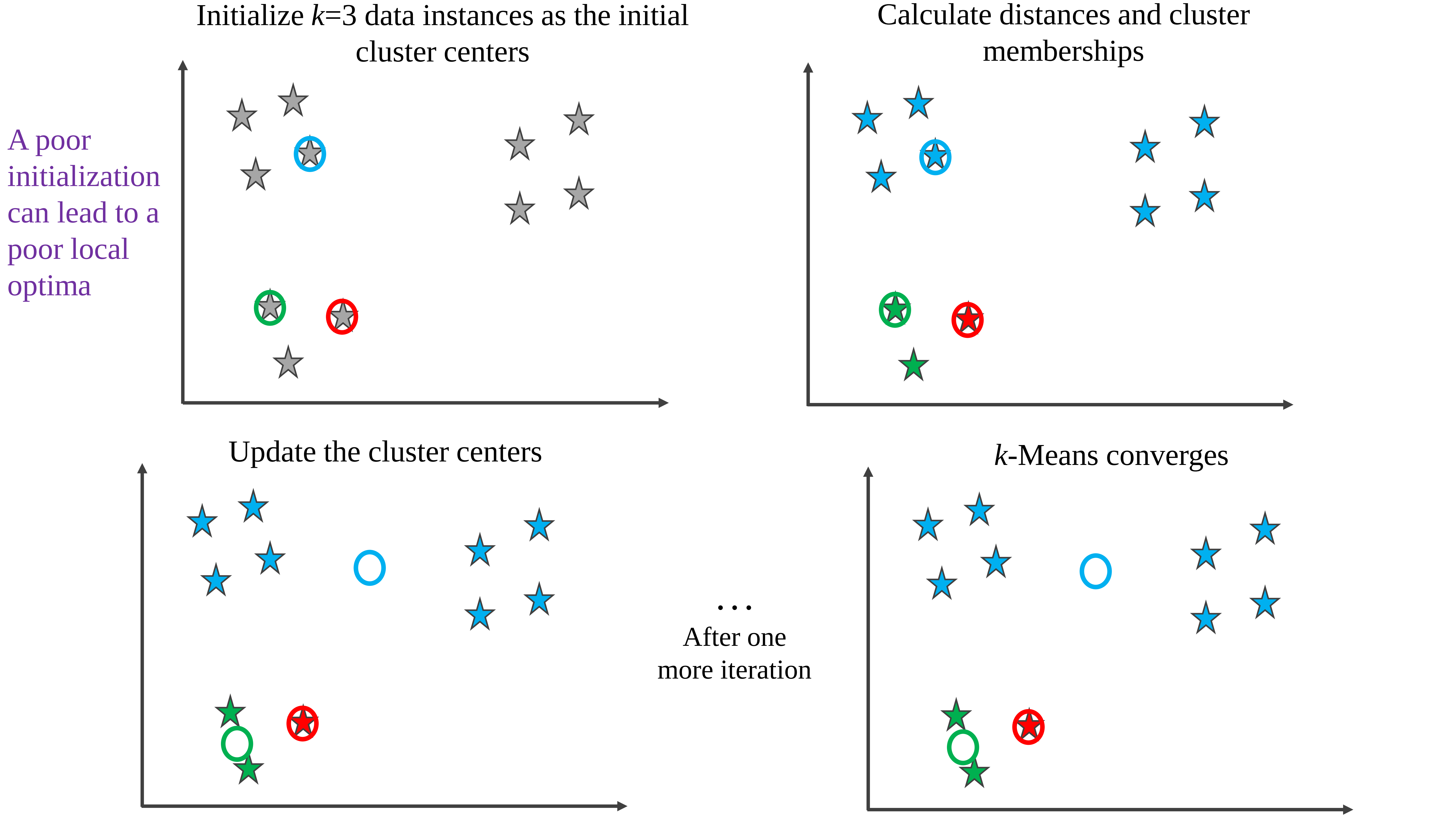
Input: The data *X*, the number of clusters to find *k*
Output: The *k* cluster centers, the cluster memberships of each data instance

1. Initialize the *k* cluster centers by randomly selecting *k* data instances
2. Repeat until convergence:

    2(a). Calculate the distance between all *n* data instances and all *k* cluster centers.

    2(b). Calculate the cluster membership of each data instance, as that cluster whose center lies at the closest distance to the data instance.

    2(c). Update the *k* cluster centers, as the mean of all data instances that have membership to that cluster.



Initialize k=3 data instances as the initial cluster centers
Calculate distances and cluster memberships
A poor initialization can lead to a poor local optima
Update the cluster centers
k-Means converges
. . .
After one more iteration

k-Means Clustering: Problem Objective

Let $X = [x_1, ..., x_n], x_i \in \mathbb{R}^d$ be a data set which we wish to cluster. $X \in \mathbb{R}^{n\times d}$. k-Means clustering aims to partition this dataset into k clusters where each cluster is represented by a center of the cluster $V = \{v_1, ..., v_k\} \in \mathbb{R}^{k\times d}, v_i \in \mathbb{R}^d$. Let the cluster membership be represented by $U = [\mu_{ij}]_{(n\times k)}, \mu_{ij} \in \{0, 1\}, \sum_{j=1}^{k} \mu_{ij} = 1$.

The k-Means clustering problem is,

$$\min_{U,V} \sum_{i=1}^{n} \sum_{j=1}^{k} \mu_{ij} ||x_i - v_j||^2$$

k-Means Clustering

The k-Means clustering problem is,

$$\min_{U,V} J_{KM} = \min_{U,V} \sum_{i=1}^{n} \sum_{j=1}^{k} \mu_{ij} ||x_i - v_j||^2$$

Estimating V: Holding U constant, equating the derivative of the objective function to zero,

$$\nabla_{v_j} J_{KM} = \sum_{i=1}^{n} \mu_{ij} 2(x_i - v_j)(-1) = 0$$

$$\implies v_j = \frac{\sum_{i=1}^{n} \mu_{ij} x_i}{\sum_{i=1}^{n} \mu_{ij}}$$

Let $C_j = \{x | x \text{ is closest to } v_j\}$.

$$v_j = \frac{\sum_{x_i \in C_j} x_i}{|C_j|}$$

k-Means Clustering

The k-Means clustering problem is,

$$\min_{U,V} J_{KM} = \min_{U,V} \sum_{i=1}^{n} \sum_{j=1}^{k} \mu_{ij} ||x_i - v_j||^2$$

Estimating U: Holding V constant:

$$\min_{U} \sum_{i=1}^{n} \sum_{j=1}^{k} \mu_{ij} d_{ij}, \;\; s.t., \;\; \mu_{ij} \in \{0, 1\}, \sum_{j=1}^{k} \mu_{ij} = 1.$$

$$\mu_{ij} = \begin{cases} 1 & , \; d_{ij} \leq d_{ij'} \; \forall j' \neq j \\ 0 & , \; o/w \end{cases}$$

k-Means Clustering

The k-Means clustering problem is,

$$\min_{U,V} J_{KM} = \min_{U,V} \sum_{i=1}^{n} \sum_{j=1}^{k} \mu_{ij} ||x_i - v_j||^2$$

Alternating Optimization: We alternately update U and V,

$$\mu_{ij} = \begin{cases} 1 & , \; ||x_i - v_j||^2 \leq ||x_i - v'_j||^2 \; \forall j' \neq j \\ 0 & , \; o/w \end{cases}$$

$$v_j = \frac{\sum_{i=1}^{n} \mu_{ij} x_i}{\sum_{i=1}^{n} \mu_{ij}}$$

This provides us the update rules for the k-Means algorithm (or LLoyd's algorithm[1]).

[1] Lloyd, Stuart P. (1982), "Least squares quantization in PCM", IEEE Transactions on Information Theory, 28 (2): 129–137.