WELCOME TO

TECH I.S.

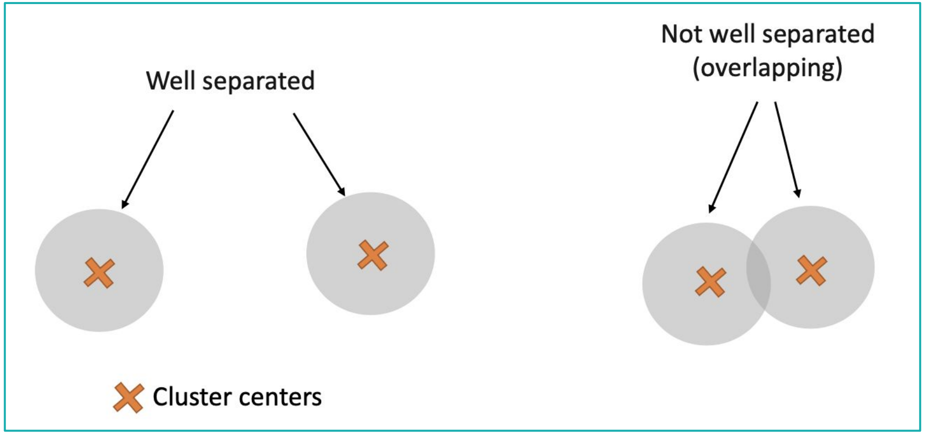
K-Means

Centroid Clustering

Partitional clustering approach where each cluster is associated with a centroid (center point).

↓

Each point is assigned to the cluster with the closest centroid and Number of clusters K, must be specified.



Centroid - Objective Function

The objective is to minimize the distances of the data points to their respective centroid.

Given a set X of n points in a d-dimensional space and an integer K group the points into K clusters C= {C1, C2,...,Ck} such that Cost(C) function based on distance is minimized.

$$Cost(C) = \sum_{i=1}^{k} \sum_{x \in C_i} dist(x, c)$$

K-means Clustering

K-means is Partitioning algorithm, also known as Lloyd's algorithm:

1. Decide on a value for K, the number of clusters.
2. Initialize the K cluster centers (randomly, if necessary).
3. Based on distance function used decide the class memberships of the N objects by assigning them to the nearest cluster center.
4. Re-estimate the K cluster centers, by assuming the memberships found above are correct.
5. Repeat 3 and 4 until none of the N objects changed membership in the last iteration.

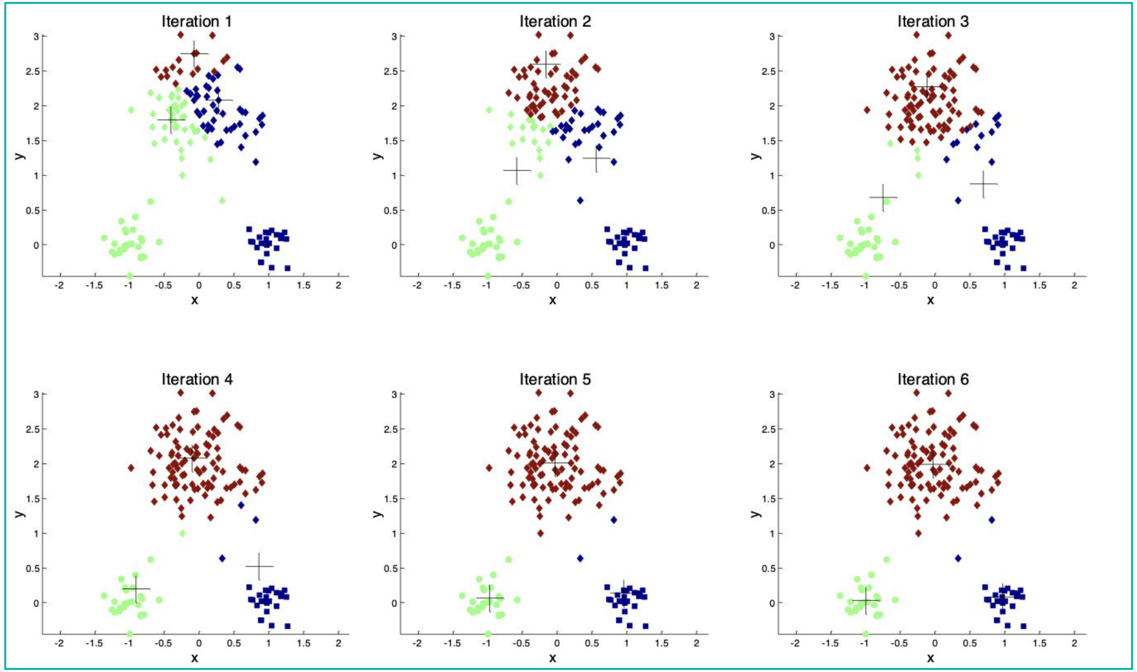
K-Means - Objective Function

Most common distance used with K-Means is with euclidean distance, minimizing the Sum of Squares Error (SSE)

$$\boxed{Cost(C) = \sum_{i=1}^{k} \sum_{x \in C_i} dist(x, c)} \longrightarrow \boxed{Cost(C) = \sum_{i=1}^{k} \sum_{x \in C_i} (x - c_i)^2}$$

K-Means Example

It is important to choose proper value for Initial Centroids k:



Pre-processing and Post-processing

Pre-processing:

- Normalize the data (e.g., scale to unit standard deviation) and Eliminate outliers

Post-processing:

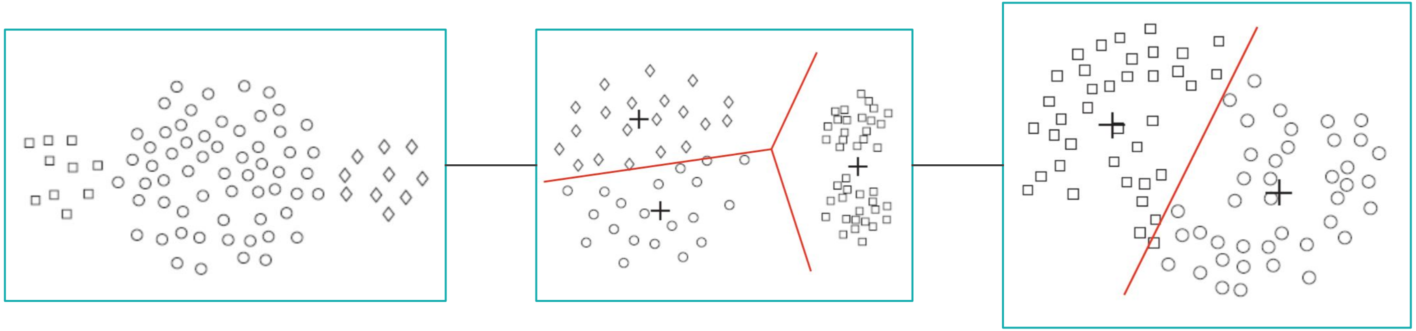
- Eliminate small clusters that may represent outliers
- Split ‘loose’ clusters, i.e., clusters with relatively high SSE
- Merge clusters that are ‘close’ and that have relatively low SSE

Use a larger number of clusters - Several clusters represent a true cluster

Limitations of K-means

K-means has problems:

- When clusters are of differing in Sizes, Densities and Non-globular shapes
- K-means has problems when the data contains outliers.



Tip: Use a larger number of clusters - Several clusters represent a true cluster

Much obliged.

TECH I.S.