UNC CHARLOTTE

The WILLIAM STATES LEE COLLEGE *of* ENGINEERING

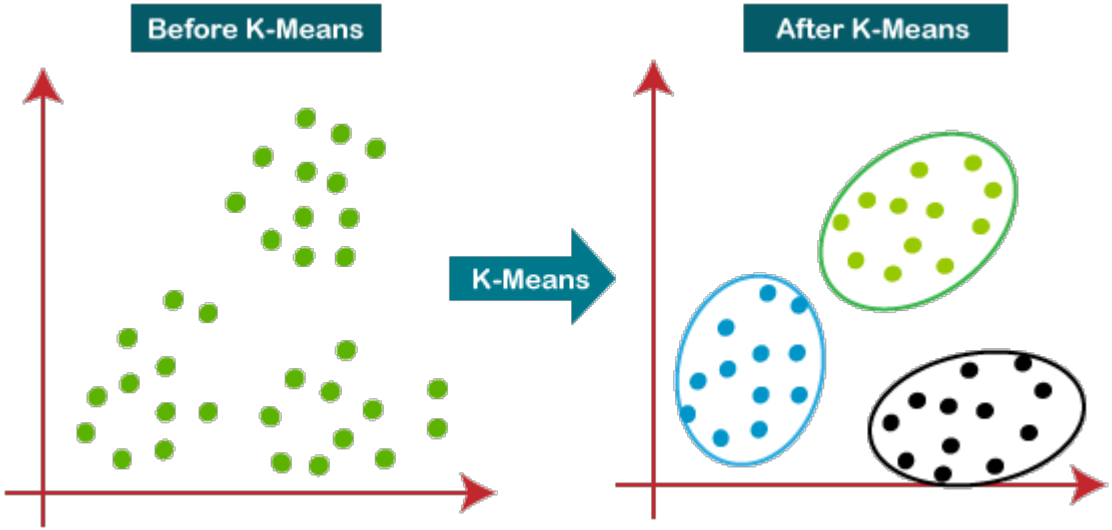
Introduction to ML
Lecture 10: K-Means Clustering (an unsupervised learning approach)

Hamed Tabkhi
Department of Electrical and Computer Engineering,
University of North Carolina Charlotte (UNCC)
htabkhiv@uncc.edu

UNC CHARLOTTE

Overview

- K means is one of the most popular Unsupervised Machine Learning Algorithms Used for Solving Classification Problems.
- ***K Means segregates the unlabeled data into various groups, called clusters, based on having similar features, common patterns.***



UNC CHARLOTTE

What is Clustering

- Suppose we have N number of Unlabeled Multivariate Datasets of various Animals like Dogs, Cats, birds etc. ***The technique to segregate Datasets into various groups, on basis of having similar features and characteristics, is being called Clustering***.
- The groups being Formed are being known as Clusters. Clustering Technique is being used in various Field such as Image recognition, Spam Filtering
- Clustering is being used in Unsupervised Learning Algorithm in Machine Learning as ***it can be segregated multivariate data into various groups, without any supervisor, on basis of common pattern hidden inside the datasets.***

K-Means Algorithm

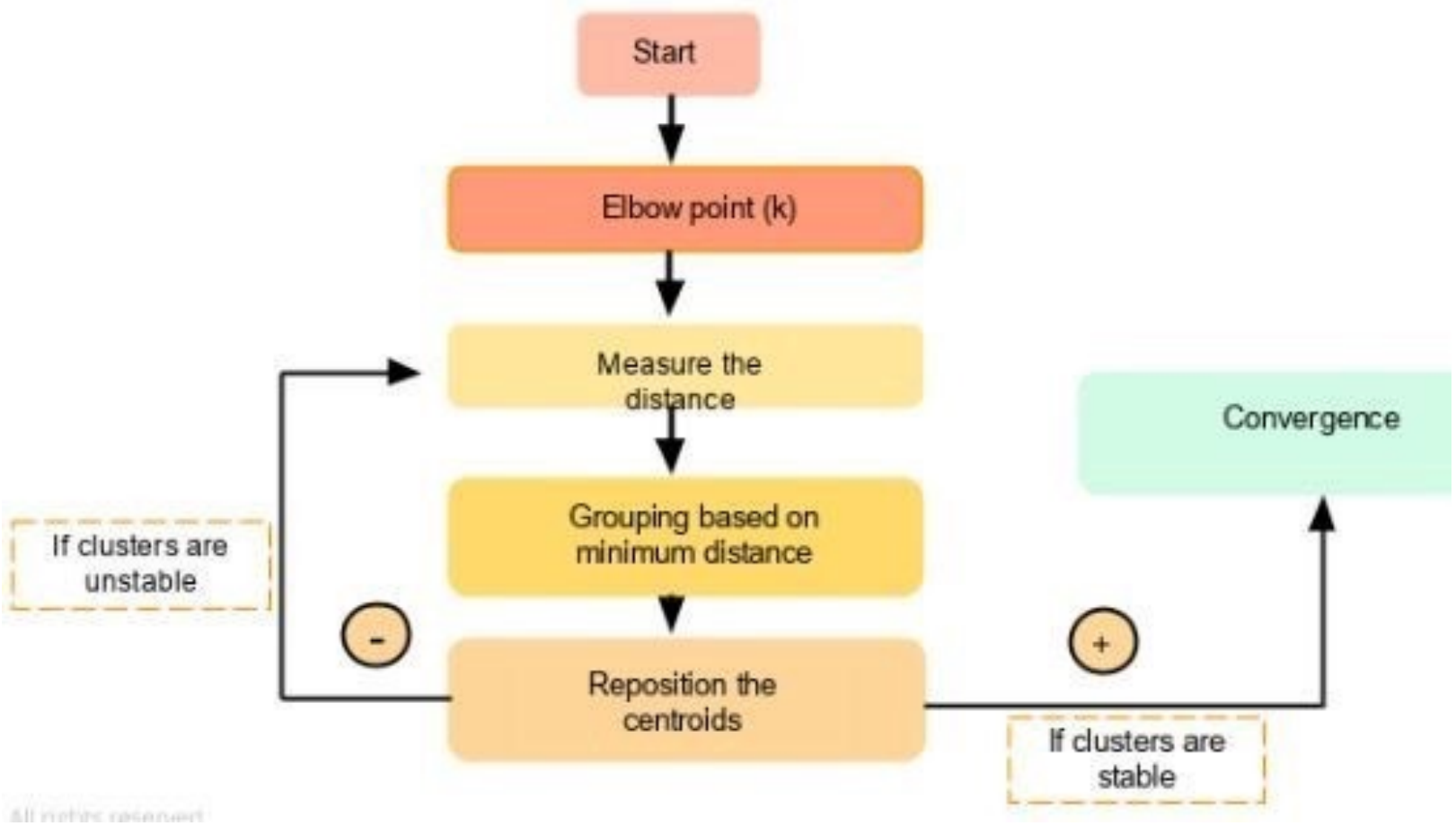
- ***Kmeans Algorithm is an Iterative algorithm that divides a group of n datasets into k subgroups /clusters based on the similarity and their mean distance from the centroid of that particular subgroup/ formed.***
- K, here is the pre-defined number of clusters to be formed by the Algorithm. If K=3, It means the number of clusters to be formed from the dataset is 3.

Algorithm Steps

The working of the K-Means algorithm is explained in the below steps:

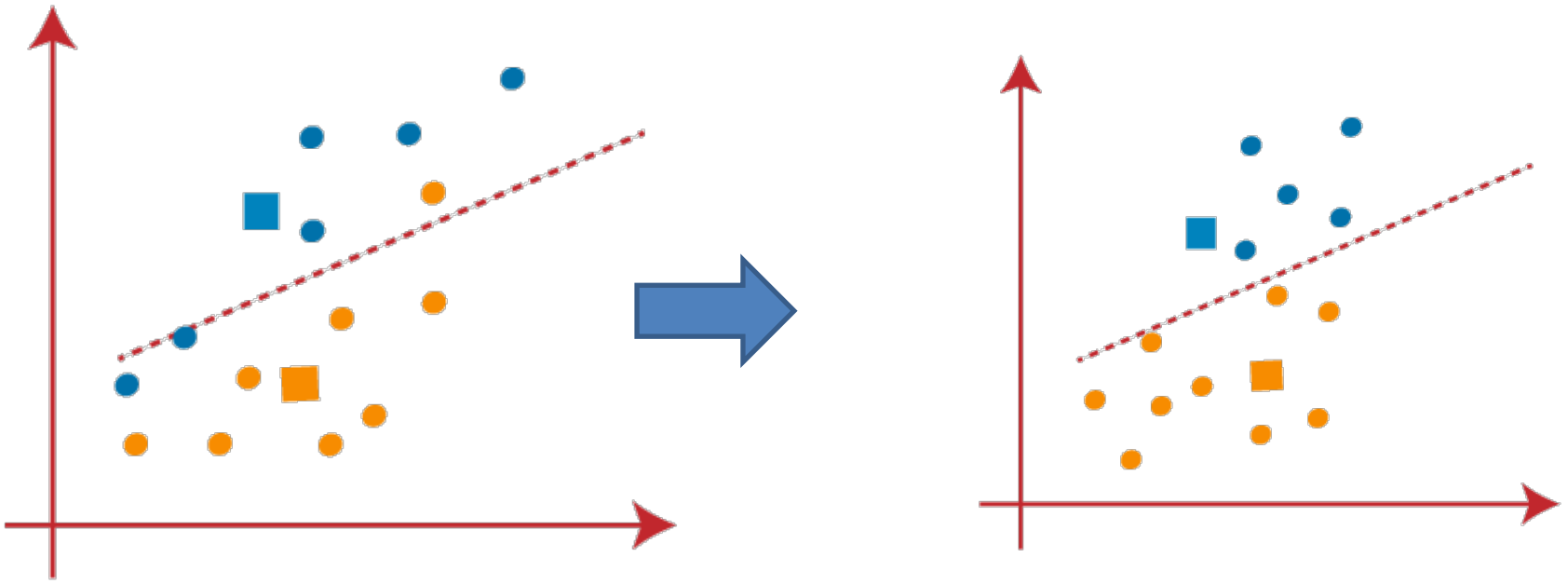
- **Step-1:** Select the value of K, to decide the number of clusters to be formed.
- **Step-2:** Select random K points which will act as centroids.
- **Step-3:** Assign each data point, based on their distance from the randomly selected points (Centroid), to the nearest/closest centroid which will form the predefined clusters.
- **Step-4:** place a new centroid of each cluster.
- **Step-5:** Repeat step no.3, which reassign each datapoint to the new closest centroid of each cluster.
- **Step-6:** If any reassignment occurs, then go to step-4 else go to Step 7.
- **Step-7**: FINISH

Algorithm Steps

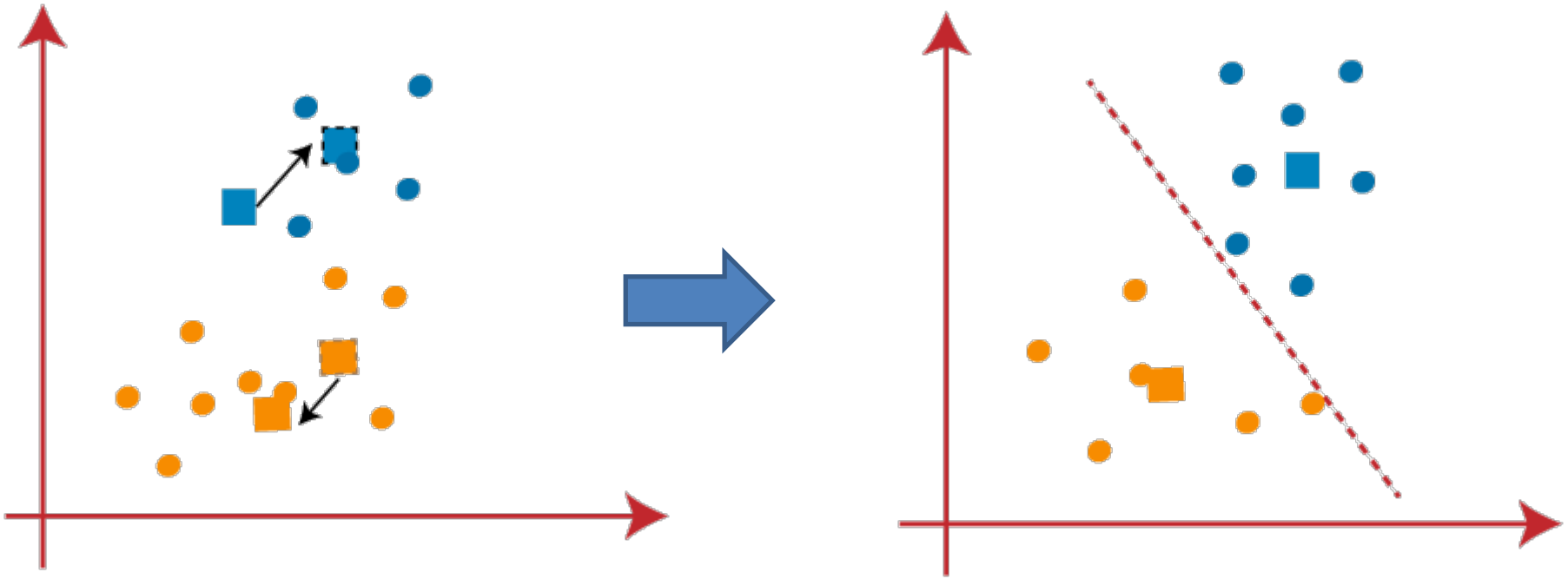


UNC CHARLOTTE

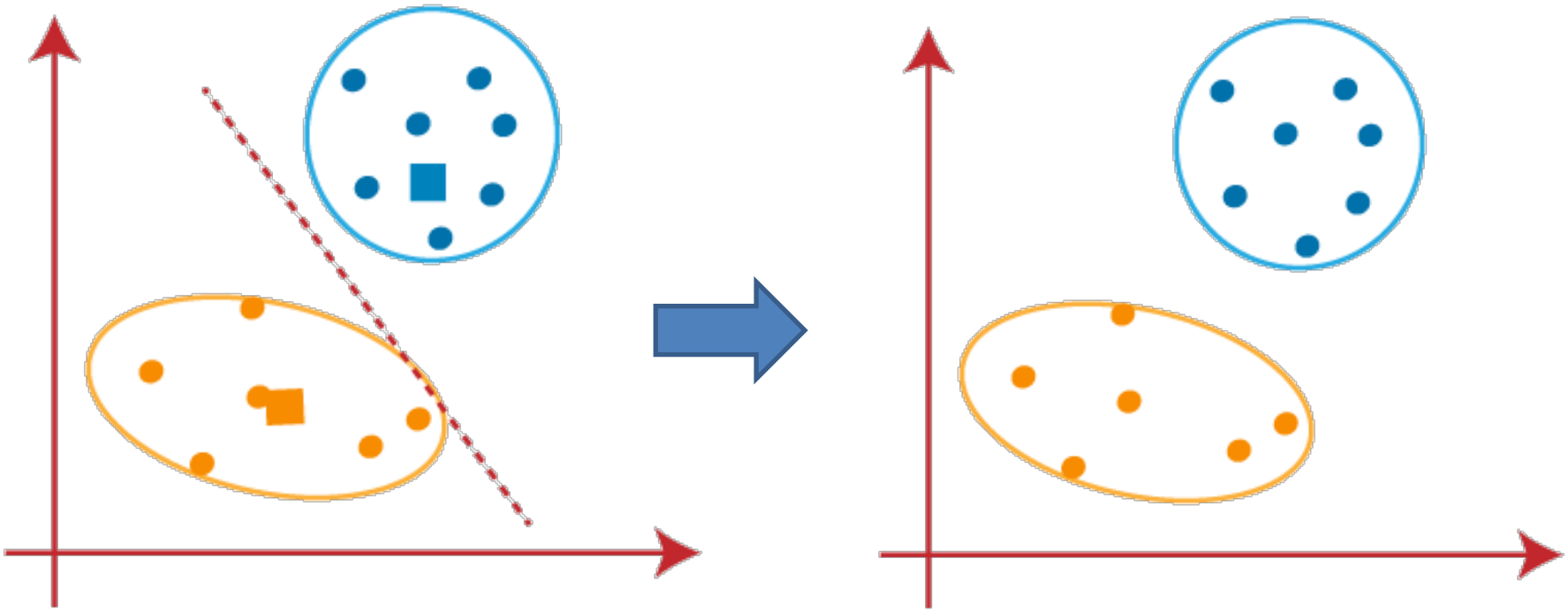
Algorithm Steps



Algorithm Steps



Algorithm Steps



Choose the right number of clusters

- ***The number of clusters that we choose for the algorithm shouldn't be random. Each and Every cluster is formed by calculating and comparing the mean distances of each data points within a cluster from its centroid.***

- We Can Choose the right number of clusters with the help of the Within-Cluster-Sum-of-Squares (WCSS) method.

- WCSS Stands for the sum of the squares of distances of the data points in each and every cluster from its centroid.

- The main idea is to minimize the distance between the data points and the centroid of the clusters. The process is iterated until we reach a minimum value for the sum of distances.

Elbow Method Steps

- 1 Execute the K-means clustering on a given dataset for different K values (ranging from 1-10).
- 2 For each value of K, calculates the WCSS value.
- 3 Plots a graph/curve between WCSS values and the respective number of clusters K.
- 4 The sharp point of bend or a point(looking like an elbow joint) of the plot like an arm, will be considered as the best/optimal value of K