CPE412 Pattern Recognition

Week 7

K-Means (Clustering)

Dr. Nehad Ramaha,
Computer Engineering Department
Karabük Universities

INTRODUCTION– What is clustering?

- **Clustering** is the classification of objects into different groups, or more precisely, the partitioning of a data set into subsets (clusters), so that the data in each subset (ideally) share some common trait – often according to some defined distance measure.

Types of clustering:

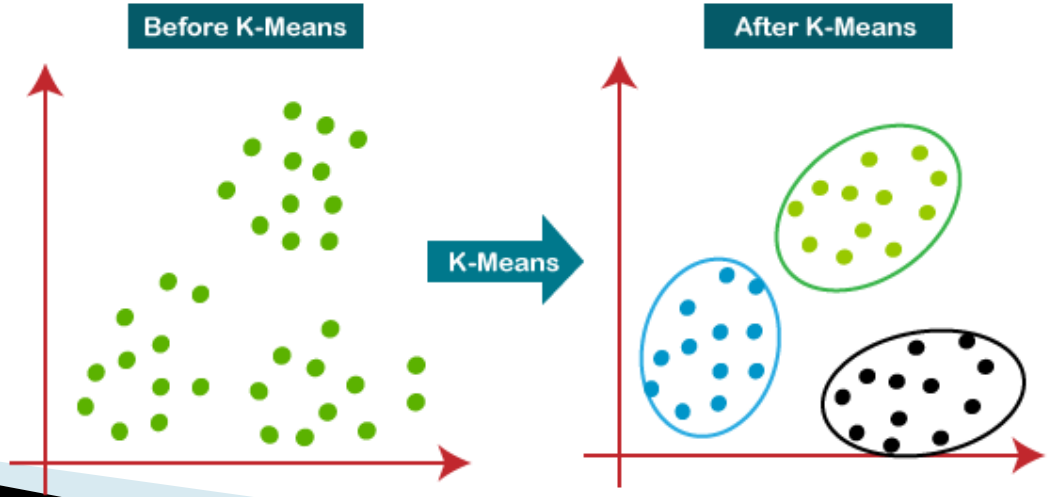
1. **Hierarchical algorithms**: these find successive clusters using previously established clusters.
   1. Agglomerative ("bottom-up"): Agglomerative algorithms begin with each element as a separate cluster and merge them into successively larger clusters.
   2. Divisive ("top-down"): Divisive algorithms begin with the whole set and proceed to divide it into successively smaller clusters.
2. **Partitional clustering**: Partitional algorithms determine all clusters at once. They include:
   - ***K*-means**
   - Fuzzy *c*-means clustering
   - QT clustering algorithm

Examples of Clustering Applications

- Marketing: Help marketers discover distinct groups in their customer bases, and then use this knowledge to develop targeted marketing programs
- Land use: Identification of areas of similar land use in an earth observation database
- Insurance: Identifying groups of motor insurance policy holders with a high average claim cost
- Urban planning: Identifying groups of houses according to their house type, value, and geographical location
- Seismology: Observed earthquake epicenters should be clustered along continent faults

K-MEANS CLUSTERING

- K-Means clustering is an unsupervised iterative clustering technique.
- The k-means algorithm is an algorithm to cluster n objects based on attributes into k partitions, where k < n.
- A cluster is defined as a collection of data points exhibiting certain similarities.



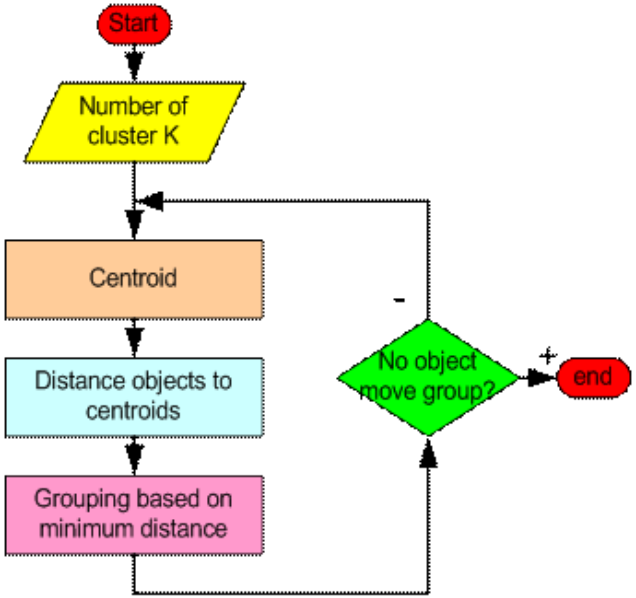
K-MEANS CLUSTERING

- Simply speaking k-means clustering is an algorithm to classify or to group the objects based on attributes/features into K number of group.
- K is positive integer number.
- Clusters based on centroids (the center of gravity or mean) .

K-MEANS CLUSTERING

- The k-means clustering algorithm mainly performs two tasks:
  - Determines the best value for K center points or centroids by an iterative process.
  - Assigns each data point to its closest k-center. Those data points which are near to the particular k-center, create a cluster.

How the K-Mean Clustering algorithm works?

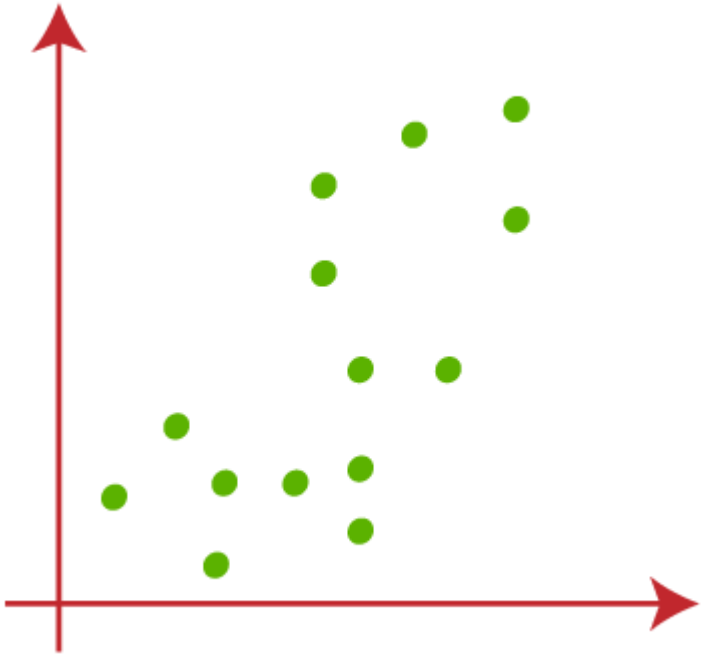


The working of the K-Means algorithm is explained in the below steps:

Step-1: Select the number K to decide the number of clusters.
Step-2: Select random K points or centroids.
Step-3: Assign each data point to their closest centroid, which will form the predefined K clusters.
Step-4: Calculate the variance and place a new centroid of each cluster.
Step-5: Repeat the third steps, which means 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 FINISH.
Step-7: The model is ready.

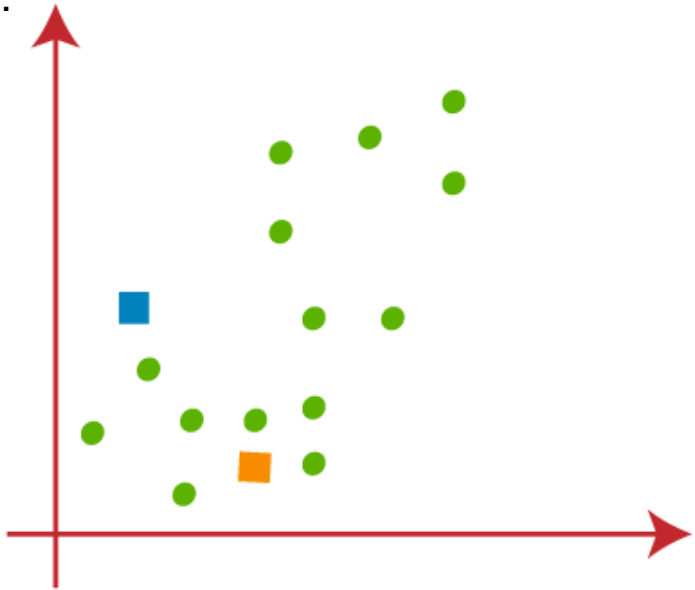
Let's understand the above steps by considering the visual plots:

- Suppose we have two variables M1 and M2. The x-y axis scatter plot of these two variables is given below:



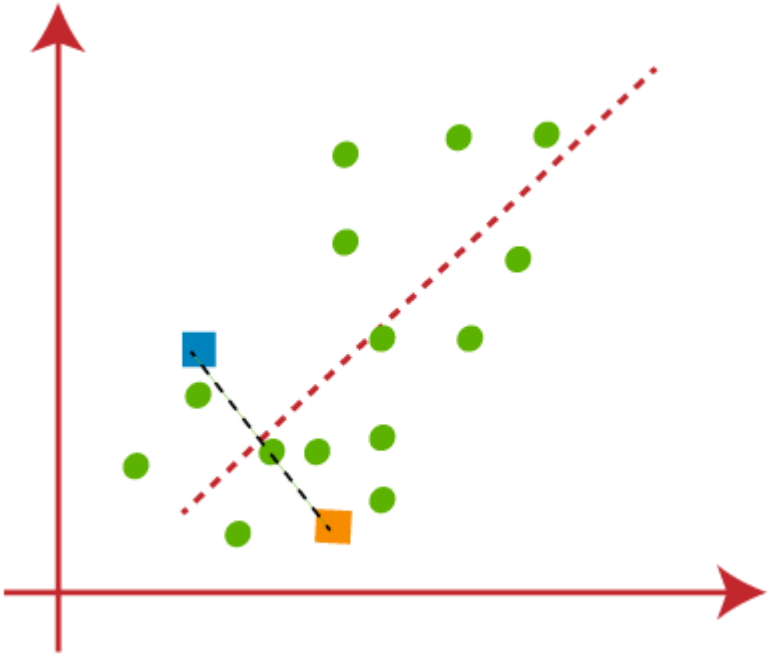
Let's understand the above steps by considering the visual plots:

- Let's take number k of clusters, i.e., K=2, to identify the dataset and to put them into different clusters. It means here we will try to group these datasets into two different clusters.
- We need to choose some random k points or centroid to form the cluster. These points can be either the points from the dataset or any other point. So, here we are selecting the below two points as k points, which are not the part of our dataset. Consider the below image:



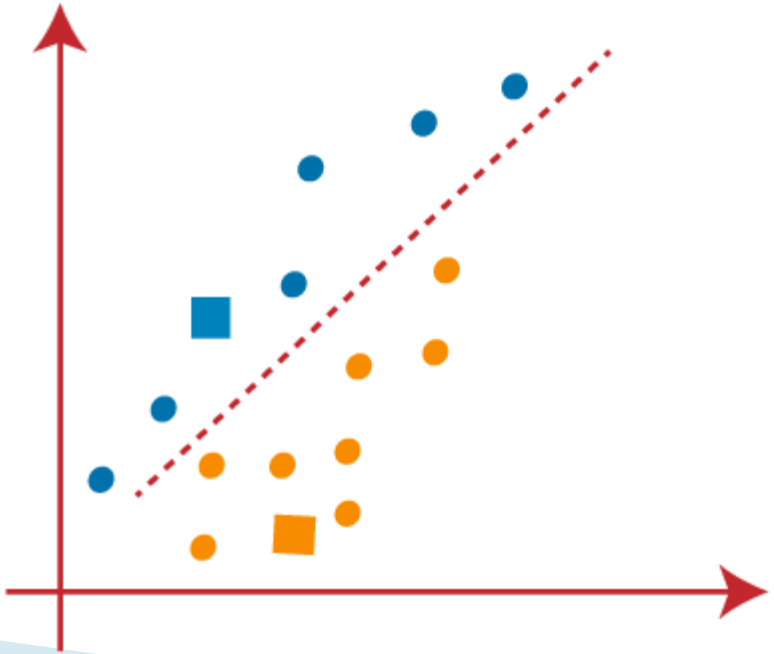
Let's understand the above steps by considering the visual plots:

- Now we will assign each data point of the scatter plot to its closest K-point or centroid. We will compute it by applying some mathematics that we have studied to calculate the distance between two points. So, we will draw a median between both the centroids. Consider the below image:



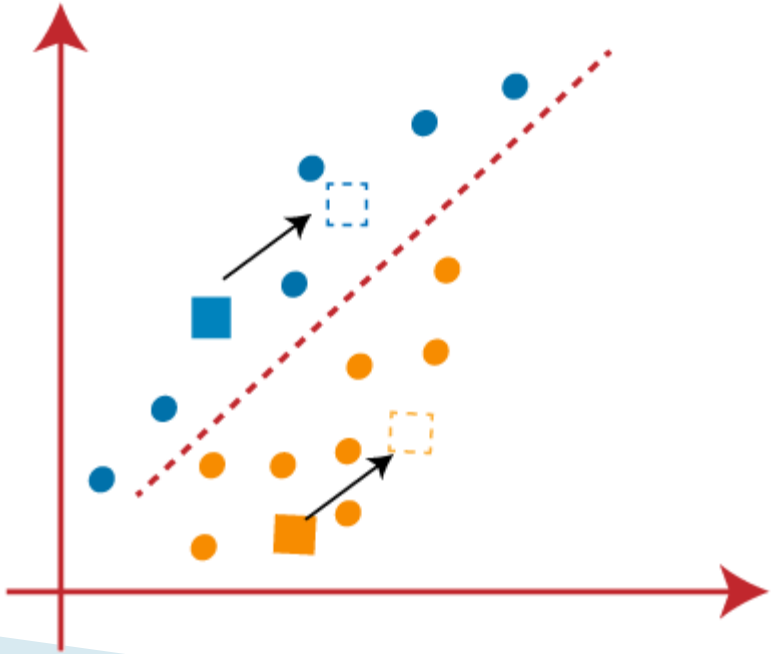
Let's understand the above steps by considering the visual plots:

- From the previous image, it is clear that points left side of the line is near to the K1 or blue centroid, and points to the right of the line are close to the yellow centroid. Let's color them as blue and yellow for clear visualization.



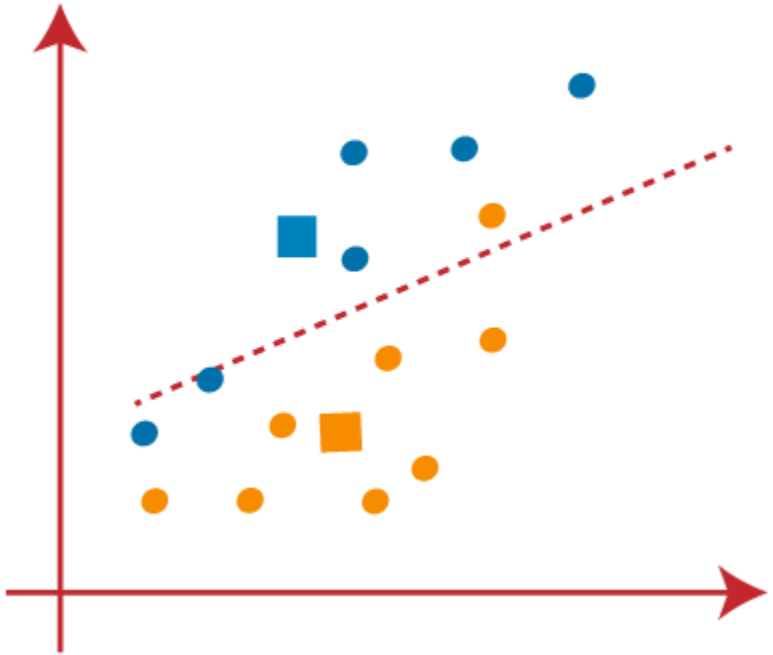
Let's understand the above steps by considering the visual plots:

- As we need to find the closest cluster, so we will repeat the process by choosing a new centroid. To choose the new centroids, we will compute the center of gravity of these centroids, and will find new centroids as below:



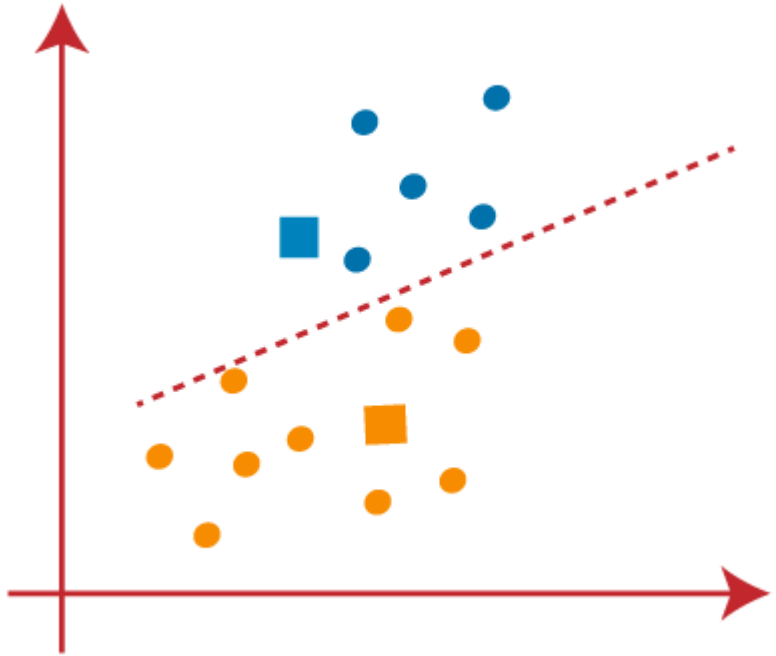
Let's understand the above steps by considering the visual plots:

- Next, we will reassign each datapoint to the new centroid. For this, we will repeat the same process of finding a median line. The median will be like below image:



Let's understand the above steps by considering the visual plots:

- From the previous image, we can see, one yellow point is on the left side of the line, and two blue points are right to the line. So, these three points will be assigned to new centroids:

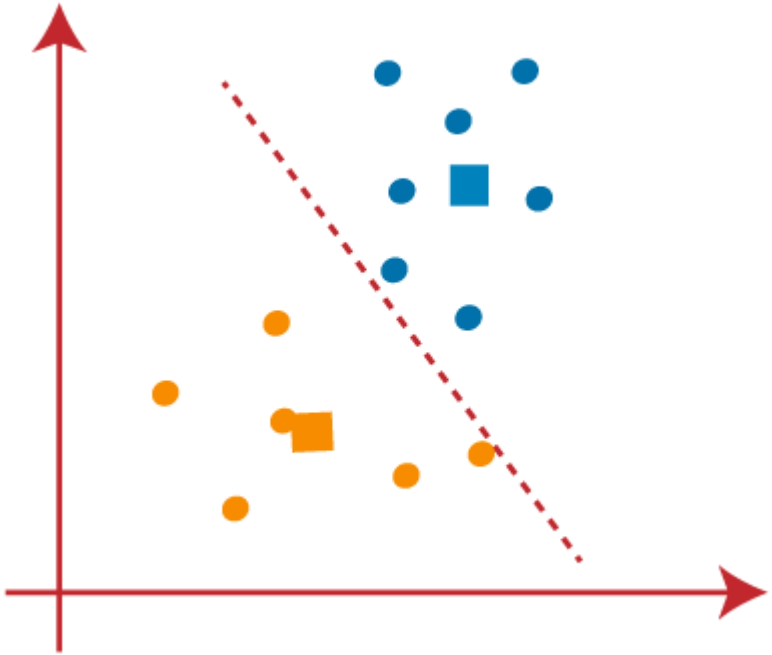


Let's understand the above steps by considering the visual plots:

- As reassignment has taken place, so we will again go to the step-4, which is finding new centroids or K-points.
- We will repeat the process by finding the center of gravity of centroids, so the new centroids will be as shown in the below image:

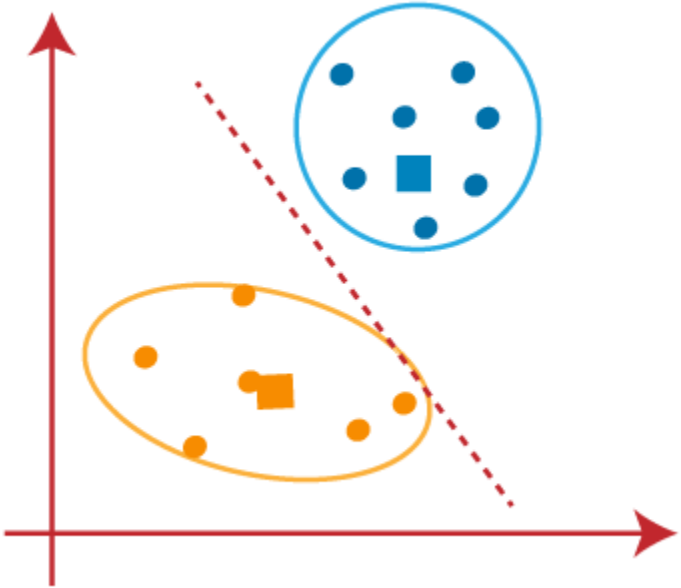
Let's understand the above steps by considering the visual plots:

- As we got the new centroids so again will draw the median line and reassign the data points. So, the image will be:



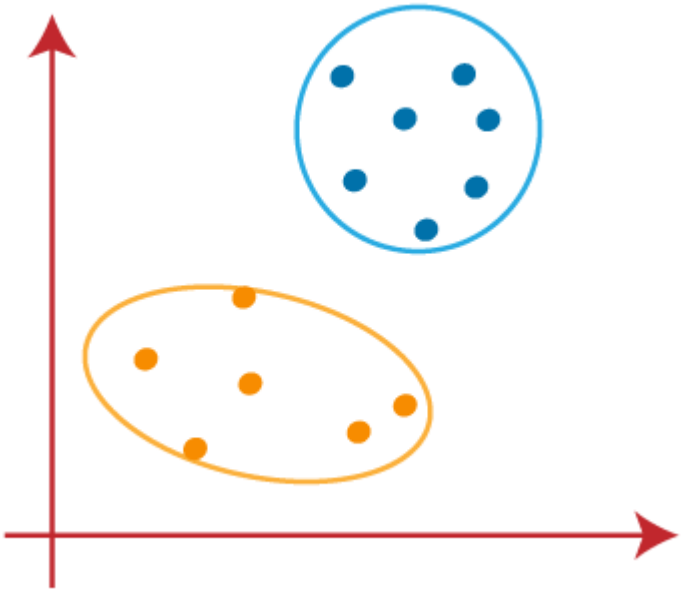
Let's understand the above steps by considering the visual plots:

- We can see in the previous image; there are no dissimilar data points on either side of the line, which means our model is formed. Consider the below image:



Let's understand the above steps by considering the visual plots:

- As our model is ready, so we can now remove the assumed centroids, and the two final clusters will be as shown in the below image:



How to choose the value of "K number of clusters" in K-means Clustering?

- The performance of the K-means clustering algorithm depends upon highly efficient clusters that it forms. But choosing the optimal number of clusters is a big task. There are some different ways to find the optimal number of clusters, the most appropriate method to find the number of clusters or value of K is “Elbow Method”.

Elbow Method

- This method uses the concept of WCSS value. WCSS stands for Within Cluster Sum of Squares, which defines the total variations within a cluster. The formula to calculate the value of WCSS (for 3 clusters) is given below:

$$WCSS= \sum_{Pi\ in\ Cluster1} distance(P_i\ C_1)^2 + \sum_{Pi\ in\ Cluster2} distance(P_i\ C_2)^2 + \sum_{Pi\ in\ CLuster3} distance(P_i\ C_3)^2$$

In the above formula of WCSS,

∑Pi in Cluster1 distance(Pi C1)2: It is the sum of the square of the distances between each data point and its centroid within a cluster1 and the same for the other two terms.

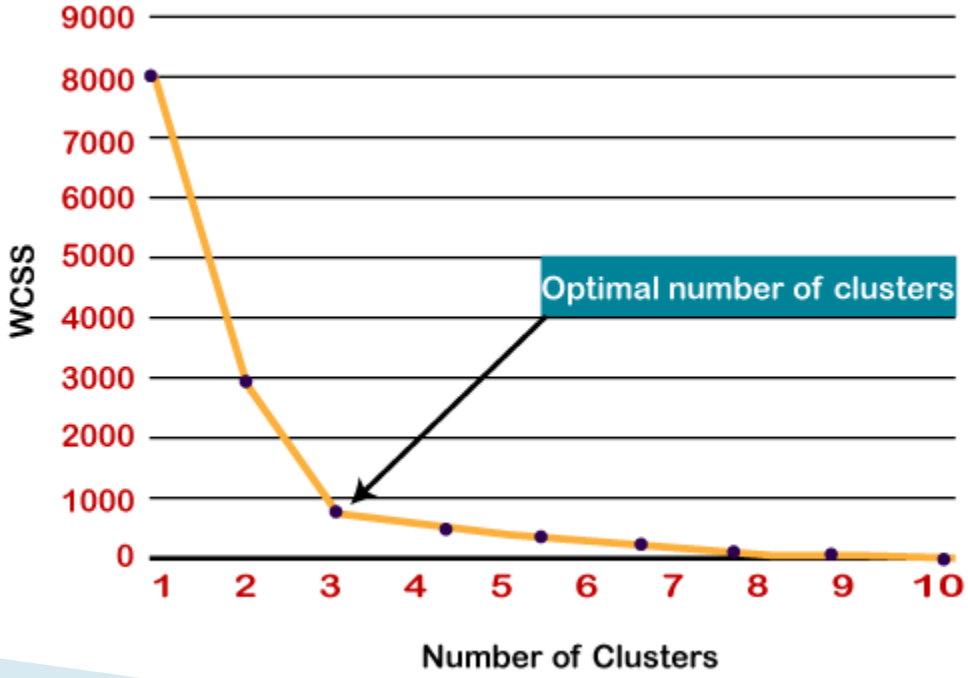
To measure the distance between data points and centroid, we can use any method such as Euclidean distance or Manhattan distance.

Elbow Method

- To find the optimal value of clusters, the elbow method follows the below steps:
  - It executes the K-means clustering on a given dataset for different K values (ranges from 1-10).
  - For each value of K, calculates the WCSS value.
  - Plots a curve between calculated WCSS values and the number of clusters K.
  - The sharp point of bend or a point of the plot looks like an arm, then that point is considered as the best value of K.

Elbow Method

- Since the graph shows the sharp bend, which looks like an elbow, hence it is known as the elbow method. The graph for the elbow method looks like the below image:



PRACTICE PROBLEMS BASED ON K-MEANS CLUSTERING ALGORITHM:

- Cluster the following eight points (with (x, y) representing locations) into three clusters:
- A1(2, 10), A2(2, 5), A3(8, 4), A4(5, 8), A5(7, 5), A6(6, 4), A7(1, 2), A8(4, 9)
- Initial cluster centers are: A1(2, 10), A4(5, 8) and A7(1, 2).
- The distance function between two points a = (x1, y1) and b = (x2, y2) is defined as:
- P(a, b) = |x2 – x1| + |y2 – y1|
- Use K-Means Algorithm to find the three cluster centers after the second iteration.

PRACTICE PROBLEMS BASED ON K-MEANS CLUSTERING ALGORITHM: Solution

- We follow the above discussed K-Means Clustering Algorithm.
- **Iteration-01:**
- We calculate the distance of each point from each of the center of the three clusters.
- The distance is calculated by using the given distance function.
- The following slides illustration shows the calculation of distance between point A1(2, 10) and each of the center of the three clusters:

Solution

- **Calculating Distance Between A1(2, 10) and C1(2, 10)**
- P(A1, C1)
- = $|x2 - x1| + |y2 - y1|$
- = $|2 - 2| + |10 - 10|$
- = 0

Solution

- **<u>Calculating Distance Between A1(2, 10) and C2(5, 8)-</u>**
- P(A1, C2)
- = |x2 – x1| + |y2 – y1|
- = |5 – 2| + |8 – 10|
- = 3 + 2
- = 5

Solution

- **Calculating Distance Between A1(2, 10) and C3(1, 2)-**
- P(A1, C3)
- = |x2 – x1| + |y2 – y1|
- = |1 – 2| + |2 – 10|
- = 1 + 8
- = 9
- In the similar manner, we calculate the distance of other points from each of the center of the three clusters.

Solution

- Next,
- We draw a table showing all the results.
- Using the table, we decide which point belongs to which cluster.
- The given point belongs to that cluster whose center is nearest to it.

Solution

Given Points	Distance from center (2, 10) of Cluster-01	Distance from center (5, 8) of Cluster-02	Distance from center (1, 2) of Cluster-03	Point belongs to Cluster
A1(2, 10)	0	5	9	C1
A2(2, 5)	5	6	4	C3
A3(8, 4)	12	7	9	C2
A4(5, 8)	5	0	10	C2
A5(7, 5)	10	5	9	C2
A6(6, 4)	10	5	7	C2
A7(1, 2)	9	10	0	C3
A8(4, 9)	3	2	10	C2

Solution

- From here, New clusters are-
- **Cluster-01:**
- First cluster contains points-
  - A1(2, 10)
- **Cluster-02:**
- Second cluster contains points-
  - A3(8, 4)
  - A4(5, 8)
  - A5(7, 5)
  - A6(6, 4)
  - A8(4, 9)
- **Cluster-03:**
  - Third cluster contains points
  - A2(2, 5)
  - A7(1, 2)

Solution

- Now,
- We re-compute the new cluster clusters.
- The new cluster center is computed by taking mean of all the points contained in that cluster.
- **For Cluster-01:**
- We have only one point A1(2, 10) in Cluster-01.
- So, cluster center remains the same.
- **For Cluster-02:**
- Center of Cluster-02
- = ((8 + 5 + 7 + 6 + 4)/5, (4 + 8 + 5 + 4 + 9)/5)
- = (6, 6)

Solution

- **For Cluster-03:**
- Center of Cluster-03
- = ((2 + 1)/2, (5 + 2)/2)
- = (1.5, 3.5)
- This is completion of Iteration-01.

Solution

- **Iteration-02:**
- We calculate the distance of each point from each of the center of the three clusters.
- The distance is calculated by using the given distance function.
- The following illustration shows the calculation of distance between point A1(2, 10) and each of the center of the three clusters-

Solution

- **Calculating Distance Between A1(2, 10) and C1(2, 10)-**
-
- P(A1, C1)
- = |x2 – x1| + |y2 – y1|
- = |2 – 2| + |10 – 10|
- = 0

Solution

- **Calculating Distance Between A1(2, 10) and C2(6, 6)-**
-
- P(A1, C2)
- = |x2 – x1| + |y2 – y1|
- = |6 – 2| + |6 – 10|
- = 4 + 4
- = 8

Solution

- **Calculating Distance Between A1(2, 10) and C3(1.5, 3.5)-**
- P(A1, C3)
- = |x2 – x1| + |y2 – y1|
- = |1.5 – 2| + |3.5 – 10|
- = 0.5 + 6.5
- = 7
- In the similar manner, we calculate the distance of other points from each of the center of the three clusters.

Solution

- Next,
- We draw a table showing all the results.
- Using the table, we decide which point belongs to which cluster.
- The given point belongs to that cluster whose center is nearest to it.

Solution

Given Points	Distance from center (2, 10) of Cluster-01	Distance from center (6, 6) of Cluster-02	Distance from center (1.5, 3.5) of Cluster-03	Point belongs to Cluster
A1(2, 10)	0	8	7	C1
A2(2, 5)	5	5	2	C3
A3(8, 4)	12	4	7	C2
A4(5, 8)	5	3	8	C2
A5(7, 5)	10	2	7	C2
A6(6, 4)	10	2	5	C2
A7(1, 2)	9	9	2	C3
A8(4, 9)	3	5	8	C1

Solution

- From here, New clusters are-
-
- **Cluster-01:**
-
- First cluster contains points-
- A1(2, 10)
- A8(4, 9)

Solution

- **Cluster-02:**
-
- Second cluster contains points-
  - A3(8, 4)
  - A4(5, 8)
  - A5(7, 5)
  - A6(6, 4)

Solution

- **Cluster-03:**
-
- Third cluster contains points-
- A2(2, 5)
- A7(1, 2)

Solution

- Now,
- We re-compute the new cluster clusters.
- The new cluster center is computed by taking mean of all the points contained in that cluster.

Solution

- **For Cluster-01:**
-
- Center of Cluster-01
- = ((2 + 4)/2, (10 + 9)/2)
- = (3, 9.5)

Solution

- **For Cluster-02:**
-
- Center of Cluster-02
- = ((8 + 5 + 7 + 6)/4, (4 + 8 + 5 + 4)/4)
- = (6.5, 5.25)

Solution

- **For Cluster-03:**
-
- Center of Cluster-03
- = ((2 + 1)/2, (5 + 2)/2)
- = (1.5, 3.5)

Solution

- This is completion of Iteration-02.
- After second iteration, the center of the three clusters are-
- C1(3, 9.5)
- C2(6.5, 5.25)
- C3(1.5, 3.5)

- Next, we go to iteration-03, iteration-04 and so on until the centers do not change anymore.

Thanks ☺