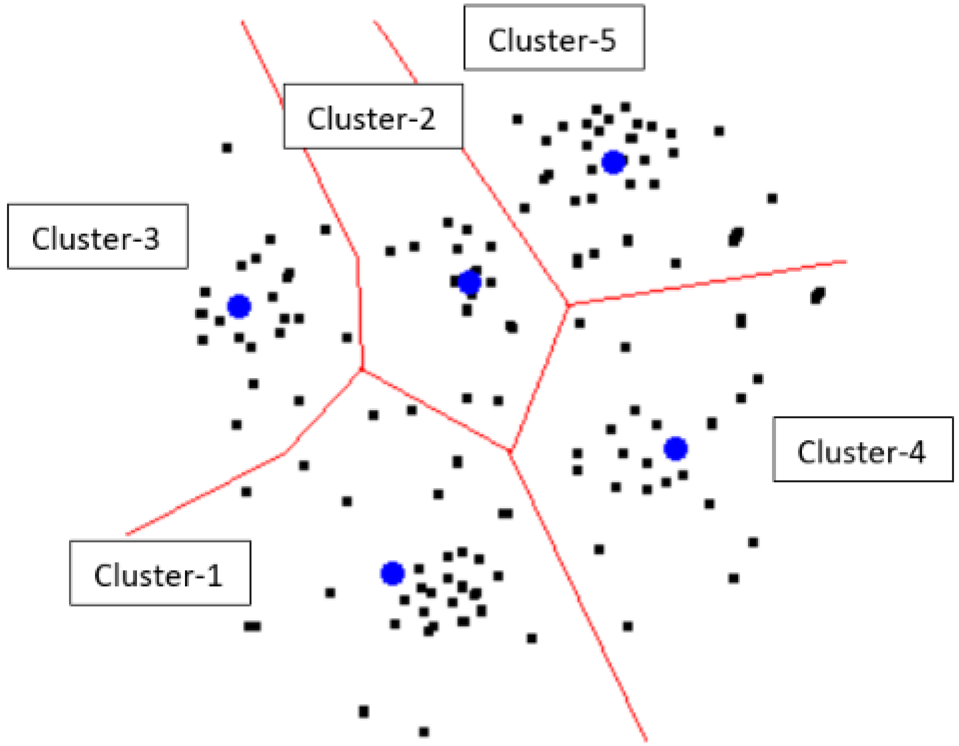
K – Means Clustering

What is k – means clustering?

k – means performs division of objects into clusters which are “similar” between them and are “dissimilar” to the objects belonging to another cluster



What is k – means clustering?

Can you explain this with example?

It can be explained with **cricket** example?

What is k – means clustering?

Task: Identify bowlers and batsmen

What is k – means clustering?

Task: Identify bowlers and batsmen

- The data contains wickets and runs gained in the last 10 matches
- So, the bowler will have more wickets and the batsman will have higher runs.

Scores

28	**92**

What is k – means clustering?

Assign data points

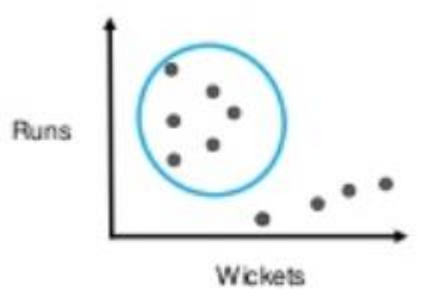
Here, we have our dataset with x and y co ordinates

Now, we want to cluster this data using k-means

What is k – means clustering?

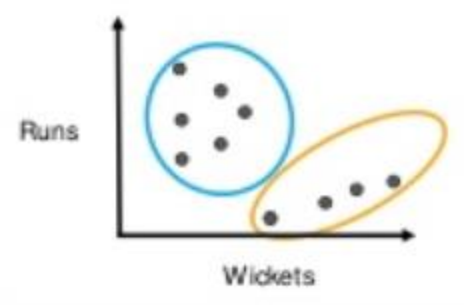
Cluster 1

We can see that this cluster has players with high runs and low wickets

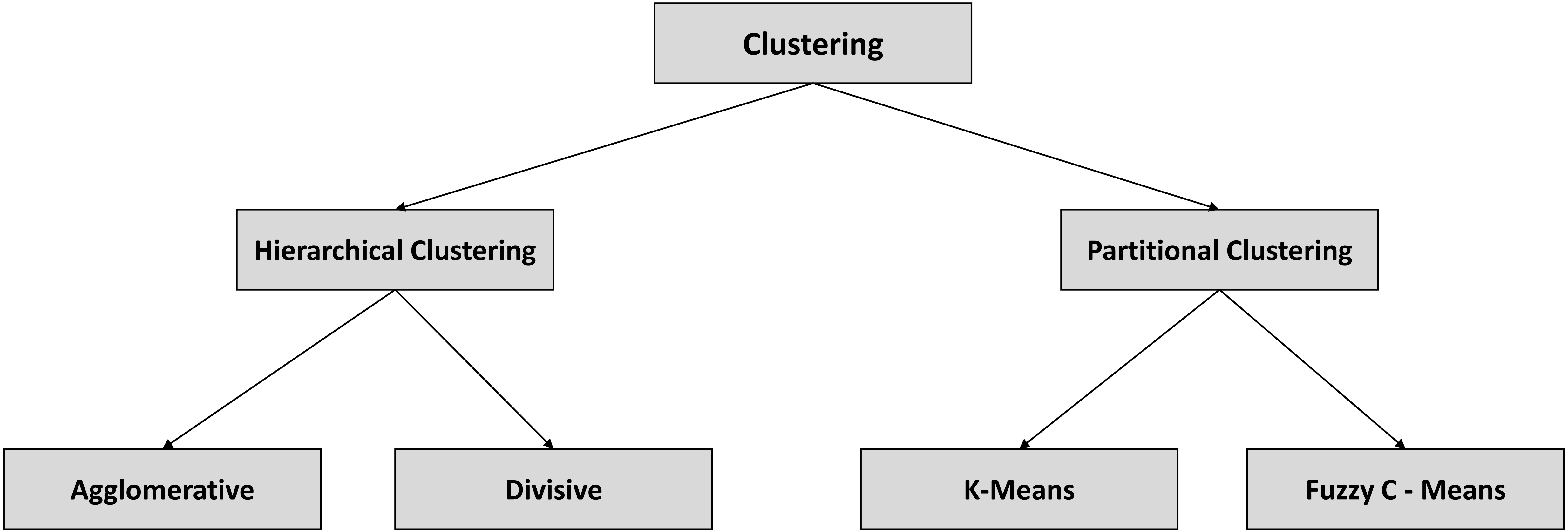


Cluster 2

And here, we can see that this cluster has players with high wickets and low wickets



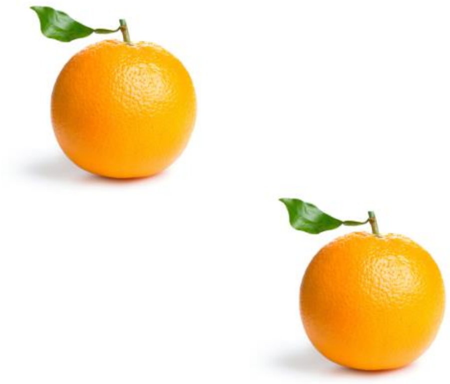
Types of clustering



Types of clustering

Clustering

Hierarchical Clustering



Clusters have a tree like structure or a parent child relationship

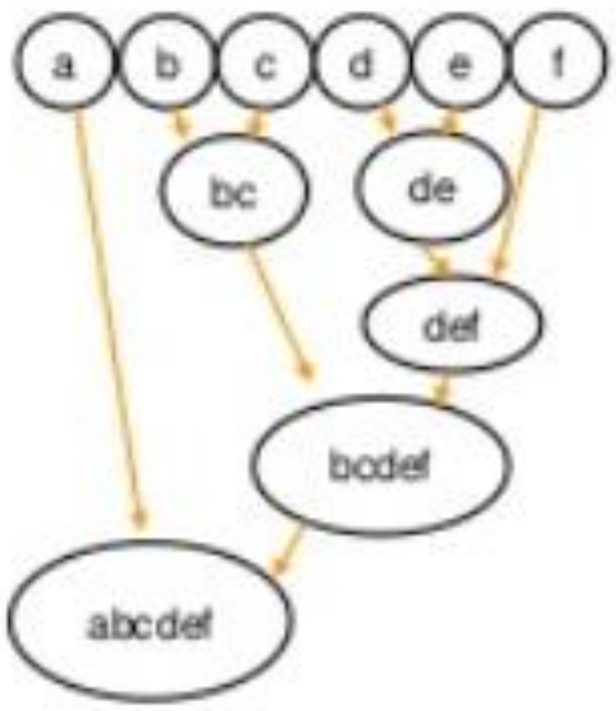
Types of clustering

Clustering

Hierarchical Clustering

Agglomerative

Divisive



Bottom up approach: Begin with each element as a separate cluster and merge them into successively large cluster

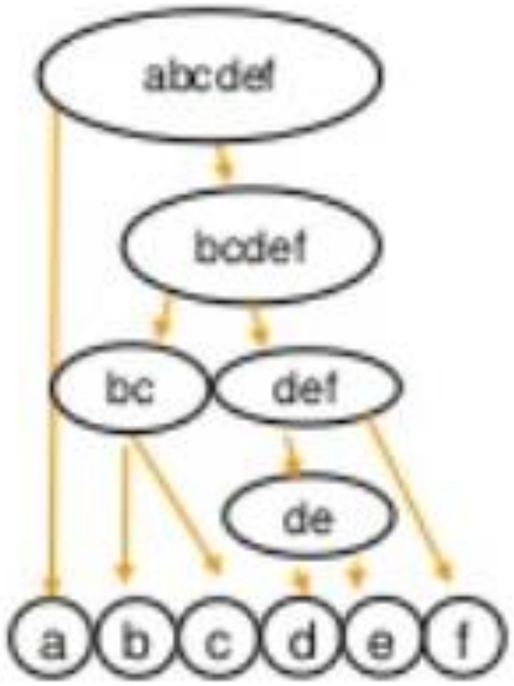
Types of clustering

Clustering

Hierarchical Clustering

Agglomerative

Divisive



Top down approach begin with whole set and proceed to divide it into successively smaller clusters

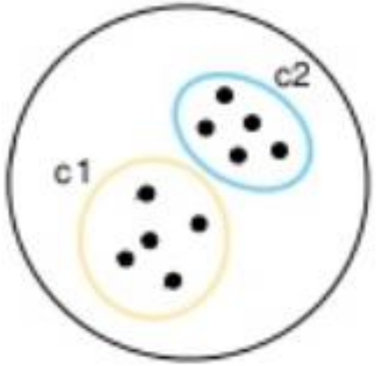
Types of clustering

Clustering

Partitional Clustering

K-Means

Fuzzy C - Means



Division of objects into clusters such that each object is in exactly one cluster, not several

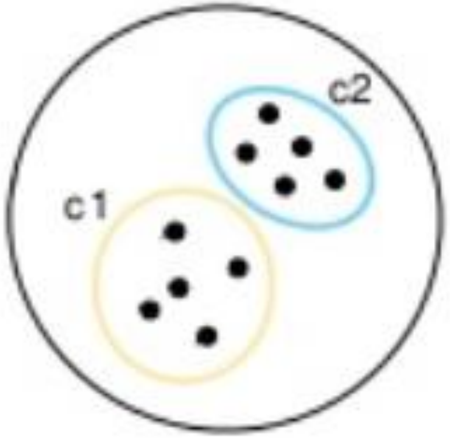
Types of clustering

Clustering

Partitional Clustering

K-Means

Fuzzy C - Means



Division of objects into clusters such that each object can belong to multiple clusters

Distance Measure

Euclidean distance measure

Manhattan distance measure

Distance measure will determine the similarity between two elements and it will influence the shape of the clusters

Squared Euclidean distance measure

Cosine distance measure

Euclidean Distance Measure

Euclidean distance measure

→

The Euclidean distance is the “ordinary” straight line.
It is the distance between two points in Euclidean space

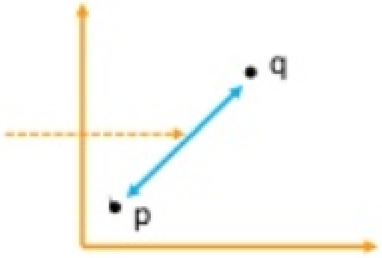
Squared Euclidean distance measure

$$d=\sqrt{\sum_{i=1}^{n}(q_i-p_i)^2}$$

Manhattan distance measure

Cosine distance measure

Euclidian distance



Squared Euclidean Distance Measure

Euclidean distance measure

Squared Euclidean distance measure →

Manhattan distance measure

Cosine distance measure

The Euclidean squared distance metric uses the same equation as the Euclidean distance metric, but does not take the square root.

$$d = \sum_{i=1}^{n} (q_i - p_i)^2$$

Manhattan Distance Measure

Euclidean distance measure

Squared Euclidean distance measure

Manhattan distance measure

Cosine distance measure

The Manhattan distance is the simple sum of the horizontal and vertical components or the distance between two points measured along axes at right angles

$$\mathbf{d}= \sum_{i=1}^{n} | q_x - p_x | + | q_y - p_y |$$

Cosine Distance Measure

Euclidean distance measure

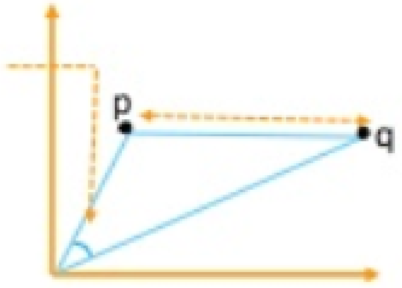
Squared Euclidean distance measure

Manhattan distance measure

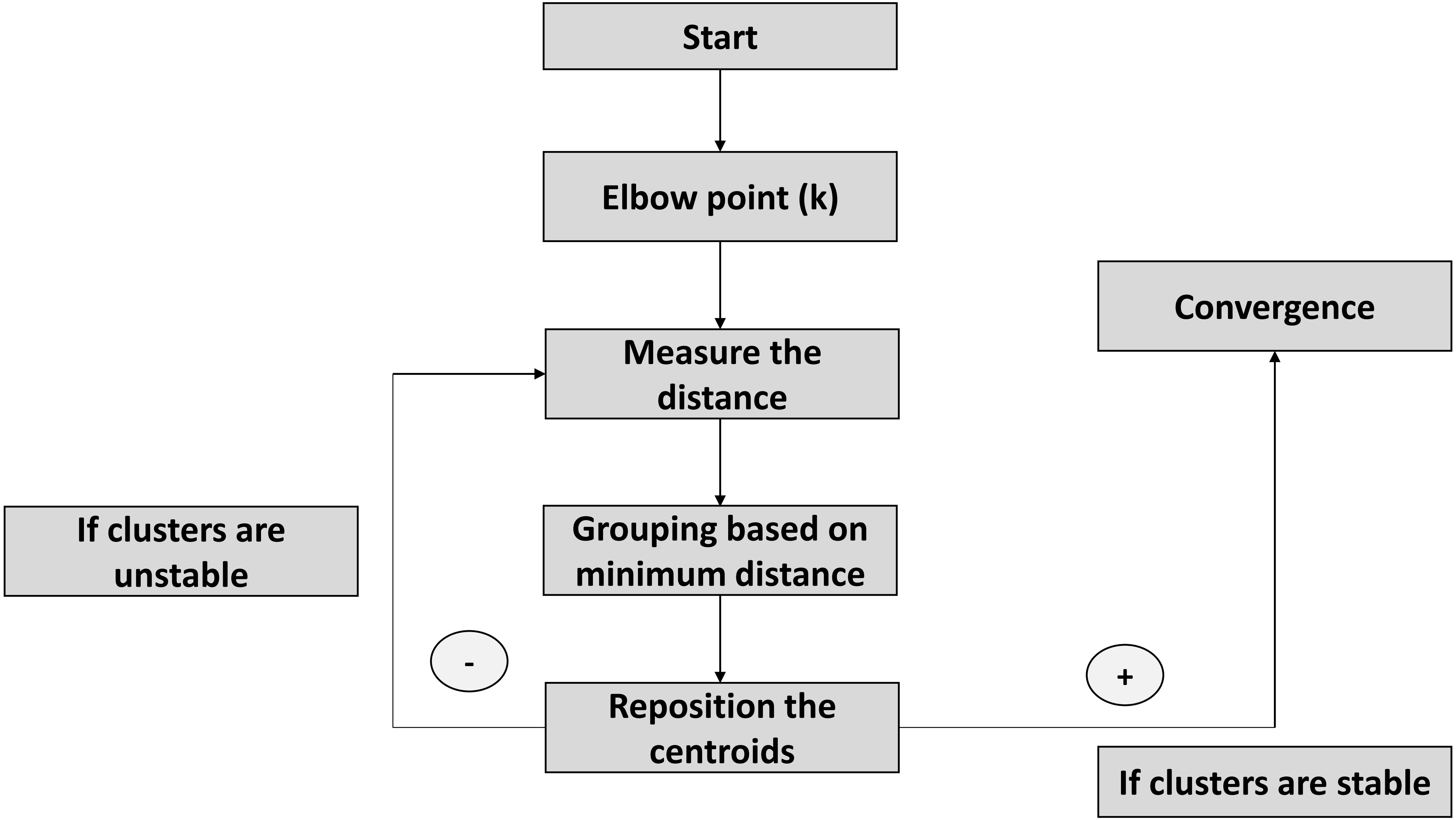
Cosine distance measure → Cosine distance

The Cosine distance similarity measures the angle between the two vectors

$$d = \frac{\sum_{i=0}^{n-1} q_i - p_x}{\sum_{i=0}^{n-1} (q_i)^2 \times \sum_{i=0}^{n-1} (p_i)^2}$$



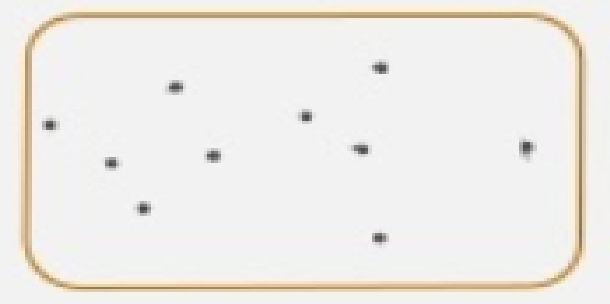
How does K – Means Clustering work?



How does K – Means Clustering work?

Elbow point (k)

- Lets say, you have a dataset for a Grocery shop



- Now, the important question is, "how would you choose the optimum number of clusters?"

How does K – Means Clustering work?

Elbow point (k)

- The best way to do this is by elbow method
- The idea of the elbow method is to run K – Means clustering on the dataset where ‘K’ is referred as number of clusters

- Within sum of squares (WSS) is defined as the sum of squared distance between each member of the cluster and its centroid

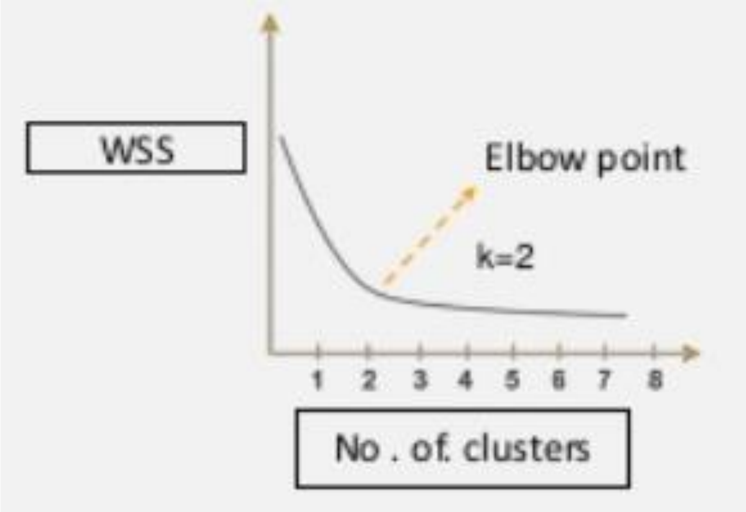
$$WSS = \sum_{i=1}^{m} (x_i - c_i)^2$$

Data point

Closest point to centroid

How does K – Means Clustering work?

Elbow point (k)

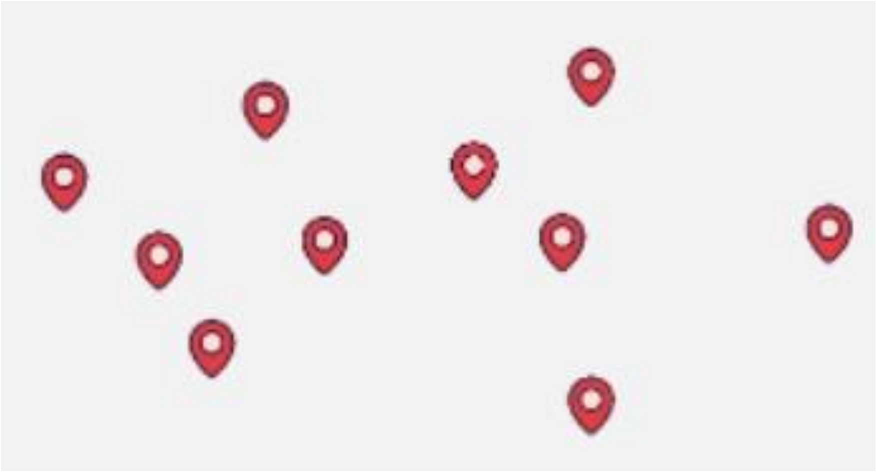


- Now, we draw a curve between WSS (within sum of squares) and the number of clusters
- Here, we can see a very slow change in the value of WSS after k = 2, so you should take that elbow point value as the final number of clusters

How does K – Means Clustering work?

Measure the distance

Step:1 Given data points below are assumed as delivery points



How does K – Means Clustering work?

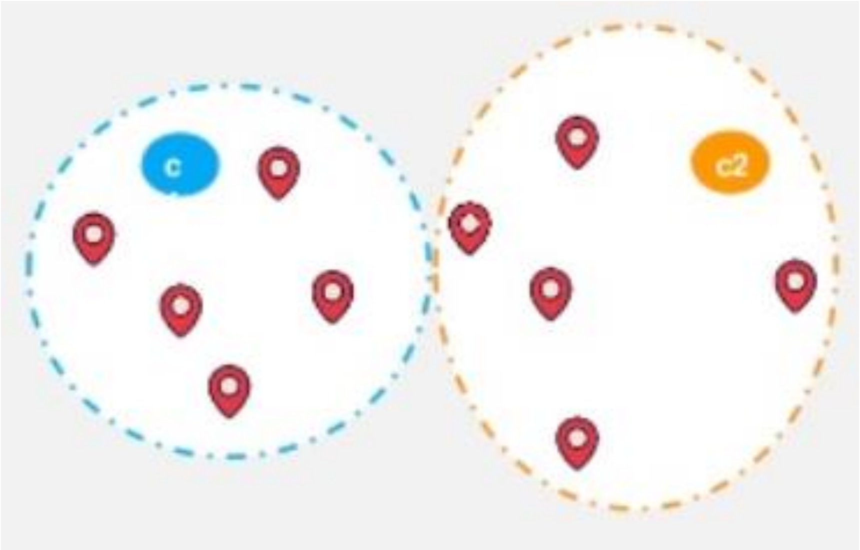
Measure the distance

Step:2 We can randomly initialize two points called the cluster centroids, Euclidean distance is a distance measure used to find out which data point is closest to our centroids.

How does K – Means Clustering work?

Grouping

Step:3 Based upon the distance from c1 and c2 centroids, the data points will group itself into clusters



How does K – Means Clustering work?

Reposition the centroids

Step 4: Compute the centroid of data points inside blue cluster
Step 5: Reposition the centroid of the blue cluster to the new centroid

How does K – Means Clustering work?

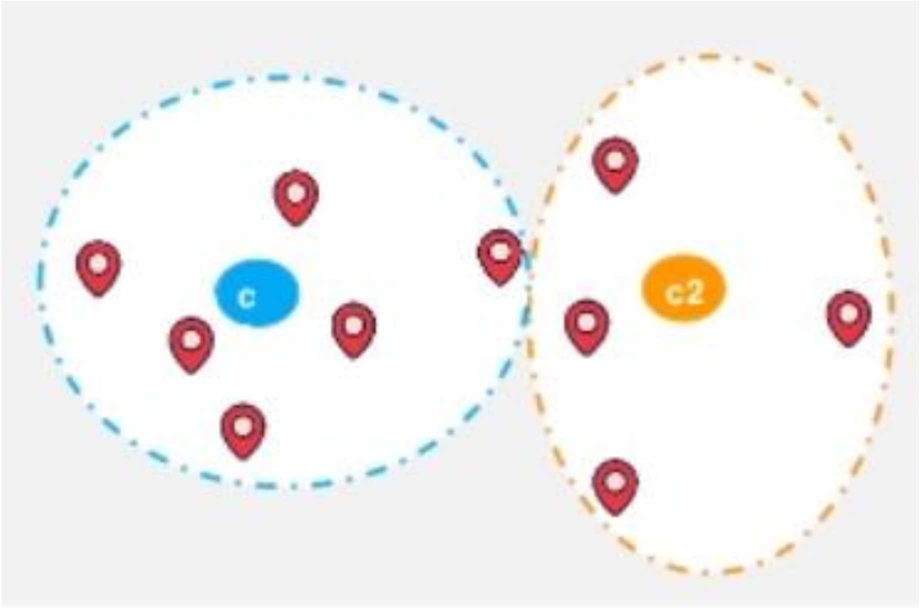
Reposition the centroids

Step 6: Compute the centroid of data points inside orange cluster
Step 7: Reposition the centroid of the orange cluster to the new centroid

How does K – Means Clustering work?

Convergence

Step 8: Once the clusters become static, K – Means clustering algorithm is said to be converged



K – Means Clustering Algorithm

Assuming we have inputs x1,x2,x3….. and value of K,

Step 1: Pick K random points as cluster centers called centroids

Step 2: Assign each xi to nearest cluster by calculating its distance to each centroid

Step 3: Find new cluster center by taking the average of the assigned points

Step 4: Repeat Step 2 and Step 3 until none of the cluster assignments change

K – Means Clustering Algorithm

Step 1: Begin with a decision on the value of k – number of clusters

Step 2: Put any initial partition that classifies the data into k clusters. You may assign the training samples randomly, or systematically as the following”

I. Take the first k training sample as single element clusters

II. Assign each of the remaining (N-k) training sample to the cluster with the nearest centroid. After each assignment, recompute the centroid of the gaining cluster

K – Means Clustering Algorithm

Step 3: Take each sample in sequence and compute its distance from the centroid of each of the clusters. If a sample is not currently in the cluster with the closest centroid, switch this sample to that cluster and update the centroid of the cluster gaining the new sample and the cluster losing the sample

Step 4: Repeat step 3 until convergence is achieved, that is until a pass through the training sample causes no new assignments.

K – Means Clustering Algorithm

Example – Implementation of k – means algorithm (using K = 2)

Individual	Variable 1	Variable 2
1	1.0	1.0
2	1.5	2.0
3	3.0	4.0
4	5.0	7.0
5	3.5	5.0
6	4.5	5.0
7	3.5	4.5

K – Means Clustering Algorithm

Step 1:

Initialization: Randomly we choose following two centroids (k=2) for 2 clusters.

In this case the 2 centroid are: (1.0,1.0) and (5.0,7.0)

Individual	Variable 1	Variable 2
1	1.0	1.0
2	1.5	2.0
3	3.0	4.0
4	5.0	7.0
5	3.5	5.0
6	4.5	5.0
7	3.5	4.5

	Individual	Mean Vector
Group 1	1	(1.0,1.0)
Group 2	4	(5.0,7.0)

K – Means Clustering Algorithm

Step – 2:
The remaining individuals are now examined in sequence and allocated to the cluster to which they are closest, in terms of Euclidean distance to the cluster mean.

	Cluster 1		**Cluster 2**	
Step	Individual	Mean Vector (Cluster 1)	Individual	Mean Vector (Cluster 2)
1	1	(1.0, 1.0)	4	(5.0, 7.0)
2	1,2	(1.2, 1.5)	4,5	(4.2, 6.0)
3	1,2,3	(1.8, 2.3)	4,5,6	(4.3, 5.7)
			4,5,6,7	(4.1, 5.4)

Now, Clusters have following characteristics:

	Individual	**Mean Vector**
Cluster 1	{1,2,3}	(1.8, 2.3)
Cluster 2	{4,5,6,7}	(4.1, 5.4)

K – Means Clustering Algorithm

Step – 3:

We compare each individual's distance to mean of Cluster 1 and Cluster 2.

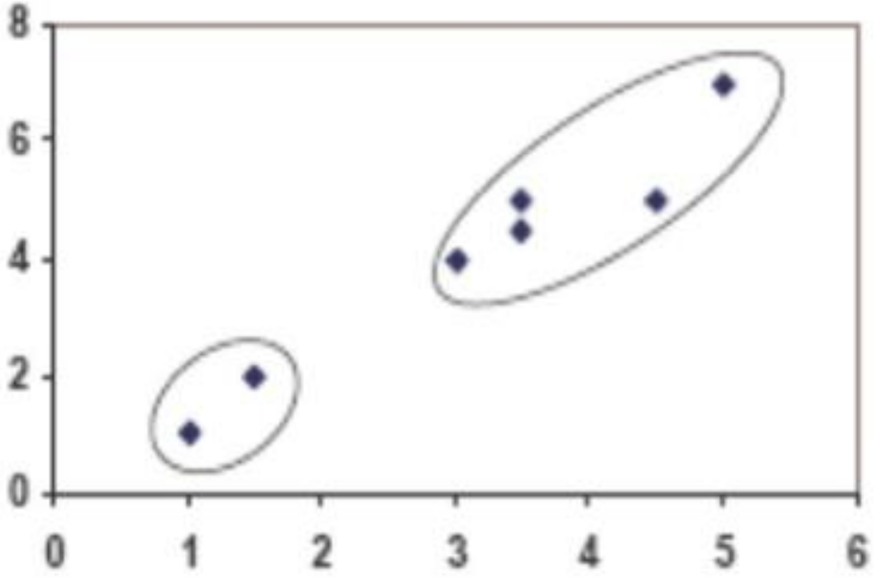
Individual	Distance Mean with Cluster 1	Distance Mean with Cluster 2
1	1.5	5.4
2	0.4	4.3
3	2.1	1.8
4	5.7	1.8
5	3.2	0.7
6	3.8	0.6
7	2.8	1.1

We allocate the individual to the cluster with least distance. The new cluster is now:

	Individual	Mean Vector
Cluster 1	{1,2}	(1.3, 1.5)
Cluster 2	{3,4,5,6,7}	(3.9, 5.1)

K – Means Clustering Algorithm

Plot



Thank You