ABDELGHAFOR'S VIRTUAL INTERNSHIP

MACHINE LEARNING PROGRAM

SESSION (4)

PREPARED BY : MARK KOSTANTINE

LECTURE OVERVIEW

- Unsupervised Learning : Get Started
- Clustering in Machine Learning
- K-Means Clustering
- Hierarchical Clustering
- Association Rule Learning
- Questions

WHAT IS UNSUPERVISED LEARNING?

unsupervised learning is a machine learning technique in which models are not supervised using training dataset. Instead, **models itself find the hidden patterns** and insights from the given data. It can be compared to learning which takes place in the human brain while learning new things. It can be defined as:

- *Unsupervised learning is a type of machine learning in which models are trained using unlabeled dataset and are allowed to act on that data without any supervision.*

Unsupervised learning cannot be directly applied to a regression or classification problem because unlike supervised learning, we have the input data ***but no corresponding output data.*** *The goal of unsupervised learning is to find the underlying structure of dataset, group that data according to similarities, and represent that dataset in a compressed format.*

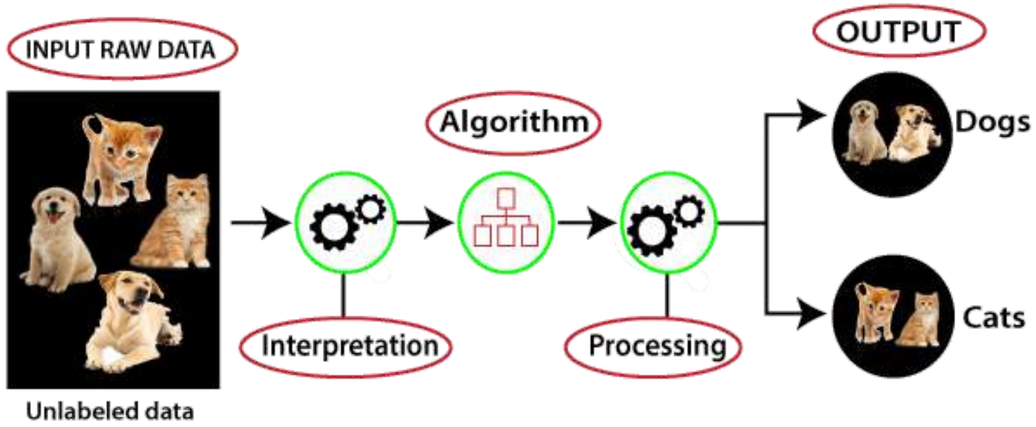
Example: *Suppose the unsupervised learning algorithm is given an input dataset containing images of different types of cats and dogs. The algorithm is never trained upon the given dataset, which means it does not have any idea about the features of the dataset. The task of the unsupervised learning algorithm is to identify the image features on their own. Unsupervised learning algorithm will perform this task by clustering the image dataset into the groups according to similarities between images.*

WHY USE UNSUPERVISED LEARNING?

- Unsupervised learning is helpful for finding useful insights from the data.
- Unsupervised learning **is much similar as** a human learns to think by their own experiences, which makes it closer to the real AI.
- Unsupervised learning works on **unlabeled and uncategorized data** which make unsupervised learning more important.
- In real-world, we do not always have input data with the corresponding output so to solve such cases, we need unsupervised learning.

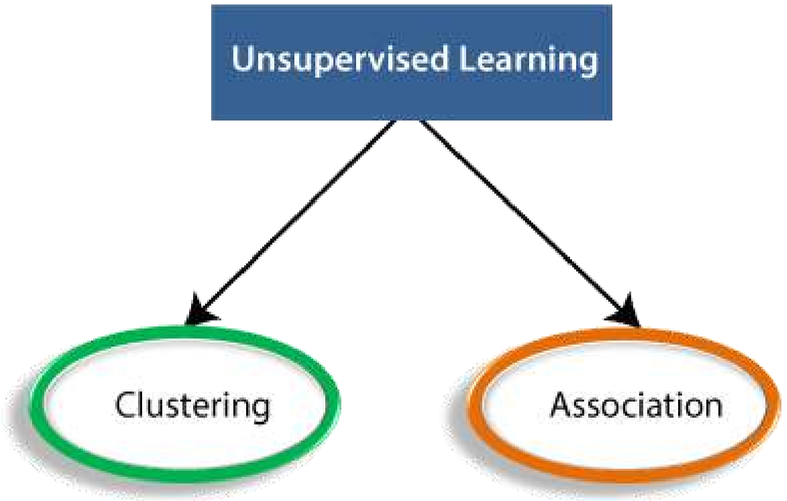
WORKING OF UNSUPERVISED LEARNING

Here, we have taken an unlabeled input data, which means it is not categorized and corresponding outputs are also not given. Now, this unlabeled input data is fed to the machine learning model in order to train it. Firstly, it will interpret the raw data to find the hidden patterns from the data and then will apply suitable algorithms such as k-means clustering



TYPES OF UNSUPERVISED LEARNING

The unsupervised learning algorithm can be further categorized into two types of problems



- **Clustering:** Clustering is a method of grouping the objects into clusters such that objects with most similarities remains into a group and has less or no similarities with the objects of another group.

- **Association**: An association rule is an unsupervised learning method which is used for finding the relationships between variables in the large database. It determines the set of items that occurs together in the dataset. Association rule makes marketing strategy more effective

UNSUPERVISED LEARNING ALGORITHMS

- Hierarchal clustering
- K-means clustering
- Anomaly detection
- Neural Networks
- Principle Component Analysis
- Independent Component Analysis
- Apriori algorithm
- Singular value decomposition

ADVANTAGES OF UNSUPERVISED LEARNING

- Unsupervised learning is used for **more complex tasks** as compared to supervised learning because, in unsupervised learning, we don't have labeled input data.
- Unsupervised learning is preferable as it is easy to get unlabeled data in comparison to labeled data.

DISADVANTAGES OF UNSUPERVISED LEARNING

- Unsupervised learning is intrinsically **more difficult** than supervised learning as it does not have corresponding output.
- The result of the unsupervised learning algorithm might be less accurate as input data is not labeled, and algorithms do not know the exact output in advance.

CLUSTERING IN MACHINE LEARNING

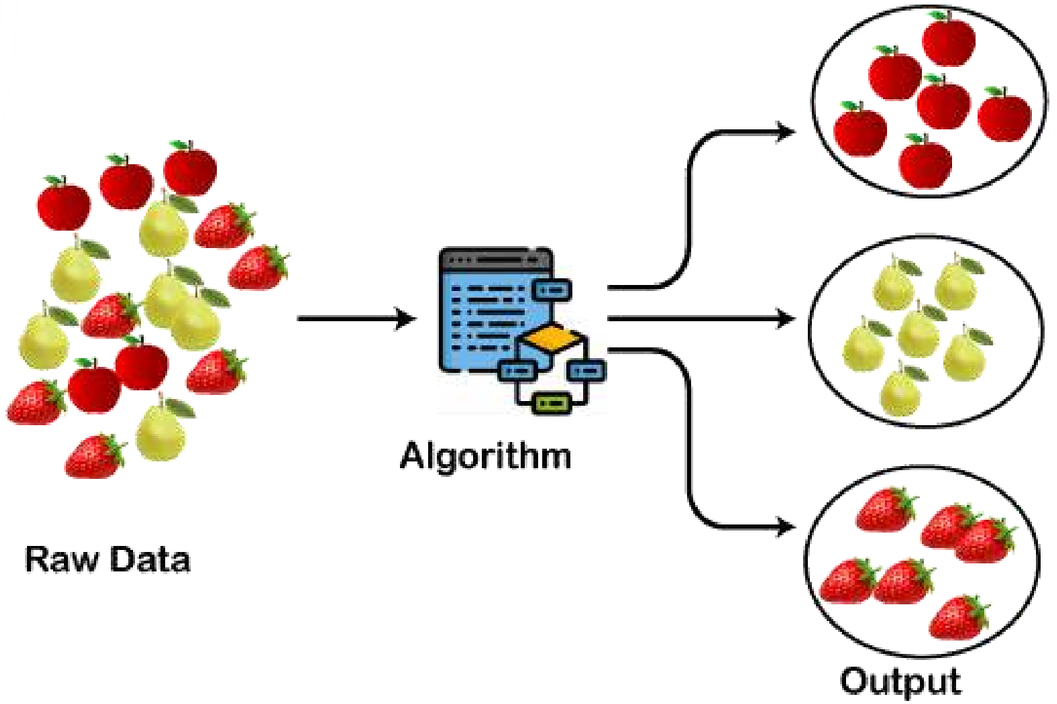
- Clustering or cluster analysis is a machine learning technique, which groups the unlabelled dataset. It can be defined as "***A way of grouping the data points into different clusters, consisting of similar data points. The objects with the possible similarities remain in a group that has less or no similarities with another group.***"
- *It does it by finding some similar patterns in the unlabelled dataset such as shape, size, color, behavior, etc., and divides them as per the presence and absence of those similar patterns.*
- *It is an* ***unsupervised learning method*** *hence no supervision is provided to the algorithm, and it deals with the unlabeled dataset.*
- *After applying this clustering technique, each cluster or group is provided with a cluster-ID. ML system can use this id to simplify the processing of large and complex datasets.*
- *The clustering technique is commonly used for statistical data analysis.*
- **Note:** Clustering is somewhere **similar to the classification algorithm**, but the difference is the type of dataset that we are using. In classification, we work with the labeled data set, whereas in clustering, we work with the unlabelled dataset.

CLUSTERING IN MACHINE LEARNING

The clustering technique can be widely used in various tasks. Some most common uses of this technique are:

- Market Segmentation
- Statistical data analysis
- Social network analysis
- Image segmentation
- Anomaly detection

Apart from these general usages, it is used by the **Amazon** in its recommendation system to provide the recommendations as per the past search of products. **Netflix** also uses this technique to recommend the movies and web-series to its users as per the watch history.



TYPES OF CLUSTERING METHODS

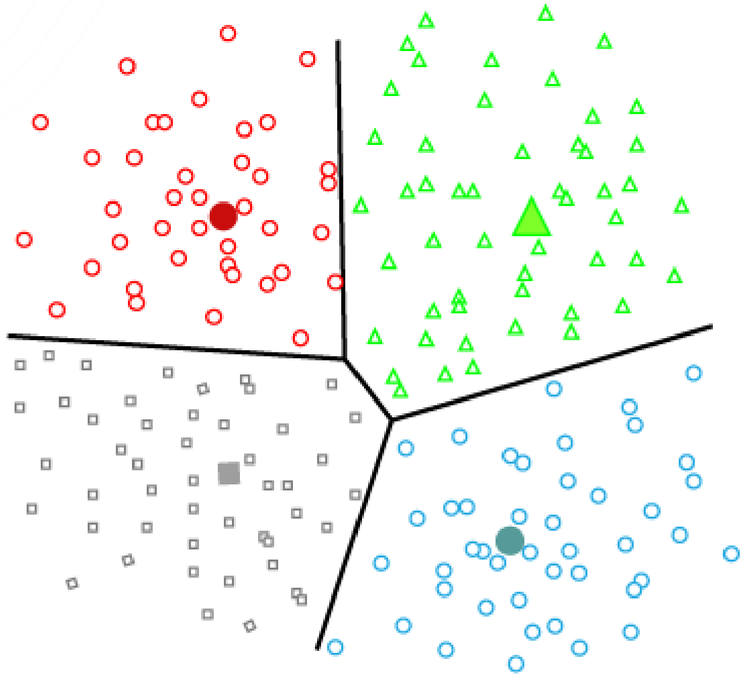
The clustering methods are broadly divided into **Hard clustering** (datapoint belongs to only one group) and **Soft Clustering** (data points can belong to another group also). But there are also other various approaches of Clustering exist. Below are the main clustering methods used in Machine learning:

- **Partitioning Clustering**
- **Density-Based Clustering**
- **Distribution Model-Based Clustering**
- **Hierarchical Clustering**

PARTITIONING CLUSTERING

It is a type of clustering that divides the data into non-hierarchical groups. It is also known as the centroid-based method. The most common example of partitioning clustering is the **K-Means Clustering algorithm.**

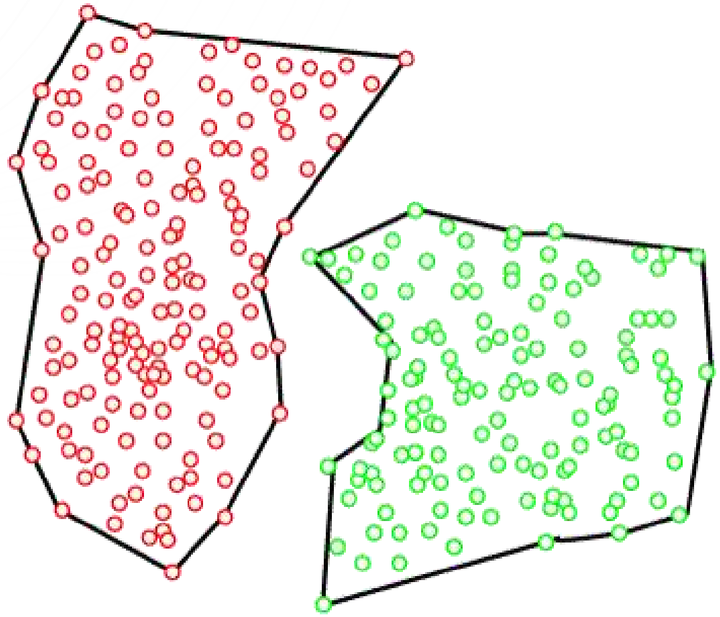
In this type, the dataset is divided into a set of k groups, where K is used to define the number of pre-defined groups. The cluster center is created in such a way that the distance between the data points of one cluster is minimum as compared to another cluster centroid.



DENSITY-BASED CLUSTERING

The density-based clustering method connects the highly-dense areas into clusters, and the arbitrarily shaped distributions are formed as long as the dense region can be connected. This algorithm does it by identifying different clusters in the dataset and connects the areas of high densities into clusters. The dense areas in data space are divided from each other by sparser areas.

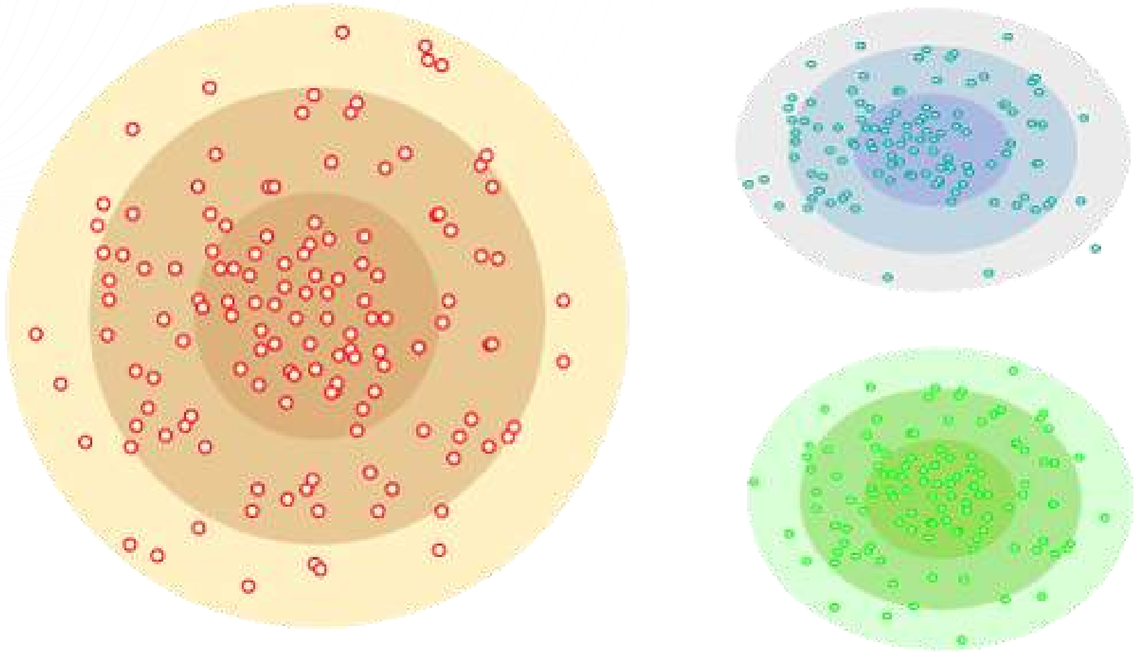
These algorithms can face difficulty in clustering the data points if the dataset has varying densities and high dimensions



DISTRIBUTION MODEL-BASED CLUSTERING

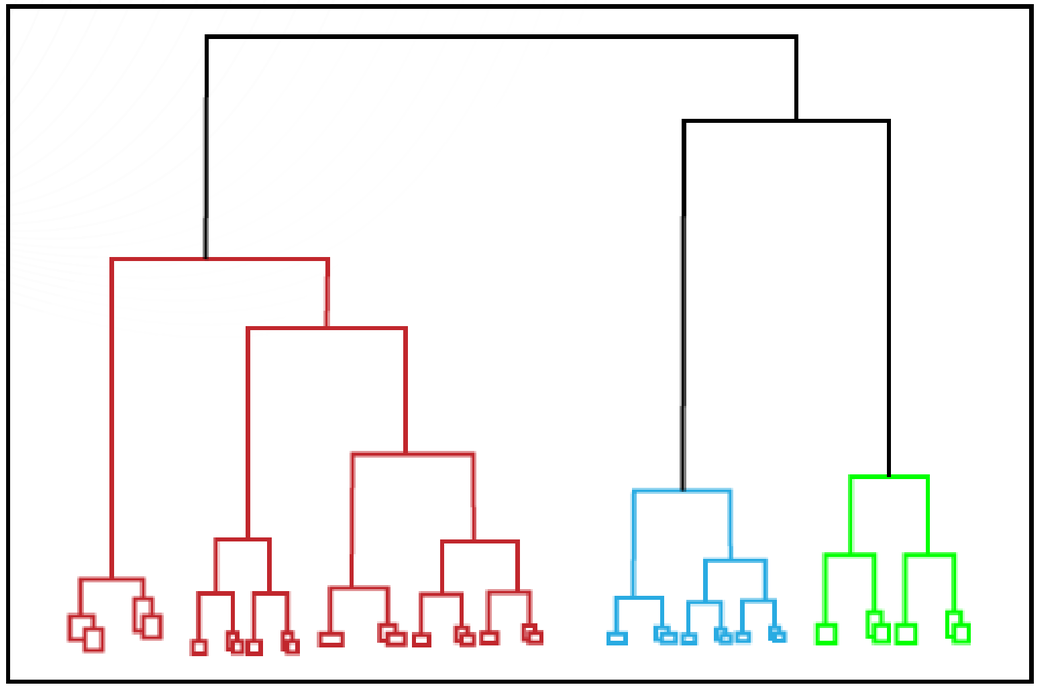
In the distribution model-based clustering method, the data is divided based on the probability of how a dataset belongs to a particular distribution. The grouping is done by assuming some distributions commonly **Gaussian Distribution.**

The example of this type is the **Expectation-Maximization Clustering algorithm** that uses Gaussian Mixture Models (GMM).



HIERARCHICAL CLUSTERING

Hierarchical clustering can be used as an alternative for the partitioned clustering as there is no requirement of pre-specifying the number of clusters to be created. In this technique, the dataset is divided into clusters to create a tree-like structure, which is also called a **dendrogram**. The observations or any number of clusters can be selected by cutting the tree at the correct level. The most common example of this method is the **Agglomerative** Hierarchical algorithm.



APPLICATIONS OF CLUSTERING

- **In Identification of Cancer Cells:** The clustering algorithms are widely used for the identification of cancerous cells. It divides the cancerous and non-cancerous data sets into different groups.
- **In Search Engines:** Search engines also work on the clustering technique. The search result appears based on the closest object to the search query. It does it by grouping similar data objects in one group that is far from the other dissimilar objects. The accurate result of a query depends on the quality of the clustering algorithm used.
- **Customer Segmentation:** It is used in market research to segment the customers based on their choice and preferences.
- **In Biology**: It is used in the biology stream to classify different species of plants and animals using the image recognition technique.
- **In Land Use:** The clustering technique is used in identifying the area of similar lands use in the GIS database. This can be very useful to find that for what purpose the particular land should be used, that means for which purpose it is more suitable.

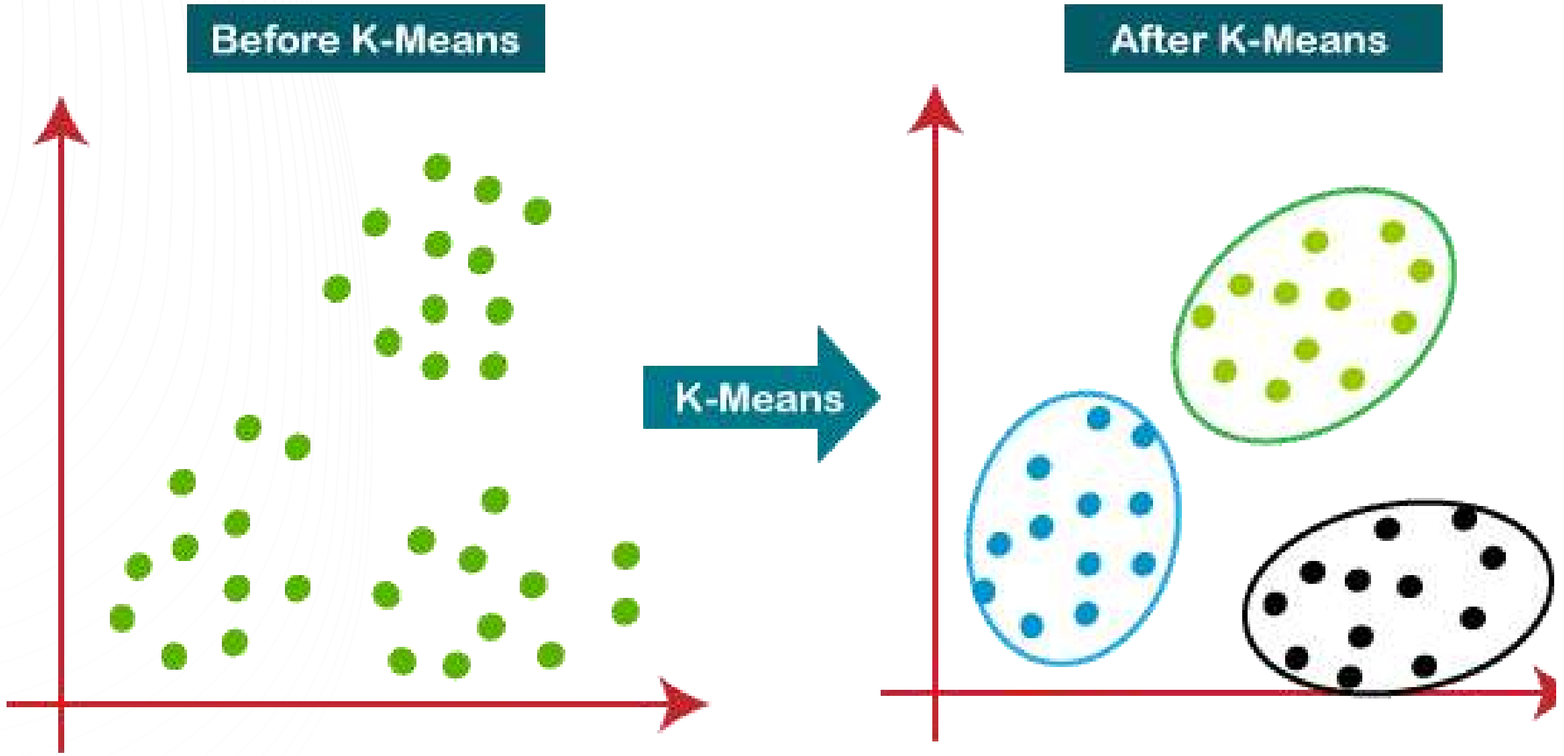
K-MEANS CLUSTERING

- K-Means Clustering is an Unsupervised Learning algorithm, which groups the unlabeled dataset into different clusters. Here K defines the **number of pre-defined clusters** that need to be created in the process, as if K=2, there will be two clusters, and for K=3, there will be three clusters, and so on.

It is an iterative algorithm that divides the unlabeled dataset into k different clusters in such a way that each dataset belongs only one group that has similar properties.

- It allows us to cluster the data into different groups and a convenient way to discover the categories of groups in the unlabeled dataset on its own without the need for any training.
- It is a **centroid-based algorithm**, where each cluster is associated with a centroid. The main aim of this algorithm is to **minimize the sum of distances between the data point and their corresponding clusters.**
- The algorithm takes the unlabeled dataset as input, divides the dataset into k-number of clusters, and repeats the process until it does not find the best clusters. The value of k should be predetermined in this algorithm.
- 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.

K-MEANS CLUSTERING



HOW DOES THE K-MEANS ALGORITHM WORK?

Step-1: Select the number K to decide the number of clusters.

Step-2: Select random K points or centroids. (It can be other from the input dataset).

Step-3: Assign each data point to their closest centroid, which will form the predefined K clusters.

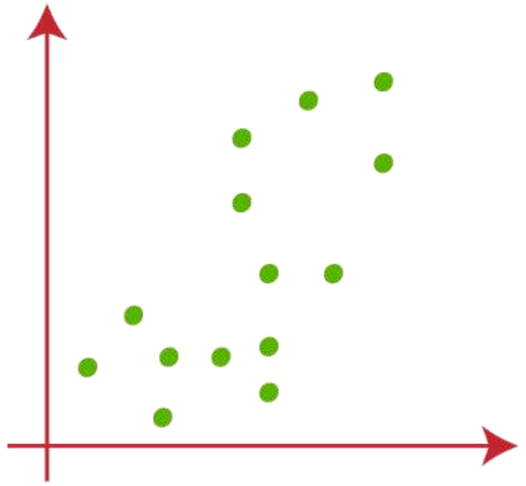
Step-4: Calculate the variance and place a new centroid of each cluster.

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.

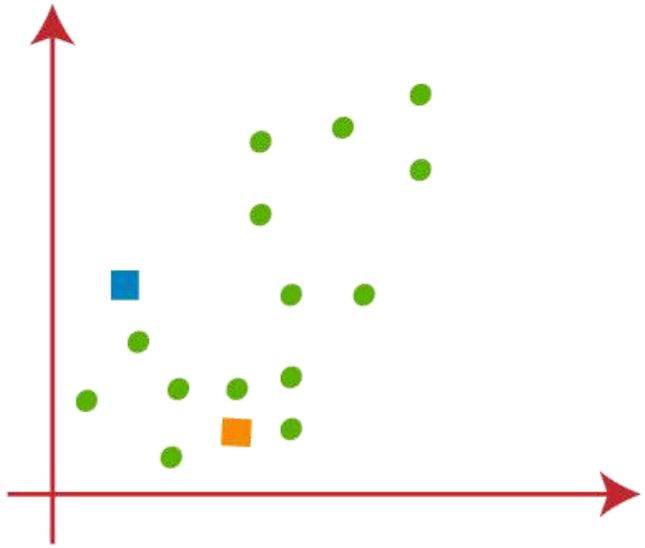
Step-7: The model is ready.

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



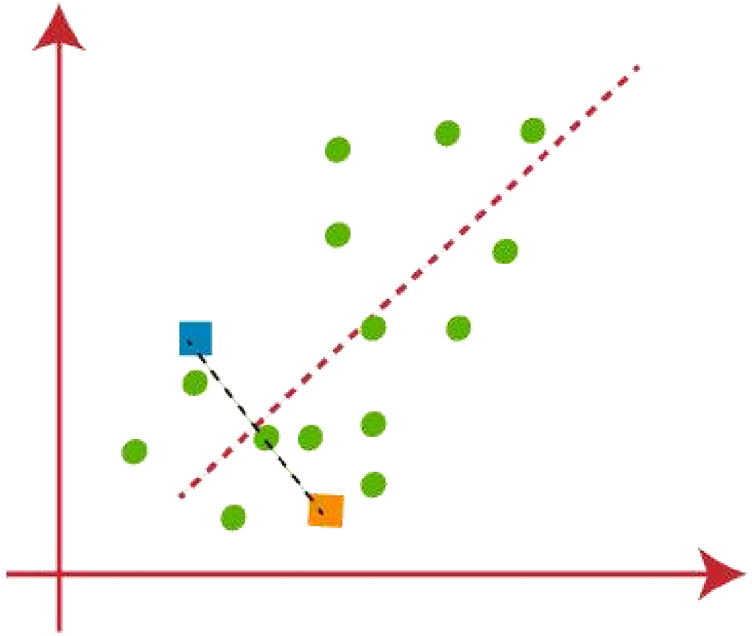
HOW DOES THE K-MEANS ALGORITHM WORK?

- 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:



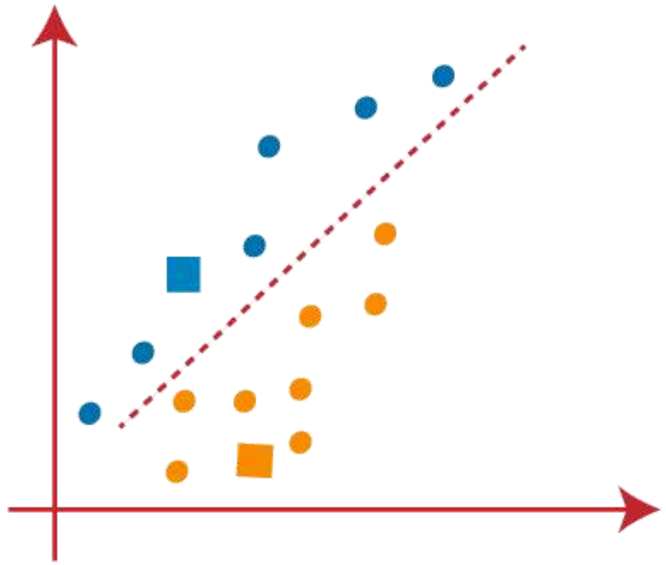
HOW DOES THE K-MEANS ALGORITHM WORK?

- 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 to calculate the distance between two points. So, we will draw a median between both the centroids. Consider the below image:



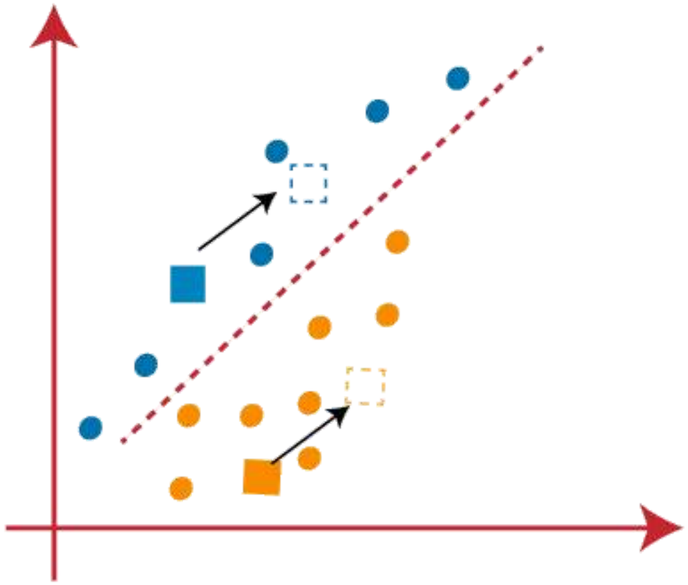
HOW DOES THE K-MEANS ALGORITHM WORK?

- From the above 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.



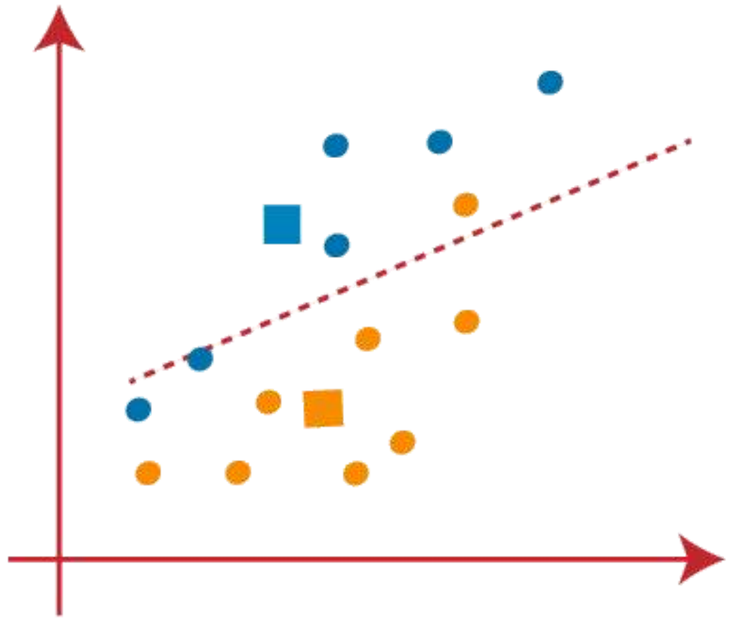
HOW DOES THE K-MEANS ALGORITHM WORK?

- 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:



HOW DOES THE K-MEANS ALGORITHM WORK?

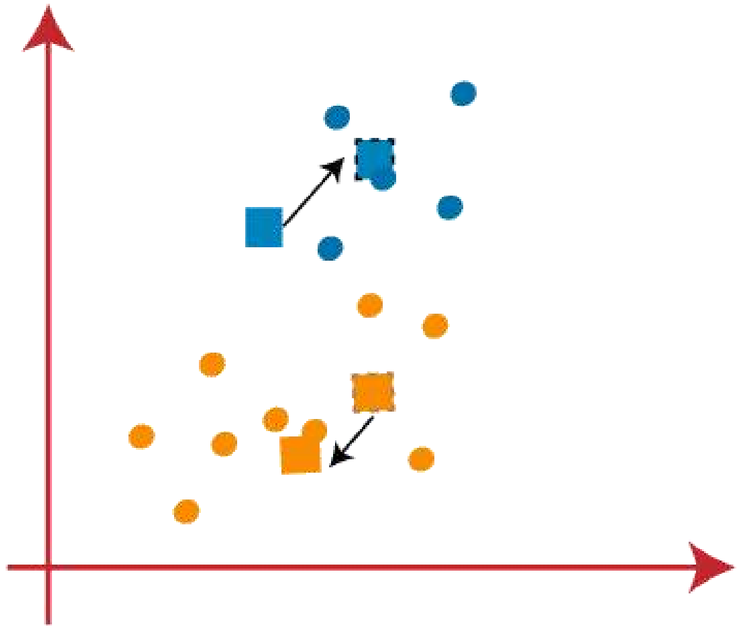
- Next, we will reassign each datapoint to the new centroid. For this, we will repeat the same process of finding a median line. The median will be like below image:



- From the above 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.

HOW DOES THE K-MEANS ALGORITHM WORK?

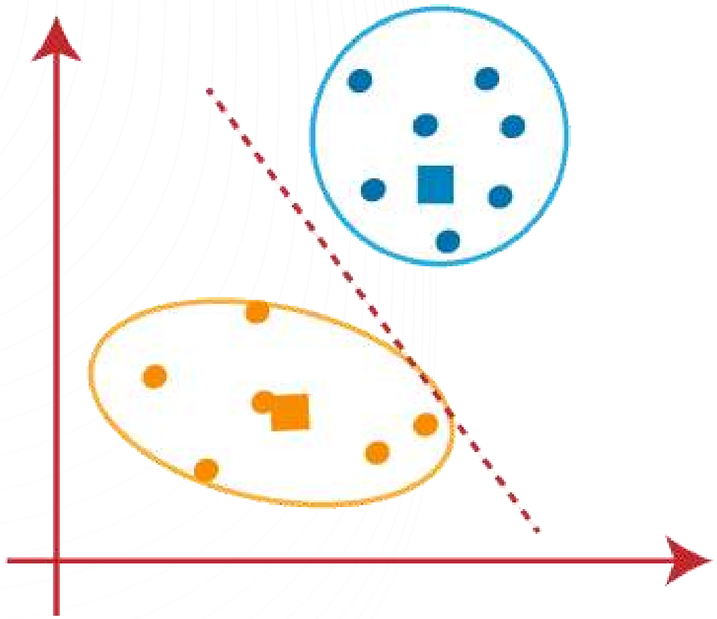
- We will repeat the process by finding the center of gravity of centroids, so the new centroids will be as shown in the below image:

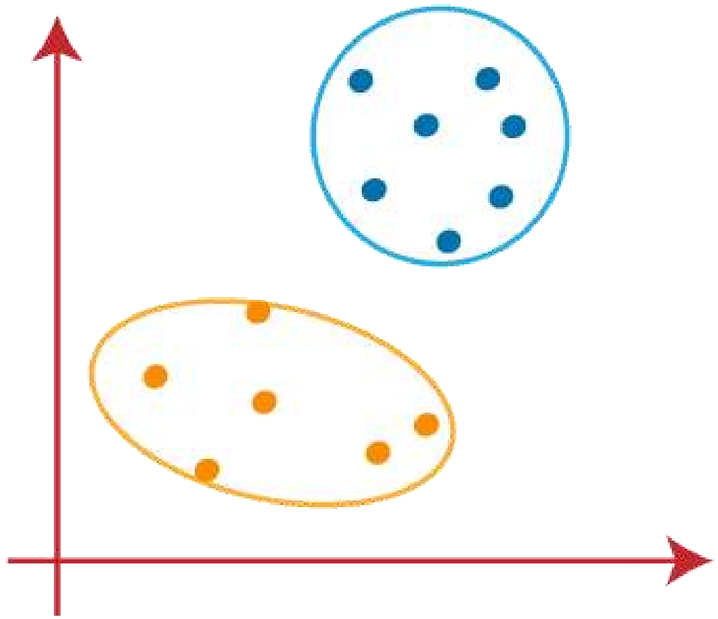


- As we got the new centroids so again will draw the median line and reassign the data points

HOW DOES THE K-MEANS ALGORITHM WORK?

- there are no dissimilar data points on either side of the line, which means our model is formed.





HOW TO CHOOSE THE VALUE OF K NUMBER OF CLUSTERS

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, but here we are discussing the most appropriate method to find the number of clusters or value of K.

ELBOW METHOD

The Elbow method is one of the most popular ways to find the optimal number of clusters. 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:

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

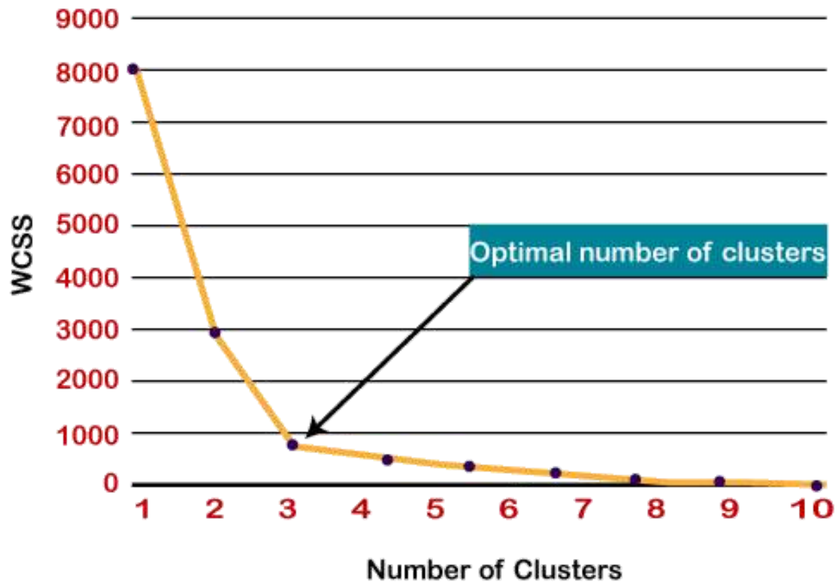
∑Pi in Cluster1 distance(Pi C1)2: It is the sum of the square of the distances between each data point and its centroid within a cluster1 and the same for the other two terms.

ELBOW METHOD

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

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.



Since the graph shows the sharp bend, which looks like an elbow, hence it is known as the elbow method.

Note: We can choose the number of clusters equal to the given data points. If we choose the number of clusters equal to the data points, then the value of WCSS becomes **zero**, and that will be the endpoint of the plot.

HIERARCHICAL CLUSTERING

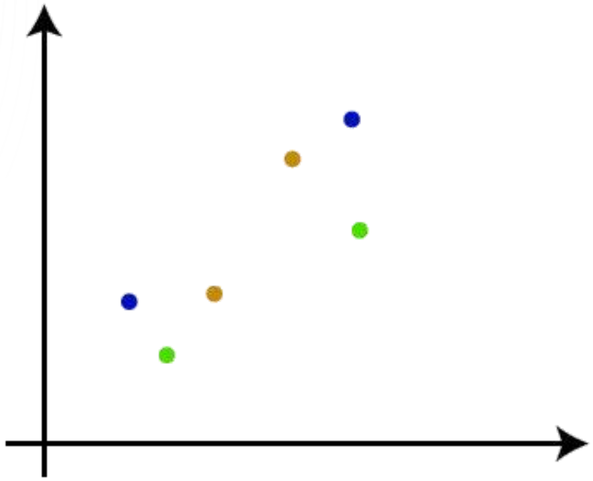
- Hierarchical clustering is unsupervised machine learning algorithm, which is used to group the unlabeled datasets into a cluster and also known as **hierarchical cluster analysis or HCA.**
- In this algorithm, we develop the hierarchy of clusters in the form of a tree, and this tree-shaped structure is known as the **dendrogram**
- **Sometimes the results of K-means clustering and hierarchical clustering may look similar**, but they both differ depending on how they work. As there is no requirement to predetermine the number of clusters as we did in the K-Means algorithm.
- The hierarchical clustering technique has two approaches:
  - **Agglomerative:** Agglomerative is **a bottom-up approach**, in which the algorithm starts with taking all data points as single clusters and merging them until one cluster is left.
  - **Divisive:** Divisive algorithm is the reverse of the agglomerative algorithm as it is **a top-down approach.**

AGGLOMERATIVE HIERARCHICAL CLUSTERING

- The agglomerative hierarchical clustering algorithm is a popular example of HCA. To group the datasets into clusters, it follows the **bottom-up approach.** It means, this algorithm considers each dataset as a single cluster at the beginning, and then start combining the closest pair of clusters together. It does this until all the clusters are merged into a single cluster that contains all the datasets.
- This hierarchy of clusters is represented in the form of the dendrogram.

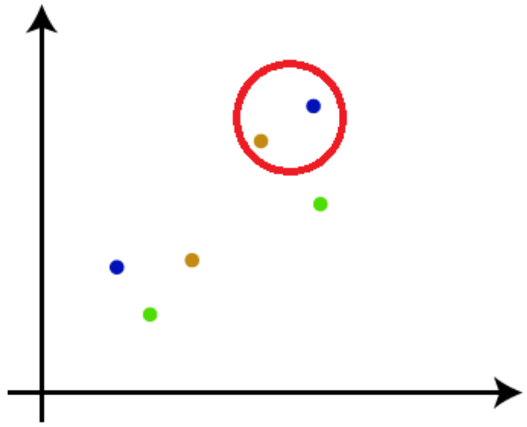
HOW THE AGGLOMERATIVE HIERARCHICAL CLUSTERING WORK?

Step-1: Create each data point as a single cluster. Let's say there are N data points, so the number of clusters will also be N.

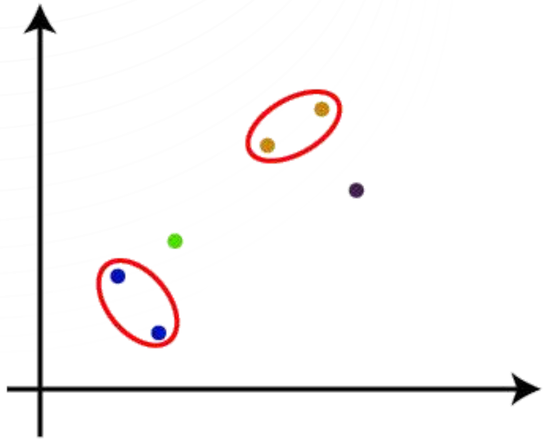


HOW THE AGGLOMERATIVE HIERARCHICAL CLUSTERING WORK?

Step-2: Take two closest data points or clusters and merge them to form one cluster. So, there will now be N-1 clusters.

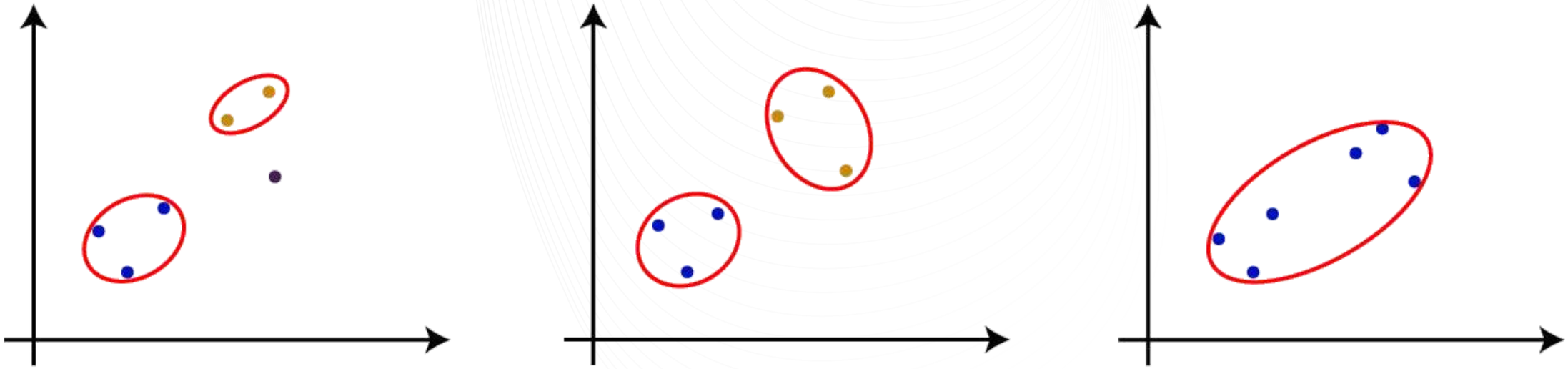


Step-3: Again, take the two closest clusters and merge them together to form one cluster. There will be N-2 clusters.



HOW THE AGGLOMERATIVE HIERARCHICAL CLUSTERING WORK?

Step-4: Repeat Step 3 until only one cluster left. So, we will get the following clusters. Consider the below images:

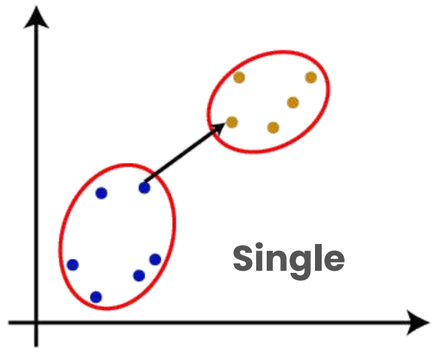


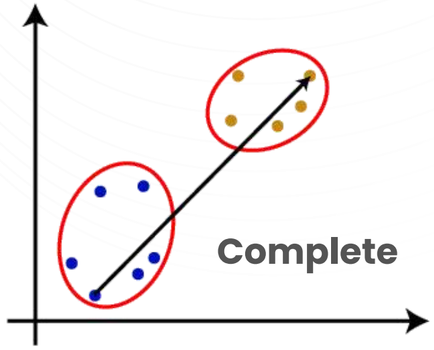
Step-5: Once all the clusters are combined into one big cluster, develop the dendrogram to divide the clusters as per the problem.

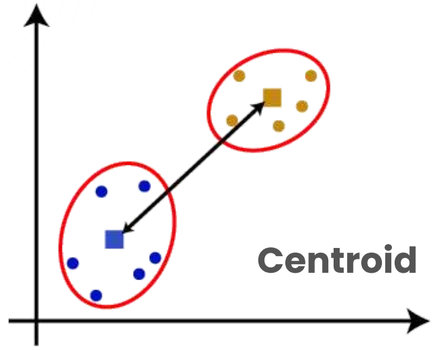
MEASURE FOR THE DISTANCE BETWEEN TWO CLUSTERS

the closest distance between the two clusters is **crucial** for the hierarchical clustering. There are various ways to calculate the distance between two clusters, and these ways decide the rule for clustering. These measures are called **Linkage methods.**

1. **Single Linkage:** It is the Shortest Distance between the closest points of the clusters.
2. **Complete Linkage:** It is the farthest distance between the two points of two different clusters. It is one of the popular linkage methods as it forms tighter clusters than single-linkage.
3. **Average Linkage:** It is the linkage method in which the distance between each pair of datasets is added up and then divided by the total number of datasets to calculate the average distance between two clusters. It is also one of the most popular linkage methods.
4. **Centroid Linkage:** It is the linkage method in which the distance between the centroid of the clusters is calculated.



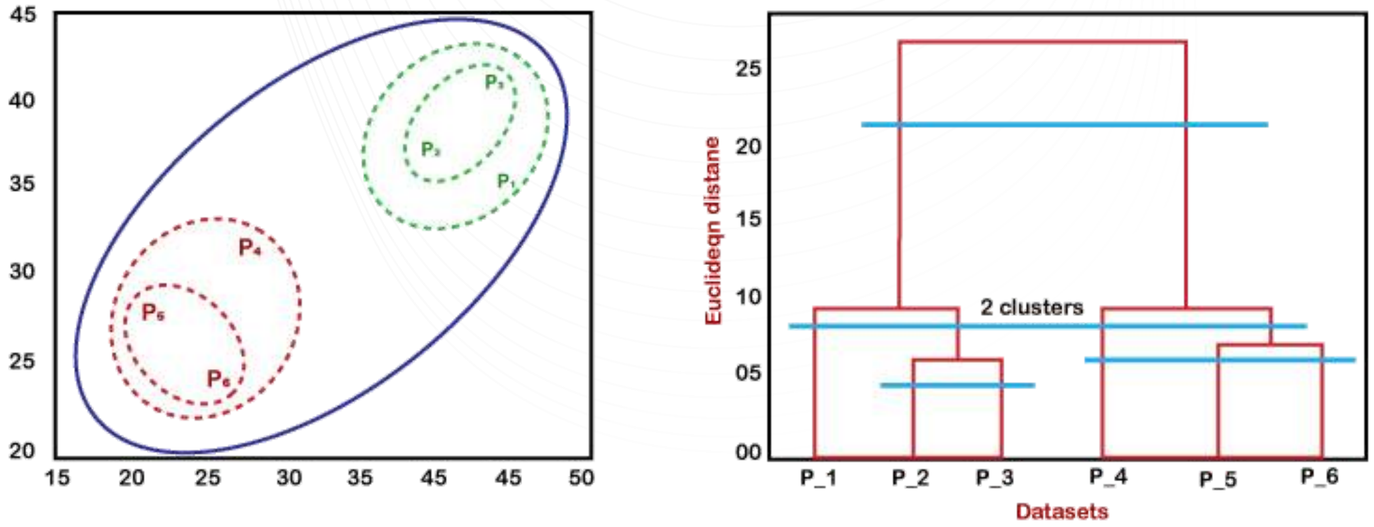




WOKING OF DENDROGRAM IN HIERARCHICAL CLUSTERING

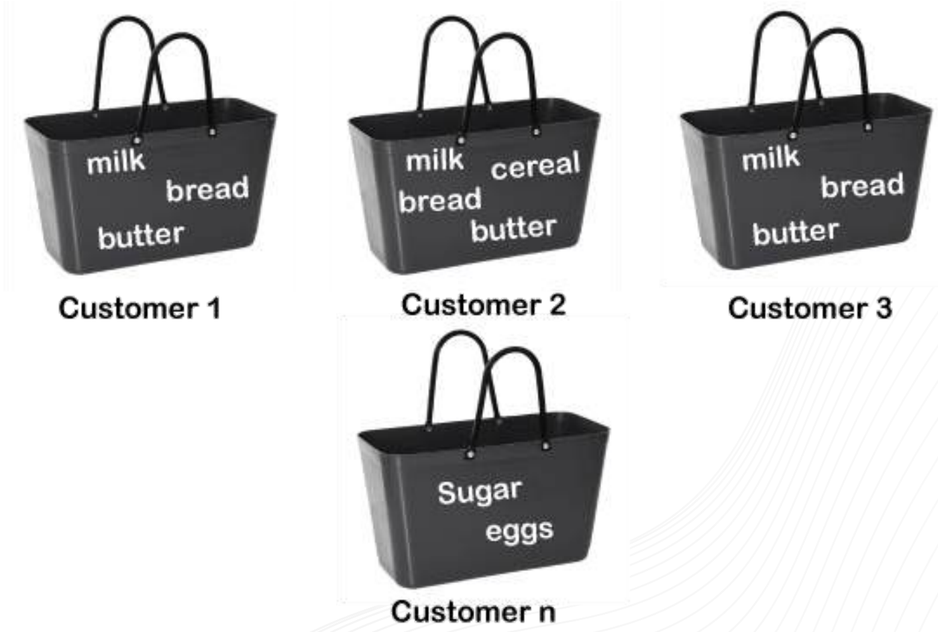
The dendrogram is a tree-like structure that is mainly used to store each step as a memory that the HC algorithm performs. In the dendrogram plot, the Y-axis shows the Euclidean distances between the data points, and the x-axis shows all the data points of the given dataset.

The working of the dendrogram can be explained using the below diagram:



ASSOCIATION RULE LEARNING

- Association rule learning is a type of unsupervised learning technique that checks for the dependency of one data item on another data item and maps accordingly so that it can be more profitable. It tries to find some interesting relations or associations among the variables of dataset. It is based on different rules to discover the interesting relations between variables in the database.
- The association rule learning is one of the very important concepts of machine learning, and it is employed in **Market Basket analysis, Web usage mining, continuous production, etc.**
- For example, if a customer buys bread, he most likely can also buy butter, eggs, or milk, so these products are stored within a shelf or mostly nearby.



ASSOCIATION RULE LEARNING

Association rule learning can be divided into three types of algorithms:

- **Apriori**
- **Eclat**
- **F-P Growth Algorithm**

HOW DOES ASSOCIATION RULE LEARNING WORK?

Association rule learning works on the concept of If and Else Statement, such as if A then B.

Here the If element is called **antecedent**, and then statement is called as **Consequent**. These types of relationships where we can find out some association or relation between two items is known as single cardinality. It is all about creating rules, and if the number of items increases, then cardinality also increases accordingly. So, to measure the associations between thousands of data items, there are several metrics:

- **Support**
- **Confidence**
- **Lift**

HOW DOES ASSOCIATION RULE LEARNING WORK?

Support

Support is the frequency of A or how frequently an item appears in the dataset. It is defined as the fraction of the transaction T that contains the itemset X. If there are X datasets, then for transactions T, it can be written as:

$$\text{Supp}(X) = \frac{Freq(X)}{T}$$

Confidence

Confidence indicates how often the rule has been found to be true. Or how often the items X and Y occur together in the dataset when the occurrence of X is already given. It is the ratio of the transaction that contains X and Y to the number of records that contain X.

$$\text{Confidence} = \frac{Freq(X,Y)}{Freq(X)}$$

HOW DOES ASSOCIATION RULE LEARNING WORK?

Lift

It is the strength of any rule, which can be defined as below formula:

$$\text{Lift}=\frac{Supp(X,Y)}{Supp(X)\times Supp(Y)}$$

It is the ratio of the observed support measure and expected support if X and Y are independent of each other. It has **three** possible values:

- **If Lift= 1:** The probability of occurrence of antecedent and consequent is independent of each other.
- **Lift>1:** It determines the degree to which the two itemsets are dependent to each other.
- **Lift<1:** It tells us that one item is a substitute for other items, which means one item has a negative effect on another.

TYPES OF ASSOCIATION RULE LERNING

Apriori Algorithm

This algorithm uses frequent datasets to generate association rules. It is designed to work on the databases that contain transactions. This algorithm uses **a breadth-first search and Hash Tree** to calculate the itemset efficiently.
It is mainly used for **market basket analysis** and helps to understand the products that can be bought together. It can also be used in the healthcare field to find drug reactions for patients.

Eclat Algorithm

Eclat algorithm stands for **Equivalence Class Transformation.** This algorithm uses **a depth-first search** technique to find frequent itemsets in a transaction database. It performs faster execution than Apriori Algorithm.

F-P Growth Algorithm

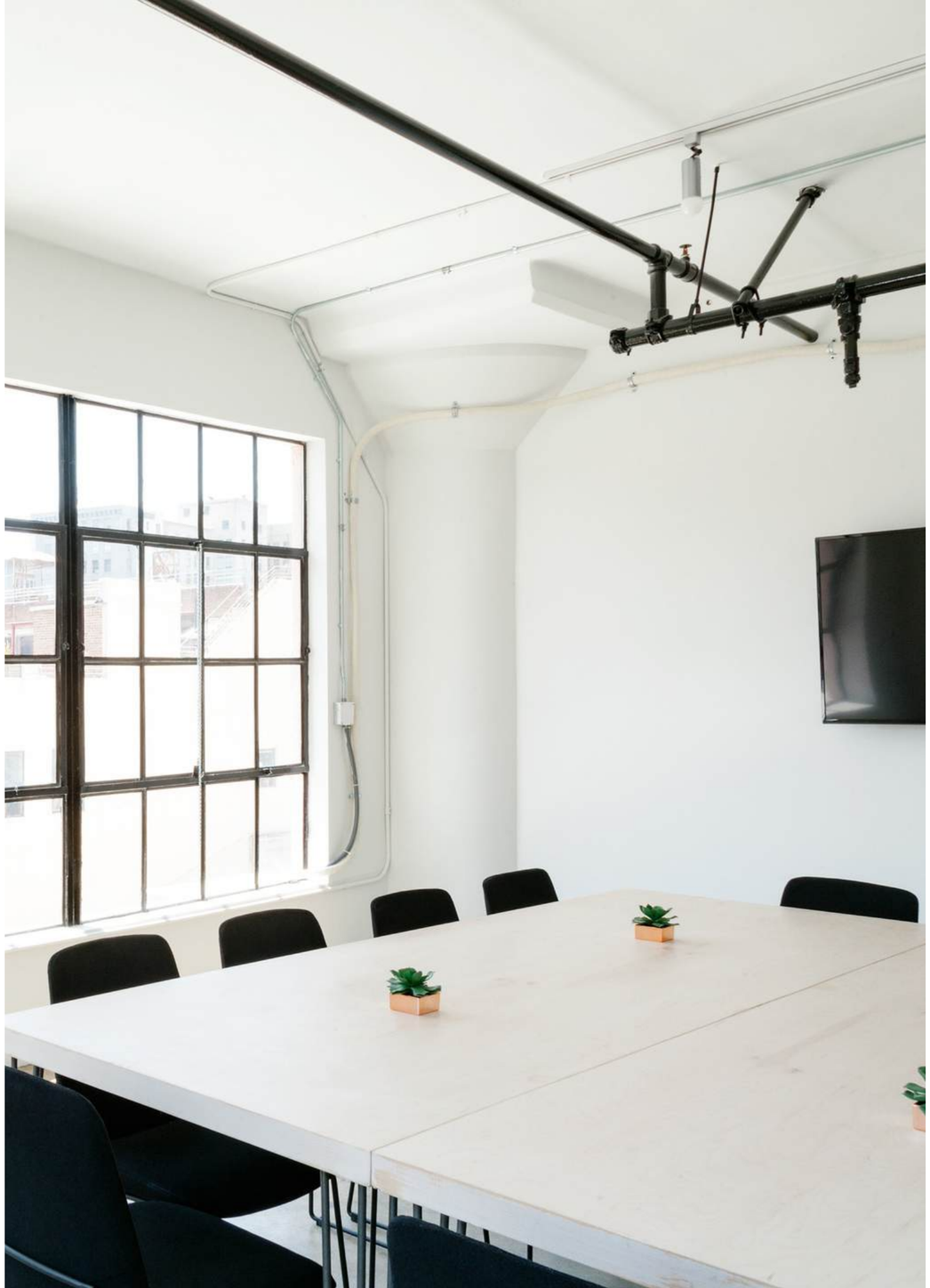
The F-P growth algorithm stands for **Frequent Pattern**, and it is the ***improved version of the Apriori Algorithm***. It represents the database in the form of a tree structure that is known as a frequent pattern or tree. The purpose of this frequent tree is to extract the most frequent patterns.

APPLICATIONS OF ASSOCIATION RULE LEARNING

- **Market Basket Analysis:** It is one of the popular examples and applications of association rule mining. This technique is commonly used by big retailers to determine the association between items.
- **Medical Diagnosis:** With the help of association rules, patients can be cured easily, as it helps in identifying the probability of illness for a particular disease.
- **Protein Sequence:** The association rules help in determining the synthesis of artificial Proteins.
- It is also used for the **Catalog Design** and **Loss-leader Analysis** and many more other applications.

ANY QUESTIONS?

Feel free to ask



MACHINE LEARNING PROGRAM

THANK YOU

UPCOMING NEXT WEEK : SESSION (5)