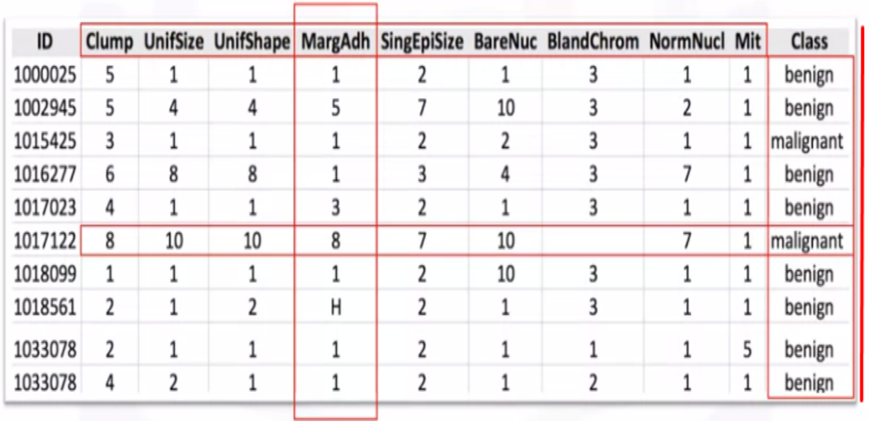
# Unsupervised learning

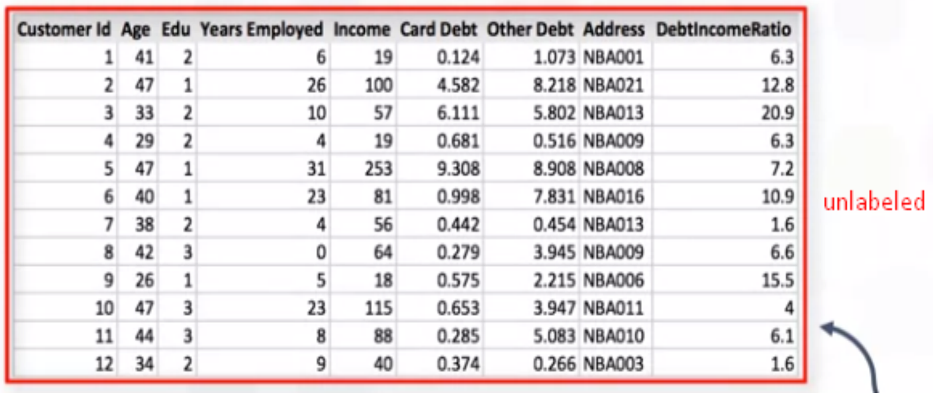
# Unsupervised learning

- Unsupervised learning, also known as unsupervised machine learning, uses machine learning algorithms to analyze and cluster unlabelled datasets.



| ID | Clump | UnifSize | UnifShape | MargAdh | SingEpiSize | BareNuc | BlandChrom | NormNucl | Mit | Class |
|---|---|---|---|---|---|---|---|---|---|---|
| 1000025 | 5 | 1 | 1 | 1 | 2 | 1 | 3 | 1 | 1 | benign |
| 1002945 | 5 | 4 | 4 | 5 | 7 | 10 | 3 | 2 | 1 | benign |
| 1015425 | 3 | 1 | 1 | 1 | 2 | 2 | 3 | 1 | 1 | malignant |
| 1016277 | 6 | 8 | 8 | 1 | 3 | 4 | 3 | 7 | 1 | benign |
| 1017023 | 4 | 1 | 1 | 3 | 2 | 1 | 3 | 1 | 1 | benign |
| 1017122 | 8 | 10 | 10 | 8 | 7 | 10 | | 7 | 1 | malignant |
| 1018099 | 1 | 1 | 1 | 1 | 2 | 10 | 3 | 1 | 1 | benign |
| 1018561 | 2 | 1 | 2 | H | 2 | 1 | 3 | 1 | 1 | benign |
| 1033078 | 2 | 1 | 1 | 1 | 2 | 1 | 1 | 1 | 5 | benign |
| 1033078 | 4 | 2 | 1 | 1 | 2 | 1 | 2 | 1 | 1 | benign |

labels



| Customer Id | Age | Edu | Years Employed | Income | Card Debt | Other Debt | Address | DebtIncomeRatio |
|---|---|---|---|---|---|---|---|---|
| 1 | 41 | 2 | 6 | 19 | 0.124 | 1.073 | NBA001 | 6.3 |
| 2 | 47 | 1 | 26 | 100 | 4.582 | 8.218 | NBA021 | 12.8 |
| 3 | 33 | 2 | 10 | 57 | 6.111 | 5.802 | NBA013 | 20.9 |
| 4 | 29 | 2 | 4 | 19 | 0.681 | 0.516 | NBA009 | 6.3 |
| 5 | 47 | 1 | 31 | 253 | 9.308 | 8.908 | NBA008 | 7.2 |
| 6 | 40 | 1 | 23 | 81 | 0.998 | 7.831 | NBA016 | 10.9 |
| 7 | 38 | 2 | 4 | 56 | 0.442 | 0.454 | NBA013 | 1.6 |
| 8 | 42 | 3 | 0 | 64 | 0.279 | 3.945 | NBA009 | 6.6 |
| 9 | 26 | 1 | 5 | 18 | 0.575 | 2.215 | NBA006 | 15.5 |
| 10 | 47 | 3 | 23 | 115 | 0.653 | 3.947 | NBA011 | 4 |
| 11 | 44 | 3 | 8 | 88 | 0.285 | 5.083 | NBA010 | 6.1 |
| 12 | 34 | 2 | 9 | 40 | 0.374 | 0.266 | NBA003 | 1.6 |

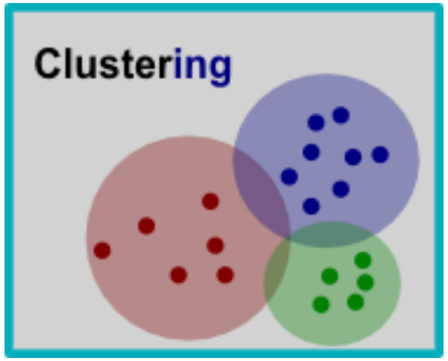
unlabeled

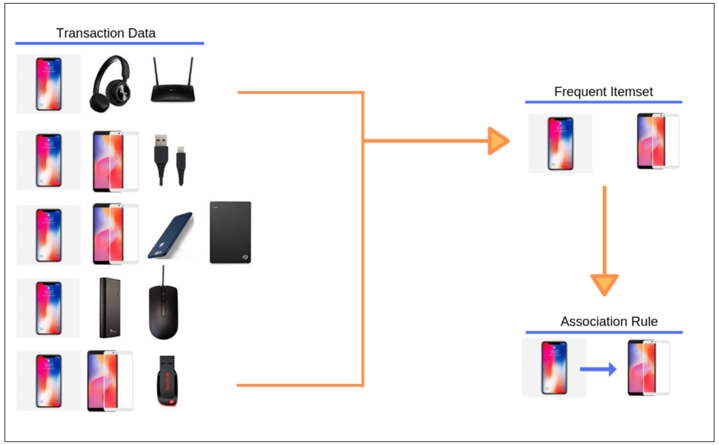
- These algorithms discover hidden patterns or data groupings without the need for human intervention.
- Its ability to discover similarities and differences in information make it the ideal solution for exploratory data analysis, cross-selling strategies, customer segmentation, and image recognition.

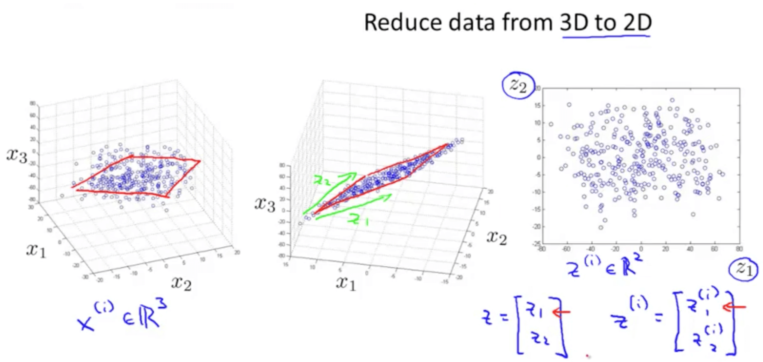
# Common unsupervised learning approaches

Unsupervised learning models are utilized for three main tasks—

- Clustering
- Association
- Dimensionality reduction







# Clustering

## Clustering

- Clustering is a data mining technique which groups unlabelled data based on their similarities or differences. Clustering algorithms are used to process raw, unclassified data objects into groups represented by structures or patterns in the information.
- Clustering algorithms can be categorized into a few types, specifically exclusive, overlapping, hierarchical, and probabilistic.

## Exclusive and Overlapping Clustering

- Exclusive clustering is a form of grouping that stipulates a data point can exist only in one cluster.
- This can also be referred to as “hard” clustering. The K-means clustering algorithm is an example of exclusive clustering.

# Clustering

## K-means clustering

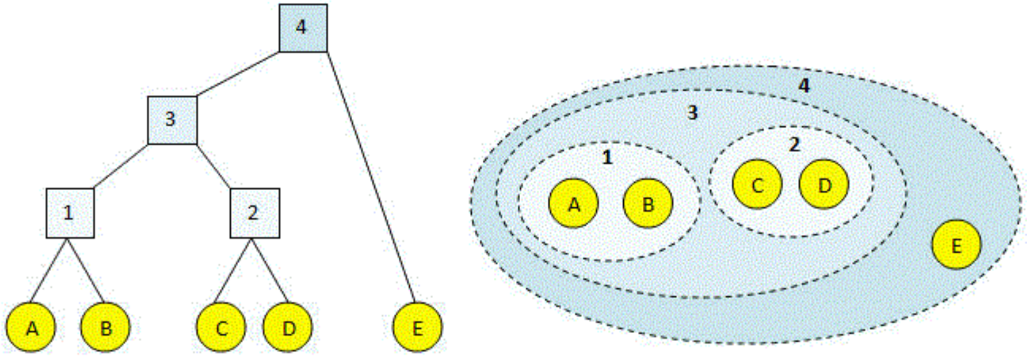
- It is a common example of an exclusive clustering method where data points are assigned into K groups, where K represents the number of clusters based on the distance from each group's centroid.
- The data points closest to a given centroid will be clustered under the same category.
- A larger K value will be indicative of smaller groupings with more granularity whereas a smaller K value will have larger groupings and less granularity.
- K-means clustering is commonly used in market segmentation, document clustering, image segmentation, and image compression.

Overlapping clusters differs from exclusive clustering in that it allows data points to belong to multiple clusters with separate degrees of membership.
"Soft" or fuzzy k-means clustering is an example of overlapping clustering.

# Clustering

## Hierarchical clustering

- Hierarchical clustering, also known as hierarchical cluster analysis (HCA), is an unsupervised clustering algorithm that can be categorized in two ways; they can be agglomerative or divisive.
- Agglomerative clustering is considered a “bottoms-up approach.” Its data points are isolated as separate groupings initially, and then they are merged together iteratively on the basis of similarity until one cluster has been achieved.



# Clustering

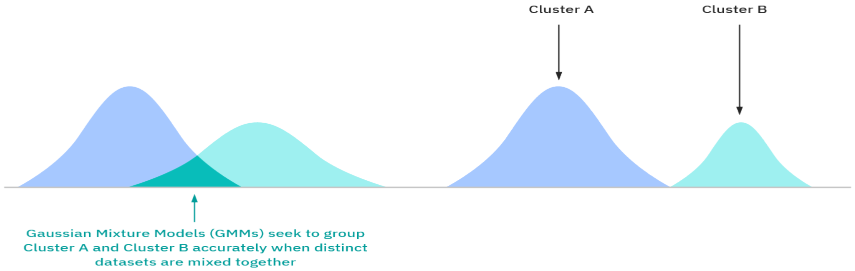
## Probabilistic clustering

- A probabilistic model is an unsupervised technique that helps us solve density estimation or "soft" clustering problems.
- In probabilistic clustering, data points are clustered based on the likelihood that they belong to a particular distribution.
- The Gaussian Mixture Model (GMM) is the one of the most commonly used probabilistic clustering methods.

## Gaussian Mixture Models

These are classified as mixture models, which means that they are made up of an unspecified number of probability distribution functions.

GMMs are primarily leveraged to determine which Gaussian, or normal, probability distribution a given data point belongs to. If the mean or variance are known, then we can determine which distribution a given data point belongs to. However, in GMMs, these variables are not known, so we assume that a latent, or hidden, variable exists to cluster data points appropriately.


Cluster A
Cluster B
Gaussian Mixture Models (GMMs) seek to group Cluster A and Cluster B accurately when distinct datasets are mixed together

# Association Rules

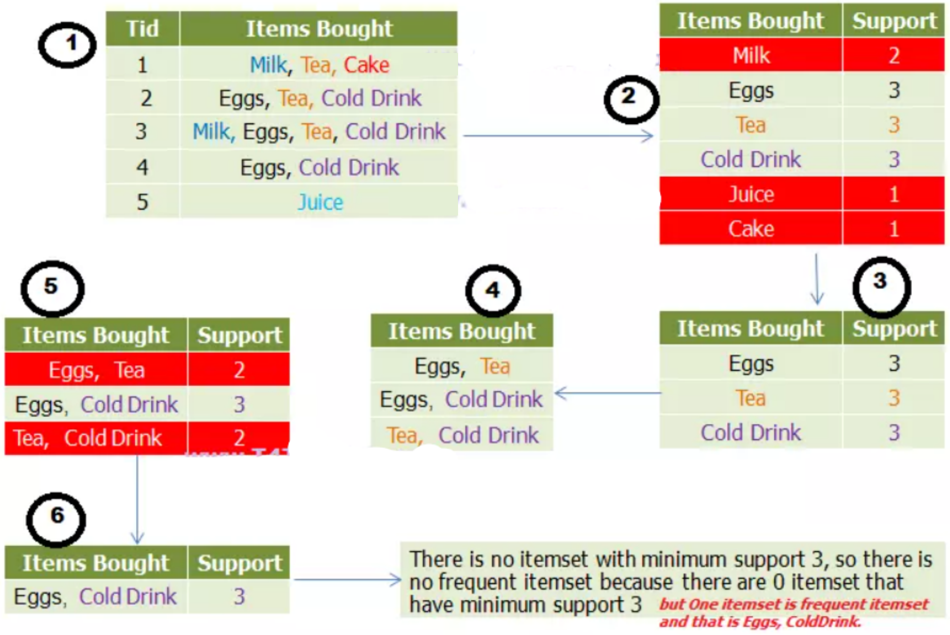
## Association Rules

- An association rule is a rule-based method for finding relationships between variables in a given dataset.
- These methods are frequently used for market basket analysis, allowing companies to better understand relationships between different products.
- Understanding consumption habits of customers enables businesses to develop better cross-selling strategies and recommendation engines.
- Examples of this can be seen in Amazon’s “Customers Who Bought This Item Also Bought” or Spotify’s "Discover Weekly" playlist. While there are a few different algorithms used to generate association rules, such as Apriori, Eclat, and FP-Growth, the Apriori algorithm is most widely used.

# Association Rules

## Apriori algorithms

- Apriori algorithms have been popularized through market basket analyses, leading to different recommendation engines for music platforms and online retailers.
- They are used within transactional datasets to identify frequent itemsets, or collections of items, to identify the likelihood of consuming a product given the consumption of another product.



1

| Tid | Items Bought |
|---|---|
| 1 | Milk, Tea, Cake |
| 2 | Eggs, Tea, Cold Drink |
| 3 | Milk, Eggs, Tea, Cold Drink |
| 4 | Eggs, Cold Drink |
| 5 | Juice |

2

| Items Bought | Support |
|---|---|
| Milk | 2 |
| Eggs | 3 |
| Tea | 3 |
| Cold Drink | 3 |
| Juice | 1 |
| Cake | 1 |

3

| Items Bought | Support |
|---|---|
| Eggs | 3 |
| Tea | 3 |
| Cold Drink | 3 |

4

| Items Bought |
|---|
| Eggs, Tea |
| Eggs, Cold Drink |
| Tea, Cold Drink |

5

| Items Bought | Support |
|---|---|
| Eggs, Tea | 2 |
| Eggs, Cold Drink | 3 |
| Tea, Cold Drink | 2 |

6

| Items Bought | Support |
|---|---|
| Eggs, Cold Drink | 3 |

There is no itemset with minimum support 3, so there is no frequent itemset because there are 0 itemset that have minimum support 3 *but One itemset is frequent itemset and that is Eggs, ColdDrink.*

# Dimensionality Reduction

## Dimensionality reduction

- While more data generally yields more accurate results, it can also impact the performance of machine learning algorithms (e.g. overfitting) and it can also make it difficult to visualize datasets.
- Dimensionality reduction is a technique used when the number of features, or dimensions, in a given dataset is too high.
- It reduces the number of data inputs to a manageable size while also preserving the integrity of the dataset as much as possible.
- It is commonly used in the preprocessing data stage.

# Dimensionality Reduction

## Principal component analysis

- Principal component analysis (PCA) is a type of dimensionality reduction algorithm which is used to reduce redundancies and to compress datasets through feature extraction.
- This method uses a linear transformation to create a new data representation, yielding a set of "principal components." The first principal component is the direction which maximizes the variance of the dataset.
- While the second principal component also finds the maximum variance in the data, it is completely uncorrelated to the first principal component, yielding a direction that is perpendicular, or orthogonal, to the first component.
- This process repeats based on the number of dimensions, where a next principal component is the direction orthogonal to the prior components with the most variance.

# Application of Clustering

Clustering is widely used in many industries. Below are some commonly known applications of clustering technique in Machine Learning:

- **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.

# Clustering Metrics

Correctly measuring the performance of Clustering algorithms is key. This is especially true as it often happens that clusters are manually and qualitatively inspected to determine whether the results are meaningful.

## Silhouette Score

- The Silhouette Score and Silhouette Plot are used to measure the separation distance between clusters.
- It displays a measure of **how close each point in a cluster is to points in the neighbouring clusters.**
- This measure has a range of [-1, 1] and is a great tool to visually inspect the similarities within clusters and differences across clusters.
- The Silhouette Score is calculated using the mean intra-cluster distance (i) and the mean nearest-cluster distance (n) for each sample.
- The Silhouette Coefficient for a sample is $\frac{(n - i)}{\max(i, n)}$
- n is the distance between each sample and the nearest cluster that the sample is not a part of while i is the mean distance within each cluster.

# Clustering Metrics

- The higher the Silhouette Coefficients (the closer to +1), the further away the cluster’s samples are from the neighbouring clusters samples.
- A value of 0 indicates that the sample is on or very close to the decision boundary between two neighbouring clusters. Negative values, instead, indicate that those samples might have been assigned to the wrong cluster.

## Rand Index

Another commonly used metric is the Rand Index. It computes a similarity measure between two clusters by considering all pairs of samples and counting pairs that are assigned in the same or different clusters in the predicted and true clustering.

The formula of the Rand Index is:

$$RI = \frac{\text{Number of Agreeing Pairs}}{\text{Number of Pairs}}$$

The RI can range from zero to 1, a perfect match.

# Clustering Metrics

## Adjusted Rand Index

The Rand Index computes a similarity measure between two clusterings by considering all pairs of samples and counting pairs that are assigned in the same or different clusters in the predicted and true clusterings.

The raw RI score is then “adjusted for chance” into the ARI score using the following scheme:

$$ARI = \frac{\text{RI - Expected RI}}{\text{Max(RI) - Expected RI}}$$

The Adjusted Rand Index, similarly to RI, ranges from zero to one, with zero equating to random labelling and one when the clusters are identical.

# Clustering Metrics

## Davies-Bouldin Index

The Davies-Bouldin Index is defined as the average similarity measure of each cluster with its most similar cluster. Similarity is the ratio of within-cluster distances to between-cluster distances. In this way, clusters which are farther apart and less dispersed will lead to a better score.

The minimum score is zero, and differently from most performance metrics, the lower values the better clustering performance.

Similarly to the Silhouette Score, the D-B Index does not require the a-priori knowledge of the ground-truth labels, but has a simpler implementation in terms of formulation than Silhouette Score.

# K Means Clustering

# K-Means Algorithm

## K-Means Clustering Algorithm

- K-Means Clustering is an unsupervised learning algorithm that is used to solve the clustering problems in machine learning or data science.
- In K-Means algorithm we group the unlabelled 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 unlabelled dataset into k different clusters in such a way that each dataset belongs only one group that has similar properties.
- 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.

# K-Means Algorithm

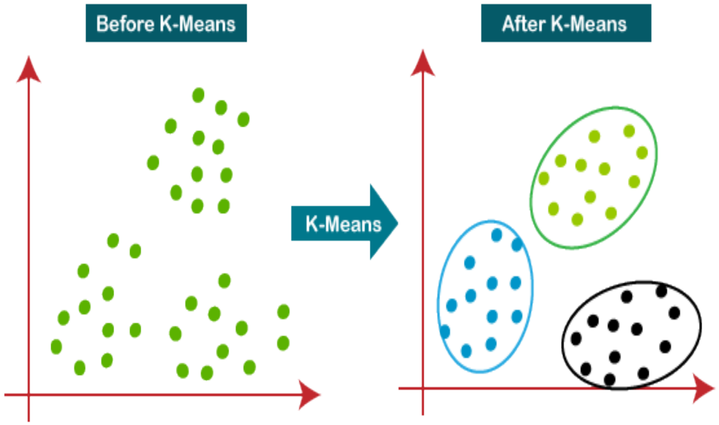
## K-Means Clustering Algorithm

- The algorithm takes the unlabelled 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.

Hence each cluster has datapoints with some commonalities, and it is away from other clusters.



# Working of K-Means Algorithm

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. (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.

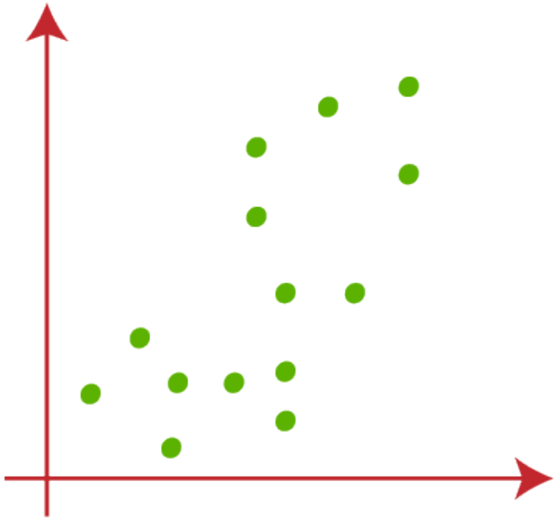
**Step-5:** Repeat the third steps, which means reassign each datapoint to the new closest centroid of each cluster.

**Step-6:** If any reassignment occurs, then go to step-4 else go to finish.

**Step-7:** The model is ready.

# Working of K-Means Algorithm with Example

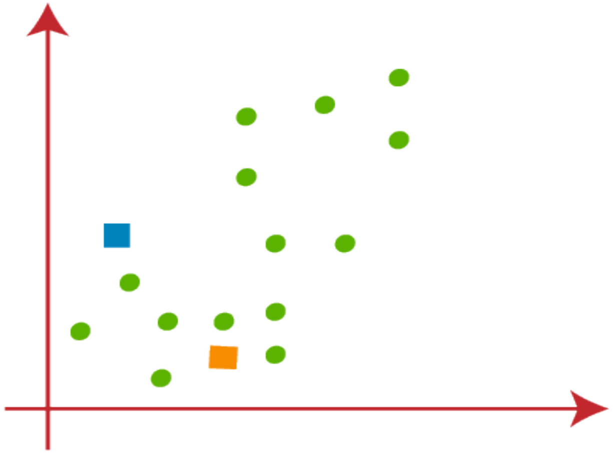
Suppose we have two variables M1 and M2. The x-y axis scatter plot of these two variables is given below:



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.

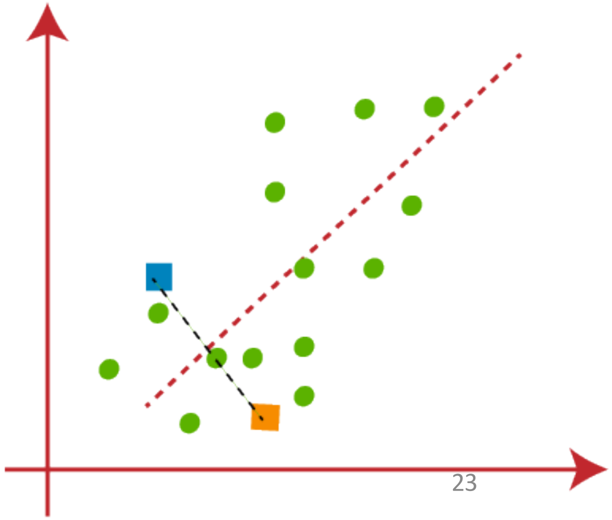
# Working of K-Means Algorithm with Example

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 two points as k points, which are not the part of our dataset.



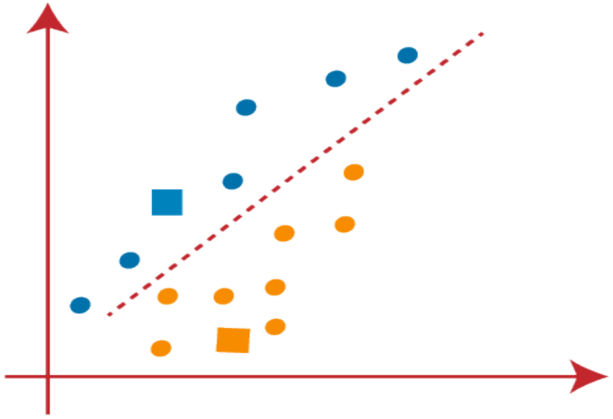
Now we will assign each data point of the scatter plot to its closest K-point or centroid. We will compute it by applying some mathematics that we have studied to calculate the distance between two points.

So, we will draw a median between both the centroids.

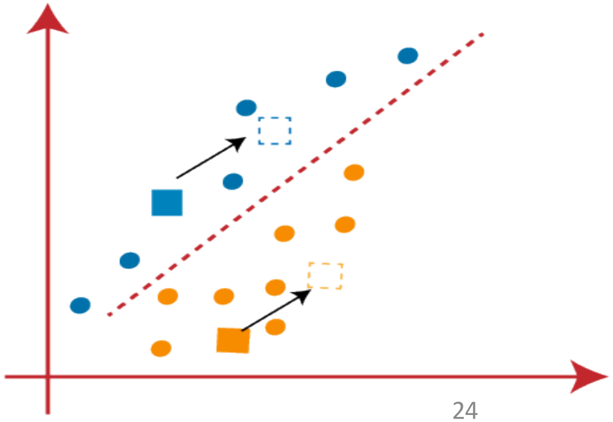


# Working of K-Means Algorithm with Example

From the above steps, 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.

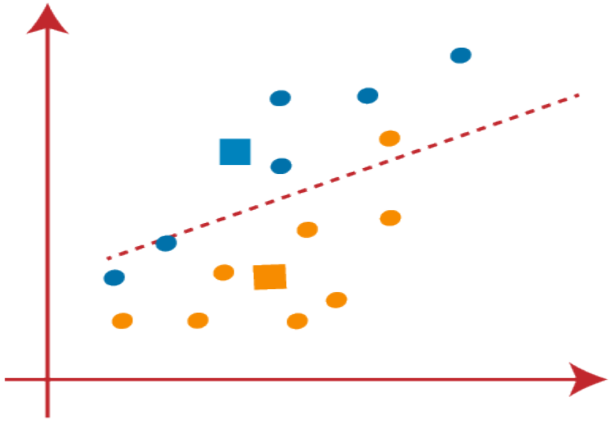


As we need to find the closest cluster, so we will repeat the process by choosing a new centroid. To choose the new centroids, we will compute the center of gravity of these centroids, and will find new centroids as below:

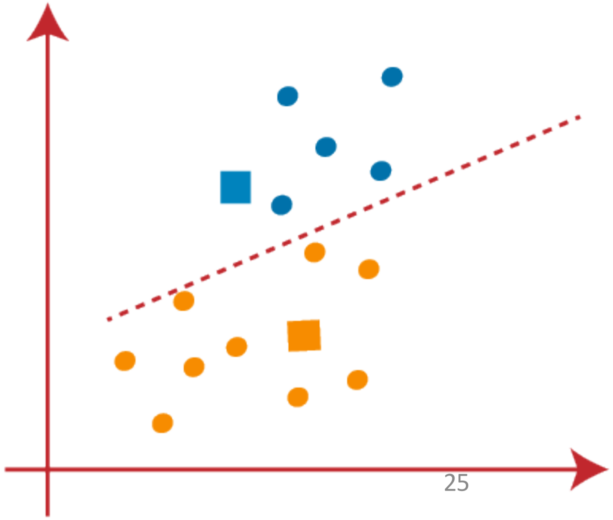


# Working of K-Means Algorithm with Example

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 adjacent image:



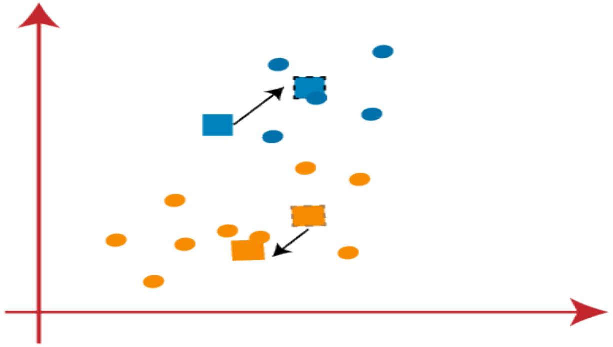
From the previous step, we can see, one yellow point is on the left side of the line, and two blue points are right to the line. So, these points will be assigned to new centroids.



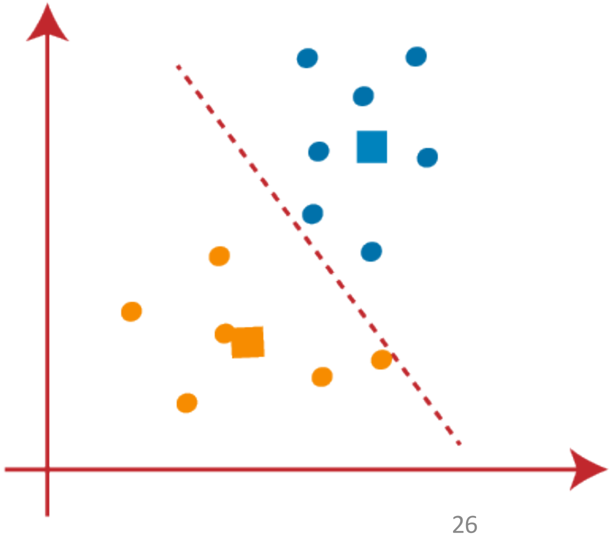
# Working of K-Means Algorithm with Example

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 adjacent image:

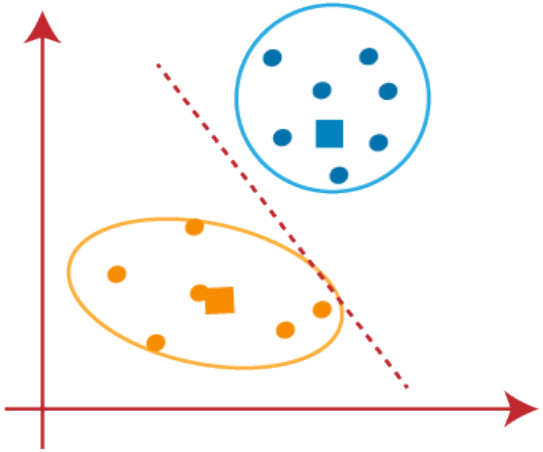


As we got the new centroids so again will draw the median line and reassign the data points. So, the image will be:

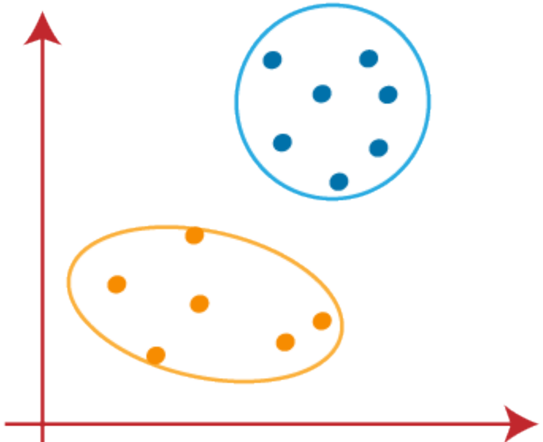


# Working of K-Means Algorithm with Example

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



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



# How to choose the value of K

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.

## 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.

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

# Steps to choose the value of K

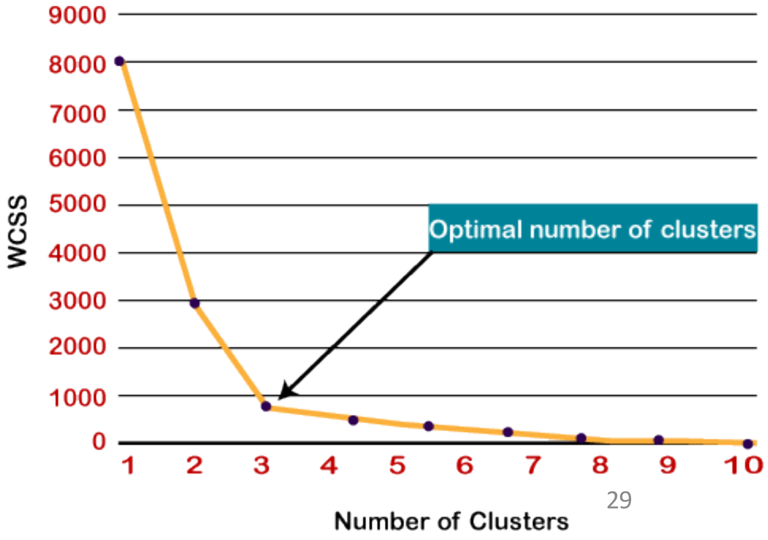
To find the optimal value of clusters, the elbow method follows the below steps:

- 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.

The graph for the elbow method looks like the adjacent image:



# Important Hyperparameters

**n_clusters : int, default=8**

The number of clusters to form as well as the number of centroids to generate.

**Init {‘k-means++’, ‘random’}**

‘k-means++’ : selects initial cluster centroids using sampling based on an empirical probability distribution of the points’ contribution to the overall inertia.

‘random’: choose n_clusters observations (rows) at random from data for the initial centroids.

**n_init :** int, default=10

Number of time the k-means algorithm will be run with different centroid seeds.

# Hierarchical Clustering: Agglomerative

# Hierarchical Clustering

Hierarchical clustering is another 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**.

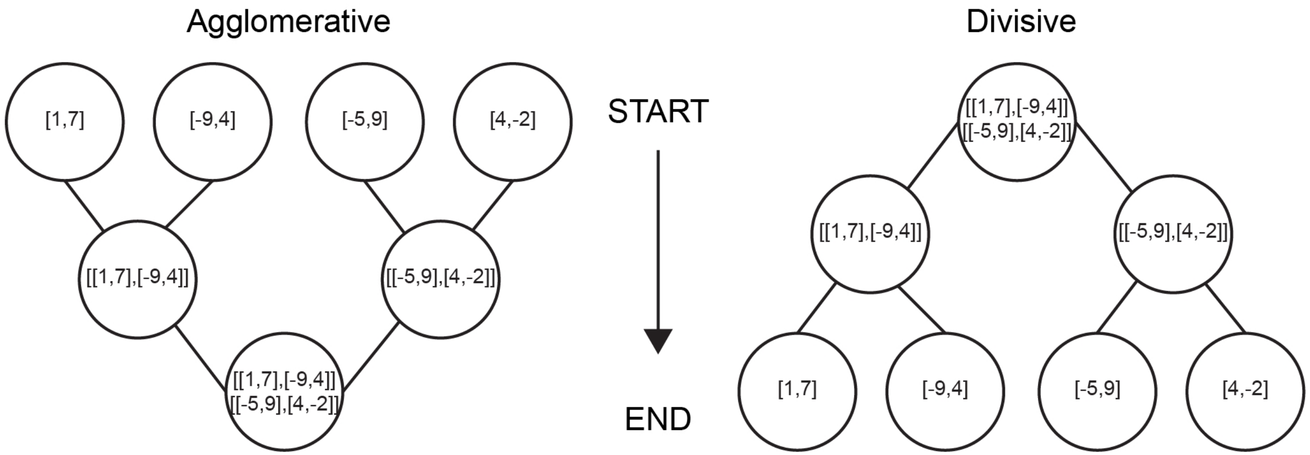
Sometimes the results of K-means clustering and hierarchical clustering may look similar, but they both differ depending on how they work. As there is no requirement to predetermine the number of clusters as we did in the K-Means algorithm.

# Hierarchical Clustering

The hierarchical clustering technique has two approaches:

**1.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.

**2.Divisive:** Divisive algorithm is the reverse of the agglomerative algorithm as it is a **top-down approach.**



# Hierarchical Clustering

## Why hierarchical clustering?

As we already have other clustering algorithms such as K-Means Clustering, then why we need hierarchical clustering?

So, as we have seen in the K-means clustering that there are some challenges with this algorithm, which are a predetermined number of clusters, and it always tries to create the clusters of the same size.

To solve these two challenges, we can opt for the hierarchical clustering algorithm because, in this algorithm, we don't need to have knowledge about the predefined number of clusters.

# 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.

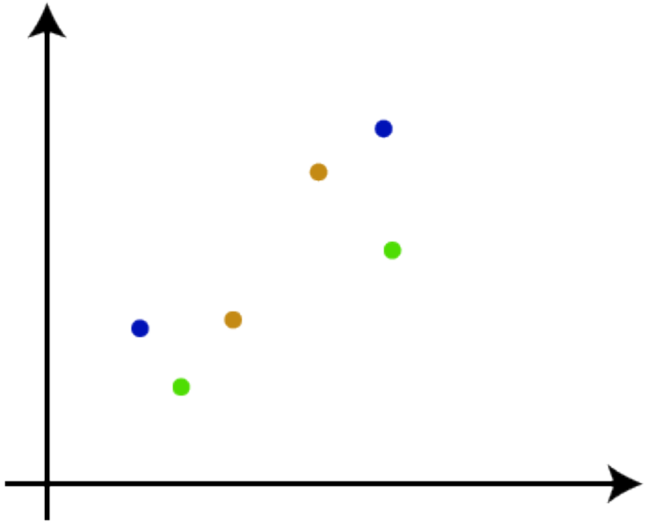
It does this until all the clusters are merged into a single cluster that contains all the datasets.

This hierarchy of clusters is represented in the form of the dendrogram.

# Steps of Agglomerative Hierarchical clustering

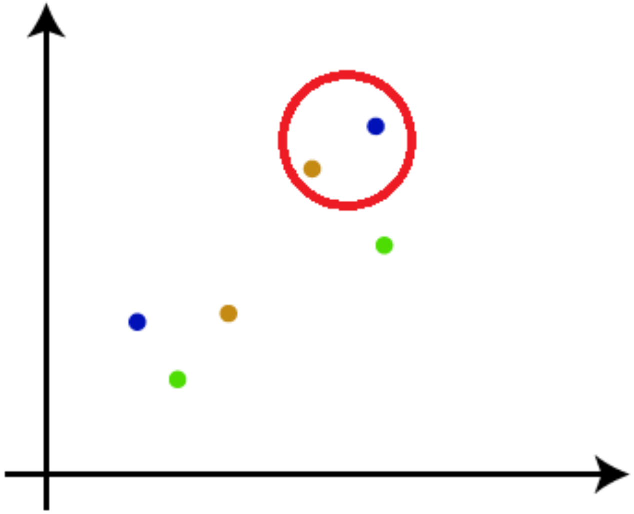
The working of the AHC algorithm can be explained using the below steps:

**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.



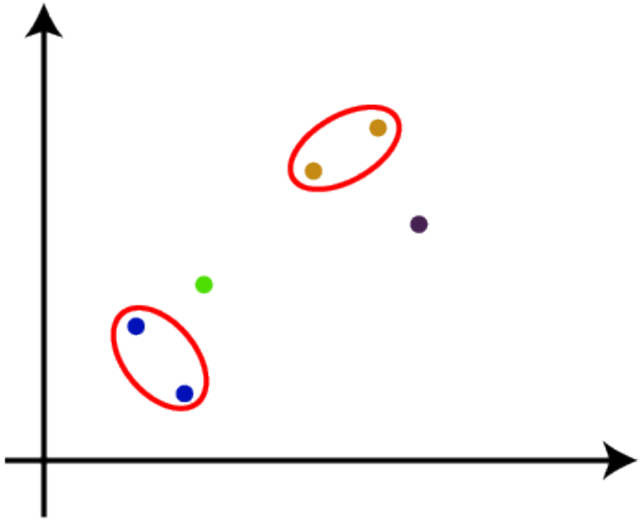
# Steps of Agglomerative Hierarchical clustering

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



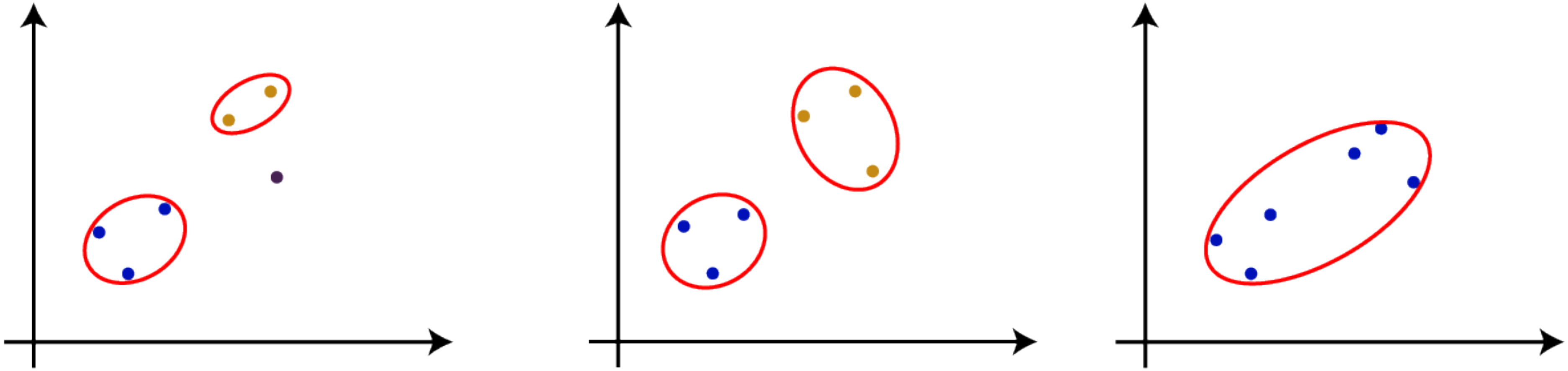
# Steps of Agglomerative Hierarchical clustering

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



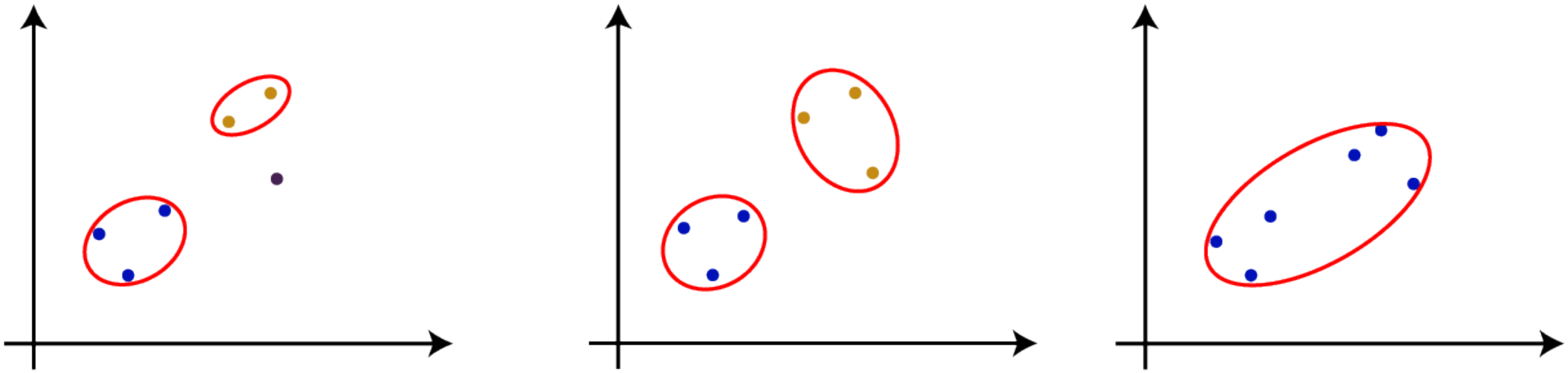
# Steps of Agglomerative Hierarchical clustering

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



# Steps of Agglomerative Hierarchical clustering

**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.

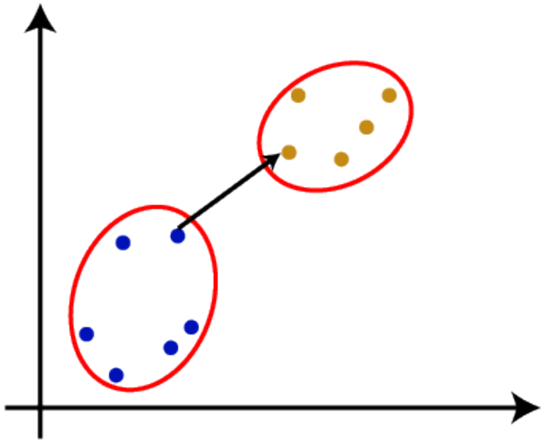
# Measure for the distance between two clusters

As we have seen, 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.**

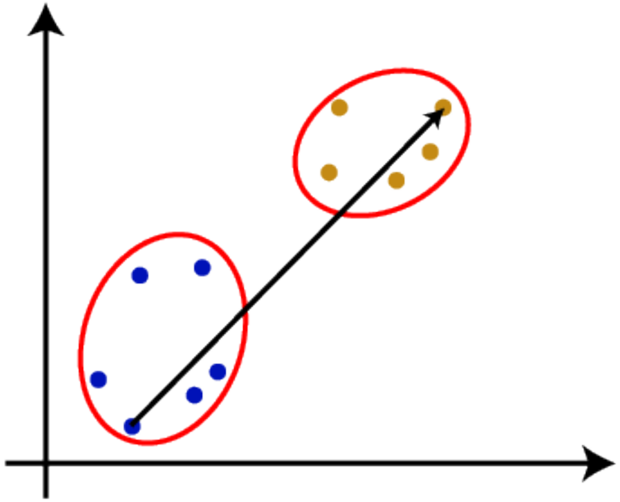
Some of the popular linkage methods are :

**Single Linkage:** It is the Shortest Distance between the closest points of the clusters. Consider the below image:



# Measure for the distance between two clusters

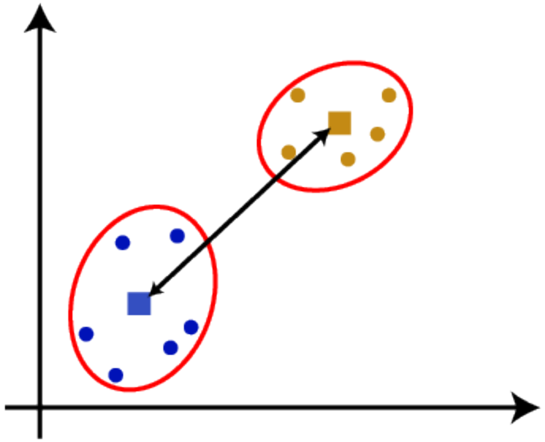
**Complete Linkage:** It is the farthest distance between the two points of two different clusters. It is one of the popular linkage methods as it forms tighter clusters than single-linkage.



# Measure for the distance between two clusters

**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.

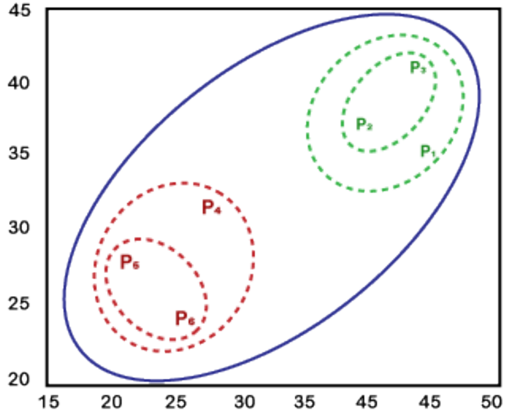
**Centroid Linkage:** It is the linkage method in which the distance between the centroid of the clusters is calculated. Consider the below image:

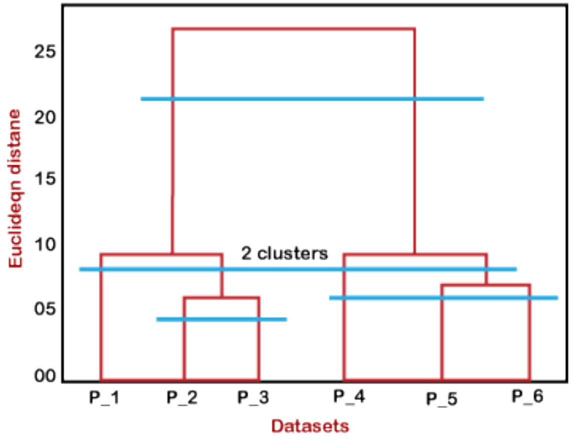


# Woking of Dendrogram in Hierarchical clustering

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.

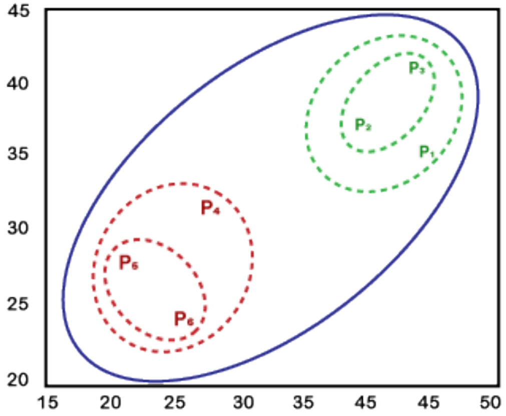
The working of the dendrogram can be explained using the below diagram:

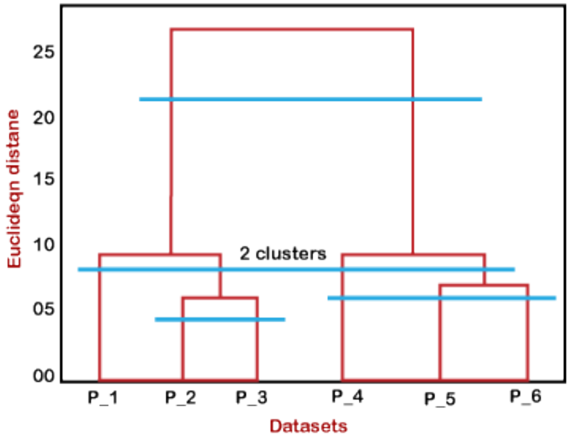




# Woking of Dendrogram in Hierarchical clustering

- As we have discussed, firstly, the datapoints P2 and P3 combine together and form a cluster, correspondingly a dendrogram is created, which connects P2 and P3 with a rectangular shape. The height is decided according to the Euclidean distance between the data points.
- In the next step, P5 and P6 form a cluster, and the corresponding dendrogram is created. It is higher than of previous, as the Euclidean distance between P5 and P6 is a little bit greater than the P2 and P3.
- Again, two new dendrograms are created that combine P1, P2, and P3 in one dendrogram, and P4, P5, and P6, in another dendrogram.
- At last, the final dendrogram is created that combines all the data points together.





**We can cut the dendrogram tree structure at any level as per our requirement.**