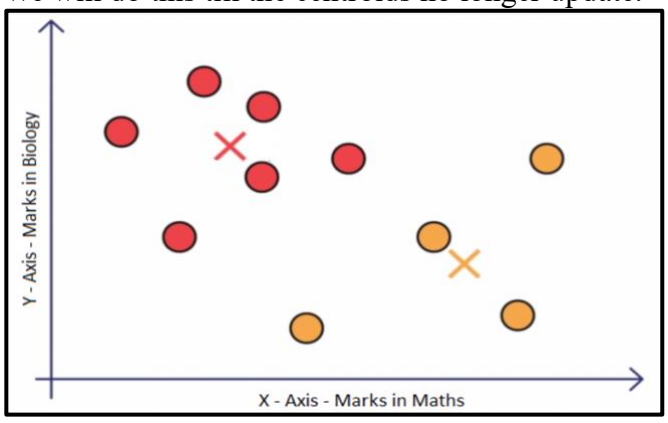
Unsupervised Machine Learning Algorithm

- In **supervised ML** we have a dependent feature.
- In **unsupervised ML** we don't have any pre-defined labels.
    - We group objects that are similar to each other

There are two types of UMLA:

- **K-Means**
    - Let's say we have N=10 points. We need to divide it into 2 clusters. i.e., k=2
    - We will pick 2 random cluster centers. This choice is completely random.
    - We will allocate each point in the data set to nearest cluster center. We calculate this by using ED and allocate the point to the centroid with least distance.
    - This will be done for every data point and we have set points allocated to each cluster.
    - Now we will re-compute the center of each of these clusters which will be the mean of individual points of each of these clusters and this will give us new cluster center.
    - We will do this till the centroids no longer update.



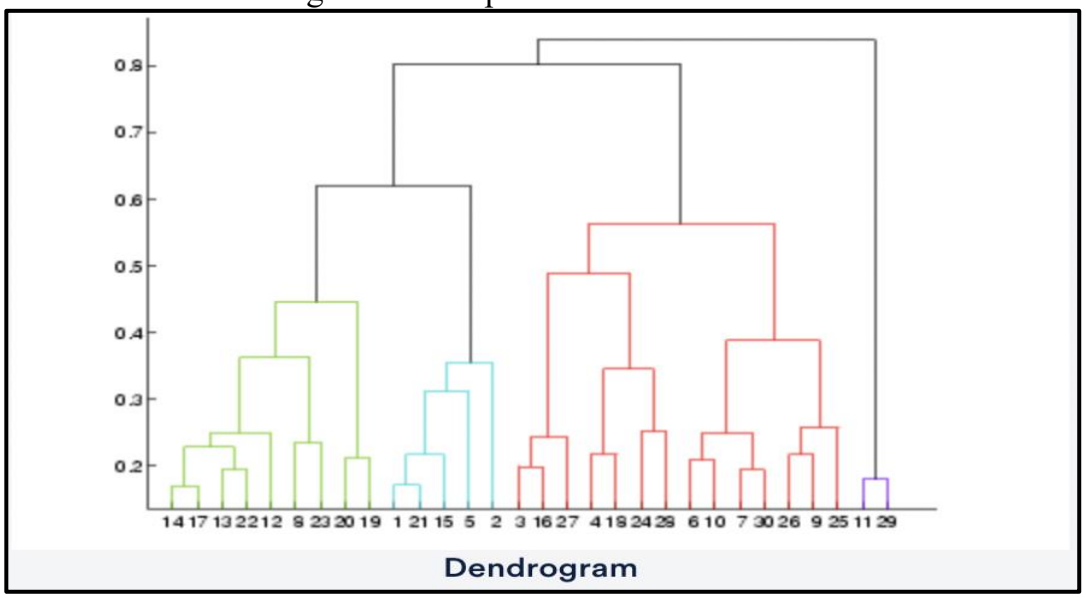
    - K-Means algorithm does not work with categorical data.

- **Deciding the number of clusters**
    - **Elbow Curve**
        - We create multiple clusters (k=2,3,4,5 etc.) and then we see the amount of information explained in each of those.
        - **SSD**-The sum of squared distances of samples to their closest cluster center.
        - At the elbow drop in SSD is significant and as we go on increasing the no. of clusters we reach a saturation point where there is no more change in SSD value.

    - **Silhouette Metric/Analysis**
        - **a(i)** - It is the average distance from own cluster (as small as possible) (**cohesion**) i.e., **points within the cluster should be close to each other**
        - **b(i)** - It is the average distance from the nearest neighbor cluster.(**as large as possible**) (Separation)
        - s(i)=a(i)-b(i)/max(b(i),a(i))
        - Max value of s(i) can be 1

Hierarchical Clustering Algorithm

- We do not decide the value of K in HC.
- We use dendrogram to interpret our results.



Process:

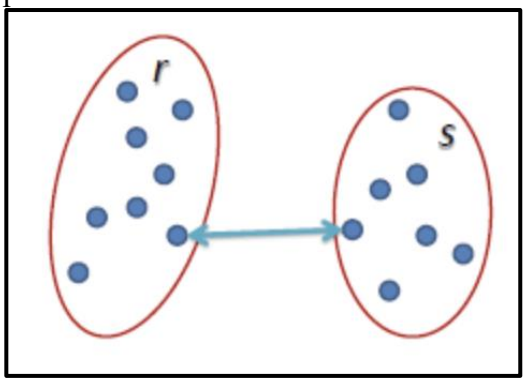
- Individual data points are considered N individual clusters. E.g. 10
- The 2 clusters that are closest to each other are then fused together to form a cluster. This is done using ED.
- Now we have N-1 clusters. 9 clusters.
- Now in order to calculate the distance b/w a group of clusters and an individual data point we use the concept of '**linkage**'.
- Suppose points 5 and 7 are in a cluster and we need to calculate the distance of this set with point 8.
- Now we would do this by calculating distance b/w 5 and 8 & 7 and 8. And then take minimum of these 2 distances as a measure of dissimilarity.
- Now we have N-2 clusters.
- Algorithm continues till all the points are fused together.
- This will form a dendrogram.

Dendrogram Interpretation

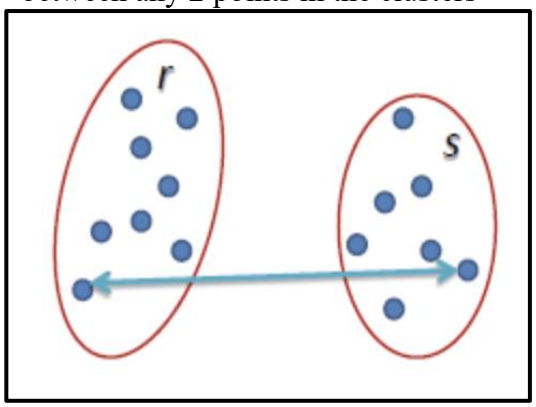
- We need to decide the height at which we will cut dendrogram.
- **Bottom Up Approach** is called-agglomerative clustering. Starting with N clusters and reaching a point where we have only 1.
- **Top Down Approach** is called-**divisive clustering**. Starting with 1 big cluster and reaching N clusters.

Type of linkages

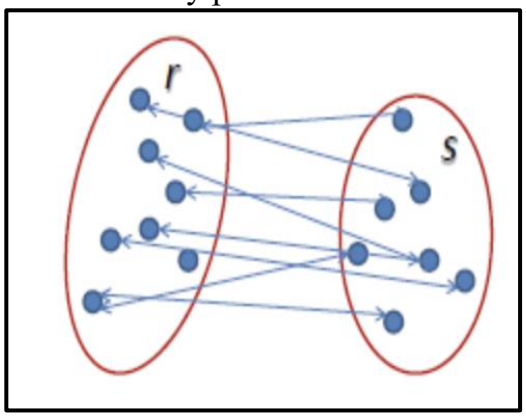
- **Single Linkage**-the distance between 2 clusters is defined as the shortest distance between points in the two clusters.



- **Complete Linkage**-the distance between 2 clusters is defined as the maximum distance between any 2 points in the clusters



- **Average linkage:** the distance between 2 clusters is defined as the average distance between every point of one cluster to every other point of the other cluster.



When to use which one?

- Whenever there is big data set we should use K-means and for small data we should use HC this is true but there is a deeper understanding to it.
- In HC, we are combining elements and step by step it starts building but at every step it becomes computationally ram heavy.
- K-Means biggest challenge is to specify the value of k.