# CLUSTERING

* It is a form of unsupervised learning
* Putting similar unlabelled data into groups
* Dimensionality reduction is a technique where we reduce the number of variables being used
* One of the most popular clustering algorithms is KMeans Algorithm

## KMEANS

### STEPS

1. Specify the number of clusters ‘K’
2. Randomly select ‘K’ datapoints and attach them to a cluster
3. Compute cluster centroids
4. Find sum of squared distance and compute centroids ,then assign datapoints to the closest cluster
5. Recompute the centroids

### PRECAUTIONS

* Standardize the data
* Use different initializations of centroids as Kmeans can stick in a local optimum rather than the global optimum

### ADVANTAGES

* Easy to understand and implement
* Faster than hierarchial clustering
* Tighter clusters
* A single instance can change the cluster

### DISADVANTAGES

* Tough to predict the optimum value of k
* Output depends strongly on k
* Output depends on the order of data
* Sensitive to rescaling
* Not good in clustering if clusters have complex geometric shape

### APPLICATIONS

* Market Segmentation
* Document Clustering
* Image Segmentation
* Image Compression
* Customer Segmentation
* Analyzing trends on dynamic data.

# HIERARCHIAL CLUSTERING

* It is an unsupervised type of learning where the unlabelled data is clustered hierarchically
* It is of two types – Agglomerative and Divisive
* In agglomerative all data points are treated as a separate cluster and then using various distance metrics they combine in a bottom-up approach to give the resulting the large cluster
* In divisive the entire dataset is treated as a large cluster and then broken down into smaller clusters in a top-down fashion.
* The clusters are in a tree form of structure called a dendrogram.

## STEPS

1. Treat every datapoint as a separate cluster
2. Join the nearby clusters to form a large cluster
3. Continue this process until we have a large cluster encompassing all the points
4. Then dendrograms can be used to divide into multiple clusters deepending on the problem.

## ADVANTAGES

* Easy to implement
* Outputs a hierarchy
* No need to find value of K beforehand

## DISADVANTAGES

* Impossible to undo the previous step
* Time complexity- not suitable for large datasets
* Initial seeds have strong impact on final results
* Order of data has strong impact
* Very sensitive to outliers

# DBSCAN

* It is a form of clustering based on the abundance or density of data
* It needs 2 inputs – eps and minpts .
* Eps refers to the minimum distance and minpts refers to minimum number of points for the creation of a cluster.
* Dbscan stands for density based spatial clustering of applications with noise

## LOGIC

* First we start with an arbitrary point.
* We consider a circle of radius=eps and find and count datapoints within that circle.
* If there are no other points then the existing point is treated as a noise point and we move on to a different arbitrary point
* But if the count > minpts then that particular point is referred to as a core point and the cluster is formed ,we move on to the datapoints in the cluster,if they cant form another whole cluster then they are referred to as a border point

## ADVANTAGES

* DBSCAN can have nonspherical clusters
* It is not that sensitive to outliers
* Number of clusters is not required