Data Science with Python : K means Clustering Algorithm #1527

Presented by: Deepthi M

Batch Number: 05

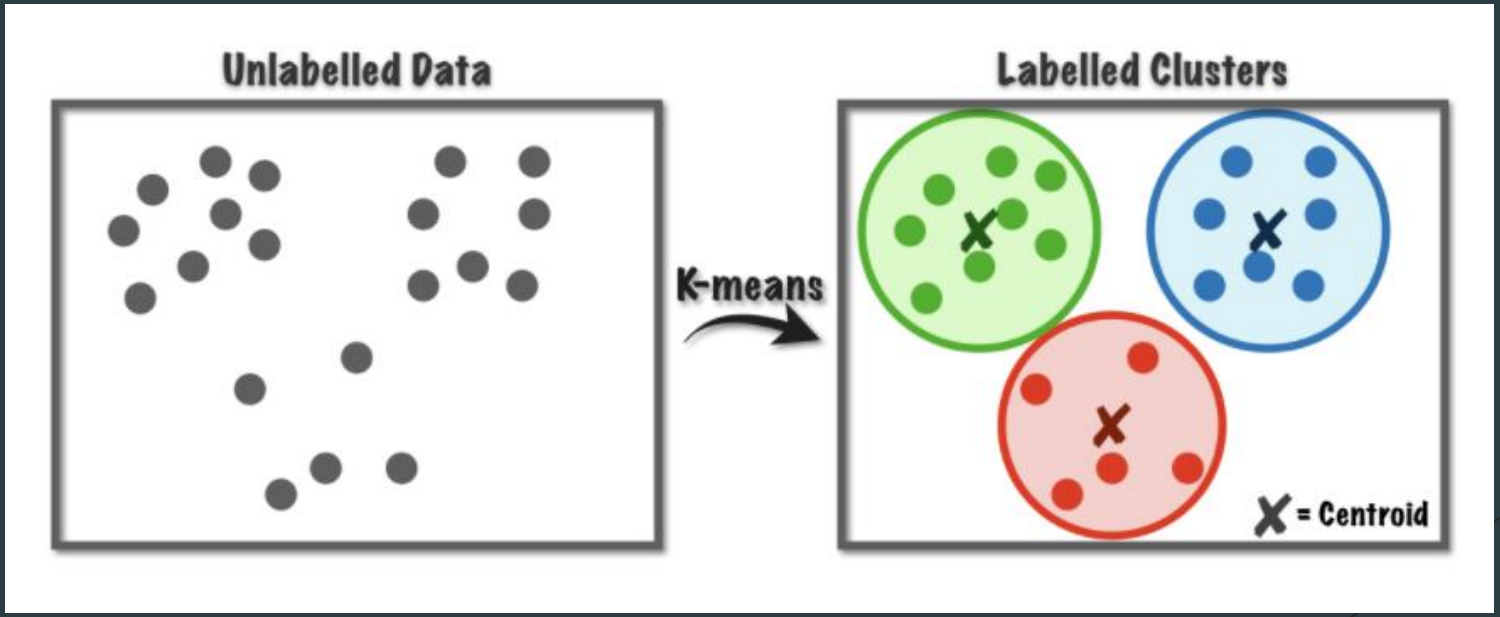
Serial Number: 172

What is clustering?

Clustering is unsupervised machine learning model, it is used to cluster the homogenous groups based on similarity.

Distance is the measure that is used to measure the similarity.

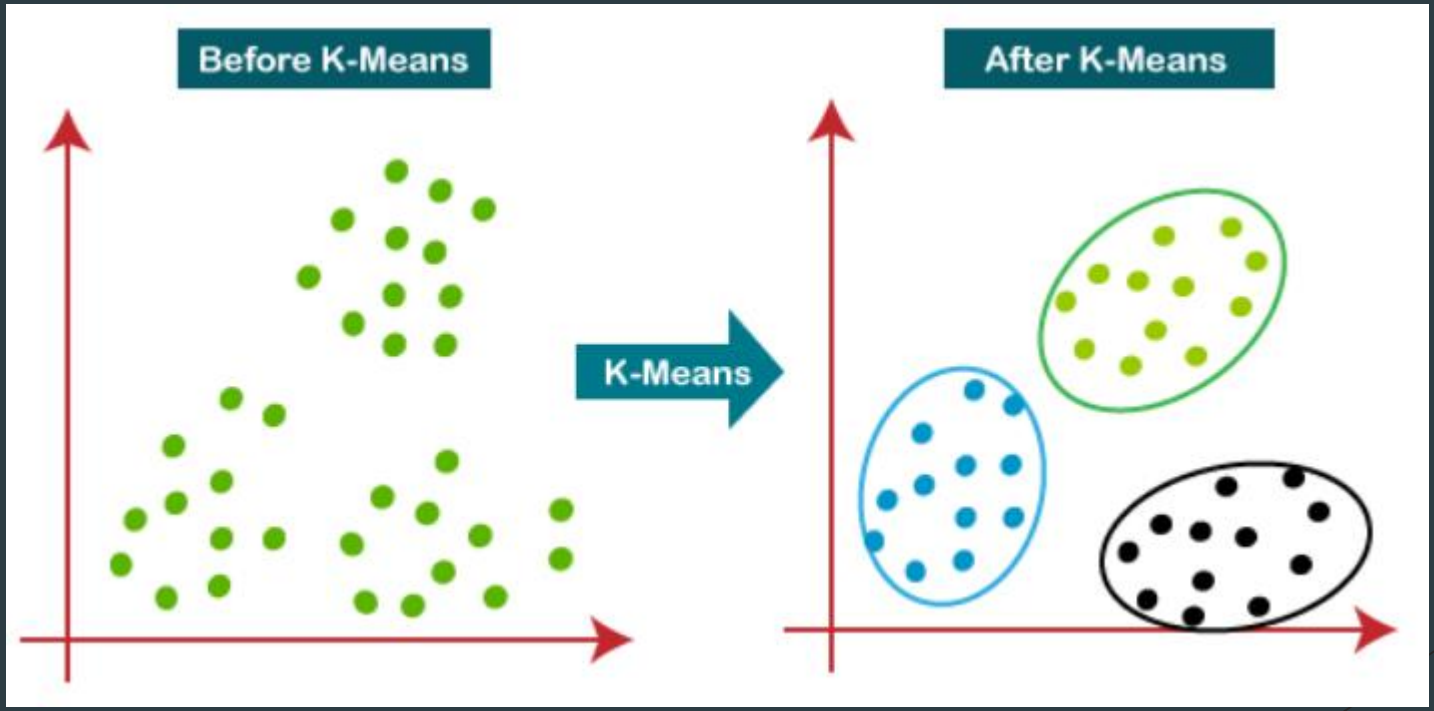
Some of the distance measures:
Euclidean distance
Manhattan distance



What is k-means clustering?

K-means clustering is a non-hierarchical clustering. It is non-hierarchical because it does not follow any hierarchy.

K-means clustering used to group the homogeneous data points.



How do k-means clustering algorithm works?

K-means clustering usually forms the clusters based on the number of clusters we pass While building the model.

It randomly chooses centroid and forms a cluster by grouping the nearest data points.

It is a non-deterministic model which changes on every execution.

Selection of k in k-mean clustering:

- Randomly assigned
- Odd number of k is chosen
- Large number of k is not preferred as it forms large number of cluster which might lose the homogeneity nature of clusters.
- Too small is not chosen as it has more prone to outliers.

K-Means Advantages :

1) If variables are huge, then K-Means most of the times computationally faster than hierarchical clustering.

2) K-Means produce tighter clusters than hierarchical clustering.

K-Means Disadvantages :

1) Difficult to predict K-Value.

2) Different initial partitions can result in different final clusters.

3) It does not work well with clusters of Different size and Different density

I hope you have got an overview on the topic:

K-means clustering

Thankyou.