



* Aim → Implement K means algorithm for clustering, to create a cluster on dataset (using Python)

* Objective → 1. To understand the concept of clustering
2. To implement K means clustering algorithm

* Theory →

K means clustering aims to partition n observations into K clusters in which each observation belongs to the ~~nearest~~ cluster with the nearest mean serving as prototype of the cluster

* Working :

The K means clustering algorithm attempts to split a given anonymous dataset into a fixed number of clusters.

Initially K number of so called centroids are chosen. They are picked up randomly in the initial stage such that all centroids are unique. These centroids are used to train the KNN classifier.

The resulting classifier is used to classify the data and thereby produce an initial randomized set of clusters. Each centroid is thereafter set to the arithmetic mean of the cluster it defines. The process of classification and centroid adjustment is repeated until the values of the centroids stabilize.

The final centroids will be used to produce the final classification clustering of the input data, effectively turning the set of initially anonymous data into a set of data, each with a class identity.



• Advantages :

1. If variables are huge, then k means most of the times is computationally faster than hierarchical cluster
2. Relatively simple to implement
3. Scales to large dataset
4. Easily adapts to new examples
5. Generalises to clusters of different shapes and sizes

• Disadvantages :

1. Choosing k manually
2. Centroids can be dragged by outliers
3. Scaling with number of dimension
4. Dependent on the initial values

* Conclusion → Thus in this assignment we learned about k means clustering algorithm and implemented it on a randomly generated dataset.