2208-CSE-6363-001-MACHINE LEARNING
INSTRUCTOR : DAJIANG ZHU

K-Means Clustering Algorithm on Iris Dataset
Project 3 – Report

Note : The path given is not relative i.e it wont be same for every system. SO please kindly change the path accordingly in-order to view the result correctly. I have displayed the screenshots of the correct result after providing that path.

SYNOPSIS - ABOUT K-MEANS CLUSTERING

Given is the famous iris dataset in which the task is to perform K-Means Cluster Algorithm. Clustering is one of the most famous and common unsupervised data analysis technique which helps users to get a basic knowledge on how is the structure of the data. So in-order to achieve this clustering tries to find a homogeneous or similar sub-groups in the whole group and then measure the distance (depending on the type of distance to be found Euclidian/correlation) to find out the similarity parameter.

So a point is said to be in a related or particular cluster when it's nearer or closer to the centroid of the cluster as compare to the others. The most interesting thing about clustering is that target isn't the main thing. But we try to gather similar observations and then make them into certain groups as a result of which it's an unsupervised learning problem. So the clustering can be performed in various ways where each way or type of clustering has it's own merits and de-merits which would depend on the data on which the clustering would be performed.

Some of the merits would be – Because of the model interpretation i.e finding the centroid, one can make changes to the cluster accordingly. Being unsupervised method, it works on unlabeled data too. It's one if the effective ways to interpret the solution, it even works well on bigger variables in the given dataset. In contrast to the merits, the de-merits would be as the result depends on the input of points that user provides, if the points provided aren't the right one compatible to model it dosen't provide accurate results

IMPLEMENTATION OF THE METHOD

So provided is a an iris dataset. According to Wikipedia, the data set consists of 50 samples from each of three species of Iris (Iris setosa, Iris virginica, Iris versicolor) and four features that describe each of them will be length and width of sepals and petals (in cm). Now our main is to perform the analysis and build a model to form them into groups using K-Means Clustering, which can be done by searching for the centroid.

K-MEANS PROCEDURE

- **Firstly the clusters count has to be chosen in-order to start the process, according to which the clustering will de done.**
  This is the first and foremost step in-order to begin the process.

- **Secondly the cluster centroids initialization i.e. choose those initial points with which the process can begin.**
  Here the initialization can be done in a way that they can be taken at random or else order or sort the dataset and then split into K parts. Now from this sorted list, points can be chosen.

- **Followed by, allocate the respective data points to the formed groups (clusters) one by one by choosing the closer centroid.**
  Till now the points have been provided and now it's for the model to compute by itself where it tries to form a group with the points assigned i.e into clusters. Then distance will be computed (different types of calculations are present like Euclidian, etc.) from the chosen points and all the centroids present. Then whichever distance is less, the model tries to cluster or group with that respective centroid.

- **Then try improving the cluster centroids by choosing another points apart from the ones that has been chosen in the beginning.**
  This step will help enhancing the model because the points which were provided in the start may or may not match with the model which wouldn't result in efficient clustering in centroid formation

- **Now continue the above process from centroid initialization up to the point where a favorable condition is met**
  Observe that clusters come closer gradually by doing the process. So the process will come to a hold when reasonable or good amount of clusters were formed and observed. So the process would stop when based on several reasons where some of them would be the cluster count doesn't change, no proper cluster formation, etc.

CODE IMPLEMENTATION

- Initially relevant libraries are imported like numpy, pandas (both for array work), matplotlib (for graph plot) and k-means with sklearn.