Shri Vile Parle Kelavani Mandal's
DWARKADAS J. SANGHVI COLLEGE OF ENGINEERING
(Autonomous College Affiliated to the University of Mumbai)
NAAC Accredited with "A" Grade (CGPA : 3.18)

Department of Computer Science and Engineering (Data Science)

Subject: Machine Learning – I (DJ19DSC402)

AY: 2022-23

JHANVI PAREKH

60009210033

CSE(Data Science)

Experiment 9

(K-Means)

Aim: Explore K means clustering with variations on different datasets. **Theory:**

The K-means clustering algorithm computes centroids and repeats until the optimal centroid is found. It is presumptively known how many clusters there are. It is also known as the flat clustering algorithm. The number of clusters found from data by the method is denoted by the letter 'K' in Kmeans.
In this method, data points are assigned to clusters in such a way that the sum of the squared distances between the data points and the centroid is as small as possible. It is essential to note that reduced diversity within clusters leads to more identical data points within the same cluster. The following stages will help us understand how the K-Means clustering technique works- ***Step 1:*** First, we need to provide the number of clusters, K, that need to be generated by this algorithm. ***Step 2:*** Next, choose K data points at random and assign each to a cluster. Briefly, categorize the data based on the number of data points. ***Step 3:*** The cluster centroids will now be computed. ***Step 4:*** Iterate the steps below until we find the ideal centroid, which is the assigning of data points to clusters that do not vary.
4.1 The sum of squared distances between data points and centroids would be calculated first.
4.2 At this point, we need to allocate each data point to the cluster that is closest to the others (centroid).
4.3 Finally, compute the centroids for the clusters by averaging all of the cluster's data points.

When using the K-means algorithm, we must keep the following points in mind:
It is suggested to normalize the data while dealing with clustering algorithms such as K-Means since such algorithms employ distance-based measurement to identify the similarity between data points.
Because of the iterative nature of K-Means and the random initialization of centroids, K-Means may become stuck in a local optimum and fail to converge to the global optimum. As a result, it is advised to employ distinct centroids' initializations.

Lab Assignments to complete in this session:

Use the given dataset and perform the following tasks:
Dataset 1: Synthetic Data (200 samples, 3 clusters and cluster_std = 2.7)
Dataset 2: Titanic dataset (http://s3.amazonaws.com/assets.datacamp.com/course/Kaggle/train.csv

Shri Vile Parle Kelavani Mandal's
DWARKADAS J. SANGHVI COLLEGE OF ENGINEERING
(Autonomous College Affiliated to the University of Mumbai)
NAAC Accredited with "A" Grade (CGPA : 3.18)

Department of Computer Science and Engineering (Data Science)

And http://s3.amazonaws.com/assets.datacamp.com/course/Kaggle/test.csv)

Task 1: Perform Kmeans clustering on Dataset 1 with random initialisation, 10 variations of initial means, 300 iteration. Find Lowest SSE value, final location of centroids and number of iterations to converge. Show the predicted labels for first 10 points.
Task 2: Perform elbow method and silhouette method to find appropriate clustering value on Dataset 1.
Task 3: Preform data cleaning and pre-processing on dataset 2. Form three clustering using Kmeans++ initialisation.
Link for all three tasks:
https://colab.research.google.com/drive/1_oWLl5R57HYp_Z5_e-dV18PiATuxKS5b?usp=sharing