Report for Project Part 2 - K-Means Clustering

The purpose of the project is to perform K-Means clustering on the given "AllSamples.mat" dataset using two different strategies.

Strategy 1 - Clustering using randomly selected centroids (Standard K-Means)

Strategy 2 - Clustering using farthest points (K-Means++)

K-Means Clustering is a type of unsupervised learning which is used when we have unlabelled data. The algorithm classsifies data into specific clusters based on some similarity measures. Some similarity measures are Eucledian distance, Cosine similarity, Shortest path distance on a graph and so on. In this project, the clustering is carried out using the Eucledian distance method.

The algorithm follows a naive approach in that it assumes the number of Clusters(K) before it performs clustering in the dataset.For this particular algorithm to work, the number of clusters has to be defined beforehand.

STRATEGY 1 - Standard K-means - Cluster using random centroids

The K-means algorithm starts by randomly choosing a centroid value for each cluster. After that the algorithm iteratively performs three steps:

(i) Find the Euclidean distance between each data instance and centroids of all the clusters

(ii) Find the nearest centroid to a data instance and assign it to that cluster

(iii) Calculate new centroids by calculating the mean of each cluster and repeat the above steps till the model converges

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

The function to calculate Eucledian distance

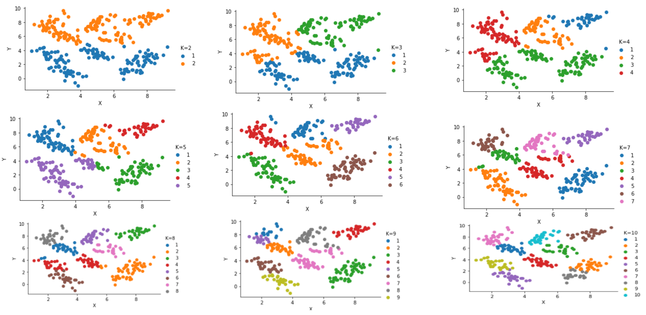
The distance is measured in python using the following command

(np.sqrt(sum((X2[j] - X1[i])**2)))

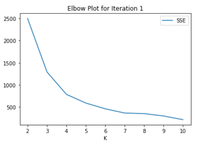
where X2 and X1 are arrays denoting the centroid and data point values respectively.

The clustering is carried out for two iterations since the centroid initialization is random.

The clustering when random centroids where picked based on the value of K starting from K=2 to K=10 is shown below



The elbow plot for the above iteration is shown below



The elbow plot is used to infer optimal number of K value.

To determine the optimal number of clusters, we have to select the value of k at the the point after which there is a linear change in the Sum of Squared error value in the plot.

The sharp drop in the line plot is observed from K=3 to K=4. This indicates that 4 is the optimal cluster value for the given dataset.

STRATEGY2 - K-Means++ - Cluster using Farthest Points

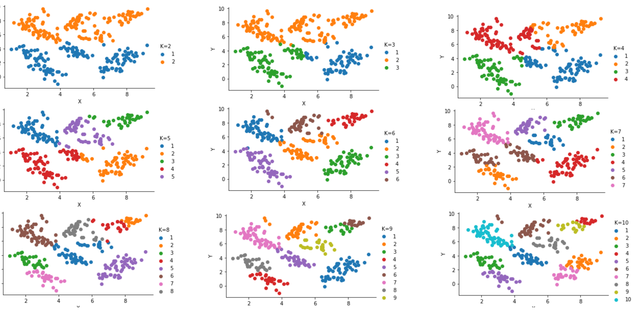
The Strategy 2 varies from Strategy 1 in the initialization of centroids. As opposed to the previous strategy, not all centroids are chosen randomly in this approach. Based on the Value of K, only the first centroid is selected randomly from the dataset. The next centroid is the point which is of farthest distance from the selected centroid point. After picking two centroids in this manner, to find the next centroid, the average value of the previous two centroids are calculated i.e. [X,Y] = [(X1+X2)/2,(Y1+Y2)/2]

The farthest datapoint from [X,Y] average value of the centroids is calculated. Likewise the process is followed till K centroids are achieved.

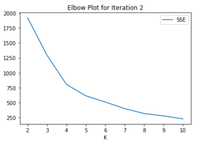
It is important to note that once a centroid is picked it is removed from further calculation of centroids. This is done to achieve unique centroids.

The init_centroid() method is used to initialize K unique centroids in our case.

Clusters from K=2 to K=10



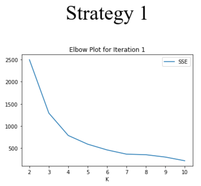
The elbow plot for the above iteration

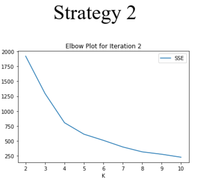


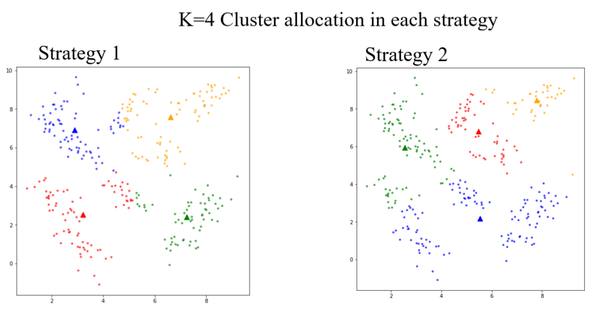
The optimal value of K = 4

Strategy 1 vs Strategy 2

The elbow plots of Strategy 1 and 2







The cluster allocation for each strategy for the optimal value K=4 is shown above. It can be evidently seen that the centroids in each strategy are different(Centroids are denoted by Triangle markers in the plot.)

It can be observed from the plot that the centroids allocated in Strategy 2 is far more distinctive than in Strategy 1.

Findings:

-> On Comparing the Strategy 1 and 2, it was observed that the model converges faster in strategy 2 i.e. relatively less number of iterations were required to cluster the dataset.

->Also, it was observed that Strategy 2 was able to locate farthest centroids there by leading to distinct clusters where as the Standard K-Means approach followed in Strategy 1 intialized random centroids meaning that there are high chances for centroids to be near each other and so the clusters may not be distinct. This was inferred from the elbow plots for each strategy. Sometimes, the elbow plot for Strategy 1 showed spikes where as the elbow plot produced by Strategy 2 was consistent for most of the iterations.