HW8 Q4

We know that in K-Means, K represent the number of clusters. KMeans Clustering is a unsupervised way of group items which are so likely to fall in same cluster as opposed to other. But how can we define what is the suitable K value for our dataset. It can be perfectly said that K valid range lies from 1 to the number of distinct class labels in the dataset. It can also be a possibility that the dataset belongs to the group of independently and identically distributed variables. If K is taken to be very large than the over all time complexity of the program increases and if K is taken very low than there may be under-fitting of the data points and appropriate clusters won't be visible. Brute force method is of making clusters for different K, and checking the Sum of squared error (SSE). If for a particular K, SSE turns to increase, we stop but we pay in terms of time complexity in this process.

We have determined K using two methods :

1. Elbow Method: This method tends to maximize V, at a point where rate of decline changes themost.
   S_list.append(sum(np.min(cdist(X,kmeans.cluster_centers_,'euclidean'),axis=1))/np.asarray(X).shape[0]).

$$V = \sum \sum d(Cj, xi)^2$$

$where\ Cj\ is\ the\ jth\ cluster\ and\ xi\ is\ the\ is\ ith\ sample. Cj\ lies\ form\ 1\ to\ K\ and\ xi\ lies\ from\ 1\ to\ length\ of\ the\ datset$

2. Silhouette Scores : these scores are obtained using the mean intra-cluster distance (X) and mean nearest cluster distance (Y) for each sample. The best value is 1 and worst value is -1 and for each K with take average of these scores over all the samples in the dataset.

$$Silhouette\ scores = \frac{(X - Y)}{\max(X, Y)}$$

Considering the time complexity of these methods, for smaller K both execute with same efficiency but with higher K Elbow methods over powers Silhouette score. The graph obtained by Elbow methods is more like negative log graph whereas silhouette graph is more like positive log graph.