CSE 575 Statistical Machine Learning

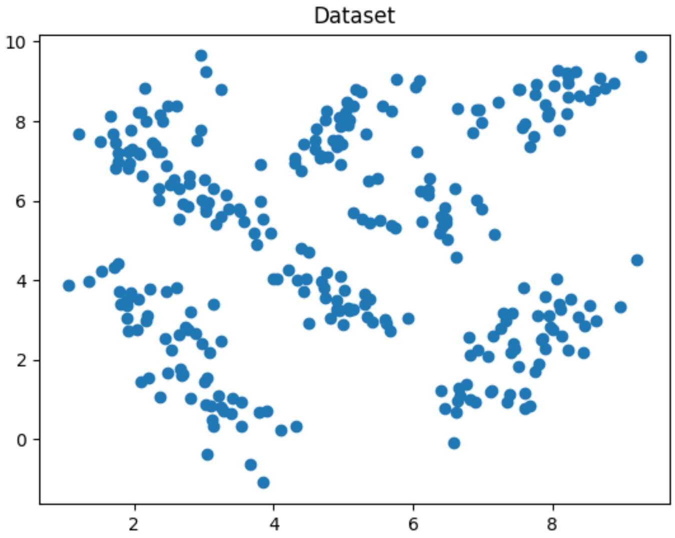
Project Part 2

Saianirud Reddy Nellore | snellore@asu.edu | ASU ID: 1219495602

The given dataset contains a set of points spread across a 2D space. We have to perform K-means clustering on the dataset using two strategies.

1. Initializing the cluster centers randomly from the given dataset.
2. Initializing the first center randomly and choosing the ith center such that average distance of this center from the previous (i-1) centers is maximum.

The dataset is loaded using SciPy and stored in a variable $dataset$. The following figure below is the scatter plot of the dataset.



To perform K-means clustering on the given dataset, we first need to calculate the initial cluster centers. Once the centers are calculated we then apply K-means algorithm on the given dataset.

K–means algorithm:

The k value, initial cluster centers and dataset are given as input to the algorithm. The similarity measure used to implement the algorithm is Euclidean distance.

1. Find the Euclidean distance between a sample(s) in the dataset and the cluster centers.

   Euclidean distance between two points $(x_1, y_1), (x_2, y_2)$ is given by,
   $$d = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

2. Assign the sample to the cluster with nearest center i.e., minimum distance from the cluster center.

   $$sample \in cluster\ with \min\left(d_{x_i}\right), where$$
   $$d_{x_i} = distance\ from\ sample\ x\ to\ center\ of\ cluster\ i.$$

3. Repeat step 2 and step 3 for all the samples in the dataset and assign them to respective clusters.
4. By now, all samples in the dataset are assigned to clusters. Calculate the mean of the samples belonging to a particular cluster and replace that cluster's center with the newly calculated mean. Do this for all clusters.

   Mean for the points $(x_1, y_1), (x_2, y_2), \dots \dots, (x_n, y_n)$ is given by,
   $$mean = (\frac{x_1 + x_2 + \cdots \dots + x_n}{n}, \frac{y_1 + y_2 + \cdots \dots + y_n}{n})$$

5. Repeat the above process (step 2 to step 5) until the cluster centers converge.

   $$convergence\ criteria\ is\ current\ cluster\ centers = \ previous\ cluster\ centers$$

The clusters obtained after the convergence of centers are considered as the final set of clusters or the output of K-means clustering.

Strategy 1:

Pick the initial cluster centers randomly from the given dataset. Then apply the above K-means algorithm on the dataset. Once you get the final clusters calculate and plot the objective function values with respect to k.

Strategy 2:

Strategy 2, unlike strategy 1 which picks initial centers randomly, picks the first center randomly and then choose the ith center such that average distance of this center from the previous (i-1) centers is maximum.

1. Pick the first center randomly from the given dataset.
2. Calculate the average of the distances between a sample in the remaining dataset and the previous cluster centers.

   Euclidean distance between two points $(x_1, y_1), (x_2, y_2)$ is given by,
   $$d = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

   Average for the values $x_1, x_2, \dots \dots, x_n$ is given by,
   $$average = \frac{x_1 + x_2 + \cdots \dots + x_n}{n}$$

3. Repeat step 2 for all the samples in the remaining dataset and assign the sample which has the maximum average as a cluster center. Remove this sample from the dataset.
4. Repeat step 2, step 3 until you find all the cluster centers.

Once we find the initial cluster centers from the above strategy, we apply the K-means algorithm on the dataset. Once you get the final clusters calculate and plot the objective function values with respect to k.

Objective Function:

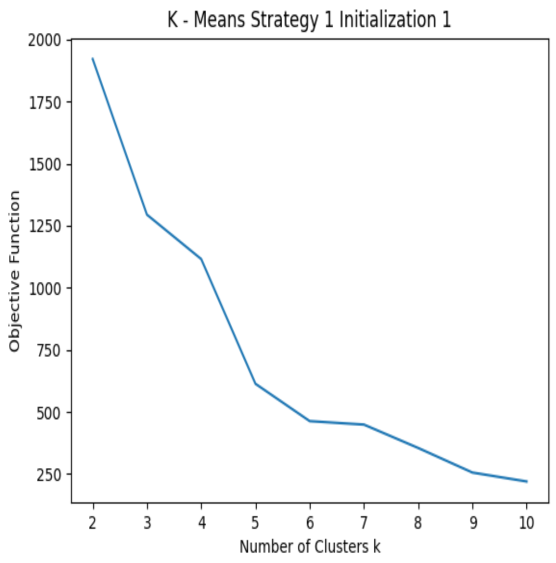
When a dataset is divided into k clusters, with $\mu_1, \mu_2, \dots \dots, \mu_k$ being the cluster centers respectively. The Objective function/sum is given by,

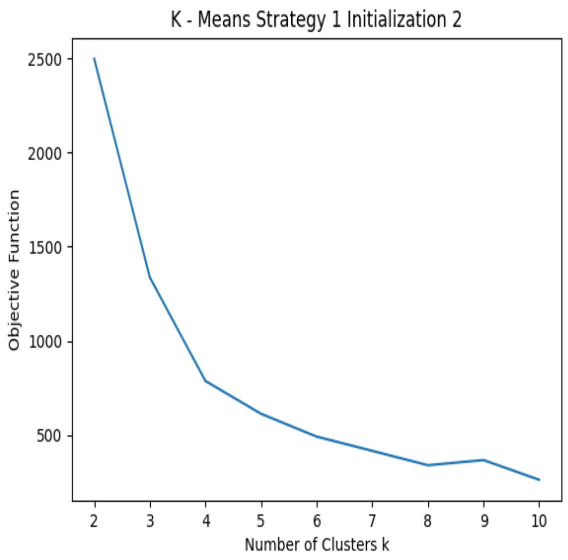
$Objective\ function = \sum_{i=1}^{k} \sum_{x \in D_i} \left|\left|x - \mu_i\right|\right|^2$, where D_i is the set of samples belonging to cluster i. $\left|\left|x - \mu_i\right|\right|$ is the Euclidian distance between points x and μ_i.

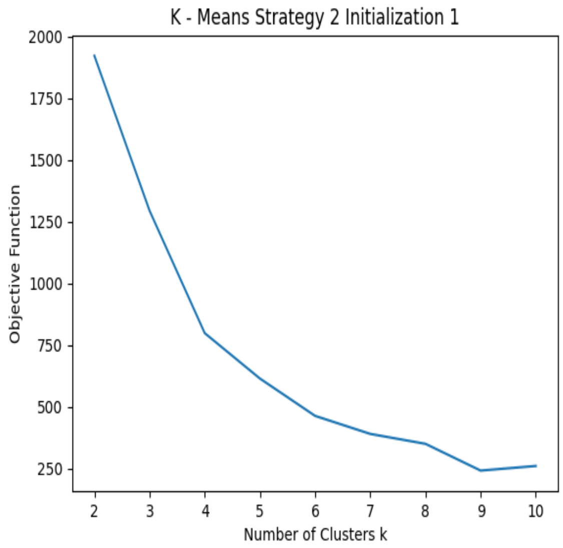
1. For a particular cluster, calculate the sum of squared distances between every sample in that cluster and the center of the cluster. Let this be $clusterSum$.
2. Repeat step 1 for all available clusters.
3. Objective function value will be sum of the $clusterSum$ values of all clusters.

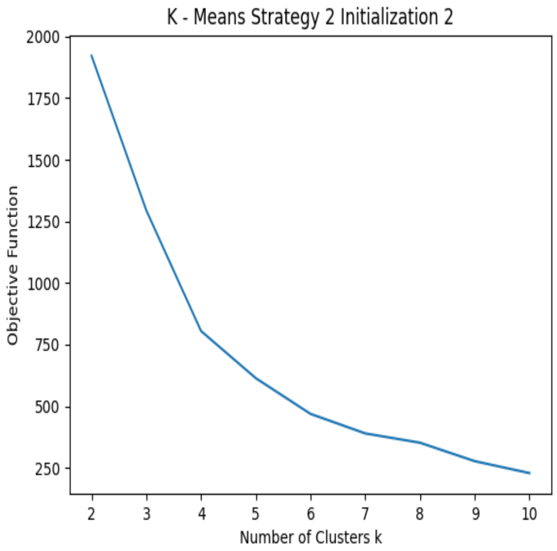
Each strategy, strategy 1 and strategy 2, is run two times and their respective objective function value is calculated for k clusters, where $k = 2, 3, 4, \dots., 10$.

Objective Function vs Number of Clusters($\boldsymbol{k}$):









The k value from which the objective function does not change drastically is generally considered as the appropriate number of clusters.

Observations:

From the above graphs, we can observe that,

The objective function value decreases most of the time as the number of clusters increases.

Strategy 1:

In this strategy, we initialize the cluster centers randomly. Random center initialization could result in different final clusters. As you can see in the above two plots related to strategy 1, the objective function value is different for same value of k i.e., the final clusters formed are different. In addition to that, outliers may affect the clusters by pulling the center towards it thereby getting its own cluster instead of being ignored. Having different plots for different initializations, will be a difficult task to predict the k value.

Strategy 2:

In this strategy, we try to push the cluster center as far as possible from the previous cluster centers. The initial centers will be spread across covering the data space as much as possible. This will lead to getting almost similar final clusters with slight variation as there is less randomness. As you can see in the above two graphs related to strategy 2, the two graphs are nearly similar, which means the final clusters formed are similar. This strategy will also help us to correctly predict the k value.