Project Part 2: Unsupervised Learning (K -means)

Overview:

To implement the k-means algorithm and apply the implementation on the given dataset, which contains a set of 2-D points.

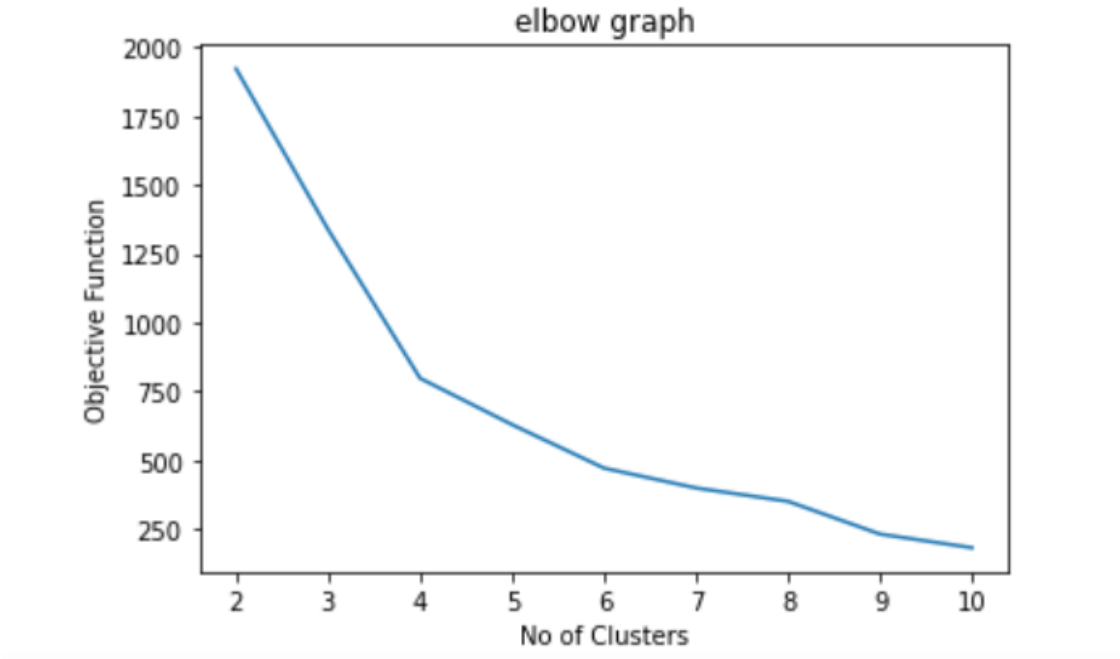
The different strategies for choosing the initial cluster centroids are:

1. randomly pick the initial centers from the given samples.
2. pick the first center randomly; for the i-th center (i >1), choose a sample (among all possible samples) such that the average distance of this chosen one to all previous (i-1) centers is maximal.

The K-Means algorithm was run using the above two strategies for clusters k 2 to 10 and for each cluster the number of iterations used are 10. The cluster centroids were constant after a certain iteration which is below 10 for all of the clusters i.e the centroids converge before 10 iterations. The algorithm is run twice for each of the two strategies and the plots obtained are shown below.

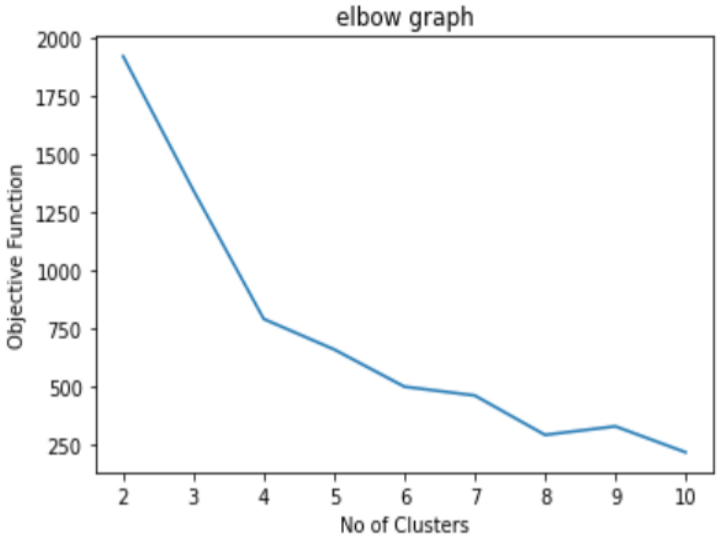
Strategy – 1 – Implementation1

The graph obtained by plotting objective function and number of clusters k when run first time for strategy 1 is



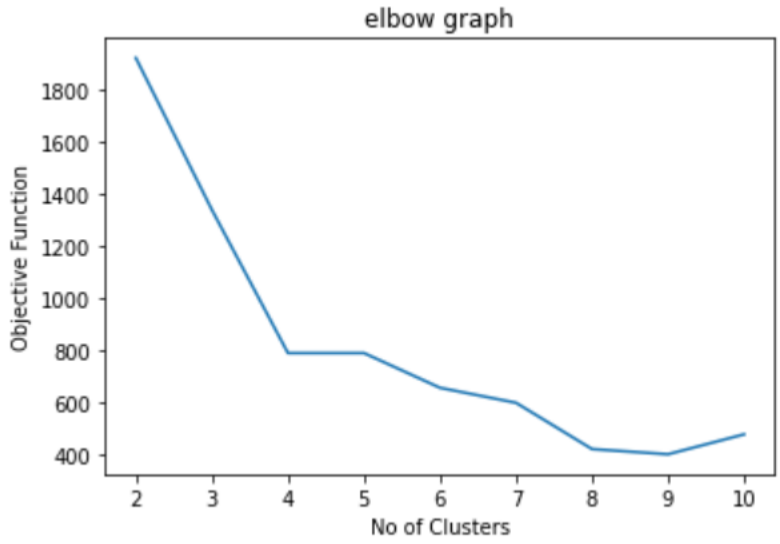
By observing the above graph we can see that there is a sharp decrease at k = 4 i.e notice an elbow and the rate of decrease of objective function is decreasing from 4. So we can consider k = 4 to be optimal number of cluster values.

Strategy 1-Implementation 2
The graph obtained by plotting objective function and number of clusters k when run second time for strategy 1 is



By observing the above graph we can see that there is a sharp decrease at k = 4 i.e notice an elbow and the rate of decrease of objective function is decreasing from 4. So we can consider k = 4 to be optimal number of cluster values.

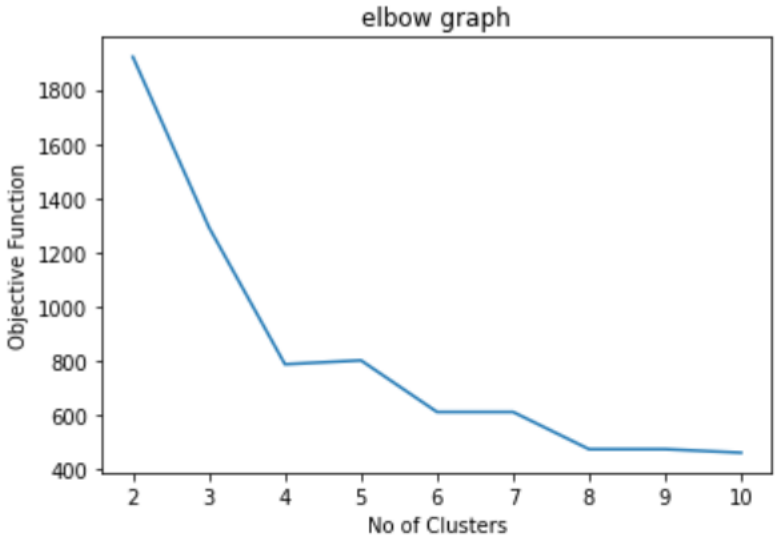
Strategy – 2-Implementation -1
The graph obtained by plotting objective function and number of clusters k when run first time for strategy 2 is



By observing the above graph we can see that there is a sharp decrease at k = 4 i.e notice an elbow and the rate of decrease of objective function is decreasing from 4. So we can consider k = 4 to be optimal number of cluster values.

Strategy-2-Implementation-2

The graph obtained by plotting objective function and number of clusters k when run second time for strategy 2 is



By observing the above graph we can see that there is a sharp decrease at k = 4 i.e notice an elbow and the rate of decrease of objective function is decreasing from 4. So we can consider k = 4 to be optimal number of cluster values.