Name: Shivang Soni SID:915623718

Problem 1. K-means clustering [65 pt]
Implement the “k-means” algorithm to cluster the dense data (“data dense.pl”) into K = 10 clusters. The k-means algorithm can be found in lecture 11, page 15 or 16 (they are equivalent). Initialize cluster centers m 1 , . . . , m 10 using 10 randomly sampled data points. Print out the k-means objective at each iteration, where C k is the data points belong to k-th cluster, and m k is the cluster center of the k-th cluster. Run the program for 40 iterations and report the objective function and running time at iteration 10, 20, 30, 40. Discuss your findings.

Solution 1.
Code is present in file name “Test.py”
Top 3 iteration for the K-means clustering objective function:
1. 107272.0
2. 61146.18715684153
3. 57943.070742836666

The time taken for iterating from 0 Iteration to 10 iteration is 15.849793195724487
Objective function at iteration 10^{th} is 54918.4560975

The time taken for iterating from 0 Iteration to 20 iteration is 30.54816246032715
Objective function at iteration 20^{th} is 54503.1222843

The time taken for iterating from 0 Iteration to 30 iteration is 46.49749207496643
Objective function at iteration 30^{th} is 54498.6888173

The time taken for iterating from 0 Iteration to 40 iteration is 62.371434450149536
Objective function at iteration 40^{th} is 54498.6888173

In every iterations the distance of the points from the centroid will decrease so we obtain our objective function value to decrease at every iteration until the points are properly classified into clusters. If there is less distance of centroid from the other the points in the cluster better will be the result of classifying the particular points into clusters. After every iteration the centroid position also changes in order to account for better classifications of the points

Name: Shivang Soni SID:915623718

Problem 2. K-means for sparse data [35 pt]
Apply the same k-means algorithm to sparse data (“data sparse E2006.pl”). Note that in this pickle file X is stored in Compressed Sparse Row (CSR) format, and you will need to modify your code accordingly to use sparse matrix (turn the data into dense matrix will be out-of-memory). Run the program for 40 iterations and report the objective function and running time at iteration 10, 20, 30, 40. Discuss your findings.

Solution 2.
Code is present in filename “TestPart2.py”
Top 3 iteration for the K-means clustering objective function using sparse data:
1. 1036.5612988889443
2. 429.0855338698723
3. 335.7292424424334

These time are calculated having less processor speed.

The time taken for iterating from 0 Iteration to 10 iteration is 703.5179858207703
Objective function at iteration 10^{th} is 201.33462249172976

The time taken for iterating from 0 Iteration to 20 iteration is 1460.4035696983337
Objective function at iteration 20^{th} is 169.7985135477643

The time taken for iterating from 0 Iteration to 30 iteration is 2132.8817121982574
Objective function at iteration 30^{th} is 163.53946795478214

The time taken for iterating from 0 Iteration to 40 iteration is 2693.5964941978455
Objective function at iteration 40^{th} is 162.5798011795414

Due to the presence of the sparse data proper conversion has to be performed to deal with sparse and dense data together for computing the euclidean distance. In every iterations the distance of the points from the centroid will decrease so we obtain our objective function value to decrease at every iteration until the points are properly classified into clusters. If there is less distance of centroid from the other the points in the cluster better will be the

result of classifying the particular points into clusters. After every iteration the centroid position also changes in order to account for better classifications of the points