CSE-6010 Assignment 2

Name: Juichang Lu GT username: jlu345

K-Mean analysis:

The algorithm was developed using the following method. First the first k points were initialized as initial centroid. Second, each point in the file was assigned to the centroid that's closest to the point. Third, the centroid was updated by calculating the centroid of all the points in the cluster. Fourth, the difference between centroid were calculated using SSE(sum of square error). If the value is under set threshold, then process two to four will be repeated.

To test if the code is working properly, a set of data was made to test the algorithm (Figure 1). By creating the data points in three clusters on purpose, the program correctly assign each point to the right cluster. This proves that the program is running properly.

Next, we move to testing different k value on the given sample data set with 766 data points and 39 dimensions. To determine the right value for k, I first the algorithm from 1 - 10. According to the elbow test, sse stop decreasing dramatically when k = 4, so this is the optimal value of k.

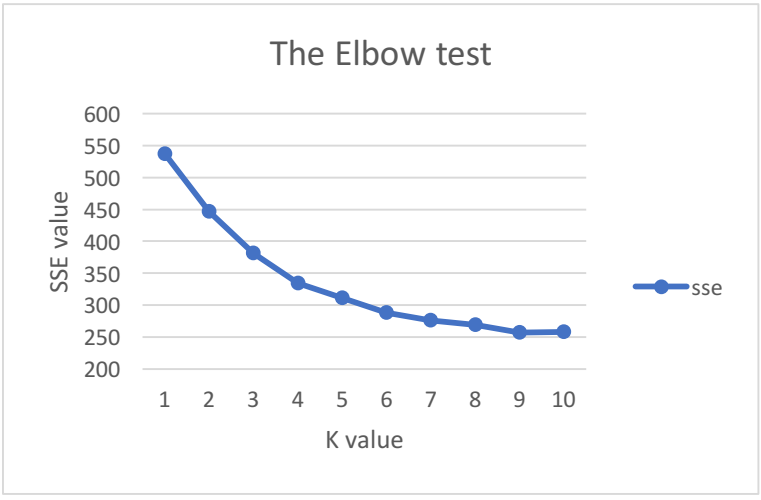

Figure 2. The elbow test. The decreasing rate was reduced when k = 4.

30 2	30 2 3
1 1	1.000000 1.000000 1.000000
2 2	2.000000 2.000000 1.000000
3 3	3.000000 3.000000 1.000000
4 4	4.000000 4.000000 1.000000
5 5	5.000000 5.000000 1.000000
2 1	2.000000 1.000000 1.000000
3 1	3.000000 1.000000 1.000000
3 2	3.000000 2.000000 1.000000
3 6	3.000000 6.000000 1.000000
1 5	1.000000 5.000000 1.000000
13 15	13.000000 15.000000 2.000000
15 18	15.000000 18.000000 2.000000
17 16	17.000000 16.000000 2.000000
19 20	19.000000 20.000000 2.000000
21 19	21.000000 19.000000 2.000000
19 30	19.000000 30.000000 2.000000
12 17	12.000000 17.000000 2.000000
29 25	29.000000 25.000000 2.000000
13 35	13.000000 35.000000 2.000000
16 24	16.000000 24.000000 2.000000
-3 -5	-3.000000 -5.000000 0.000000
-6 -7	-6.000000 -7.000000 0.000000
-6 -3	-6.000000 -3.000000 0.000000
-3 -6	-3.000000 -6.000000 0.000000
-2 -3	-2.000000 -3.000000 0.000000
-1 -3	-1.000000 -3.000000 0.000000
-5 -2	-5.000000 -2.000000 0.000000
-5 -3	-5.000000 -3.000000 0.000000
-5 -8	-5.000000 -8.000000 0.000000
-2 -10	-2.000000 -10.000000 0.000000

Figure 1. the input data and the output result of the synthetic data. The input file format follows the format that's described by the assignment. The output file has an additional k value in the header line and an additional column is added at last.