**ECE 595 Homework 8 Due: April 17, 8 PM**

Consider the unlabeled dataset ***X*** from the attached file, *HW8.mat*. The two columns of data show the annual income (in thousands of dollars) of a sample of adults visiting a mall and their spending score. Plot the data and observe that there appear to be five distinct clusters that form the data in . (a) Using K = 5, apply K-means algorithm and develop MATLAB code to arrive at the five cluster centroids with initial centroids of . From these initial centroids, iteratively find the cluster assignment for each training example  , *i = 1, 2, …, m* by finding the minimum squared Euclidean distance (squared norm) between each example and the five centroids. That is, for each example , its cluster assignment  is given by

 such that is the minimum, where  is the index of the cluster centroid closest to  and  is the *jth* centroid.

After the cluster assignment to all *m* examples, recalculate the centroids based on the current clusters as

, where is the set of examples that are assigned to centroid *k*.

If, for example, , that is, are assigned to centroid *k = 2*, then the updated centroid is given by



You may terminate the iterative process when D, the sum of all the distances in each cluster from the examples that are assigned to that cluster, is a minimum (choose, for example, D = 0.05).  (b) Plot the cost function D after each epoch as a function of epoch. Note that cluster assignment and centroid modification for all K clusters together constitute one epoch. If the error does not converge after 500 epochs, try raising D to 0.5 or so. What are the final centroids? (c) Plot the final centroid locations along with the data points. How many iterations does it take to converge? (d) Instead of initializing centroids with fixed values, randomize the input data (note that the income column is in an increasing order) using the function randperm and pick the first five observations as initial centroids. Run the code three times ensuring each run has a different set of centroids. Terminate the iteration when none of the centroids change significantly between two successive iterations. That is, stop the process when

 for all k = 1, 2, …, 5.

Print the initial and final centroids and show the plot with centroids for each run.