Homework 4

Due Nov 9

The MNIST dataset is a dataset of 28 by 28 images of hand-written digits. (you only really need the training images and labels though). Since the dataset is quite large, restrict yourself to the first 1000 images, and their labels. You can read in the file mnist\_train.csv which is the processed file having 1000 rows and 785 cols

1. Store the cols 2:785 of this matrix as digits and the first col as labels. Digits should be a 1000 by 784 matrix
2. Write a function mykmeans to perform a k-means clustering of the 1000 images of

digits. Use Euclidean distance as your distance measure between images (which can be viewed as vectors in a 784 dimensional space). Your function should take 3 arguments, the matrix

digits , the number of clusters K and the number of initializations N.

You code should consist of 3 nested loops.

**The outermost (from 1 to N) cycles over random cluster initializations (i.e. you will call k-means N times with different initializations).**

**The second loop (this could be a for or while loop) is the actual k-means algorithm for that initialization, and cycles over the iterations of k-means. Inside this are the actual iterations of k-means. Each iteration can have 2 successive loops from 1 to K**:

* the first assigns observations to each cluster and
* the second recalculates the means of each cluster.

These should not require further loops. (You will probably encounter empty clusters. It is possible to deal with these in clever ways, but here it is sufficient to assign empty clusters a random mean (just like you initialized them)).

Your function should return:

1. the cluster parameters ( centroid ) and cluster assignments( class assignments ) for the best solution
2. the objective/loss-function over k-means iterations for the best solution (this should

be non-increasing)

c) Run your code on the 1000 digits for K= 5; 10; 20; 25. Plot K values on the x-axis. For each setting of K , plot the best solution of objective/loss function on y axis.

Copy the output of your code and the plot from c) to a docx file and submit to Dropbox hw4.