ECE 532 Final Project Update #1
Owen Seymour, Section 002
November 16, 2020

Project GitHub page: https://github.com/oseymour/ECE_532_Final_Project.git

First, my project proposal accidentally stated I would be developing a KNN classifier. This was incorrect, I meant to say that I would develop a K-means classifier.

Since publishing the project proposal, the K-means algorithm has been developed and initial testing on the test data set has been done. The code allows for some modification of the running algorithm. The user is able to select from one of two centroid initiation methods, the number of centroids, and the number of iterations to use.

- Centroid initiation: the two methods for centroid initiation are randomly chosen centroids from the entire training set or randomly chosen centroids from each digit group in the training set. In the first method, k images are randomly chosen from the training set as the starting centroids. This has been termed "random initiation". In the second method, the user passes a ten element array to the function, where the number at each index indicates the number of starting centroids to choose from the corresponding subset of images that are labelled with that index. For example, if the second index is three, three images will randomly be chosen from all images in the training set that are labelled as the digit two. This method has been termed "selected initiation".
- Number of centroids: this is a necessary part of any K-means algorithm implementation and allows the user to choose any number of centroids from the given training subset.
- Number of iterations: the user can also choose how many iterations to run to allow the centroids to converge. The default is ten, since during preliminary troubleshooting it was found that it rarely, if ever, takes more iterations for the centroids to converge.

Initial results on the test data are not very great, admittedly. So far, testing has been done with four configurations and results are shown in the table below. After the centroids were finalized, each centroid was printed out and labeled as a number. Whichever centroid a given image in the test data set was nearest to was the label that test image was given.

Error rates	Random Initiation	Selected Initiation
10 centroids	0.4431	0.3888
20 centroids	0.3019	0.2994

For the ten selectively initiated centroids, one from each class was chosen. For the 20 selectively chosen centroids, the number from each class was as follows, in order: 2 zeros, 1 one, 1 two, 1 three, 1 four, 3 fives, 3 sixes, 3 sevens, 4 eights, 1 nine. Some numbers were chosen more times since they were much closer to other centroids. For example, the centroids of six and eight are very close together, since those digits are so similar.

Moving forward, the following goals still need to be accomplished:

- Try using a different norm besides the Frobenius norm, which is the default in np.linalg.norm(). The one-norm and two-norm will be tested.
- Develop the truncated SVD linear classifier
- Test and refine the truncated SVD classifier
- Develop the neural network classifier
- Test and refine the neural network classifier

The next deadline of note is December 1, 2020, when the next project update will be published. If all goes well, the K-means algorithm will allow for choosing the norm used in calculating distances between matrices and the truncated SVD linear classifier will be completed by this update.