HW3 Project Problem B for Math 420

In this problem you will analyze a collection of handwritten digits collected by USPS, see, e.g., the data at the page of Prof. S. Rowe is:

`http://www.cs.nyu.edu/~roweis/data/usps_all.mat`

[This time, you **can** simply cut and paste this URL because the printed symbol $\sim$ above **is** a single ASCII character.]

This dataset is formed by 1100 examples of images of each of 10 digits (0, 1, 2, 3, 4, 5, 6, 7, 8, 9) in the 8-bit grayscale format.

(i) Load the data, and divide into two equal sets of 550 images per digit: training set and testing set. Make sure your sets are properly labeled. Plot and verify some examples.

(ii) Represent each of these 16x16 images as a vector with 256 coefficients. Utilizing the Euclidean distance as defined in class ($\|x\|_2 = \sqrt{\sum_{j=1}^{d} |x_j|^2}$, where d represents the number of coefficients of x), compute pairwise distances between the elements of your training data. For a fixed integer K, and for each element of your training data set find its K nearest neighbors, i.e., K digits with smallest distances from the given data element. Try to do this in a way that will not overload memory of your computer :)

(iii) Use the labels of the K nearest neighbors to classify each element of the training set by voting. Determine and explicitly state your own rules to break ties in the voting procedure.

(iv) Find the value of K which yields optimal performance of classification from part (iii) on the training data.

(v) Classify the testing data by computing, for each element of the testing set separately, its K nearest neighbors within the training dataset, and using your voting procedure. Use the K from part (iv). Set the result to be true if it matches the actual label of the digit, and false otherwise. Report your success percentages for each digit separately, as well as globally.

(vi) Will your results change if you replace the Euclidean distance with vector angle or cosine of the vector angle?