MATH 420, HW 5 (Machine Learning Project 1 HW 1), SPRING 2015, Due: March 10, 2015

Download the handwritten digits collected by USPS dataset, and divide randomly into two sets: the training set of 100 images per digit, and the testing set of 1000 images per digit. Make sure your sets are properly labeled. Plot and verify some examples.

Represent each of these 16×16 images as a vector with 256 coefficients. Consider 2 basic schemes for doing this: reading the images row by row, from left to right; and reading the images column by column, from top to bottom.

For each of the 2 vectorization strategies, utilize the Euclidean distance as defined in class to compute pairwise distances between the elements of your training data. For $k = 3, 5, 7, 9, 11$, 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. (Note that there is no need to sort your distances for each k separately.)

Use the labels of the $k = 3, 5, 7, 9, 11$ 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. (Note that as there are 10 classes, simply having an odd number of votes is not sufficient to avoid ties.)

Find the value of k which yields best performance of classification from the previous part on the training data. (Note that you must define how you measure what is the best performance.)

For each of the 2 vectorization strategies, classify the testing data by computing, for each element of the testing set separately, its optimal k nearest neighbors (determined in the preceding part) within the training dataset, and using your voting procedure. 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.

Compare the 2 vectorization strategies.

Will your results change if you replace the Euclidean distance with vector angle or cosine of the vector angle, or some other metric?