AMSC/MATH 420, Project 4, SPRING 2015

Oral Presentation Due: March 24, 2015

Written Presentation Due: March 26, 2015

1) Download the dataset of handwritten digits collected by USPS and divide them into two equal sets of 5500 images: the training set consisting of 1000 examples of each of the digits 0 through 4 and 100 examples of each of the digits 5 through 9; and the testing set consisting of 100 examples of each of the digits 0 through 4 and 1000 examples of each of the digits 5 through 9. The goals of this project are to develop and test methods for classification of the handwritten data (to be discussed further in class), which are optimized on the training set and then applied to the testing set to assess the performance of the developed methodology.

Use the nearest neighbors classification in the standard Euclidean metric with $k = 20$ (to be introduced in class), to verify the success rate of this classification scheme. Then optimize the representation of your data to maximize the global success rate. Analyze the role of the structure of the training data in your classification.