AMSC/MATH 420, Project 3A, SPRING 2016

Oral Presentation Due: TBA

Written Presentation Due: TBA

1) Download the dataset of handwritten digits collected by USPS and divide them into two sets: the training set and the testing set. Construct the training set to contain 100 examples of some of the digits and 550 examples of the remaining digits. Use the remainder of the data to be the testing set. The goals of this project are as follows:

- 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.
- Analyze the impact of the structure of the training set on the performance of the classification scheme.

Use the nearest neighbors classification scheme in the standard Euclidean metric with $k = 20$ (to be introduced in class), or any other suitable classifier that you may already know, to verify the success rate of your classifications. Analyze the role of the structure of the training set on your classification results. Which digits require larger training sets and why?