Final Project Proposal

Jeffrey Cheng and Kevin Shao

Dataset

We wish to use the MNIST dataset, which provides 28x28 pixel gray-scale images, along with their classification as a digit, 0-9. The dataset is rather large, having 60,000 training observations and 10,000 test observations.

Predictors

Of the 784 predictors (28x28), which range from 0-255 based on brightness of that pixel, we plan on using all of them. If this proves computationally expensive or inefficient, we will try to cluster or subsample.

Response

The response variable is simply a number, 0-9, corresponding to the digit that the image represents. This is categorical, since the number has no correspondence with any value, only a category.

Interest

The MNIST dataset represents a baseline in machine vision, and is fairly clean. Running multiple classification models against such a dataset will prove valuable. While the data is very clean (no missing values, all values within 0-255), completely different people wrote the digits. When one considers the different ways to write numbers (fat vs. skinny 0, crossing 7’s, vertical vs. slanted 1s, etc.), the feat of classifying handwritten digits becomes seems far more impressive.