MNIST

Elijah Verdoorn

November 30, 2016

Introduction

The MNIST dataset is a popular dataset for research in machine learning. The dataset consists of 70,000 images, each a digitization of a handwritten digit. The standard for working with this data is to use 60,000 of the image for training, reserving the other 10,000 for testing and validating the predictive models. Each image in MNIST is 28 pixels by 28 pixels, yielding 784 distinct values. These values range from 0 to 255 in base 10, each representing the amount of black in that pixel. Since the images are grayscale, we can represent them with only one matrix rather than more complicated color images, which can either be represented with three distinct matrices (one for each of the color channels) or by a single matrix with all values in base 16. This reduces the amount of computation needed to work with the dataset, with the discarded color information being classified as out of scope for the project. Predictive results on the data set by various groups across the world can be found on the internet, sorted by type of classification algorithm used to make predictions. Simple models such as linear classifiers have error rates as high as 12.0%, while the most complicated models listed are able to drive the error rate as low as .23%.