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%. Some of the classification models use pre-processing to make the data more suitible for use with the chosen method. Especially popular are normalization techniques such as deskewing, Since the dataset has 784 predictors, feature selection methods such as least absolute shrinkage and selection operators (LASSO) are frequently used to eliminate variables that do not contribute to the final suolution. Examples of such variables in the case of the MNIST data would be the value in the very first pixel - it is totally white the vast majority of the time. Such a value does not improve our ability to predict what class our data belongs to, so we can safely ignore it.

Derivation of LASSO Feature Selection

LASSO feature selection can be derived from linear algebra principles. The

Compact Matrix Decomposition

In an effort to improve computation speed, specifically around the training of new models, optimization techniques are applied. These techniques range in effectiveness and complexity, and are a very active area of mathematic research. One recently developed technique is the Compact Matrix Decomposition (CMD). The CMD was developed at Carnegie Mellon University, and first published in the paper “Less is More: Compact Matrix Decomposition for Large Sparse Graphs”. This paper has been cited 94 times since its publication in 2007 according to Google Scholar, increasing interest in the topic. The paper claims that the CMD requires less than 1/10 of the space of the more traditional Singular Value Decomposition (SVD) and is 10x more computationally efficient, while maintaning the same level of reconstruction accuracy. Here I seek to apply te CMD to the