Dalya Adams
MSDS 422 section 59

MNST Classification

Up until recently, the computing power needed to process large datasets was extremely expensive. This meant that if an organization wanted to analyze large amounts of data at a reasonable price, the dataset needed to be trimmed to decrease the time used to analyze the dataset. Removing valuable observations is not ideal for any data scientist, which is why dimensionality reduction methods have maintained such prominence in the data analysis world. By reducing the dimensionality of a dataset, it can be processed using less computing power, which means cheaper and faster analysis.

As times have changed, and computing power is much cheaper, the need to reduce the dimensionality of the dataset is not as essential. Despite the lack of monetary incentive of dimensionality reduction, the discussion of the speed with which data can be processed and an answer gleaned is still ever present. When implementing models for computer vision or image classification, it is inevitable that the dataset will be very large. When considering a simple image of a number, sized 28x28, we can expect 784 variables, one for each pixel. As images become larger or the quantity of images increases, the speed with which a classification can be reached slows down. This is where the argument for incorporating dimensionality reduction retains its importance. By reducing the dimensionality of a dataset, using techniques like Principle Component Analysis, the speed with which is can be processed increases. Some of the accuracy is lost, which raises the question which is more important, speed of execution or accuracy?

The purpose of this research is to determine if the use of principle component analysis (PCA) decreases the time to actionable insight on a large dataset outweighing the possible decrease in accuracy. In reviewing the dataset, it is composed of 70,000 observations and 784 variables. These variables are numerical attributes of each pixel making up the 28x 28 image for each observation. The large number of variables contained in this dataset makes the use of PCA seem reasonable. Each observation is a handwritten number between 0 and 9, with some being very clear representations of the number written and others being less clear, as would be expected when looking at the handwriting of different people. When viewing the descriptive stats of a subset of the 784 variable, we immediately notice that the scale of the variables differs greatly. This extreme scale, ranging from a max value of 0 to 255, can decrease the effectiveness of PCA as well as machine learning and deep learning algorithms.

In performing the PCA, we first scale the dataset, using both a standard scaler and a min max scaler. The use of two different scalers is to ensure the best scaler for accuracy is selected. Next ,we fit and transform the dataset using the SciKit Learn's PCA function set to catch 95% of variability contained in the dataset. This decreases our dataset from 784 variables down to 331 variables. Finally running a random forest on the dataset to predict the value of each observations based on the reduced dataset.

For this dataset, the use of PCA is unnecessary. We did not see a decrease in time between the random forest with PCA to the random forest run on all the data. In reviewing the results, the time to run the PCA was 16 seconds and 16 seconds to run the random forest, providing us an F1 score of .889. The random forest run on the 784 variables took 6 seconds and resulted in an F1 score of .948. As computing power is relatively inexpensive, the use of the random forest on the unscaled and full dataset is preferred. It capitalizes on both the accuracy of the random forest and the speed of input to insight. These same results may not hold true for other datasets and as new datasets, with differing complexity, are encountered, reevaluating the applicability of these results is essential to ensuring the accurate and timely results hold.