Diego Valdes

IST 718

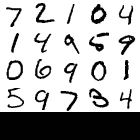
Lab 2

Aug 9, 2019

**Data Exploration**

Data was obtained from <https://github.com/zalandoresearch/fashion-mnist> and required processing requiring a program (also obtained from the repository). Data was divided into 2 groups, train and test, with each group containing 2 files, a list of arrays representing a number on a 28x28 grid (images) and an array with numbers from 0-9 (labels). The nth entry in the image list was the number data in the corresponding nth digit in the labels array.

Image data for each number was 784 elements with values ranging from 0 – 255. If formatted in a 28x28 grid it was possible to display the images.



The data was formatted into a csv file with each row representing one digits image data and a final column with the label.

In a previous analysis using machine learning to distinguish numbers, a second file was created where any value above 0 was converted to 1. This was done to eliminate the values below 255 which represented shading from black to white and provide the models with more distinct images to analyze and process. This was not done for this analysis as the results of that previous conversion didn’t result in a significant difference in results between the two datasets.

**Analysis and Results**

**Naïve Bayes**

The Naïve Bayes model performed at 56% accuracy on the training and test set. Model creation and evaluation performance was fast, 6.29 s for model generation and 833 ms to run predictions. Model creation was straight forward and didn’t require experimentation with parameters.

**Gradient Boosting**

Gradient Boosting model performed at 100% test accuracy on the training set and 85% on the test set. Model creation required parameter adjustments to raise the accuracy on the test set. With the increase of max\_depth model generation time increased. The final parameters chosen resulted in the model creation time of 3 min 40 s.

**Naïve Bayes vs Gradient Boosting**

While generating the model for gradient boosting required more computing power and time, the performance of the model justifies the trade off when compared to Naïve Bayes. Gradient Boosting was able to be 35% more accurate than Naïve Bayes.