IST 707: Homework 7

**Introduction**

Technology has accelerated the global economy in the past several decades. One of the most revolutionary advances has been the advent of eCommerce. Consumers no longer need to visit brick and mortar locations to obtain essential items, opting instead to order items online.

The lack of necessity for visiting physical locations across the majority industries has presented new challenges for businesses. Those most impacted by the growth of eCommerce have been those that previously relied on superior customer service for competitive advantage. Establishments such as local banks and credit unions previously relied on the inviting, ‘home-town’ feel to attract customers; however, now customers are more prone to choose larger banks for their increased accessibility online.

In order to stay competitive with the major banks, the Syracuse Family Credit Union (SFCU) is seeking to improve its mobile check deposit process. They rushed the feature to market to stay relevant, which led to the use of two full time employees to examine mobile check deposits and verify the handwritten information and signatures. They are hoping to use a machine learning method to automate the process—improving efficiency and reducing customer wait time for mobile funds availability. They previously sponsored a study to help them decide which model is best for recognizing handwritten information, however, the models did not offer complete automation of the process. Therefore, they have initiated a second study to find a more viable option.

**Analysis and Models**

**About the Data**

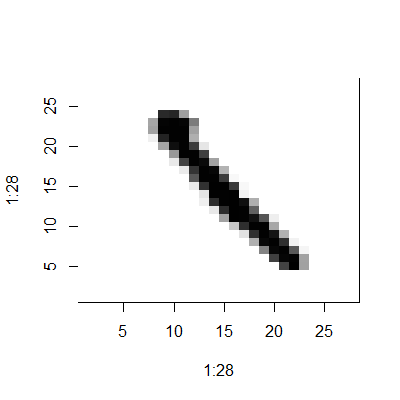
The data for this study consisted of images of handwritten digits (numbers) from Kaggle (https://www.kaggle.com/c/digit-recognizer/data). This data was divided into two separate sets for training and testing. The training data set consisted of 31500 digits and the testing data set contained 10500 digits.

For both the train and test data set, each row of data represented a vectorized version of a handwritten digit. Each digit had 28x28pixels (784 total pixels) that comprised the image, and each row had the individual pixels as attributes. The values for each pixel represented the darkness of the pixel. The datasets also contained a label at the beginning of each row denoting what number was depicted in the image.

To enable modeling on the data, all the pixel values were converted to numeric values and all the labels for the training dataset were converted to a factor ranging from 0-9.

Figure 1 (below) depicts the first row from the training dataset rendered as an image. The correct label for this image is the number ‘1’.

**Figure 1: Image 1 from the Training Digit Data (Number ‘1’)**



**SVM Model**

Initially, a Support Vector Machines (SVM) Model was applied to the digits data. This model attempts to divide the data with vectors so that digits on either side of a vector are classified as one label or another. Due to the non-linear nature of such high dimensional data, a polynomial kernel transformation was used to make the data divisible by linear vectors. Once trained, this model was used to predict the digit labels of the test data.

**kNN Model**

Additionally, a k Nearest Neighbor (kNN) Model was used on the digit data. This model chooses a number of (k) center points within the data. Once it has determined the centers, it calculates the distance from the center points to determine which center a data point is nearest to classify it. After initial centers and mean distances of digits in a cluster are determined, the model will recalculate until the mean distances are minimized around each center. Due to this iterative nature, this model may take longer to train. Once this model was trained, it was also used to predict the label for the handwritten test digit data.

**Random Forest Model**

Furthermore, a Random Forest Model was applied the data sets. This model uses a collection of decision trees to properly classify the labels for the data. It will keep the best results to determine the optimal classifications for each data level. This model may have computational implications because it is iterating over the data so many times to create multiple decision trees. Once this model is trained, it will be tested on the test digit data.

**Results**

**SVM Model Results**

This model was able to correctly predict the train data digit label with a 100.00% accuracy and the test data digit label with 97.48% accuracy. The confusion matrix (with correct predictions in green) from using the model to predict the test data is:

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Actual Digit** | | | | | | | | | | |
| **Predicted Digit** |  | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| 0 | 984 | 0 | 3 | 1 | 2 | 0 | 6 | 1 | 3 | 2 |
| 1 | 0 | 1169 | 3 | 1 | 2 | 1 | 0 | 3 | 4 | 1 |
| 2 | 4 | 1 | 1006 | 8 | 4 | 0 | 5 | 5 | 5 | 2 |
| 3 | 0 | 1 | 5 | 1049 | 1 | 6 | 0 | 1 | 8 | 2 |
| 4 | 3 | 3 | 3 | 0 | 1016 | 1 | 3 | 2 | 3 | 8 |
| 5 | 5 | 0 | 5 | 12 | 1 | 944 | 8 | 0 | 4 | 2 |
| 6 | 5 | 0 | 2 | 1 | 3 | 6 | 1017 | 0 | 2 | 0 |
| 7 | 0 | 1 | 7 | 5 | 3 | 1 | 0 | 1094 | 2 | 9 |
| 8 | 0 | 2 | 6 | 8 | 2 | 0 | 4 | 2 | 955 | 6 |
| 9 | 0 | 0 | 2 | 6 | 10 | 3 | 0 | 12 | 1 | 1001 |

This model appears to be nearly perfect at predicting handwritten digits. This model would certainly eliminate the need for SFCU to have any full-time employees assigned to mobile deposit verification. There may be an occasional need for verification on the 2.5% of mobile deposits, but this would certainly be a feasible option for full automation.

**kNN Model Results**

This model was relatively less accurate than the SVM Model, but still had very strong prediction results. It achieved a 91.26% accuracy when predicting the training data digit labels and a 91.32% accuracy when predicting the test data digit labels. The confusion matrix (with correct predictions in green) from using the model to predict the test data is:

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Actual Digit** | | | | | | | | | | |
| **Predicted Digit** |  | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| 0 | 965 | 0 | 18 | 1 | 1 | 3 | 17 | 1 | 7 | 3 |
| 1 | 1 | 1172 | 88 | 37 | 28 | 29 | 19 | 50 | 50 | 11 |
| 2 | 0 | 1 | 848 | 4 | 0 | 0 | 0 | 0 | 0 | 1 |
| 3 | 0 | 1 | 4 | 989 | 0 | 30 | 0 | 0 | 35 | 5 |
| 4 | 2 | 1 | 15 | 2 | 927 | 4 | 1 | 3 | 9 | 10 |
| 5 | 9 | 0 | 7 | 18 | 1 | 857 | 6 | 0 | 25 | 0 |
| 6 | 18 | 1 | 8 | 4 | 10 | 13 | 999 | 1 | 10 | 0 |
| 7 | 2 | 0 | 34 | 15 | 3 | 2 | 1 | 1041 | 6 | 26 |
| 8 | 1 | 1 | 15 | 8 | 1 | 1 | 0 | 0 | 814 | 0 |
| 9 | 3 | 0 | 5 | 13 | 73 | 23 | 0 | 24 | 31 | 977 |

Although this model was quite accurate, the computing performance was far worse than the SVM Model. If SFCU were to opt for using the kNN Model to verify its mobile check deposits, it would also greatly reduce the needed manpower; however, it may result in slower responses due to high compute requirements. If SFCU wanted to use this model, it may be best to invest in a dedicated server with sufficient processing power to execute the model in a timely manner.

**Random Forest Model Results**

This model, similar to the kNN Model, had computational difficulties. The model was unable to parse through the full train and test data sets with the provided resources for this study. As a result, the data sets were reduced to half their size. This model was able to correctly predict the reduced train digit labels with 81.55% accuracy and the test digit labels with 91.29% accuracy. The confusion matrix (with correct predictions in green) from using the model to predict the test data is:

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Actual Digit** | | | | | | | | | | |
| **Predicted Digit** |  | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| 0 | 404 | 0 | 4 | 0 | 0 | 4 | 5 | 2 | 1 | 3 |
| 1 | 0 | 443 | 4 | 1 | 2 | 1 | 1 | 3 | 1 | 1 |
| 2 | 2 | 3 | 350 | 7 | 3 | 1 | 5 | 5 | 8 | 1 |
| 3 | 3 | 0 | 4 | 343 | 0 | 13 | 0 | 2 | 11 | 10 |
| 4 | 0 | 0 | 1 | 3 | 344 | 4 | 1 | 9 | 5 | 16 |
| 5 | 1 | 4 | 1 | 16 | 2 | 297 | 5 | 1 | 8 | 4 |
| 6 | 3 | 2 | 4 | 1 | 2 | 5 | 376 | 0 | 2 | 2 |
| 7 | 3 | 0 | 9 | 4 | 2 | 1 | 1 | 381 | 2 | 10 |
| 8 | 3 | 1 | 6 | 18 | 0 | 10 | 2 | 1 | 347 | 9 |
| 9 | 1 | 0 | 1 | 9 | 18 | 4 | 0 | 9 | 6 | 310 |

This model would offer manpower reduction for SFCU but may have issues with larger volumes of mobile check deposits. As with the kNN model, additional computational resources (i.e. a new high-powered server) would likely be required.

**Conclusions**

In summation, technology has contributed exponentially to the global economy as of late. Perhaps the most significant of these technological contributions has been the rapid growth of eCommerce. Consumers no longer need face-to-face interactions in physical locations to obtain the goods and services they need.

Formerly customer service centric industries such as banking have had to adapt their business strategy to be more accessible to customers in online platforms. The Syracuse Family Credit Union (SFCU) rushed the launch of its mobile check depositing system to keep pace with the bigger banks in industry. As a result, they are inefficiently using two full-time employees to validate the handwriting on mobile deposits. In order to help automate this process, SFCU previously sponsored a study to determine the best machine learning method to employ. Although the results of this study (using Naïve Bayes and Decision Tree Models) offered some manpower reduction, it did not provide a solution that would support full automation of the process. In response they initiated a second study.

From the second study, two mostly automated options were identified (kNN and Random Forest Models); however, both options would likely require the acquisition of additional server resources due to high computational requirements. The most viable option for SFCU is to use an SVM model which will not require additional investment and will lead to near full automation with a high level of accuracy.