Handwriting Analysis

APAN 4335 Machine Learning Midterm Project

|  |  |  |
| --- | --- | --- |
| Nischitha Rao  Applied Analytics  nr2605@columbia.edu | Qimeng Deng  Applied Analytics  qd2133@columbia.edu | Ray Anthony Roderos  Applied Analytics  rdr2137@columbia.edu |

**Project Topic and Problem Statement:**

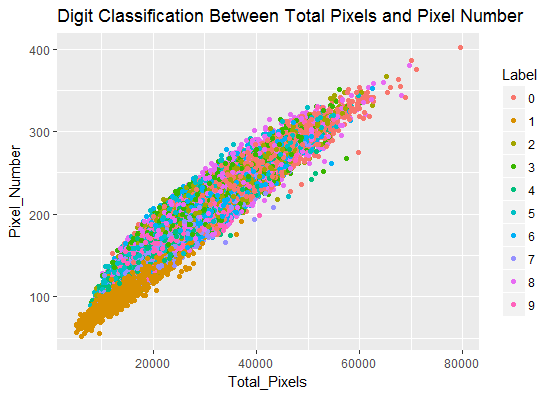
The project topic the group has chosen is Handwriting Recognition and the problem statement is the following:

Which is the best classifier algorithm to recognize handwritten digits provided by the MNIST database of handwritten digits before and after engineering the data?

**Dataset, Features and Software:**

The dataset that was used is the MNIST database of handwritten digits. Initially an MNIST file but converted into CSV. It has a training set of 60,000 examples, and a test set of 10,000 examples. It has 785 features - 1 classifier variable categorized between 0 to 9 and 784 predictor variables. The predictor variables are pixels with values from an 8-bit grayscale (0-256) describing the intensity of the shade of gray (from white to black). If the predictors are arranged in a 28x28 matrix and plotted according to their grayscale value, it will create an image of the handwritten digit. The software that was used to analyse the data is the latest version of R and RStudio as of March 2017.

**Methodology:**



Classification Plot between Total Value of Pixels and Number of Pixels with Non-Zero Values

For the methodology, we first ran the following algorithms: Multiple Linear Regression, Ridge Regression, Lasso Regression, Support Vector Machines (SVM), Random Forest, K-Means and Decision Trees, using the original data from the training set. Once the model has been created from the training set, the group used the predict() function with the model on the test set which will create the predicted y values. The predicted y values will then be compared with the true y-values of the test set to have an accuracy rate. Then the datasets will be engineered to add more predictors from the original that will extract more insight for the algorithms. Then the process of training the model, predicting the test set’s y-values and comparing it with the true y-values will be executed with the new predictors only and the original and the new predictors combined.

The milestones of the group is to create record the accuracy rates of the algorithms using the original dataset, dataset with the new predictors only and the dataset combining the original and the new predictors. Another milestone is to engineer the dataset to be able to create new features from the original dataset. And lastly, for the bonus points - to determine if the k-means cluster analysis is able to successfully cluster the right labels and if the group is able to successfully recreate the same accuracy rate with a smaller training set.

**Machine Learning Classifier Approaches:**

**Multiple Linear Regression:** The group used a multiple linear regression model to get a quick start in creating a classifier and seeing the overview of the model

**Advantages:** It represents a least-square fit of the response to the data. It chooses the hypothesis by minimizing the cost function. It is one of the fastest in terms of processing time and the relationship between the predictors and the response are clear.

**Disadvantages:** Due to the large number of independent variables and the highly variance relationship among different rows of data, the linear relationship is not very reliable because it focuses too much on formalizing the data into one simple relationship, thus it is very prone to under fit. This can be seen as the accuracy only ranges from 18.69% to 22.1% when the group executed it with 3 different dataset combinations.

**Ridge and Lasso Regression:** The group used these classification models to try to improve the accuracy. Multinomial classification is the mode for classifying instances into one of the more than two classes. In this case, the group classified them into 10 different classes (label 0 to 9).

**Advantages:** Ridge and Lasso regression are good ways to regularize linear regressions and filter out the most relevant parameters. In the ridge, the coefficients of the linear transformation are normal distributed and in the lasso they are Laplace distributed. In the lasso, this makes it easier for the coefficients to be zero and therefore easier to eliminate some variable not contributing to the output. The ridge is easier to implement and faster to compute than the lasso.

**Disadvantages:** Both models does not seem to fit for this case, especially the ridge, whose accuracy is only around 10%. The lasso performs better than the ridge and has 30.47% accuracy when building model on training dataset. Therefore, both of them may not be the best models to predict the labels. Or it is because handwriting case does not fit for traditional and standardized models.

**Support Vector Machines (SVM):** SVM is supervised learning model with associated learning algorithms that analyses data used for classification and regression analysis with a hyperplane. So the group used it to perform non-linear classification.

**Advantages:** SVM has a regularisation parameter and thus avoids over-fitting. It is an approximation to a bound on the test error rate, so it is efficient and useful on classifying dataset.

**Disadvantages:** It takes an unreasonably long amount of time to train a SVM model and it does not generate very satisfying results. The group spent more than two days to train a SVM model on original training dataset but the accuracy is just around 10%. And because of the technical constraints, the group only used 10 thousand observations and just half of the features. Due to the low accuracy rate and the overlapping features, the SVM tended to under fit because there was no potential boundary. Visually, it is harder to understand for higher dimension models.

**K-Means Cluster Analysis:** K-Means clustering divides the observation into K clusters in which each observation belongs to a group of clusters with the nearest mean. An R code was written in which the cluster with the highest percentage of a certain classification is then applied that label and any classification without an assigned label will then be assigned to the cluster with the highest percentage thus there were some clusters with multiple labels assigned. Accuracy was then measured after.

**Advantages:** It is the simplest unsupervised learning algorithm. K-means are computationally faster for datasets that have a high number of features.

**Disadvantages:** It is difficult to find the optimum k-value. Since this unsupervised learning, more data does not necessarily means more accuracy. And since the starting points are random, each iteration gives a different value requiring several runs to make sure the combination of features are robust.

**Decision Trees:** Decision trees is a predictive model that maps observations in a dataset. It is used to classify larger data and facilitate decision-making.

**Advantages:** Decision trees implicitly perform variable screening. There’s a clear-cut yes/no. Decision trees require relatively little effort for data preparation and it is visually easy to understand.

**Disadvantages:** It is time consuming. The results can be easily misinterpreted. Pruning is almost always required or there is will be a tendency of overfitting.

**Random Forest:** In random forest, the team chose to grow 100 trees for their Random Forest Model because of the computationally long time required with a number higher than that.

**Advantages:** It runs efficiently on large databases. It can handle thousands of input variables without variable deletion. And visually it is, is similar to the decision trees.

**Disadvantages:** However, because of the multiple trees that are generated, the time required is long for the machine to compute. It also tends to over fit the data as seen in the very low accuracy rate. Tuning/ Pruning is also required.

**Data Engineering:**

The data transformation is based on manipulating the original features. Many of the engineering is combining the pixels that are near each other to make a bigger pixel and getting the average of the gray scale. This was done with several dimensions for data variety and also creating a feature that counts the number of pixels with non-zero values. The following on the right are the new features that were transformed:

**Name and Descriptions of New Features Added**

|  |  |
| --- | --- |
| Name of the Feature | Description of the Feature |
| Total\_Pixel | Total value of all the pixels are added |
| Left\_Half, Right\_Half, Upper\_Half, Lower\_Half | Values of the pixels in opposite halves of the 28x28 matrix are added to create these 4 features |
| Qudrant\_1 - Quadrant\_4 | Values of the pixels in the 4 quadrants of the 28x28 matrix are added creating these 4 features |
| Qudrant1Quartile1 - Quadrant4Quartile4 | Values of the pixels combined in 7x7 subset the comprise the whole image with the sums creating the 16 features |
| Subset1 - Subset49 | Values of the pixels combined in smaller 4x4 subsets that comprise the whole image creating 49 new features |
| Pixel\_Number | Total number of pixels with non-zero values |

**Results:**

In this project, we implemented 7 different ML algorithms to determine the best predictor handwriting analysis. From the results, we can conclude that K-Means and Decision trees were the standard algorithms with the highest accuracy rate to classify the correct label for handwritten digits. K-means and Decision trees gave the highest accuracy for the 3 scenarios of data. Linear Regression models also had a tendency to predict poorly as new features were added and combinations of the new and the old features were used. SVMs, aside from the technical limitations to analyse all the features and all the observations, plotting the training set on various combinations of features, examples from more than 1 label always mix together, and it shows no potential boundary. Random Forest with 100 trees tend to over fit the data as seen from the less than 1% of the accuracy rate. There were also technical limitations as Ridge, Lasso, SVM and Random Forest analysis took too long if all 60 thousand observations and 785 features were inputted or were not computed at all. From our analysis, it would look like that K-Means and Decision trees, simpler models were the best.

**Accuracy Tables of the ML Algorithms for Handwriting Analysis**

|  |  |  |  |
| --- | --- | --- | --- |
| Algorithm | Accuracy  (Original Features only) | Accuracy  (New Features only) | Accuracy  (Original + New) |
| Multiple Linear Regression | 22.09% | 18.69% | 22.1% |
| Ridge | 9.8% | 10.32% | 9.8% |
| Lasso | 30.47% | 7.53% | 11.5% |
| SVM  (10,000 obs.) | 11.35% | 9.82% | 9.82% |
| Random Forest  (trees = 100) | 0.28% | 0.12% | 0.29% |
| K-Means  (k=10) | 66.82% | 39.24% | 38.9% |
| Decision Trees | 61.96% | 63.72% | 66.43% |

**Bonus 1**

The team tried analysed the training set with K-Mean with K = 7 up to 15, trying to determine if the unsupervised model is able to see the 10 clusters. From our accuracy formula, we were able to determine that, aside from the 10 clusters, a 7-cluster K-Means gave the highest accuracy with 43.56%. As we delve deeper into it. It seems that this K-Means were able to group similar digits together with 3 clusters grouping 3 visually similar digits each (such as 9,3 and 5). Other K-Means also tend to do the same, but this combination offered the highest accuracy. For more details, please refer to the Annex.

**Bonus 2**

The team tried to predict the testing set with a smaller training set. The team started with 50 thousand but the team saw more than 10% decline in all the training models. And as the team used smaller and smaller training data, the accuracy rate continued declining. Because if the relatively high number of features and from plotting, the team concluded that certain models tended to over fit the data such as Random Forest and Decision Trees while regression analysis tended to under fit because of the overlapping features. In future work, regularization and use of fewer features are recommended.

**Team Contribution:**

Nischitha Rao: Analysis on K-means and Random Forests, Bonus question

Qimeng Deng: Analysis on Regression - Multiple Linear, Ridge and Lasso and SVM, Bonus question

Ray Anthony Roderos: Analysis on Decision Trees and Data engineering, Bonus question

**Sources:**

Victoria Catterson. (2013, December 23). Understanding data science: clustering with k-means in R. Retrieved from <http://cowlet.org/2013/12/23/understanding-data-science-clustering-with-k-means-in-r.html>

HeRen,&Quan Yang. (n.d.). Predicting and Evaluating the Popularity of Online News. Stanford University. Retrieved from <http://cs229.stanford.edu/proj2015/328_report.pdf>

Analytics Vidhya, Wikipedia – Definitions of Machine learning Approaches.

**Supplementary Materials:**

1. R- code File
2. Annexure