IJIRR

International Journal of Information Research and Review
Vol. 05, Issue, 05, pp.5526-5527, May, 2018

REVIEW ARTICLE

HANDWRITING RECOGNITION AND PREDICTION USING STOCHASTIC LOGISTIC REGRESSION

***Dr. Andy W. Chen**

University of British Columbia, 2053 Main Mall, Vancouver, BC, V6T 1Z2, Canada

ARTICLE INFO

Article History:

Received 27th February, 2018
Received in revised form
20th March, 2018
Accepted 18th April, 2018
Published online 30th May, 2018

Keywords:

Machine Learning,
Digit Identification,
Stochastic Gradient Descent,
Classification Models.

ABSTRACT

In this paper, I explore the use of logistic regression for automatic digit recognition, which is a common method for improving efficiency in many areas. I train and test the models on two pairs of digits: 0 and 1, 3 and 5. The models are estimated using stochastic gradient descent. I also vary the sample size of the training and test data and observe the accuracies of the models. I find that the prediction accuracies stay consistent across all sample sizes for both pairs of digits. This could be attributed to the comprehensiveness of the data set which includes a great variety of handwritten digits for both training and testing.

INTRODUCTION

There is a need for automatic digit recognition in many areas. For example, companies may use digit recognition to read accounting reports. Map websites may use digit recognition to read street numbers of houses in photos taken by cameras. Machine learning is a common approach to solving the problem of digit recognition. There is a variety of machine learning models such as logistic regression, support vector machine, decision tree, and ensembles of different models. Research in this area has utilized a variety of models. For example, Saxena *et al.* (1997) explore digit recognition using adaptive optical multilayer perceptrons (MLPs). Khan (2017) propose a new Multiple-Cell Size (MCS) model with Histogram of Oriented Gradient (HOG) features trained using support vector machines (SVM). McDonnell (2015) builds an Extreme Learning Machine (ELM) capable of fast training time around 10 minutes and achieving error rates lower than 1%. Chibani and Nemmour (Chibani and Nemmour, 2010) proposes a model based on a cascade of neural networks and support vector machines (SVMs). This approach uses binary support vector machines for its ability of high linear separation between target categories. Neural networks are used for its ability to do multi-class classification. Cardoso and Wichert (Cardoso and Wichert, 2013) build a model using the outputs of a biologically inspired model as features.

****Corresponding author:*** **Dr. Andy W. Chen**
University of British Columbia, 2053 Main Mall, Vancouver, BC, V6T 1Z2, Canada

In this paper, I present a logistic regression model for digit recognition. The model is estimated using stochastic gradient descent. I also explore the effect of sample size on the training and test accuracies of the models.

METHODS

The data used in the paper is the MNIST (Modified National Institute of Standards and Technology) data. It is a database of 60,000 training samples and 10,000 test samples of digits from 0 to 9. These digits saved in images of size 28 by 28. These images are mainly written by high school students and U.S. Census Bureau employees. Each image is represented by a binary vector of 784 elements. I use a subset of the data: digits 0, 1, 3, and 5. I then train and test two binary logistic regression models using stochastic gradient descent. One model is for distinguishing 0 and 1, while the other is for distinguishing 3 and 5. I build multiple logistic regression models. I vary the sample size of the training set each time from 5% to 100%, incrementing by 5% each time. I then use the trained model on the test set and compare how the training and testing accuracies change as sample size varies. The training set used in the model for distinguishing 0 and 1 contains 12665 images in total, with 5923 images of the digit 0 and 6742 images of the digit 1. The test set contains 2115 images in total, with 980 images of the digit 0 and 1135 images of the digit 1. For the other model, the training set used in the model for distinguishing 3 and 5 contains 11552 images in total, with 6131 images of the digit 3 and 5421 images of the digit 5. The test set contains 1902

images in total, with 1010 images of the digit 3 and 892 images of the digit 5. I discuss briefly the algorithm of stochastic gradient below. Stochastic gradient descent is the algorithm for estimating parameters in the logistic regression. Each parameter is a weight of a part of the image of each digit, which has a size of 28 by 28. Therefore, there are 784 parameters to be estimated. Let each parameter be, be the be the features of the data point, and be the value of the digit for the data point. Stochastic gradient starts with an initial value for each parameter to be estimated. Then it calculates the gradient at each data point and aims to update the value of the parameters by subtracting the gradient multiplied by a learning rate from the current value. This is repeated until the values of all the estimated parameters have converged. In mathematical equations, the gradient can be expressed as

At each iteration, the parameter is updated using the equation below with as the learning rate.

Below is a pseudo-code for running the stochastic gradient descent.

1. Initialize parameters to random values.
2. Shuffle data points 1 to n randomly.
3. While values have not converged:

for i from 1 to n: #This is a data point taken from the randomized list of data points
for j from 1 to d: #Once a data point is chosen, update each dimension
Calculate the gradient using
Update parameter using

In each epoch, stochastic gradient descent iterates through all training samples, using 1 sample at a time to update weights for features. Each feature requires linear time to update, so it takes time in the inner loop. The outer loop takes time. Overall, the running time for each epoch is. Using this algorithm, I explore the effect of sample size on training and test accuracies of the logistic regression.

RESULTS AND DISCUSSION

Figure 1 below shows how the accuracy of the model for distinguishing between digits 0 and 1 changes with respect to sample size. Figure 2 shows how the accuracy of the model for distinguishing between digits 3 and 5 changes with respect to sample size. The accuracy of the model for distinguishing 0 and 1 is between 0.97 and 0.99 for all sample sizes. On the other hand, the accuracy of the model for distinguishing 3 and 5 is between 0.85 and 0.94 for all sample sizes. There does not seem to be a large impact in training and test accuracies as sample size changes. The accuracies for the distinguishing 0 and 1 are slightly higher than the accuracies for distinguishing 3 and 5. Overall, the results show high training and test accuracies for both models. The high accuracies could be attributed to the cleanness and completeness of the data. That is, the MNIST data collection contains sufficient varieties of handwritten digits, so the models are able to predict the test set with high accuracies even with smaller sample sizes. Taking a closer at the graphs shows the test accuracies are lower for sample sizes 30% or lower of the total data available. This is expected as the smaller sample sizes contain fewer varieties of handwritten digits, so there is less information in the trained model, making it more difficult to accurately predict the test set or new data.

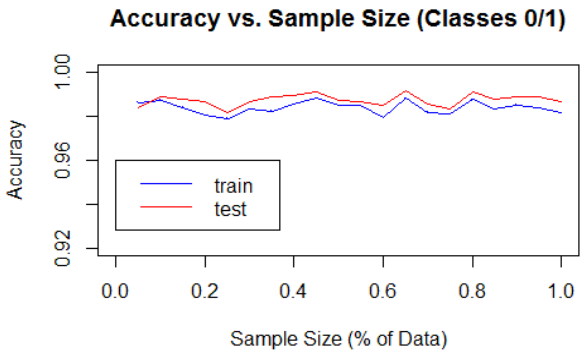

Figure 1. Training and test accuracies vs. sample size (digits 0 and 1)

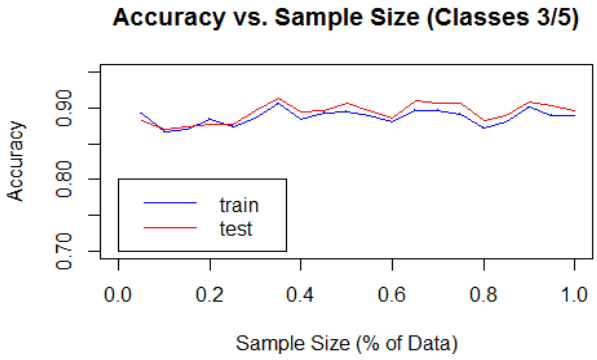

Figure 2. Training and test accuracies vs. sample size (digits 3 and 5)

Conclusion

This paper shows that logistic regression with stochastic gradient descent can achieve high training and test accuracies for the MNIST handwritten digit dataset. The accuracies for distinguishing 0 and 1 are slightly higher than the accuracies for distinguishing 3 and 5. For small training samples (30% or lower of the data available), the accuracies are lower than the accuracies obtained using larger samples. The MNIST dataset is a comprehensive collection of a large variety of handwritten digits, so the accuracies are at least over 85% even in the worst cases. As a future extension, it would be interesting to replicate this project using other datasets to see if a greater impact of sample size on accuracies can be obtained.

REFERENCES

Cardoso A, Wichert A. 2013. Handwritten digit recognition using biologically inspired features. *Neurocomputing*, 99:575-580.

Chibani Y, Nemmour H. 2010. Handwritten digit recognition based on a neural-SVM combination. *International Journal of Computers and Applications,* 32(1):104-109.

Khan, HA. 2017. MCS HOG features and SVM based handwritten digit recognition system. *Journal of Intelligent Learning Systems and Applications,* 9(2):21-33.

McDonnell MD, Tissera, MD, Vladusich T, Van Schaik A, Tapson J. 2015. Fast, simple and accurate handwritten digit classification by training shallow neural network classifiers with the 'extreme learning machine' algorithm. PLOS One.

Saxena I, Moerland P, Fiesler E, Pourznad A. 1997. Handwritten digit recognition with binary optical perceptron. Proceedings of the International Conference on Artificial Neural Networks (ICANN '97), 1253-1258.

****