Handwritten Digits Recognition - Analysis 5

Shashi Roshan (NetID : sroshan2)

Abstract

Statistical learning models were applied to hand written digits data set in order to predict the digit. A variety of learning techniques were explored and validated, and ultimately the predictive power of the models appears to be good. Recognizing hand written digits using statistical learning is an old problem. In this analysis, we compare the performance of different statistical models, rather than trying the replicate the statistical models in practice.

Introduction

We use the MNIST data set to build statistical models, which can recognize hand written digits. This can find applications in automation of reading hand written digits, eg. recognizing zipcode on postcards, etc. The data is in image format, out of which numerical features have been extracted. We try various statistical models, namely Random Forest, Support Vector Machines, Gradient Boosting and K Nearest Neighbors, and report the performance on test dataset. The original dataset has 60000 training examples, and 10000 test examples. We use a subset of this whole dataset.

Methods

Data

The MNIST data is a very popular dataset for hand written recognition. The samples from NIST's original dataset were re-mixed to create the MNIST dataset. NIST's training dataset was collected from the American Census Bureau employees. The testing dataset for NIST was collected from American high school students. The original data set consists of 60000 training samples, and 10000 test samples. We use a subset of this data to save computation time.

Modelling

In order predict the hand written digits, four different statistical models were used :

- Random forest model was trained using all available predictors. The choice of mtry was chosen using 5-fold cross validation. The accuracy metric was Out of bag error rate.
- K Nearest Neighbors model was trained using all available predictors. The choice of k was chosen using 5-fold cross validation.
- Gradient Boosting model was trained using all available predictors. The choice of n.trees and interaction depth was chosen using 5-fold cross validation.
- Support Vector Machine model was trained using all available predictors. The choice of degree, scale and c was chosen using 5-fold cross validation.

Evaluation

To evaluate the performance of different models, we split the data into train and test set. On the training set, we did 5-fold cross validation to find the optimal set of parameters. To select the best model while performing hyper parameter tuning, Accuracy was used. Once all the statistical models' performance were recorded using cross-validation results, we picked the best statistical model. Then, the best statistical model was tested on the test data set.

Results

Statistical Models	Cross Validation Accuracy	Model Description
Support Vector Machines (with Polynomial Kernel)	0.94	degree = 2, scale = 0.001 and C = 0.25
K Nearest Neighbor	0.92	k = 5
Gradient Boosting Model	0.90	n.trees = 150, interaction.depth = 3, shrinkage = 0.1
Random Forest	0.93	mtry = 39

	Class-wise performance of SVM Model (on test data)	
Digit	Sensitivity	Specificity
0	0.97	1.00
1	0.97	1.00
2	0.97	0.99
3	0.95	0.99
4	0.94	1.00
5	0.90	0.99

	Class-wise performance of SVM Model (on test data)	
Digit	**Sensitivity**	**Specificity**
6	0.95	0.99
7	0.94	1.00
8	0.94	1.00
9	0.99	1.00

Discussion

We found that the statistical models performed really well in case of classifying hand written digits. Support Vector Machines (with Polynomial Kernel) gave the best performance. The accuracy on test data set was 95.4 %. The sensitivity and specificity of the best performing model seems really good. K Nearest neighbors and Random forest model also performed well on the test data set.

We can conclude that the statistical models used in this analysis are providing good results. These models can be used in real life prediction scenarios, to test their performance.

Future work can include creation of more complex models. Also, the current analysis is performed on a subset of data set. Larger data set can be used, and models should be refit to them. Future work can also look at how the models are performing in predicting each digit.