On comparing 5 classification algorithms - Decision Trees, Boosted Trees, Random Forest, Support Vector Machines and Neural Networks

*Note: with dataset from machine learning course at coursera week 5

Prashant Kalyani
AITS Internship 2019 Batch 5
AI Tech and Systems
kalyaniprashant7@gmail.com
www.ai-techsystems.com

***Abstract*—The work is developed on a digit recognition data set which includes 5000 examples , with each example having 400 features.The dataset was taken from coursera course.The dataset was part of week 5 programming assignment.This report simply compares 5 different classification algorithm on the basis of thier accuracy on cross validation and test set**

***Index Terms*—machine learning,neural network, classification algorithm**

I. Introduction

This project is based on performance of five different algorithm on a digit recognition data set.The performance of these algorithm was evaluated on the basis of cross validation set.The data set was a part of machine learning exercise from coursera.All of the model developed for comparison are classification model

II. LIBRARIES

A. Scikit Learn

Scikit-learn provides a range of supervised and unsupervised learning algorithms via a consistent interface in Python.Different classification model were imported from scikit learn in order to train the model on the data set.

B. MatplotLib

Matplotlib.pyplot was used to visualize the data set as it is the most important part in any machile learning model building process.Matplot lib with seaborn was used to visualize the error in the data set with the help of confusion matrix

III. PREPARE DATASET

The work is developed on a digit recognition data set which includes 5000 examples , with each example having 400 features.The dataset was taken from coursera course.The dataset was part of week 5 programming assignment.Each image in data set was represented as 20 by 20 pixel.This image was the rolled out to give 400 features with each training example

A. Loading and Splitting Dataset

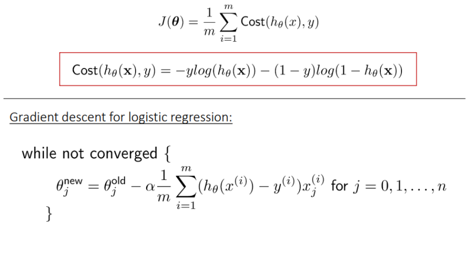
The data set was originally a matlab file which was loaded with the help of scikit learn library.Each training example label was converted to one hot vector encoding.The data was divided into 3 parts: cross-validation set, test set, training set.

IV. Different Training Model

Five different classification algorithm : SVM,Neural net,Decision Tree,Boosted Tree,Random Forest.The algorithm were trained on training data set and the best among them was selected on the basis of the algorithm performance on cross validation set.

A. Neural Net

Neural Networks is one of the most popular machine learning algorithms at present. It has been decisively proven over time that neural networks outperform other algorithms in accuracy and speed. With various variants like CNN (Convolutional Neural Networks), RNN(Recurrent Neural Networks), AutoEncoders, Deep Learning etc. The neural net developed had one layer with 25 neurons.Different activation function were tested but the best turned out to be relu.This was because other activation function such as tanh and sigmoid were responsible for slowing down the learning.For small values of parameters tanh and sigmoid derivative was equal to zero which eventually reduced the pace of learning Different values of learning rate and regularization parameter were tested and



$$J(\boldsymbol{\theta}) = \frac{1}{m}\sum_{i=1}^{m}\text{Cost}(h_\theta(x), y)$$

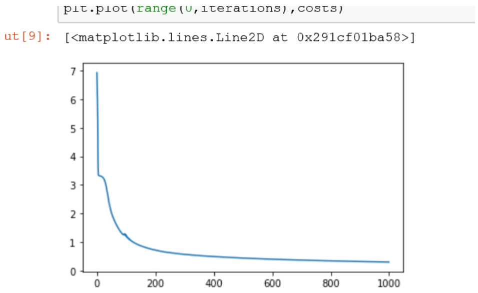
$$\boxed{\text{Cost}(h_\theta(\mathbf{x}), y) = -ylog(h_\theta(\mathbf{x})) - (1-y)log(1-h_\theta(\mathbf{x}))}$$

Gradient descent for logistic regression:

```
while not converged {
    $\theta_j^{\text{new}} = \theta_j^{\text{old}} - \alpha\frac{1}{m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})x_j^{(i)}$ for $j = 0, 1, \ldots, n$
}
```

Fig. 1. Cost Function

Identify applicable funding agency here. If none, delete this.



```
plt.plot(range(0,iterations),costs)
ut[9]: [<matplotlib.lines.Line2D at 0x291cf01ba58>]
```

Fig. 2. Iteration vs Error Graph

the best turned out to be 0.3 and 1 The lowest error the optimization algorithm achieved was 0.3 Gradient Descent optimaztion algorithm was used as the number of training example were very small.

B. Support Vector Machine

Sklearn library was used to import SVM model.The support vector machine is a classification algorithm which can classifiy data based on trained parameter.SVM model is a supervised learning approach which require label while training phase.The model had an accuracy of 94 when evaluated on cross validation set and was most among the other trained model.Based on the results of cross validation SVM was selected as best approach and was finally evaluated on test set .The accuracy on test set was 94.9

$$k(x,y) = \exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right)$$

Fig. 3. Equation Of Gaussian Kernal

C. Decision Tree Model

Sklearn library was used to import the model.The decision tree algorithm will develop a tree where its node will be different features of the provided data.The tree will try to include features as node which inturn lead to minimum entropy.This method will allow the tree to then predict the outcome of a paticular input by looking at different leaf node under a parent node.The parent as said earlier is the feature and leaf node are example which fall under that particular feature or grouping.The decesion tree had an accuracy of 100 on the train set because it generally overfits the data and the accuracy on cross validation set was calculated to be 74.8.The accuracy on cross validation set is low because it is overfitting on train data

D. Ensemble Learning

Ensemble learning is generally used for classification and regression problems.The strategy helps in prediction by including various other model rather than just one model.It has various strategy to ensemble different model such as Boosting,Bagging.This strategy generally has better accuracy as compared to single model.Random Forest is an ensemble of decision tree.The reason to use ensemble model is that provides higher accuracy , avoids overfitting and reduces bias variance error

1) Random Forest Algorithm: Random forest algorithm is basically based on decision tree and is a part of ensemble learning models.In this aproach various decision tree are developed and the most common answer predicted by different tree is considered as the final outcome.For the higher values of n estimator parameter the accuracy for test set was equal to 100 as the values of the prameter were reduced the model had less variance problem .The accuracy of the algorithm on cross validation set was equal to 87.3

2) Boosted Tree Algorithm: Boosting tree model is part of ensemble learning.It gives more emphasis on the data points which give wrong prediction , so that the accuracy can be increased.The first step is we select some random data points from training set and train the model.After the model is trained we calculate accuracy on test data set and extract data points from test set which give wrong prediction.These data points are then included with the next subset of training data.This is done in order to help model increase accuracy.The accuracy on cross validation was equal to 74.3

V. Evaluation

According to cross validation set best accuracy was given by SVM model.Decision Tree and Boosted tree overfit the training data hence result in low accuracy on cross validation set.The high variance problem is the reason for low accuracy.Heat map displays the distribution where the svm model was not able to predict the correct number.

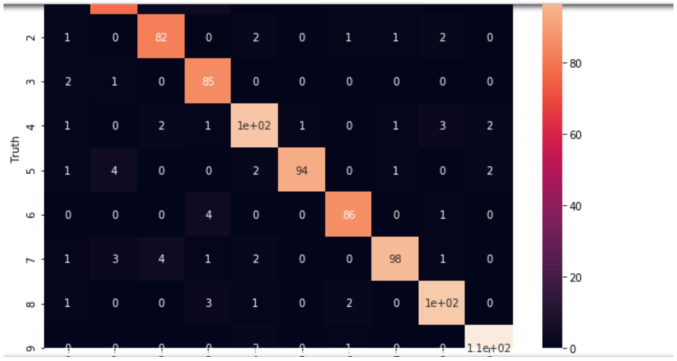


Fig. 4. Test Accuracy Of SVM model(94.9)

TABLE I
Cross Validation Performance

Case	Model	Accuracy
1	Neural Net	91.7
2	Support Vector Machine	94.1
3	Random Forest	87.3
4	Decision Tree	74.8
5	Boosted Tree	74.3

References

[1] https://www.coursera.org/learn/machine-learning/home/week/5
[2] https://www.coursera.org/learn/machine-learning/home/week/6
[3] https://www.coursera.org/learn/machine-learning/home/week/7
[4] https://www.coursera.org/learn/neural-networks-deep-learning/home/welcome