A Project Based Report

on

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

Submitted to AI Technology and Systems

by

Prashant Kalyani

(Batch 4)

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.Different model were developed using the sklearn library and were tuned for the best accuracy.RandomForest and SVM were the algorithm with the highest accuracy

Chapter 1

Introduction

1.1 INTRODUCTION TO PROJECT TOPIC

1.1.1 Introduction to Project

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 result of evaluation concluded that SVM and RandomForest were the best as compared to other as thier accuracy was same.

1.1.2 Aim and Objective(s) of the work

The aim of the project was to compare five different classification algorithm on a one particular database.The project helped to conclude that which algorithm would be best for a paticular dataset

Chapter 2

Performance Of Different Model

2.1 Neural Networks

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.

2.1.1 Training Through Neural Network

A single hidden layer neural net was coded with 25 neurons or hidden unit.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 the best turned out to be 0.3 and 1 The lowest error the optimization algorithm achieved was 0.3

2.1.2 Advantages

Neural net have the ability to learn and model non-linear and complex relationships, which is really important because in real-life, many of the relationships between inputs and outputs are non-linear as well as complex. ANNs can generalize After learning from the initial inputs and their relationships, it can infer unseen relationships on unseen data as well, thus making the model generalize and predict on unseen data.

2.2 Support Vector Machine

2.2.1 SVM model

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 93.6

2.3 Decision Tree Model

2.3.1 Decision Tree

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

2.4 Ensemble Learning

2.4.1 Information

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

2.4.2 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

2.4.3 Bossting Decision Tree

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

Chapter 3

Conclusion

3.1 Conclusion

The svm algorithm had the best accuracy on cross validation set and therefore was selected to predict the output for test set.The accuracy on test set was equal to 93.1