Project Report: Assess Learners

Younes EL BOUZEKRAOUI
ybouzekraoui3@gatech.edu

Abstract—Supervised learning is defined by its use of labeled data sets to train algorithms to do classification and regression and predict a target. Classification and Regression Trees (CARTs) are one of the predictive modelling approaches used in supervised learning. In this Project using a dataset of the returns of multiple worldwide indexes for several days in history, four CART regression algorithms a Decision Tree learner, a Random Tree learner, a Bag learner and an Insane Learner was implemented and evaluated in terms of their performance with respect to some parameters

1 INTRODUCTION

In this project we implemented 4 Classification and Regression Trees learning algorithms using python. The Dataset and arrays are handled using the Numpy library of python, the learners are implemented in the files TLearner.py DTLearner.py InsaneLearner.py BagLearner.py.

The file testlearner.py is performing some tests and experiences the a dataset Istanbul.csv using the learners, the figures are saved on the Figures directory

2 METHODS AND DISCUSSION

2.1 Experiment 1

This experiment consists in testing the DT learner that we implemented and tuning the Leaf size parameter in order to find the best value that minimizes the Root mean squared error (rmse) of the testing data.

In order to evaluate better the model, the experiment was performed multiple times with data shuffling each time and we calculate the mean Rmse over all the rounds.

The leraning curves with respect to the leaf size are shown in the figure below

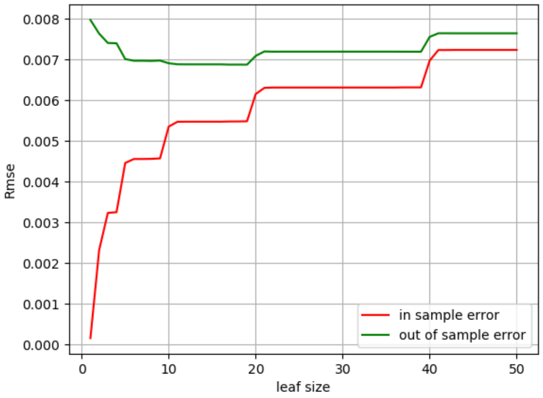

Figure 1—Experiment 1 : Learning curves

One can observe that the in sample error is increasing , and this is due to the fact that by increasing the leaf size the model becomes less attached and dependent to the training data. In contrast, the out of sample error is decreasing until the leaf size value of 19 and starts increasing this means that before this value the model was over-fitting and after this value the model is starting to under-fit. Thus the best value for the size value is 19 to have good fitting of the data.

2.2 Experiment 2

This experiment consists in discussing the use of bagging and its effect on over-fitting We tested bagging by tuning the Leaf size parameter in order to find the best value that minimizes the Root mean squared error (rmse) of the testing data.

We choose a fixed number of bags 20 and variate leaf size in order to evaluate to evaluate.

To evaluate better the model, the experiment was performed multiple times with data shuffling each time and we calculate the mean Rmse over all the rounds.

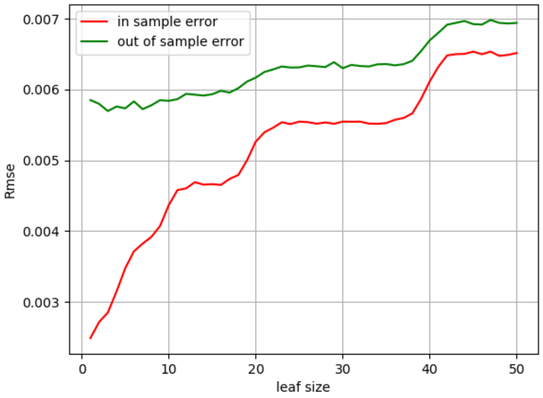

Figure 2—Experiment 2 : Learning curves

On can observe that using bagging the learner is performing better than the previous experience (rmse of testing data < 0.007). We can see that the out of sample error is slightly decreasing at the beginning until the leaf size value of 3 and then increasing.

The area on the left of that value (leaf size =3) is corresponding to the over-fitting area and through the use of bagging we were able the reduce it.

2.3 Experiment 3

The goal of this experiment is to compare the performance of the DT and RT learners. we used the MAE metric and the fitting time to compare the learners.

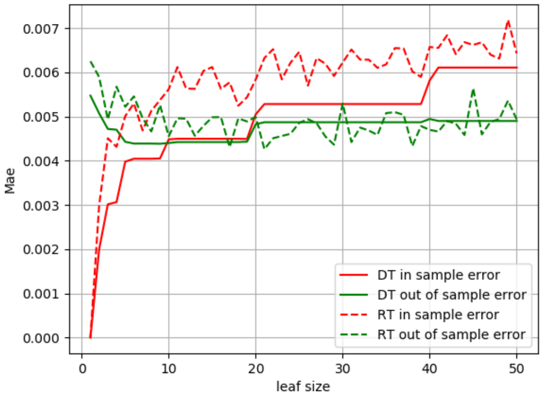

Figure 3—Experiment 3 : Learning curves (MAE)

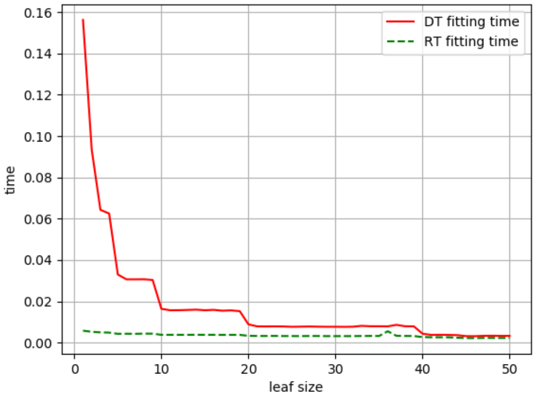

Figure 4—Experiment 3 : Learning curves (time)

We can see the in term of time performance the RT is better but in term of MAE score the DT is performing better. This makes sense because the RT is a choosing the best feature randomly, in contrast the DT is performing calculus to find the best feature to split on

2.4 Summary

In this project we implemented , tested and tuned different Classification and Regression Trees learning algorithms, we saw the benefit of using bagging and how it can reduce or avoid overfitting , we also saw that random learners perform well regarding the time but are not good enough regarding Mae and RMSe metrics.