CS7641 – Machine Learning

Assignment 1 – Supervised Learning

Mason J. Kelchner  
mason.kelchner@gatech.edu

# 

# INTRODUCTION

A variety of supervised machine learning algorithms were applied to two classification data set problems to investigate how they perform under various circumstances, how tuning the algorithms for training affects their performs, and benefits and drawbacks of each as they are applied to a particular dataset. The techniques investigated were Decision Trees with pruning, Boosted Decision Trees, Neural Networks implemented using a Multi-Layer Perceptron structure, Support Vector Machine, and K-Nearest Neighbors. All algorithms were implemented using the Python open source module sklearn.

The two chosen classification problem datasets were for a prediction of credit default and for mushroom classification. These are problems faced in industry today as credit card companies or any company that provides loans faces the risk of their lendees default on their credit, causing a financial loss to the lender. It would be important to implement algorithms to predict whether some parameters describing a lendee or credit applicant would be a higher risk associated with default. This would inform either approving or declining credit or applying a particular interest rate to a loan. Secondly, the mushroom classification problem is very important to foragers and suppliers of mushrooms for a variety of their purposes. In particular, some mushrooms are edible and can be very valuable while other mushrooms which can look very similar to their popular and edible relatives are in fact not edible and in some cases very poisonous to humans. It would advantageous for those interested in obtaining mushrooms to implement supervised machine learning algorithms which can identify mushroom based on various parameters associated with each. Both of these problems are well suited for supervised machine learning classification. The benefits and drawbacks of each algorithm from above as they are applied to both dataset will be discussed in this analysis.

# Experiments

Each classification problem described previously for credit default prediction and mushroom identification were used for evaluating the performance of each supervised learning algorithm. Each algorithm was trained on a subset of the entire data set, the training set, and tested without learning on a test size. For all models, a test size of 20% of the data set was set aside while the models learned on 80% of the data set. Each algorithm had hyperparameters which were varied to understand their effects on the models performance. Model performance for each data set was evaluated by plotting learning curves, validation curves, and training times for each model configuration.

## Decision Tree with Pruning

## Boosted Decision Tree

## K-Nearest Neighbors

## Support Vector Machines

## Neural Network – Multi-Layer Perceptron

# Summary

# References

There are no sources in the current document.