John Joyce & Robert Simari
Machine Learning
Challenge Assignment Report
May 6, 2018

Background

The performance results of each classifier tested on the testing set can be viewed in the Jupyter Notebook that we have submitted.

In summary, we used scikit-learn's built in classifiers to fit both the regression and classification datasets. For regression, we compared a simple linear regressor to a decision tree regressor. For classification, we compared a decision tree classifier to a neural network classifier.

To determine the superior model, we compared the 10 fold cross validation mean score of the models being compared, taking that with the higher performance as our final model. We found this to be a useful metric in that it robustly assesses the accuracy of each model across the entirety of the dataset in a focused manner with high granularity.

Results

Interestingly, the decision tree models showed superior performance on the training data in both regression and classification. Below are the final 10 fold cross validation results for both datasets.

> *Classification*:
> Neural Network Cross Val Score: 0.78 (+/- 0.10)
> Decision Tree Cross Val Score: 0.81 (+/- 0.04)
>
> *Regression:*
> Linear Regression Cross Val Score: 0.72 (+/- 0.12)
> Decision Tree Regression Cross Val Score: 0.96 (+/- 0.03)

Thus, we chose to use the Decision Tree models to predict the testing set in both cases.

We were particularly surprised to find that the decision tree outperformed the neural network classifier, as we had previously believed the neural network to be superior at the task of classification in most cases. We realize now that this is perhaps not the case. It is of note that neither classifier was able to classify the dataset very accurately. The highest cross validation score we saw was 0.81. We believe that this could be higher had we chosen to use alternative types of models.

In the case of the regression dataset, it is certainly easier to understand how a decision tree model could outperform a simpler linear regression model. The decision tree's ability to repeatedly split on

features, honing in on the valid decision space, seems to be better at predicting this dataset than the less malleable, more rigid, linear regressor model.

Conclusions

The predictions of our selected classifiers can be found in the flat files that were submitted.

This was an interesting experiment that shed light on the vast differences that can be seen between different types of machine learning models. It is clear that finding the best model is no trivial task. It is one that depends not only on the implementation of the models itself, but also the type of problem at hand as well as the nature of the datasets.

Final thoughts - this was a super interesting, insightful class that gave us a great introduction into the vast world of machine learning. There is no doubt that learnings from the class will be applicable beyond the university and into our careers.