Competitive Project: Final Report

Dudley Irish

CS6350

December 13, 2024

I. Introduction

In this project we are participating in the Kaggle competition, ML2024F, Income level prediction for mortgage.

II. Experimental Method

The procedure is as specified by the Kaggle competition terms. There are two data files with training and test data. The training data is labeled and test data has an identifying number for each case. The procedure is to train your model on the training data then generate a submission using the test data. Details are available on the Kaggle website.

The data was prepared for use by converting all category fields to number ranges and replacing all continuous fields with binary thresholds. The thresholds were determined by taking the mode of the training data.

Two models were constructed and run using a library Copyright (C) 2013 Quan Wang ¡wangq10@rpi.edu¿, Signal Analysis and Machine Perception Laboratory, Department of Electrical, Computer, and Systems Engineering, Rensselaer Polytechnic Institute, Troy, NY 12180, USA. This library is written in C++ and can be called from either MATLAB or Octave. More information is available at Decision Tree and Decision Forest

The next experiment was with an averaged perceptron. This was implemented as an Octave function.

III. Analysis

At this point in the project we have made three submissions:

(a) The model was a decision tree with 23 levels. The resulting score was 0.73315.

(b) The model was a decision forest with each tree having 23 levels and the forest having 10 trees. The resulting score was 0.73957.

(c) The model was a averaged perceptron with 50 epochs and a learning rate of 1.

The performance of the decision trees has been disappointing. There are a number of meta-parameters that could be adjusted to improve the performance but these initial numbers suggest that we will not use decision trees for our final submission.

The performance of the perceptron fell between the two decision tree models. The number of errors on the training data suggests that the data is not linearly separable since a perceptron is guaranteed to find a border if one exists.

After the perceptron submission the contest deadline passed before any more submissions could be made.

Work on the SVM submission followed the advice given in *A Practical Guid to Support Vector Classification* by Chih-Wei Hsu, Chih-Chung Chang, and Chih-Jen Lin. First the categorical features were converted into multiple columns where only the column corrisponding to the feature value is set to one. They authers say that their experience indicates that this coding is more stable as long as the number of values is not too large. Unfortunately they did not give any guidence as to what counted a "too large." This conversion took the data from fifteen columns to one hundred and eight.

The next advice the authors gave was to scale the data. Care was taken to apply the same scale adjustment to both the training and test data. The authors also sugested started with the RBF model and performing a grid search. The grid search did not result in useful results and I determined that the problem was that the model that was generated simply predected a '-1' for all the training inputs. The polynominal and sigmoid models had a similar failure.

My assumption is that the categorical feature conversion was at the root of the problem, but I did not have enough time to track it down and resolve the issue.

IV. Further Work

The most obvious further work would be to find and correct the problem with the SVM model and to try a neural network model. In both cases there are many parameters to be tuned in order to improve the results.

Another issue is the poor performance of the decision tree model. The error rates that both the decision tree and decision forest sugest that the model was simple predicting '-1' just like the SVM model. After resolving this issue for the SVM model I would apply this same fix and re-try the decision tree/forest models. With some tuning, I would think the the decision forest model is likely to get the best results.

V. Conclusion

The plan for the project was to perform initial experiments with each machine learning technique through out the semester and the concentrate on the approach that seems to work best for this data. Unfortunately, limited resources prevented a full exploration of the plan and leaves us with incomplete results.