Adam Kiehl
4/10/2023
STAA 578

Homework 1 (Income Prediction)

The goal of this modeling exercise was to predict whether an individual's annual income was more or less than $50,000 using various demographic and descriptive predictors. A min/max scaler was applied to all numeric predictors and factor level encoding was applied to all categorical predictors. Ultimately, 100 predictors were used for modeling. The training data were split into training and validation sets to tune and select models for final consideration.

Before any deep learning techniques were employed, K-Nearest Neighbors and Random Forest models were fit. The latter notably achieved a validation accuracy of 85.94%. Next, several MLP structures were explored with varying widths and depths. It seemed that the widest and deepest networks tended to overfit the training data which led to a simple three hidden layer design (256, 128, and 64 neurons) being chosen for further study. Several regularization methods were considered for study. Chiefly, an early stopping criterion based on minimizing validation loss was established for all neural networks. Additionally, normalizing penalties and dropout techniques were tried in various combinations and with various hyperparameters. It was found that both the normalizing penalty (L2 with penalty 0.001) and dropout (rate 0.25) techniques produced robust results in combination with early stopping. The three hidden layer MLPs with those regularization techniques achieved 85.77% and 85.51% validation accuracy, respectively. The regularized neural networks and random forest models all achieved similar validation accuracies, so the random forest model and neural network with normalizing penalties were ultimately chosen for final testing.

The random forsest model chosen considered 22 features at each split and included 1,000 trees. Some of the most important predictors identified by the model were age, capital gains, and marital status which all make sense in a practical context. A test submission was uploaded to Kaggle and achieved 84.41% accuracy. This was a slightly disappointing score but not terribly different from the accuracies observed throughout the analysis. A second test submission was submitted using the neural network with normalizing penalties and early stopping after 14 epochs. This submission achieved a slightly higher 84.63% accuracy. Both models proved to perform nearly equally well on multiple validation sets and were equally viable. The final submission to Kaggle was chosen to be the neural network in the spirit of the topic of the course.