Homework 4 EDA/Classification

For homework 4, we had to train 7 different sklearn classifiers in Python on a provided dataset. The provided dataset is called the ‘Adult Data Set’, or the ‘Census Income’ dataset. The data includes continuous and nominal features to predict whether the person’s income exceeded $50K/year at that time (binomial classification). There are 14 attributes in the dataset. The only missing values seem to be in occupation, which is conveyed with a question mark.

Diagram, schematic

Description automatically generatedFor data exploration analysis, I checked what kinds of data I was working with. There are 6 columns which had numeric data, with the rest being classification. I plotted a scatter matrix of the numeric data, and it seemed that there were common numbers among the 32561 points. The numeric data was all integers, and most of the features seemed to have an even spread of points. There seemed to be only one pair of correlated features: the final weight feature (fnlwgt) negatively correlated with age.

To prepare the data for training, I converted all the classes to a number by the order they appeared in the dataset. I then min-max scaled the entire dataset so that each feature would have a minimum of 0 and a maximum of 1. With this data, the scores were all between 80% and 85% accuracy on test data. The best classifier ended up being AdaBoost, with 85.086% accuracy at the time of running.

I then wanted to see if instead of squishing the classification data between 0 and 1, I could use one-hot dummy features. I prepared dummy columns for each classification feature and then trained each of the 7 classifiers. The overall score for the classifiers was low, however, the AdaBoost classifier had an accuracy of 86.026% on the test data.

I then performed a grid search to find the best hyper-parameters for AdaBoost, however the best hyperparameters ended up being the default parameters.