Layton Rosenfeld

Decision Tree Write-Up

In this project I used a type of decision tree learning, ID3, to build a tree and predict the category of a given example using a series of “learning” cases. I began by using a data set with 14 examples to help someone decide if they should or should not play tennis given the outlook, temperature, humidity, and wind. Each example had a category of “yes” (play tennis) or “no” (do not play tennis). I built a decision using all 14 examples as training data due to the small size of the data set. To make sure my test code was working, I ran the test function using the tennis data after creating the tree and got a percent accuracy of 100% because the testing data was identical to the training data. This tree was only 5 nodes deep at its deepest points, due to the small number of attributes for each example.

I also used a data set containing descriptions of different mushrooms and whether they were edible or poisonous. This data set was much larger than the tennis data, containing 8124 examples, which enabled me use only 15% of the data as training, while still getting high percentages of accuracy with the testing data. The data is randomly shuffled every time the program is run, meaning that different examples are used as training data each time. Therefor, the percent accuracy on testing data slightly differs each time it is run. When I used 15% of the examples used for training, 1218 training examples and 6906 testing examples, I got a percent accuracy of 99.768%, 100% and 99.652%. When I used 1% of the mushroom data for training, the testing accuracy was 97.911%. The testing accuracy fell to 85% when I only used .2% of the data for training. Decreasing the number of training examples decreased the accuracy of the tests because there were less examples for the program to “learn” from. At its greatest depth, the mushroom tree was 11 nodes deep because of the many attributes for each mushroom.

Additionally, I also used a data set that identified the location of a patient’s tumor given a series of medical observations. Even when using 90% of the data for training (305 for training and 34 for testing) the testing accuracy was only 32%, 29%, and 26% due to the number of examples and more importantly, the breadth of possible tumor locations, like: lung, stomach, pancreas, kidney, and many more. Although this number is not very accurate compared to the results from the mushroom dataset, there is a much higher chance of accuracy compared to simply guessing the tumor location. The tree created when running 80% of the data as training has a greatest depth of 31, because it has many attributes for each example and lots of possible categories.