Machine Learning Reflection Hannah Fisher

**Responding to Questions on Haiku**

While writing my program, I periodically tested different aspects of it using the tennis data set. After I thought that the entire program was working, I tested almost exclusively on the mushroom data set, although I used the congress data set for a bit of testing. I chose to use the mushroom data set for testing because it had significantly more data than the congress data set. I avoided the medical data because it kind of freaked me out to try making predictions based on actual people’s lives. Once my program was fully functional, I experimented with varying the number of training and testing examples. To demonstrate these results, I created the following graph of the relationship between the accuracy of the classification and the percentage of the data used for testing.



As can be seen in this graph, when the percentage of data was 0.0045% (for the mushroom data set), the accuracy was consistently above 95%. When looking at the trees created by the ID3 algorithm when the percentage of data used for training was incredibly low, there were significantly fewer nodes in the resultant decision tree, causing the classification accuracy to be lower.

**Problems in the Process**

While working on this project, I had a few setbacks and difficulties. On the first night after the project had been assigned, while trying to finish the function to read the data into my program, I decided to start over. After reviewing the code that I had written in class, and thinking about the big-picture goals of the project, I realized that the way in which I had planned to store the data would be entirely unhelpful. I realized that it would be both faster and easier for me to start over than to try to alter what I currently had. Although it was hard for me to decide to delete all of what I had created, I realized that it was the best option, and it helped me move forward in the project.

My next issue arose after I wrote the function to calculate entropy and information gain. After finishing these two tasks, I tested them, only to find that they didn’t work at all. Upon rereading both the instructions on haiku and my own code, I realized that I had completely misunderstood the fundamental method by which I needed to calculate these two numbers. Once again, I had to delete a large portion of my work and rewrite it from scratch.

After getting all of my code working, I tested on the mushroom data set. While varying the percentage of data used for testing, I was thoroughly surprised. When 60% of the data was used for testing, there was a 100% classification accuracy. However, when 50% of the data was used for testing, there was a 3% classification accuracy. I continued playing around with the percentage of data used in testing, and I determined that there must simply be one precise number at which the accuracy changed from below 10% to almost 100%. Andrew told me that I had a bug causing this problem, and that I should be able to get 100% accuracy with less than 10% of the data used for testing. I decided to open the mushroom data file and see if there were any patterns I could see, even though this seemed pretty ridiculous due to the large amount of data. What I noticed was that the edible mushrooms were concentrated near the top of the file, and the poisonous ones were near the bottom. The solution that I found was to shuffle the list of examples before splitting it into a training and a testing list. After adding this simple shuffling to my program, my classification accuracy shot way up, and the number of examples needed to achieve this high accuracy decreased significantly.