Kaley Findley
CS 3600 Project 4a

Question 6: Performance Analysis

The dummy dataset 1 and 2 fell victim to the curse of dimensionality. With 10 attributes it requires at least 2^10 examples to develop an accurate classification model. With only 20 examples that was impossible. Dummy dataset 1 was just lucky that the 20 example snapshot was similar to the test set with a 1.0 classification rate. The second dummy dataset was not so lucky with a 0.65 classification rate. You can see the overfitting with the tree sizes. Dummy dataset 1 had a size of 3; it didn't have enough examples to figure out what the important questions were. Dummy dataset 2 had a bigger tree with size 11, but it still lacked the examples to find the important attributes.

The car dataset didn't run into the issues as described above and the tree size was reasonable at 406 (for mini metric) and 408 (for entropy and info gain). Based on the number of values each of the 6 attributes could take, the total number of ways that these attribute values could be combined is 576. With tree sizes around 400 the gain functions weeded out the irrelevant attribute assignments to trim down the size of the tree. Throughout the 20 runs with different test sets it averaged a classification rate of 0.947, which is very high.

The Connect4 dataset fell victim to the same problems as the dummy datasets (required 2^42 examples). This overfitting was shown through the classification rate of 0.752 (for gini) and 0.755 (for entropy and info gain). The tree sizes varied a lot between gain functions: 41695 for gini and 41071 for entropy and info gain. Clearly using entropy and information gain illustrated the importance of the difference attribute values better than using gini.

Question 7: Applications

The car dataset describes the quality of cars in terms of a variety of different characteristics (attributes). A dataset similar to this, say a dataset that describes the quality of other products besides cars, can be used in websites that sell those products. The decision tree can be used to organize the products on their website so the consumer can find what they're looking for. It can also be used when they introduce new products to their website and using the DT to figure out where to showcase their new products.

The Connect4 dataset decision tree classifier can be used in BFS to search for subsequent moves to make depending on the situation you're in. Passing in the decision tree to BFS would create a fast and effective Connect4 playing bot. It not only would have domain knowledge from the decision tree and quick decision making.