Vikram Somu

Analysis

For training set 1 and testing set 1, had the following statistics: 100% of the testing data was categorized correctly, so there were no false positives or negatives for both the gini and infogain methods. The size of the tree run on training set 1 was three nodes, and the depth was two for both gini and infogain methods. We were able to classify the data with a relatively small and shallow tree. This means our hypothesis (decision tree) was quite concise, and therefore less susceptible to irrelevant attributes and noise in the data. The training data contained no noise in set 1 and had more of a pattern, it allowed us to create a very accurate hypothesis, which in turn allowed us to categorize our data quite accurately in the testing set.

For training set 2 and testing set 2, the statistics were as follow: 65% of the test data was categorized correctly. However, 50% of positives were false positives and 20% of negatives were false negatives. The size of the tree run on training set 2 was 11 nodes and the depth was 4. These statistics were the same for both gini and infogain methods. Our decision tree hypothesis for this set of data was clearly larger than the size of our hypothesis from the first data set, and also less accurate. Our hypothesis in this case was more susceptible to irrelevant attributes. The training data contained a weaker pattern in this case, which caused us to get not as high of a correct classification rate as in data set 1. If we had a larger sample, there is a much better chance that we would have picked up on a more accurate pattern.