Decision Tree Analysis

Problem 2: Analysis

Part A

The accuracy of the resulting decision tree on the same data is a 100%. Recall that decision-tree-learning builds a tree by selecting the fewest number of attributes that classifies all the examples in the training set. For any example used to train the decision tree, if we know the values of all the attributes that were selected in the decision tree, we can know its classification. While testing an example on the decision tree, we have the value of all of its attributes. So, trivially, we also know the value of the attributes that were selected in the decision tree. Thus, if our algorithm for decision-tree-learning is correct, then the accuracy of the resulting decision tree on the same data must necessarily be 1.

Part B

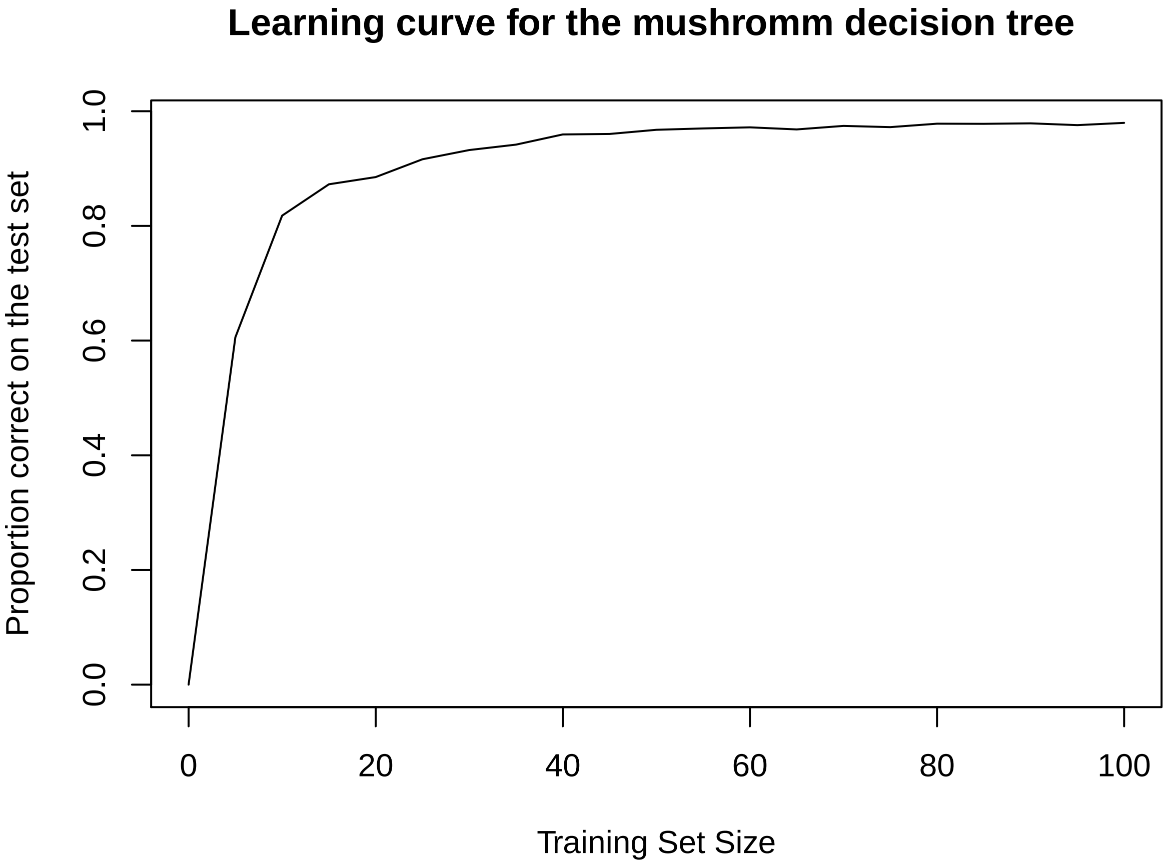

Figure 1: Data consists of classifications of whether a mushroom is poisonous or edible based on various attributes like odor, gill-size, cap-size, cap-color and so on. (Source: Jerod Weinman)

Comments:

The purpose of this experiment is to access the “generalizability” of decision-tree-learning. We use the mushroom data that was provided for Lab-9 to test the algorithm.

First, we divide the data into two disjoint sets. One of those sets is used to train a decision-tree whereas the other set is used to test the resulting decision-tree. We experiment with 21 different lengths of the training set: 0,5,10,15,20,25,. . . ,85,90,95,100. The data for the training set is randomly selected and the reaming data is used to test the resulting decision-tree. For each size of the training set, 50 independent tests are carried out by randomly selecting data for the training set and the average accuracy of the 50 tests is measured.

As expected, the accuracy of the learned decision tree improves as we increase the size of the training set. Even with a training set size of 100, the accuracy of the learned decision tree is very close to 1. This is a good result as the mushroom data consists of around 3000 examples!