University of Colorado at Colorado Springs

Home Work Assignment 2

Due 11-11-04

Write a program in a language of your choice that learns to build a decision tree.

Dataset

You will use a couple of the datasets available at http://www.ics.uci.edu/ mlearn/MLSummary.html. This Web page summarizes the datasets available at the University of California, Irvine, Machine Learning Data Repository. Download two of the following datasets to use in your program:

- Cylinder Bands Database,
- Mushroom Database,
- Glass Identification Database,
- Dermatology Database,and
- Echocardiogram Database

Things to Do

You will write programs for the following. You can do more if you want for extra credit.

1. Write a decision tree builder. To build even a minimal tree building program, you need to make decisions such as how to split the nodes, when to stop, etc.
2. Use the splitting criterion given in Mitchell's handout I gave in the class.
3. Use a pruning method to prune the tree you produce.
4. Using a program, convert the decision tree into a set of rules.
5. There are several approaches to dealing with attributes with missing values. These are discussed in page 75 of Mitchell's text. Use the probability-based approach discussed.
6. There are several approaches to handling continuous-valued attributes. These are discussed on page 72 of Mitchell's text. Think of an appropriate way for you to handle the data you have.
7. Divide the data into randomly into three parts. Use two parts for training and one part for testing.

What to hand in

You will submit a small paper with a title and your name. In this paper, you will have a short section with an appropriate heading for each of the most important steps in your program. Describe in this section how you carry out the step in question. You will also have a section called *Results*. Here, you will carry out the tree-building and testing experiments 10 times for each dataset, and report the results in the form of tables and graphs, as appropriate.

You should report the results of classification in terms of *precision*, *recall*, *F Measure*, and any other measures discussed in Salton's handout.

In addition, please provide a print-out of your code. You will be required to do a demo during my office hours.