University of Colorado at Colorado Springs

Home Work Assignment 1

Due 03-16-2010

1 Decision Tree Learning

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

1.1 Dataset

You will use a dataset 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 one of the following (suggested) datasets to use in your program:

- Cylinder Bands Database,
- Mushroom Database,
- Glass Identification Database,
- Dermatology Database, or
- Echocardiogram Database.

You can choose to use any other data set from the UCI archive or from any other source if you like, instead of one of the above. The data set should have several attributes; it also must have missing attribute values and continuous-valued attributes.

1.2 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 a node (set) splitting criterion of your choice. You can look at Mitchell's [2] and Alpyadin's [1] texts. You can also look at the survey paper by Rokach and Maimon [3] on decision tree learning that I handed out in class. Another paper you can read is the CART paper by Steinberg [4]. You need to discuss which method you use and why.
3. There are several approaches to dealing with attributes with missing values. Some of these are discussed in page 75 of Mitchell's text. Others are discussed in the survey paper and the CART paper. Use any method you deem appropriate and discuss what you do and why.

4. There are several approaches to handling continuous-valued attributes. Some of these are discussed on page 72 of Mitchell's text. Others can be found in the survey paper and the CART paper. Think of an appropriate way for you to handle the data you have.

5. Divide the data into randomly into three parts. Use two parts for training and one part for testing. Repeat the training and testing ten times, each time selecting training and test sets randomly.

6. Use a pruning method to prune the tree you produce. Document if pruning improves the classification done by the tree.

1.3 What to Hand in

You will submit a 3-5 page paper with a title and your name. Use the IEEE Author style you have been using for the semester project papers you have been writing for the class. 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 you think is appropriate, such as ROC curves.

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

References

[1] Aplaydin, Ethem. 2010. *Introduction to Machine Learning*, Second Edition, The MIT Press, Cambridge, MA.

[2] Mitchell, Tom. M. 1997. *Machine Learning*, WCB McGraw-Hill, Boston, MA.

[3] Rokach, Lior, and Oded Maimon. 2005. Top-Down Induction of Decision Tree Classifiers–A Survey, *IEEE Transactions on Systems, Man, and Cybernetics–Part C: Applications and Reviews*, Volume 35, No. 4, November 2005, pp. 476-487.

[4] Steinberg, Dan. 2009. CART: Classification and Regression Trees, pp. 179-201, Taylor and Francis Group.