CS 550 -- Machine Learning
Homework #1

Due: 15:40 (class time), March 16, 2017

1. Implement a decision tree classifier that uses prepruning approach. In your implementation, you will use your selected splitting criterion and prepruning technique. You should give the details of your selection. You may use any programming language you would like.

   You will conduct your experiments on the "Thyroid data set", which is taken from the UCI repository and available on the course web page. The details of this data set are given as follows:

   - It contains separate training ("ann-train.data") and test ("ann-test.data") sets.
   - The training set contains 3772 instances and the test set contains 3428 instances.
   - There are a total of 3 classes. Note that this dataset has unbalanced class distributions. Thus, you may want to consider this in training your decision tree classifier.
   - In the data files, each line corresponds to an instance that has 21 features (15 binary and 6 continuous features) and 1 class label.

   In this part,

   (a) Draw the decision tree that you learn on the training instances.

   (b) Obtain training and test set accuracies using your implementation. Report the class-based accuracies as well as the confusion matrices for the training and test sets.

   (c) Explain how you manage the situation of having unbalanced class distributions. Also explain if you use any normalization on the features.

   (d) If any step of your implementation (including its prepruning technique) consists of the selection of a parameter(s), list all the parameter(s) and select them using 3-fold cross validation on the training instances. List the parameter values you consider in 3-fold cross validation and report the selected values.

   (e) Make sensitivity analysis on the test instances. That is, explore how test set accuracies (class-based accuracies as well as the overall accuracy) change as a function of your parameter(s).

2. Use a machine learning toolbox (e.g., PRTools, Weka) for a decision tree classifier. In this part, you will explore decision tree classifiers with different options, which are provided by your selected toolbox. Conduct your experiments again for the Thyroid data set, give a list of what you try, and report your results on the training and test sets (class-based accuracies and confusion matrices).

3. Select an application and find at least two research papers on this application that use decision trees as a part of their proposed methods. Write a brief summary of how these papers use the decision trees. Give the citations of your selected papers.

You are expected to write your report neatly and properly. The format, structure, and writing style of your report as well as the quality of the tables and figures will be a part of your grade. Additionally, you should follow the following instructions to prepare your report.

- **Part 1:** You should explain the details of your implementation and your experimental findings. You SHOULD NOT give the screen shots or outputs of your program but you should summarize what you have found at the end of your runs. Do not forget to address the questions asked through the items (a)-(e). This part should be a maximum of 3 pages.

  Additionally you need to email the source code of your implementation. The subject line of your email should CS 550: HW1. Do not submit the printout of your source code.

- **Part 2:** You should explain the details of how you use the machine learning toolbox and your experimental findings. You SHOULD NOT give the screen shots or outputs of the toolbox but you should summarize what you have observed when you use this toolbox. This part should be a maximum of 2 pages.

- **Part 3:** You should provide a summary between 250-300 words (excluding citations).

- **All parts:** You should use reasonable font sizes, spacing, margin sizes, etc. You may submit either a one-column or a double-column document.

Please submit the hardcopy of your report before the deadline. DO NOT submit the printout of your source code.