CS 550 -- Machine Learning
Homework #1

Due: 23:59, November 23, 2020

In this homework, you will use and implement a decision tree classifier. 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:

- This data set 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.
- 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 the third part of this homework, you will construct a decision tree also considering the cost of using (extracting) the features. The cost of using each feature is given in another file ("ann-thyroid.cost"). It does not include the cost of the 21st feature because it is a combination of the other features.
- The 21st feature is defined using the 19th and 20th features. This means that you do not need to pay for this feature if the 19th and 20th features have already been extracted. Otherwise, you have to pay for the cost of the unextracted feature(s).

Part 1: 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. Using this toolbox,

- Draw the decision tree that you will have learned on the training instances (with the best configuration of the parameters that you will have selected).
- Obtain training and test set accuracies. Report the class-based accuracies as well as the confusion matrices for the training and test sets (with the best configuration of the parameters that you will have selected).
- Compare the training and test set class-based accuracies when pruning and no pruning is used. Indicate which pruning method that you will have used and report the value of its parameter, if it has any.
- Compare the training and test set class-based accuracies when normalization and no normalization is applied on the feature values. Is there any difference? Interpret these results.
- The training dataset has unbalanced class distributions. Compare the training and test set class-based accuracies when you use the training dataset as it is and when you balance the number of classes in the training dataset. Is there any difference? Interpret these results.

Part 2: Implement your own decision tree classifier that uses prepruning. In your implementation, you will use your selected splitting criterion and prepruning technique. Give the details of your selection. Using your implementation,

- Draw the decision tree that you will have learned on the training instances (with the best configuration of the parameters that you will have selected).
- Obtain training and test set accuracies. Report the class-based accuracies as well as the confusion matrices for the training and test sets (with the best configuration of the parameters that you will have selected).

Part 3: Extend your implementation in Part 2 such that now it also considers the cost of using a feature as a splitting criterion (of course together with the purity of a split). Give the explicit form of your new splitting criterion. Using this extended implementation,

- Draw the decision tree that you will have learned on the training instances (with the best configuration of the parameters that you will have selected).
- Obtain training and test set accuracies. Report the class-based accuracies as well as the confusion matrices for the training and test sets (with the best configuration of the parameters that you will have selected).
- Compute the cost of classifying each instance with this decision tree. On the test set, report the average cost of classifying an instance, separately for each class. (Take the average considering the actual classes of instances.)

The first part of this homework asks you to use a toolbox but its second and third parts ask you to implement a decision tree classifier by **writing your own codes**. Thus, in the second and third parts, you are not allowed using any machine learning package. In your implementation, you may use any programming language you would like.

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. Use reasonable font sizes, spacing, margin sizes, etc. You may submit either a one-column or a double-column document. In your report, do not give any screen shots. Do not forget to address the questions specifically asked to you. Your report should be a maximum of 5 pages.

Please email the pdf of your report and the source code of your implementation before the deadline. The subject line of your email should CS 550: HW1.