CS550-Machine Learning
Homework Assignment #1
Due: 23:59, March 30, 2023

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 provided to you via Moodle. The details of this dataset are as follows.

- The 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 21^{st} feature is defined using the 19^{th} and 20^{th} features. This means that you do not incur additional costs for this feature if the 19^{th} and 20^{th} 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 of your choice (e.g., PRTools, Weka, etc.) to construct 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 learn on the training instances (with the best configuration of the parameters that you 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 are used. Indicate which pruning method you have used and report the value of its parameter if it has any.
- Compare the training and test set class-based accuracies when normalisation and no normalisation are applied to the feature sets. Is there any difference? Interpret these results.
- The training dataset has imbalanced 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 pre-pruning. In your implementation, you will use your selected splitting criterion and pre-pruning technique. Give the details of your selection. Using your implementation:

- Draw the decision tree that you have learned on the training instances (with the best configuration of the parameters that you will select).
- 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 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 have learned on the training instances (with the best configuration of the parameters that you will select).
- 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 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 the 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 to use any machine

learning package. In your implementation, you may use any programming language you 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. The IEEE manuscript templates are preferred. In your report, do not give any screenshots. Do not forget to address the questions specifically asked in each part. Your report should not exceed 5 pages. Submit the pdf of your report along with the source code of your implementation to Moodle by the deadline.