DSCI 440 Implementation Assignment 3 (100 points)
Due 11:59 pm March 23, 2023

General instructions.

This is an individual assignment. You will only need to submit one copy of the source code and report. Please, type your report in LaTeX.
2. Your source code and report must be submitted through the Moodle site.
3. Be sure to answer all the questions in your report. You will be graded based on your code as well as the report. So please write your report in a clear and concise manner. Clearly label your figures, legends, and tables.

Decision Stump and Decision Tree

For this assignment you will implement (1) the decision stump learning algorithm and (2) the decision tree learning algorithm with early stopping. You will test your implementation on the SPECT data sets, with 22 binary features. You will train your classifies using the SPECT-train.csv file, and test on the SPECT-test.csv file. The first column of each data set contains the class label, the remaining columns are the features.

The assignment has three parts:

1. Decision stump:
Implement the algorithm for learning decision stump, i.e., a decision tree with a single test. To build a decision stump, simply apply the top down decision tree induction algorithm to select the root test and then stop and label each of the branches with its majority class label. Please use the information gain as the selection criterion for building the decision stump. You may find it useful to implement a separate function that takes a feature as a parameter and returns the information gain for that feature. Please report the information gain of each binary feature, the learned decision stump, and the training and testing error rates of the learned decision stump (as a percentage of misclassified examples).

2. Decision tree with early stopping (at depth 3):
Implement the top-down greedy induction algorithm for learning decision tree with depth $d = 3$ (where level 1 is a root node of the tree and level 3 contains the leaves). Please use the information gain as the selection criterion for building the decision tree. Please provide your learned decision tree and for each test node its information gain. Provide the training and testing error rates of the learned tree.

3. Comparison:
Compare decision stump rates with decision tree rates. What behavior do you observe? Discuss.

Your report should have the following structure:

(a) You full name and assignment number.

(b) Introduction (Briefly state the problem you are solving).

(c) For the decision stump please report the information gain of each binary feature (as a table), the learned decision stump, and the training and testing error rates of the learned decision stump (as a percentage of misclassified examples).

(d) For the decision tree please report your learned decision tree and for each test node its information gain. Provide the training and testing error rates of the learned tree.

(e) You discussion for part 3.