Machine Learning: Programming Assignment 2: Decision Trees

Problem Statement:

Write a program to learn a decision tree and use it to predict class labels of test data. Decision tree learning should use *information gain* as the criterion for choosing the attribute for splitting. If there is a tie among two attributes, the lower attribute number should be chosen (e.g., if there is a tie between x2 and x5; x2 should be chosen). *Tree pruning should not be performed*. The learned tree should be tested on test instances with unknown class labels, and the predicted class labels for the test instances should be printed as output.

Data Set Description:

Training Data Filename: *data2.csv*

Training Data File Format: Boolean input attributes (x1, x2, … , x8) in first 8 columns. The last (9^{th}) column represents the Boolean class label (y). Each row is a training instance. There are 24 training instances.

Test Data Filename: *test2.csv*

Test Data File Format: Boolean input attributes (x1, x2, … , x8) in each of the 8 columns. Note that, there is no class label column. Each row is a test instance. There are 4 test instances. The row number corresponds to the instance number of the test instances.

Please STRICTLY follow the program input/output format specified below.

Input Format: Assume the data files data2.csv and test2.csv is present in the same directory and contains the training and test data. Thus, your program should not require any input from user and should read from these files. Strictly use these filenames only.

Output Format: Predicted class labels (0/1) for the test data exactly in the order in which the test instances are present in the test file. Put a blank space between printed the class labels. (e.g., output 0 0 1 1, if the predicted class labels are - Test Instance 1: 0, Test Instance 2: 0, Test Instance 3: 1, Test Instance 4: 1). Output, in above format, should be printed to the file: *rollnumber_2.out* (e.g., 14CS10001_2.out). Strictly use this filename format.

Submission Guidelines:

You may use one of the following languages: c/C++/Java/Python. You should name your file as <rollnumber_2.extension> (e.g., 14CS10001_2.c). Your program should be standalone and should not use any *special purpose* library. numpy may be used. You should submit the program file only and not the output/input file. The submitted single program file *should* have the following header comments:

Roll # Name # Assignment number # Specific compilation/execution flags (if required)

*Please submit the program in moodle by **August 28, 2018 midnight** (hard deadline). Copying from friends/web will lead to strict penalties.*