CS 5593 - Fall 2019 - Dr. Le Gruenwald

HOMEWORK 2

Assigned: 9/30/2019; Due: 10/9/2019 by 12:00 PM (NOON) to the class website on Canvas

Maximum Points: 100 points

Notes:

- **Homework answers must be typed and submitted by 12:00 PM (NOON) to the class website on Canvas.**
- **Homework is individual work; it must be done by you only, no collaboration with anyone else is allowed.**
- **Late homework will be accepted on the class website on Canvas until 11:59 PM on the date following the due date with 5% (of the maximum points) penalty. Late homework submitted after this time will not be graded.**

Problem 1 (30 points): Using the following dataset where the class attribute is "Survived" and using the Decision Tree Induction Algorithm 3.1 given on Page 137 in the textbook, answer the following questions:

1.1. Show your construction of a decision tree using the information gain for the attribute split test condition and the following stopping condition for a node: either all records in the node have the same class label or the same attribute values or the number of records in the node is less than 3. Show your work (including the information gain calculation) at each split step by step so that we understand how you have constructed the tree. If you show only the final tree, you will get zero credit for this question.

1.2. The same as Question (1.1), but use the gain in Gini index for the attribute split test condition.

1.3. Using the final tree that you have constructed for Question (1.1), compute the generalization error rate of the tree using the pessimistic approach assuming that the penalty term associated with each leaf node is 0.5.

Instance	Gender	Age	Ticket Class	Survived
1	Male	45	Second	**Yes**
2	Female	8	First	**Yes**
3	Male	32	Second	**No**
4	Male	26	Third	**No**
5	Female	55	Second	**Yes**
6	Female	47	Third	**No**
7	Male	20	First	**No**
8	Female	24	First	**Yes**
9	Female	43	Second	**No**
10	Male	12	First	**Yes**
11	Male	34	Second	**No**
12	Female	65	Third	**No**

Problem 2 (70 points):

2.1. C5.0 and CART are two well-known decision tree algorithms. Read the published literature about these two algorithms and answer the following question: for each algorithm, provide an overview describing how the algorithm works, discuss the impurity measure it uses for the attribute split test condition, and discuss one advantage and one disadvantage of the algorithm. Provide the references to the published literature to justify your answers.

2.2. Write an R program to perform the following tasks (a)-(g) on the Leaf dataset from the UCI Machine Learning repository (https://archive.ics.uci.edu/ml/datasets/Leaf):

a. Using the Boxplot visualization method, in a single figure, draw a boxplot of each of the following attributes: Aspect Ratio, Elongation, and Solidity.

b. From the boxplots of the three attributes of Task (a), identify which attributes have outliers, which attribute values are outliers, and justify your answers. If there are outliers, write your R code to remove the entire tuples containing the outliers from the dataset, and print the dataset after those tuples have been removed.

c. Using the preprocessed dataset obtained from Task (b), repeat Task (a) and provide your interpretation of the new boxplots.

d. Using the preprocessed dataset obtained from Task (b) and using the C5.0 algorithm (available from the package C5.0), build a decision tree that classifies the tuples based on the class attribute "class" in the dataset. Print the resulting decision tree in the textual format and the graphical format. Then evaluate the error rate using k-fold cross-validation with k = 3. For each fold, print the confusion matrix to standard output, then calculate, print, and store the error rate.

e. Repeat Task (d) using the CART algorithm (available from the package 'rpart').

f. Once you have carried out the above tasks (a)-(e), use hypothesis testing as discussed in Chapter 3 in the textbook to determine whether or not the error rate difference between the two classification algorithms is statistically significant given the confidence level of 98%. Your R program must print the confidence level, calculate and print the confidence interval of the error difference, and print a message to indicate whether or not the error rate difference is significant based on the calculated confidence interval and which model (the tree produced by C5.0 or the tree produced by CART) is your selected model. Note that this question asks for a two-sided confidence interval, not a one-sided one, so be careful when reading the probability table or using the appropriate R command.

g. For predictions of class labels of future tuples, extend your R program so that it can accept a tuple as input, traverse the tree that you have selected in Task (f) to find out the class label of the tuple, and print the tuple together with its predicted class label. Conduct testing of your R code for this question by running your R program three times with three different input tuples.

Notes on Submission: Submit one complete PDF document that contains the answers to all the two problems; for Problem 2, this complete document needs to also contain the R program including the R statements to load the dataset, screenshots/scripts of your R program executions, and the required output with appropriate labels. The R program must include appropriate in-line comments for documentation. In addition, besides this complete PDF document, submit a separate .r text file containing your R program for Problem 2 as we will test your program for correctness. Failure to submit this R program file will result in a zero grade for your Problem 2. DO NOT SUBMIT ZIP FILES.

Notes on References: An additional reference on R:

Larry Pace, "Beginning R: An Introduction to Statistical Programming," APress, 2011 (available online on OU Library Website).