CS 5593 –Section 001 - Fall 2020 - Dr. Le Gruenwald
HOMEWORK 2
Assigned: 10/5/2020; Due: 10/14/2020 by 2:00 PM to the class website on Canvas
Maximum Points: 100 points

Notes:

- **Homework answers must be submitted by 2:00 PM to the class website.**
- **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 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 (40 points) Using the following dataset where the class attribute is "Treatment Applied" and using the Decision Tree Induction Algorithm 3.1 given on Page 137 in the textbook, answer the following questions:

a) 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 4. 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.

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

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

d) Compute the total description length of each tree generated for Questions (a) and (b) according to the minimum description length (MDL) principle described on Page 189 in Exercise 10 in Chapter 3 of the textbook. Answer which decision tree is better based on the MDL principle.

The Dataset for Problem 1:

Instance	Did the horse have surgery?	Pulse (beats per minute)	Abdominal distension	Treatment Applied
1	No	92	None	**Level 3**
2	Yes	88	Severe	**Level 2**
3	Yes	64	Severe	**Level 2**
4	No	48	Slight	**Level 1**
5	No	76	Slight	**Level 4**
6	Yes	76	None	**Level 1**
7	Yes	88	Severe	**Level 3**
8	No	48	Severe	**Level 1**
9	No	92	Severe	**Level 4**
10	No	48	Slight	**Level 1**
11	Yes	64	Slight	**Level 1**
12	Yes	64	Slight	**Level 4**

Problem 2 (60 points):

1. ID3 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, discuss the impurity measure it uses for the attribute split test condition and discuss one advantage and one disadvantage of the algorithm. Provide references to the published literature to justify your answers.
2. Write an R program to perform the following tasks (a)-(g) on the Wine dataset from the UCI Machine Learning repository (https://archive.ics.uci.edu/ml/datasets/Wine):
    a. In a single figure, draw a boxplot of each of the following attributes: "Magnesium", "Color Intensity" and "Malic Acid."
    b. From the boxplots of the attributes of Task (a), identify which attributes have outliers, which attribute values are outliers, and justify your answers. If there are outliers, write 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 CART algorithm (available from the package "rpart" in R which you need to install), build a decision tree that classifies the tuples based on the class attribute of the dataset. Print the resulting decision tree using the graphical format. Then evaluate its error rate using k-fold cross-validation with k = 5. For each fold, print the confusion matrix to standard output, then calculate, print and store the error rate.
    e. Repeat Task (d) using the ID3 algorithm (available from the "simpleEducationalLearningSchemes" package in the "RWeka" package which you need to install; see the code for installing and using this algorithm in R given in the last page of this homework).
    f. Once you have carried out the above Tasks (d) and (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 99%. 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 CART or the tree produced by ID3) 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: your answers for Problem 1 can be typed or hand-written; if they are hand-written, their scanned copy must be readable; if we cannot grade them because they are not readable, you will get a zero credit for your answers. Your answers for Problem 2 must be typed. Submit one complete PDF document that contains the answers to BOTH problems; for Problem 2, this complete PDF 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).

Code for Installing and Using the algorithm ID3 in the "simpleEducationalLearningSchemes" package in the "RWeka" package in R (https://stat.ethz.ch/pipermail/r-help//2014-September/421343.html)

```
## install RWeka package
Install.packages("RWeka")

## code to use the algorithm ID3 in the RWeka package

library(RWeka)
## look for a package providing id3
WPM("refresh-cache")
WPM("list-packages", "available") ## look for id3
## install package providing id3
WPM("install-package", "simpleEducationalLearningSchemes")
## load the package
WPM("load-package", "simpleEducationalLearningSchemes")
## make classifier
ID3 <- make_Weka_classifier("weka/classifiers/trees/Id3")
## test it out
DF2 <- read.arff(system.file("arff", "contact-lenses.arff",
package = "RWeka"))

DF2
x = ID3(`contact-lenses` ~ ., data = DF2)
```