CS 5593 – Section 001 - Fall 2023 - Dr. Le Gruenwald

HOMEWORK 2

Assigned: 10/2/2023; Due: 10/11/2023 by 3:00 PM to the class website on Canvas

Maximum Points: 100 points

Notes

- **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.**
- **ACADEMIC INTEGRITY: Homework is individual work; it must be done by you only; no collaboration with anyone is allowed. The use of generative AI tools (including ChatGPT, Bard, Bing Chat, and other AI writing and coding assistants) is not allowed in all homework questions, except for those questions where the use of these tools is explicitly required. Violations of any of these rules will be considered academic misconduct and will result in action as specified in the Academic Integrity Code at The University of Oklahoma: http://www.ou.edu/integrity. Consult also the following web page for a Student's Guide to Academic Integrity at The University of Oklahoma: http://www.ou.edu/integrity/students.**

Problem 1 (15 points): Imagine that you are working for a company that has collected the email messages it received in the past five years and has manually labeled each of them to be either "spam" or "not spam." Your boss now asks you to use Decision Tree Induction to automatically classify new email messages (let's hope that your boss knows what that means!) Use ChatGPT to help you with this task. Write your prompts to ask ChatGPT to tell you how you should approach solving the problem. You can try at most three prompts. Then use the final answer that ChatGPT gave and the knowledge you gained from our course including the textbook and the lectures, explain in detail the following: a) which parts of the ChatGPT's answer are correct, b) which parts are wrong, c) which parts need to be elaborated, and d) what is missing. For each of (a-d), you need to justify your answers in detail. In addition, for (c), you also need to provide your elaboration, and for (d), you also need to provide what you would add to fill in the missing parts. Be precise in your answers. Show your prompts, ChatGPT's answers, and your answers.

Problem 2 (15 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, show your manual construction of a decision tree using Gini index for the attribute split test condition. For the continuous attribute, follow the split test procedure for the example given on Pages 132-134 in the textbook to identify the best split value. Use 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 or there is no more remaining attribute to further partitioning. Show your work (including the Gini index calculation) at each split step by step so that we understand how you have constructed the tree manually. If you show only the final tree, or if you have used a program to construct the tree for you instead of constructing it manually step-by-step yourself, you will get zero credit for this question.

Instance	Gender	Age	Survived
1	Male	45	**Yes**
2	Female	8	**Yes**
3	Male	32	**No**
4	Male	26	**No**
5	Female	55	**No**

Problem 3 (70 points):

3.1. 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.

3.2. Write an R program to perform the following tasks (a)-(g) on the Wholesale Customer dataset from the UCI Machine Learning repository (https://archive.ics.uci.edu/dataset/292/wholesale+customers):

a. Using the Boxplot visualization method, in a single figure, draw a boxplot of each of the following attributes: Milk, Fresh, and Delicatessen.

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 "Region" in the dataset. Print the resulting decision tree in 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: your answers for Problem 2 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 Problems 1 and 3 must be typed. Submit one complete PDF document that contains the answers to the THREE problems; for Problem 3, 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 3 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).