BUAN 6357 (Johnston)
Homework 3 (20230207)
Code Due: 18 February 2023 (6PM)
Part B Due: 19 February 2023

Points available: 150. (Part A)

This assignment is about evaluating both the classification accuracy and the risk of misclassification across 3 modeling strategies and multiple training sample sizes (number of replications per digit). Use the "03j_noisy_segments_multiclassif" file from UTDbox>pre_recorded_materials>demo_03_classification as a starting point. Part A uses 2 sample sizes per digit: 25 and 50. Part B extends the number of sample sizes. Seed the RNG with 975773311 . Each sample must start with the same RNG value as they are considered continuations of the earlier (smaller) samples. Be prepared to generate classification tables (confusion matrix) and calculate both Bayes Risk and percent correct on globally (overall), conditionally, and individually (observation level) for each sample size.

For this assignment you will need the package "partykit" and may use the packages "tidyverse" and "data.table", in that order. You should not use any additional packages. You should use only the "require()" or "library()" statement in your code. Any use of the install.packages() function in submitted code will result in a score of 0 for that submission. Multiple instances of concurrently submitted code will result in only the most recent code being evaluated and scored.

The first commands of your code submitted for grading to eLearning MUST be:

setwd("c:/data/BUAN6357/HW_3"); source("prep.txt", echo=T)

and the last command of your code MUST be:

source("validate.txt", echo=T)

Be careful with the quote characters as they must ALL be the same at the beginning and end of a string. (Use the single or double quote character from the key close to "Enter".) Inclusion of these lines is required BEFORE your code will be tested.

Submit the code to eLearning as an ASCII file which can be copied directly into R.

You may submit this assignment as many times as needed until you get full credit. When multiple submits for a single assignment are waiting to be evaluated ("in the queue"), only the last one will be evaluated.

Generate each sample from a re-started RNG using the specified seed value.

Deliverables (all classification probabilities are normalized, component vectors are inside the data.frame (or compatible) named for the number of digit replications, e.g. "s25" is the result of running the 25 replications scenario so for Part A (this assignment) there are 2 data.frames required but the listed components will each be validated separately:

1. seed — random number generator seed
2. s25$digit — (component vector) actual digits
3. s25$lCl — (component vector) 10 logit classifications
4. s25$lPr — (component vector) 10 logit classification probabilities
5. s25$t10Cl — (component vector) 10 tree classifications
6. s25$t10Pr — (component vector) 10 tree classification probabilities
7. s25$t1Cl — (component vector) 1 tree classifications
8. s25$t1Pr — (component vector) 1 tree classification probabilities
9. s50$digit — (component vector) actual digits
10. s50$lCl — (component vector) 10 logit classifications
11. s50$lPr — (component vector) 10 logit classification probabilities
12. s50$t10Cl — (component vector) 10 tree classifications
13. s50$t10Pr — (component vector) 10 tree classification probabilities
14. s50$t1Cl — (component vector) 1 tree classifications
15. s50$t1Pr — (component vector) 1 tree classification probabilities

Note: these deliverables are intended to demonstrate your ability to successfully modify the original code, extract collections of summary values, and manage results.

Part B of HW 3 will direct you to explore both the intermediate results, the deliverables from Part A and answer questions about each of them. You may submit answers to HW 3 part B as many times as you wish but only the score for the last submitted code will be retained.

For Part B, use additional digit repetition values of 100, 250, 500, 1000, 2500, and 5000. You may need to perform some additional analysis to answer all assessment questions.

Hint: The demo code runs models for a single sample size and collects most of the required information. How can you re-use this code and manage the results?

Hint: Be familiar with all materials (spreadsheet and videos) in UTDbox>pre_recorded_materials>05b_classification_accuracy_and_risk.

Comment: This assignment has multiple goals. The calculation of accuracy and Bayes Risk values is a large portion but the management of this information for easy access is also important. The use of accuracy measures for each sample size to select an optimal sample size for each of the classification techniques is also important. For a similar application in another context, consider the use of elbow plots in selecting a number of clusters for either kmeans() or hclust().