Assignment 2 Instructions

Purpose

The purpose of this assignment is to assist you with mastering the topics in Modules 4 and 5.

Learning Outcomes

This assignment aligns to the following module learning outcomes:

- Module 4
  - Explain the key idea behind bootstrap aggregation. (CLO 1, 2, 3)
  - Describe what it means by ensemble modeling. (CLO 1, 2, 3)
- Module 5
  - Explain how decision trees work. (CLO 1, 2, 3)
  - Explain the key idea behind random forest models. (CLO 1, 2, 3)
  - Explain how boosting tree algorithm work. (CLO 1, 2, 3)
  - Demonstrate how random forest and bosting tree models can be implemented. (CLO 4)

Instructions

Part A

Please read the following questions carefully and answer each question.

QA1. What is the key idea behind bagging? Can bagging deal both with high variance (overfitting) and high bias (underfitting)? (10% of total points)

QA2. Why bagging models are computationally more efficient when compared to boosting models with the same number of weak learners? (5% of total points)

QA3. James is thinking of creating an ensemble mode to predict whether a given stock will go up or down in the next week. He has trained several decision tree models but each model is not performing any better than a random model. The models are also very similar to each other. Do you think creating an ensemble model by combining these tree models can boost the performance? Discuss your answer. (5% of total points)

QA4. Consider the following Table that classifies some objects into two classes of edible (+) and non- edible (-), based on some characteristics such as the object color, size and shape. What would be the Information gain for splitting the dataset based on the "Size" attribute? (15% of total points)

Color	Size	Shape	Edible?
Yellow	Small	Round	++
Yellow	Small	Round	– –

Color	Size	Shape	Edible?
Green	Small	Irregular	++
Green	Large	Irregular	−−
Yellow	Large	Round	++
Yellow	Small	Round	++
Yellow	Small	Round	++
Yellow	Small	Round	++
Green	Small	Round	−−
Yellow	Large	Round	−−
Yellow	Large	Round	++
Yellow	Large	Round	−−
Yellow	Large	Round	−−
Yellow	Large	Round	−−
Yellow	Small	Irregular	++
Yellow	Large	Irregular	++

QA5. Why is it important that the m parameter (number of attributes available at each split) to be optimally set in random forest models? Discuss the implications of setting this parameter too small or too large. (5% of total points)

Part B

This part of the assignment involves building decision tree and random forest models to answer a number of questions. We will use the Carseats dataset that is part of the ISLR package (you need to install and load the library). We may also need the following packages: caret, dplyr and glmnet

Let's start by loading these libraries:

```
library(ISLR)
## Warning: package 'ISLR' was built under R version 4.0.3
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.0.2
## Warning: replacing previous import 'vctrs::data_frame' by 'tibble::data_frame'
## when loading 'dplyr'
##
## Attaching package: 'dplyr'
## The following objects are masked from 'package:stats':
##
##     filter, lag
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
library(glmnet)
## Warning: package 'glmnet' was built under R version 4.0.2
## Loading required package: Matrix
## Loaded glmnet 4.0-2
library(caret)
## Warning: package 'caret' was built under R version 4.0.3
## Loading required package: lattice
## Loading required package: ggplot2
## Warning: package 'ggplot2' was built under R version 4.0.2
```

For this assignment, we only need the following attributes: "Sales", "Price", "Advertising", "Population", "Age", "Income" and "Education". The goal of the assignment is to build models to predict the sales of the carseats ("Sales" attribute) using the other attributes.

We can use the dplyr select function to select these attributes.

```
Carseats_Filtered <- Carseats %>% select("Sales", "Price",
"Advertising","Population","Age","Income","Education")
```

QB1. Build a decision tree regression model to predict Sales based on all other attributes ("Price", "Advertising", "Population", "Age", "Income" and "Education"). Which attribute is used at the top of the tree (the root node) for splitting? Hint: you can either plot () and text() functions or use the summary() function to see the decision tree rules. (15% of total points)

QB2. Consider the following input:

- Sales=9
- Price=6.54
- Population=124
- Advertising=0
- Age=76
- Income= 110
- Education=10

What will be the estimated Sales for this record using the decision tree model? (15% of total points)

QB3. Use the caret function to train a random forest (method='rf') for the same dataset. Use the caret default settings. By default, caret will examine the "mtry" values of 2,4, and 6. Recall that mtry is the number of attributes available for splitting at each splitting node.

Which mtry value gives the best performance?

(Make sure to set the random number generator seed to 123) (15% of total points)

QB4. Customize the search grid by checking the model's performance for mtry values of 2, 3 and 5 using 3 repeats of 5-fold cross validation. (15% of total points)

General Submission Instructions:

All work must be your own. Copying other people's work or from the Internet is a form of plagiarism and will be prosecuted as such.

You may submit a Microsoft Word (.doc/.docx) document as an attachment using the Canvas Assignment tool, or you may copy and paste your answer into the provided box within the Assignment tool. If you attach a document for your assignment, be sure to include your name in the text of the document and in the name of the document.

- You can only submit once, so make sure you are completely finished before submitting and that you attach the correct word .doc/.docx file.
- Submissions sent by email will NOT be accepted.

Due dates are listed in the Assignment Schedule document.