Problem Set: Regression Trees

BUAD 5082 – Spring 2019

1. Objectives

The purpose of this problem set is to provide you with an opportunity to practice the kinds of skills that I expect you to be able to perform on an exam.

2. What You Will Need

- Access to a Windows computer with R

3. Solutions

Solutions to these problems will be posted several days after this Problem Set is posted.

4. Preliminaries:

https://www.kaggle.com/mohansacharya/graduate-admissions

Problem 1:

Predict the percent chance of admission for graduate students using the other variables in the Admission_Predict.csv file from the zip folder.

a) Install ‘rpart’ and ‘rattle’ and set the seed to 527

b) Bring in the data from the file Admission_Predict.csv and discard unwanted columns

c) Create training vector and test data frame (90% training, 10% test)

d) Build the regression tree (using the default parameters) with the chance of admission as the value being predicted

e) Print the complexity parameters (cp’s). Which cp is associated with the lowest x error?

f) Plot the model with appropriate labels

g) Create a maximal tree

h) Plot this new model and compare it to the previous model. What did changing the tuning parameters do?

i) Prune the new tree to the value within one standard deviation of the lowest x error.

j) Print the rules for this model

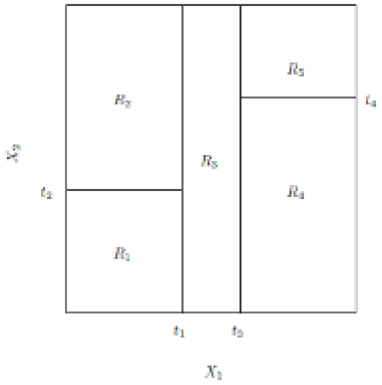
k) Predict using the test set and plot the points with a regression line

l) Compute the MSE

Problem 2

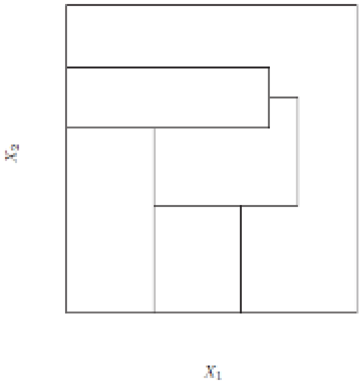
Answer the following true/false questions.

1.



The feature space above is partitioned using recursive binary splitting? **T/F**

2.



The feature space above is partitioned using recursive binary splitting? **T/F**

3.

It is better to build the regression tree as large as possible then prune it back based on the CP. **T/F**

4.

The process of nodes repeatedly splitting into branches is known as “enumerated partitioning”. **T/F**

5.

Greedy algorithms increase the computational complexity of search heuristics such as what is used in regression trees. **T/F**

6.

Decision trees are often utilized because they are simple to view and explain but at a loss of performance. **T/F**