Untitled

Elena Damiano Xierui Aharon

12/1/2020

Contents

Assignment 2 Multivariate Statistics

This document ...

```
asd asd
```

Task 1

Task 2

1. Introduction

For task 2, we are going to deal with a dataset containing information about 4601 webmails. We have 48 variables describing the frequency of some specific words like "remove" in each observation, 6 variables describing the frequency of some specific chars like "$" in one observation, and three variables, **capital_run_length_longest**, **capital_run_length_average** and **capital_run_length_total**, describing the length of the longest uninterrupted sequence of capital letters, the average length of uninterrupted sequences of capital letters, and the total number of capital letters in each observation respectively. We also have a variable called spam, which indicates whether this webmail is a spam with 0 and 1, where 1 for spam, and 0 for not spam. Here all our variables are numeric type.

Our task is to use these 57 attribute variables to classify whether a webmail is spam.

2. Methodology

In order to validate the accuracy of our methods, we firstly divide our dataset into a train set, which contains 2500 observations, and a test set, which contains 2101 observations. We use the train set to train our models, and then apply it to the test set to validate it's accuracy.

3. Results

In this part, we are going to discuss the results obtained by complex tree model and pruned tree model.

1. Classification Trees

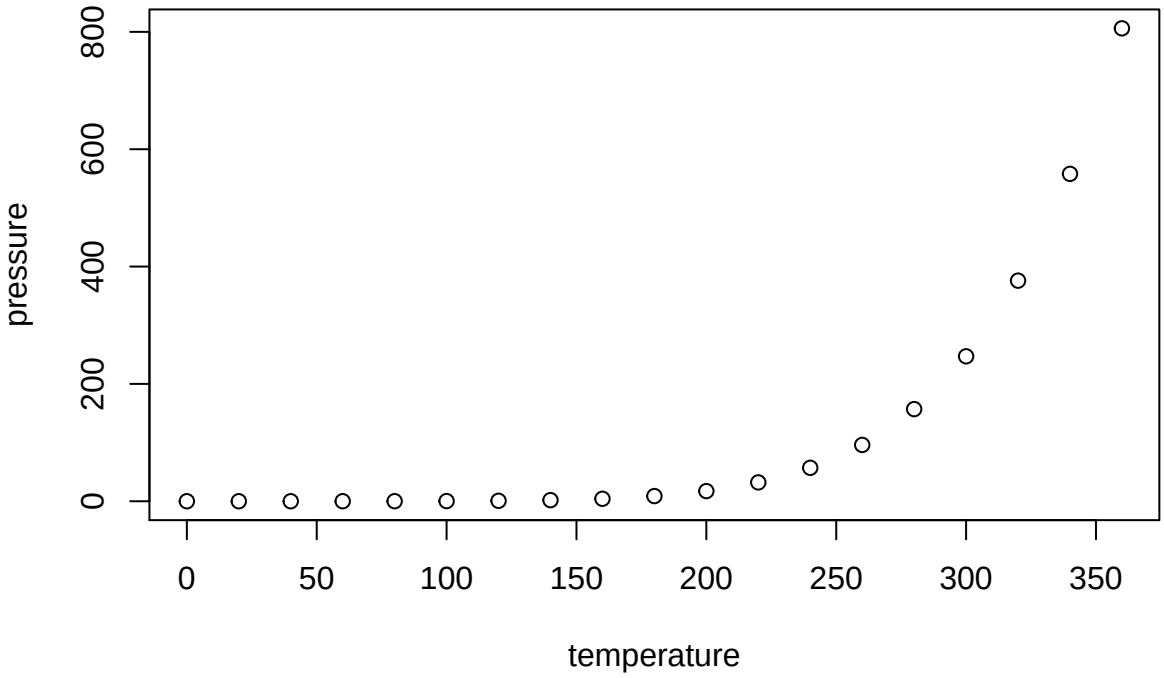
We begin with construct a complex tree model by dividing our observation into small non-overlapping regions according to some numerical criterias. Here we split our dataset until each leaf of our classification tree contains only less then 2 observations. The method used here is recursive binary splitting.

```
summary(cars)
```

```
##      speed           dist
##  Min.   : 4.0   Min.   :  2.00
##  1st Qu.:12.0   1st Qu.: 26.00
##  Median :15.0   Median : 36.00
##  Mean   :15.4   Mean   : 42.98
##  3rd Qu.:19.0   3rd Qu.: 56.00
##  Max.   :25.0   Max.   :120.00
```

Including Plots

You can also embed plots, for example:



Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.