Assignment 5 - Decision Trees

Charles Lang

November 9, 2016

For this assignment we will be using data from the Assistments Intelligent Tutoring system. This system gives students hints based on how they perform on math problems.

Install & call libraries

```
library(rpart)
```

```
## Warning: package 'rpart' was built under R version 3.5.2
```

```
library(party)
```

```
## Warning: package 'party' was built under R version 3.5.2

## Loading required package: grid

## Loading required package: mvtnorm

## Warning: package 'mvtnorm' was built under R version 3.5.2

## Loading required package: modeltools

## Loading required package: stats4

## Loading required package: strucchange

## Warning: package 'strucchange' was built under R version 3.5.2

## Loading required package: zoo

## Warning: package 'zoo' was built under R version 3.5.2

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric

## Loading required package: sandwich

## Warning: package 'sandwich' was built under R version 3.5.2
```

Part I

```
D1 <- read.csv("~/Desktop/master fall/hudk4050/assignment5/intelligent_tutor.csv")
```

Classification Tree

First we will build a classification tree to predict which students ask a teacher for help, which start a new session, or which give up, based on whether or not the student completed a session (D1*complete*)*andwhetherornottheyaskedforhints*(*D*1hint.y).

```
c.tree <- rpart(action ~ hint.y + complete, method="class", data=D1) #Notice the standard R notion for 

#Look at the error of this tree
printcp(c.tree)
```

```
##
## Classification tree:
## rpart(formula = action ~ hint.y + complete, data = D1, method = "class")
##
## Variables actually used in tree construction:
## [1] complete hint.y
##
## Root node error: 250/378 = 0.66138
##
## n= 378
##
##      CP nsplit rel error xerror     xstd
## 1 0.052      0     1.000  1.112 0.034303
## 2 0.012      1     0.948  1.092 0.034833
## 3 0.010      2     0.936  1.040 0.036036
```

```
#Plot the tree
post(c.tree, file = "tree.ps", title = "Session Completion Action: 1 - Ask teacher, 2 - Start new sessi
```

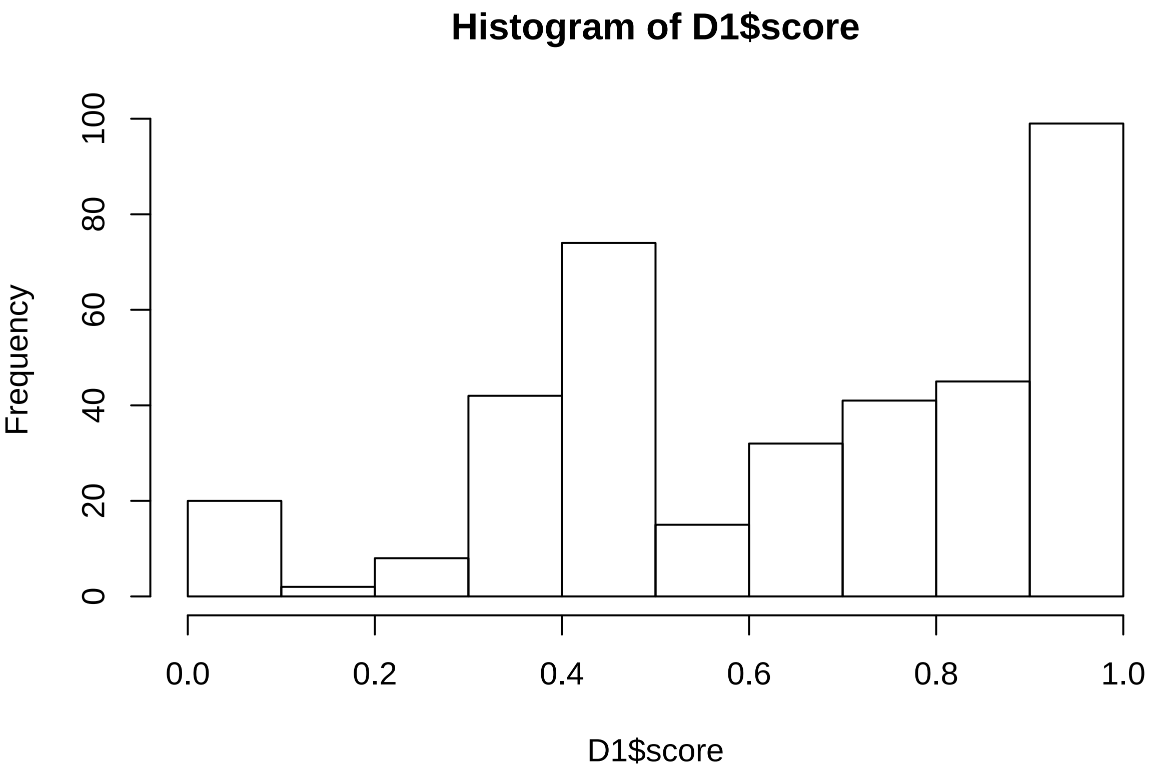
Part II

Regression Tree

We want to see if we can build a decision tree to help teachers decide which students to follow up with, based on students' performance in Assistments. We will create three groups ("teacher should intervene", "teacher should monitor student progress" and "no action") based on students' previous use of the system and how many hints they use. To do this we will be building a decision tree using the "party" package. The party package builds decision trees based on a set of statistical stopping rules.

Visualize our outcome variable "score"

```
hist(D1$score)
```



Create a categorical outcome variable based on student score to advise the teacher using an "ifelse" statement

```
#set the level of student score:
#if score between 0-0.4, should be level:"teacher should intervene",in short "intervene"
#if score between 0.4-0.7, should be level:"teacher should monitor student progress", in short "monitor
#if score greater than 0.7, should be level :"no action"
D1$advice <- ifelse (D1$score >= 0.4, ifelse(D1$score >= 0.7,"no action","monitor"),"intervene" )
```

Build a decision tree that predicts "advice" based on how many problems students have answered before, the percentage of those problems they got correct and how many hints they required

```
score_ctree <- rpart(factor(advice) ~ prior_prob_count + prior_percent_correct+hints, method="class", da
```

Plot tree

```
post(score_ctree, file = "score_tree.ps", title = " three groups:1:teacher should intervene, 2:teacher
```

Please interpret the tree, which two behaviors do you think the teacher should most closely pay attemtion to? i think the following two types of students that the teacher should most closely pay attemtion :

students who require hints greater than 12.5

students who require hints smaller than 12.5 and greater than 0.5, percentage of those problems they got correct smaller than 40%.

for these two types of students, they are mostly the student need to be monitored or intervened.

#Test Tree

Upload the data “intelligent_tutor_new.csv”. This is a data set of a differnt sample of students doing the same problems in the same system. We can use the tree we built for the previous data set to try to predict the “advice” we should give the teacher about these new students.

```
#Upload new data

D2 <- read.csv("~/Desktop/master fall/hudk4050/assignment5/intelligent_tutor_new.csv")

#Generate predicted advice using the predict() command for new students based on tree generated from ol

D2$prediction <- predict(score_ctree,D2,type = "class")
D2$prediction
```

```
##   [1] no action monitor   intervene monitor   intervene monitor   monitor  
##   [8] no action intervene no action no action monitor   no action intervene
##  [15] no action monitor   monitor   no action no action intervene no action
##  [22] intervene no action intervene intervene monitor   no action intervene
##  [29] intervene no action no action intervene no action intervene intervene
##  [36] intervene intervene monitor   intervene intervene no action monitor  
##  [43] intervene no action no action no action monitor   intervene no action
##  [50] no action intervene no action intervene no action no action intervene
##  [57] intervene no action no action no action no action no action no action
##  [64] no action no action no action no action intervene no action monitor  
##  [71] no action intervene intervene no action intervene no action no action
##  [78] no action intervene no action intervene no action intervene no action
##  [85] intervene no action intervene no action intervene no action intervene
##  [92] intervene monitor   intervene no action intervene no action intervene
##  [99] monitor   no action intervene intervene intervene no action no action
## [106] intervene no action no action no action intervene intervene intervene
## [113] intervene no action monitor   no action intervene intervene no action
## [120] no action intervene intervene no action intervene intervene monitor  
## [127] intervene monitor   intervene no action monitor   no action monitor  
## [134] intervene no action monitor   no action intervene intervene no action
## [141] intervene intervene monitor   no action intervene intervene no action
## [148] intervene no action intervene no action no action no action intervene
## [155] monitor   monitor   no action no action no action no action no action
## [162] intervene intervene no action intervene intervene intervene monitor  
## [169] intervene intervene no action intervene monitor   no action intervene
## [176] no action intervene no action no action no action monitor   no action
## [183] no action no action intervene intervene no action intervene intervene
## [190] no action no action intervene intervene monitor   intervene intervene
## [197] intervene intervene intervene monitor  
## Levels: intervene monitor no action
```

Part III

Compare the predicted advice with the actual advice that these students recieved. What is the difference between the observed and predicted results?

```r
library(caret)
```

```
## Warning: package 'caret' was built under R version 3.5.2

## Loading required package: lattice

## Loading required package: ggplot2

## Warning: package 'ggplot2' was built under R version 3.5.2
```

```r
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 3.5.2

##
## 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
```

```r
D2$advice <- ifelse (D2$score >= 0.4, ifelse(D2$score >= 0.7,'no action','monitor'),'intervene' )
comb <- select(D2,advice,prediction)
confMat <- table(comb)
confMat
```

```
##            prediction
## advice      intervene monitor no action
##   no action        85      28        87
```

```r
accuracy <- sum(diag(confMat))/sum(confMat)
accuracy
```

```
## [1] 0.425
```

the accuracy is 42.5% which means the prediction only make the 42.5% corret predicts of total students. from the new data set we can see that the score of all student is 1, so all student under the level of no action, but by prediction, there are 42.5% of student should under level of no action.

for prediction value, the student that should be monitored is 14%, but the actual value is 0%. for prediction value, the student that should be intervened is 42.5%, but the actual value is 0%. the difference between prediction and actual value is huge.

To Submit Your Assignment

Please submit your assignment by first "knitting" your RMarkdown document into an html file and then commit, push and pull request both the RMarkdown file and the html file.