Stat 471/571/701 HW 4

Overview / Instructions

This homework will be due on Monday **Dec. 3, by 11:59 PM** on Canvas.

Problem 1: This problem is conceptual. You will get familiar with the building blocks for decision trees. To save your time, a couple of R chunks are included here to generate the data and to make some plots. Feel free to use your own code to make any plots. For uniformity don't change the data2. The diagram is attached as an appendix so that you can check to make sure the data2 you generate is same as what I have.

1) A small data set is generated and it is stored in data2. It consists of four variables

Y1: a continuous response
Y2: a binary response and
X1, X2 two continuous explanatory variables.

Run the following r-chunk to generate data2

```
##########################
set.seed(1)
x.temp <- ceiling(runif(40, min=0, max=100))
data1<- matrix(x.temp,ncol=2, byrow=TRUE )
y <- round(rexp(nrow(data1), rate=2), 2)
data1 <- data.frame(data1, y)
names(data1)  <- c("X1", "X2", "Y1")

data2 <- data1
set.seed(1)
data2$Y2 <- ifelse((data1$X1+data1$X2 > 70), rbinom(1,1,.62), rbinom(1,1, .31))
data2
############################
```

Look at the data to see what's there. Make sure it matches the plot in the appendix.

2) A diagram is drawn to partition X1 and X2 into R1, … R6 regions by using the following R chunk.

```
######################################################
#### Get a diagram in two dimensions with 6 Regions R_1 to R_6

# Set up an empty plot
```

```
plot(NA, NA, type = "n", xlim = c(0,100), ylim = c(0,100), xlab = "X1", ylab = "X2")
# Draw some horizontal and vertical lines to divide the space into 6 regions
lines(x = c(40,40), y = c(0,100))
lines(x = c(0,40), y = c(75,75))
lines(x = c(75,75), y = c(0,100))
lines(x = c(20,20), y = c(0,75))
lines(x = c(75,100), y = c(25,25))

# Label the regions
text(x = (40+75)/2, y = 50, labels = c("R1"))
text(x = 20, y = (100+75)/2, labels = c("R2"))
text(x = (75+100)/2, y = (100+25)/2, labels = c("R3"))
text(x = (75+100)/2, y = 25/2, labels = c("R4"))
text(x = 30, y = 75/2, labels = c("R5"))
text(x = 10, y = 75/2, labels = c("R6"))
#### End of diagram
#######################################################
```

i) Is this a top-down, recursive tree?

Use this tree with data2 as the training data. Give the following predicted values of Y1 on the end nodes using X1 and X2:

ii) Predicted Y1 for x1=60, x2=30.
iii) Predicted Y1 for x1=90, x2=10.

3) Use tree() to produce a best decision tree for Y1. Display the tree. Is this tree very different from the decision tree given in the diagram above?

4) Let us now concentrate on classification decision trees. The event of interests is Y2=1.

i) Overlay the labels Y2 for each subject in the original tree of the appendix.

Data2 will be again the training data and we use sample proportion to estimate the probability of Y2=1 in each region.

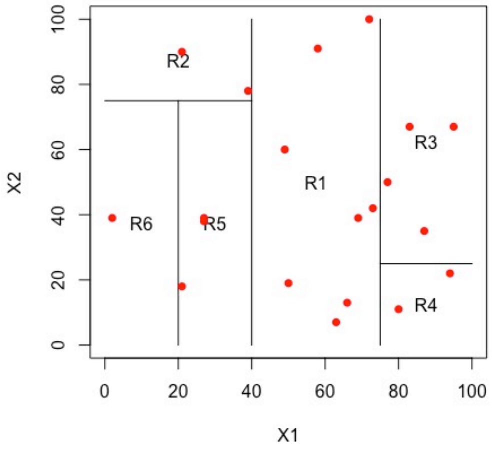
ii) Predicted Prob(Y2=1) for x1=60, x2=30.
iii) Give Y2's label for x1=60, x2=30 by majority vote.

5) Apply rpart() by default to produce a decision classification tree and plot it. Is this tree different from our original tree?

Notice: rpart() uses Gini index as default to grow a tree. It is a good package to produce a single tree. We do however have a choice to specify the splitting criterion as parms=list(split="information") to grow a tree using a different form of deviance.

We can also prune a tree using prune(fit.rpart). We did not go through the details about this topic.

Appendix: The diagram of data2:



Problem 2: Intelligence, Successes

We continue to analyze the IQ.Full.csv. Recall that this data set contains about 2600 individuals from the 1979 National Longitudinal Study of Youth (NLSY79) survey. Those subjects were re-interviewed in 2006, who had paying jobs in 2005, and who had complete values for the variables listed below.

Personal Demographic Variables:

Race 1 = Hispanic, 2 = Black, 3 = Not Hispanic or Black
Gender: a factor with levels **"female"** and **"male"**
Educ: years of education completed by 2006

Household Environment:

Imagazine: a variable taking on the value 1 if anyone in the respondent's household regularly read magazines in 1979, otherwise 0
Inewspaper: a variable taking on the value 1 if anyone in the respondent's household regularly read newspapers in 1979, otherwise 0
Ilibrary: a variable taking on the value 1 if anyone in the respondent's household had a library card in 1979, otherwise 0
MotherEd: mother's years of education
FatherEd: father's years of education

Variables Related to ASVAB test Scores in 1981

AFQT: percentile score on the AFQT intelligence test in 1981
Coding: score on the Coding Speed test in 1981
Auto: score on the Automotive and Shop test in 1981
Mechanic: score on the Mechanic test in 1981
Elec: score on the Electronics Information test in 1981

Science: score on the General Science test in 1981
Math: score on the Math test in 1981
Arith: score on the Arithmetic Reasoning test in 1981
Word: score on the Word Knowledge Test in 1981
Parag: score on the Paragraph Comprehension test in 1981
Numer: score on the Numerical Operations test in 1981

Variable Related to Life Success in 2006

Income2005: total annual income from wages and salary in 2005. We will use a natural log transformation to evaluate this data!

The following 10 questions are answered as 1: strongly agree, 2: agree, 3: disagree, 4: strongly disagree

Esteem 1: "I am a person of worth"
Esteem 2: "I have a number of good qualities"
Esteem 3: "I am inclined to feel like a failure"
Esteem 4: "I do things as well as others"
Esteem 5: "I do not have much to be proud of"
Esteem 6: "I take a positive attitude towards myself and others"
Esteem 7: "I am satisfied with myself"
Esteem 8: "I wish I could have more respect for myself"
Esteem 9: "I feel useless at times"
Esteem 10: "I think I am no good at all"

This exercise is designed partially for you to understand PCA conceptually. The models suggested might not be the most sensible way to analyze the data from a practical point of view.

For question 1) and 2): Use a subset of 100 subjects here. Make sure to **use set.seed(10)** when sampling.

1) Let us first use PCA to summarize the ***ASVAB*** tests. Run prcomp over all the tests in ASVAB. We should center and scale all the tests.

a) Report the PC1 and PC2 loadings. Are they unit vectors? Are they uncorrelated?
b) How is the PC1 score obtained for each subject? Write down the correction.
c) Are PC1 scores and PC2 scores in the data uncorrelated?
d) Plot PVE (Proportion of Variance Explained) with an explanation.
e) Also plot CPVE (Cumulative Proportion of Variance Explained). What proportion of the variance in the data is explained by the first two principal components?
f) PC's provide us with a low dimensional view of the ASVAB. Use a biplot to display the data, using the first two principal components. Give an interpretation from the plot.

g) Repeat the above biplot but label points with different colors, according to their Gender. Do you see a systematic separation between Male and Female in the biplot? Write a brief summary about your findings.

2) We next will try to summarize the 10 Esteem measurement by PCA

a) First, notice that Esteem 1, 2, 4, 6, and 7 need to be reversed prior to scoring in order for a higher score to designate higher self-esteem.
b) What are the PC1 loadings?
c) How much variance is explained by using the PC1? Provide both PVE and CPVE plots.
d) Combine c) and the biplot of the PC1 and PC2 write a brief summary about Esteem scores.

Note: To reverse the esteem score, you may try this. Say data.esteem has all the 10 esteem scores.
data.esteem[, c(1, 2, 4, 6, 7)] <- 5- data.esteem[, c(1, 2, 4, 6, 7)]

3) How well can we predict 'success' based on Intelligence?

To answer this question, we use Income <- log(Income2005) as a measure of one's success.

a) Why is it important to create a logarithmic transformation of income?
b) Run prcomp over ***ASVAB*** tests first.
c) fit1: Income ~ PC1; fit2: Income ~ PC1+PC2+PC3. Notice the LS estimates of PC1 in both fit1 and fit2 are identical. Why is this so? Are the leading PC's of ***ASVAB*** significant variables to predict Income? (You may use the elbow rule to determine how many PCs are to be included in fit2. In the scree plot of CPVE, take the leading PC's when there is a sharp change in the plot.)
d) Controlling for Personal Demographic Variables and Household Environment, are the leading PC's of ***ASVAB*** significant variables to predict Income at .01 level? Give a brief summary of your findings.

3. Case study: Yelp review (for more information check their website: http://www.yelp.com/dataset_challenge)

It is unlikely we will win the $40,000 prize posted but we get to use their data for free. We have done a detailed analysis in our lecture. This exercise is designed for you to get hands on the whole process.

The goals are 1) Try to identify important words associated with positive ratings and negative ratings. Collectively we have a sentiment analysis. 2) To predict ratings and 3) To get familiar with RTextTools

1) Take a random sample of 20000 reviews (set.seed(1)) from our original data set. Extract document term matrix for texts to keep words appearing at least 2% of the time among all 20000 documents. Go through the similar process of cleansing as we did in the lecture.

i) Briefly explain what does this matrix record? What is the cell number at row 100 and column 405? What does it represent?

ii) What is the sparsity of the dtm obtained here? What does that mean?

2) Set the stars as a two category response variable called rating to be "1" = 5,4 and "0"= 1,2,3. Combine the variable rating with the dtm as a data frame called data2.

Get a training data with 15000 reviews and the rest 5000 reserved as the testing data.

3) Use the training data to get Lasso fit. Choose lambda.1se. Keep the result here.

4) Feed the output from Lasso above, get a logistic regression.

i) Pull out all the positive coefficients and the corresponding words. Rank the coefficients in a decreasing order. Report the leading 2 words and the coefficients. Describe briefly the interpretation for those two coefficients.

ii) Make a word cloud with the top 100 positive words according to their coefficients. Interpret the cloud briefly.

iii) Repeat i) and ii) for the bag of negative words.

iv) Summarize the findings.

5) Using majority votes find the testing errors

i) From Lasso fit in 3)

ii) From logistic regression in 4)

iii) Which one is smaller?

6) Now we will apply RTextTools using the same Training and Testing data as we have reserved. Run

i) Logistic Reg. Is the testing error obtained here same as the one we got in 5) ii)?
ii) RF. Get the testing error
iii) SVM. Get the testing error
iv) Boosting. Get the testing error

Which classifier(s) seem to produce the least testing error? Are you surprised?

7) For the purpose of prediction, comment on how would you predict a rating if you are given a review using our final model?