CART

MATH 4322 Lab 15

Cathy Poliak

Fall 2022

Instructions

1) You can print this out and write on this or write/type on a seperate sheet. I also provide a Rmarkdown version of this if you desire.
2) Upload your answers in BlackBoard as you do with your homework.
3) The questions are in red.
4) This is for Decsion Trees in R.

Fitting Classification Trees

We first use classification trees to analyze the Carseats data set.
This is part of the ISLR library.
We will attempt to predict the **high** sales in 400 locations based on a number of predictors.

To investigate further:

```
library(ISLR)
?Carseats
```

Question 1: How many variables are in this dataset?

11

Question 2: Are there any variables that are categorical? If so write down the names.

ShelveLoc
Urban
US

We want to put Sales as a binary variable (categorical with two categories). We will use the ifelse() function to create a variable called High, which takes on the value of Yes if the Sales variable exceeds 8, and takes on a value of No otherwise.

Type in the following:

```
High = ifelse(Carseats$Sales <= 8, "No","Yes")
High = as.factor(High)
Carseats = data.frame(Carseats,High) #merge High with the rest of the Carseats data.
```

We now use the `tree()` function to fit a classification tree in order to predict `High` using all variables but `Sales`. Type and run the following in `R`.

```
library(tree)
```

```
## Warning: package 'tree' was built under R version 4.2.1
```

```
tree.carseats = tree(High~. -Sales,Carseats)
summary(tree.carseats)
```

Question 3: How many nodes are produced?

27

Question 4: What is the training error rate?

0.09 (9%)

To get the graphical display of these trees type and run the following in `R`.

```
plot(tree.carseats)
text(tree.carseats,pretty = 0)
```

Question 5: What is the variable of the first branch? How is that branch split?

Shelve Loc

Bad or medium => left

Good => right

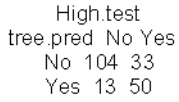
The first branch is the most important indicator of the response.

In order to properly evaluate the performance of a clasification tree on these data, we will split that observations inot a training set and a test set. Type and run the following in `R`.

```
set.seed(2)
train = sample(1:nrow(Carseats),200)
Carseats.test = Carseats[-train,]
```

```
High.test = High[-train]
tree.carseats = tree(High ~ . -Sales, Carseats,subset = train)
tree.pred = predict(tree.carseats,Carseats.test,type = "class")
table(tree.pred,High.test)
```

Question 6: What is the test error rate?



```
        High.test
tree.pred  No Yes
      No  104  33
      Yes  13  50
```

$46/200 = 23\%$

We can prune the tree to see if it leads to better results. Type and run the following.

```
set.seed(3)
cv.carseats = cv.tree(tree.carseats,FUN = prune.misclass)
cv.carseats
```

The dev corresponds to the cross-validation error rate.

$-2\sum_m\sum_k \hat{p}_{mk}\log(\hat{p}_{mk})$

Question 7: What is the lowest cross-validation error rate?

$74 \Rightarrow |T_\alpha| = 21$

$75 \Rightarrow |T_\alpha| = 8$

Run the following

```
plot(cv.carseats$size,cv.carseats$dev,type = "b")
```

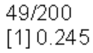
Question 8: What value corresponds to the lowest cross-validation error rate?

$|T_\alpha| = 21$ but we may want to prune to $|T_\alpha| = 8$

We now apply the prune.misclass() function in order to prune the tree.

```
prune.carseats = prune.misclass(tree.carseats,best = 8)
plot(prune.carseats)
text(prune.carseats,pretty = 0)
tree.pred = predict(prune.carseats,Carseats.test,type = "class")
table(tree.pred,High.test)
```

Question 9: What is the test error rate for the pruned tree?



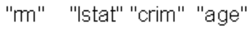
```
49/200
[1] 0.245
```

Fitting Regression Trees

Here we fit a regression tree to the `Boston` data set.
First create a test and training data.

```
library(MASS)
set.seed(1)
train = sample(1:nrow(Boston),nrow(Boston)/2)
tree.boston = tree(medv ~.,Boston,subset = train)
summary(tree.boston)
```

Question 10: What variables were used to construct this tree?



```
"rm"   "lstat" "crim"  "age"
```

Question 11: How many nodes are used to contsrtuct this tree?

7

Plot the tree

```
plot(tree.boston)
text(tree.boston,pretty = 0)
```

Question 12: What is the predicted medain house price for medium sized homes ($6.9595 \leq \text{rm} < 7.553$)?

$ 33,420

Now we will use the `cv.tree()` function to see whether pruning the tree will improve performance.

```
cv.boston = cv.tree(tree.boston)
plot(cv.boston$size,cv.boston$dev,type = "b")
```

Question 13: How many nodes would be best to use?

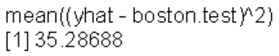
Now prune the tree.

```
prune.boston = prune.tree(tree.boston,best = 5)
plot(prune.boston)
text(prune.boston,pretty = 0)
```

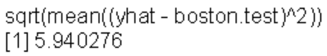
In keeping with the cross-validation results, we use the unpruned tree to make predictions on the test set.

```
yhat = predict(tree.boston,newdata = Boston[-train,])
boston.test = Boston[-train,"medv"]
plot(yhat,boston.test)
abline(0,1)
mean((yhat - boston.test)^2)
```

Question 14: What is the test set MSE associated with the regression tree?



```
mean((yhat - boston.test)^2)
[1] 35.28688
```



```
sqrt(mean((yhat - boston.test)^2))
[1] 5.940276
```