MATH 4322 Lab 15

Phu Nguyen

11/4/2021

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 variables

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

- Yes; ShelveLoc, Urban and 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 predcit `High` using all variables except `Sales`. Type and run the following in `R`.

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

```
##
## Classification tree:
## tree(formula = High ~ . - Sales, data = Carseats)
## Variables actually used in tree construction:
```

```
## [1] "ShelveLoc"   "Price"       "Income"      "CompPrice"   "Population"
## [6] "Advertising" "Age"         "US"
## Number of terminal nodes:  27
## Residual mean deviance:  0.4575 = 170.7 / 373
## Misclassification error rate: 0.09 = 36 / 400
```

Question 3: How many nodes are produced?

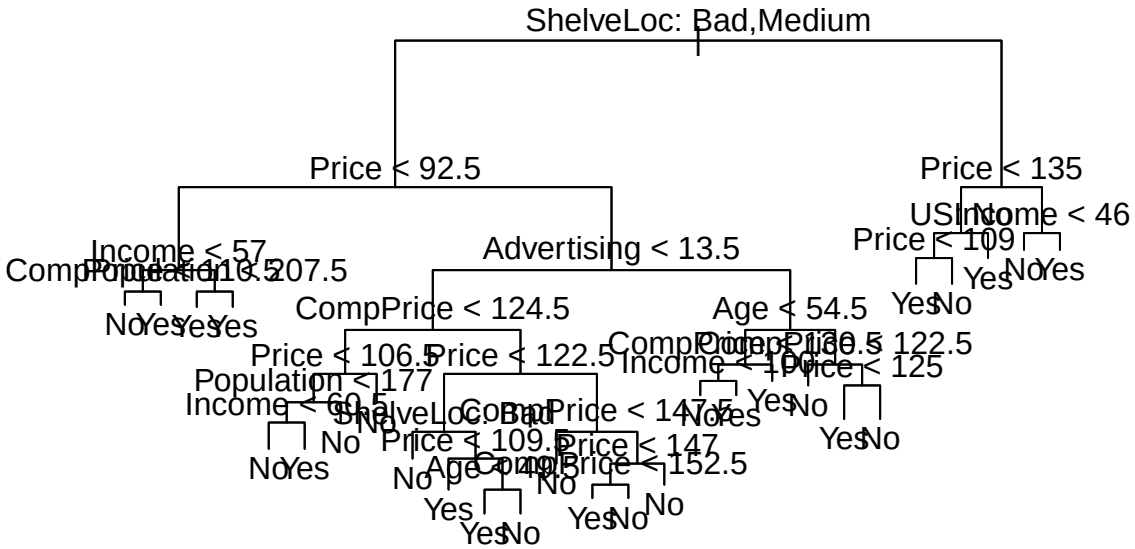
- 27 nodes

Question 4: What is the training error rate?

- 0.09 - 9%. This means that there's 9% that we missclassfied if the sale is high or low

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?

- ShelveLoc: Bad, Medium(left side) : Good (right side)

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. Overfitting.

```
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)
```

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

Question 6: What is the test error rate?

```
(33+13)/(104+33+13+50)
```

```
## [1] 0.23
```

- 23% off is pretty high, we are probably overfitting something.

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
```

```
## $size
## [1] 21 19 14  9  8  5  3  2  1
##
## $dev
## [1] 74 76 81 81 75 77 78 85 81
##
## $k
## [1] -Inf  0.0  1.0  1.4  2.0  3.0  4.0  9.0 18.0
##
## $method
## [1] "misclass"
##
## attr(,"class")
## [1] "prune"         "tree.sequence"
```

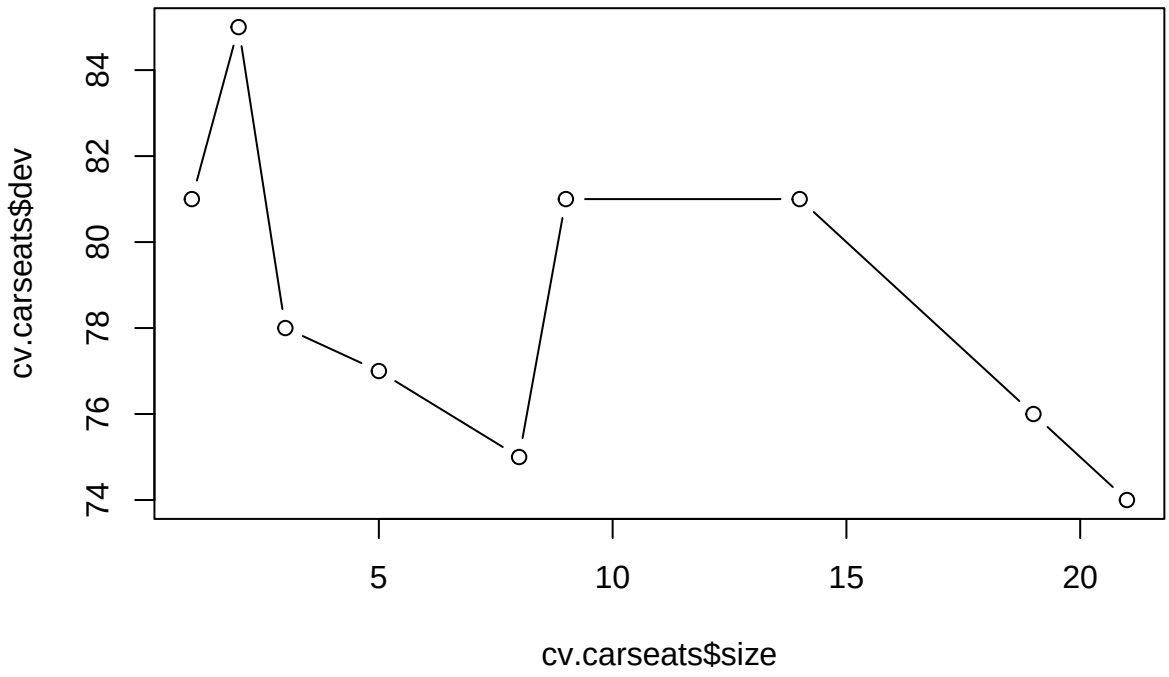
The dev corresponds to the cross-validation error rate.

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

- 74 with number of nodes is 21

Run the following

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

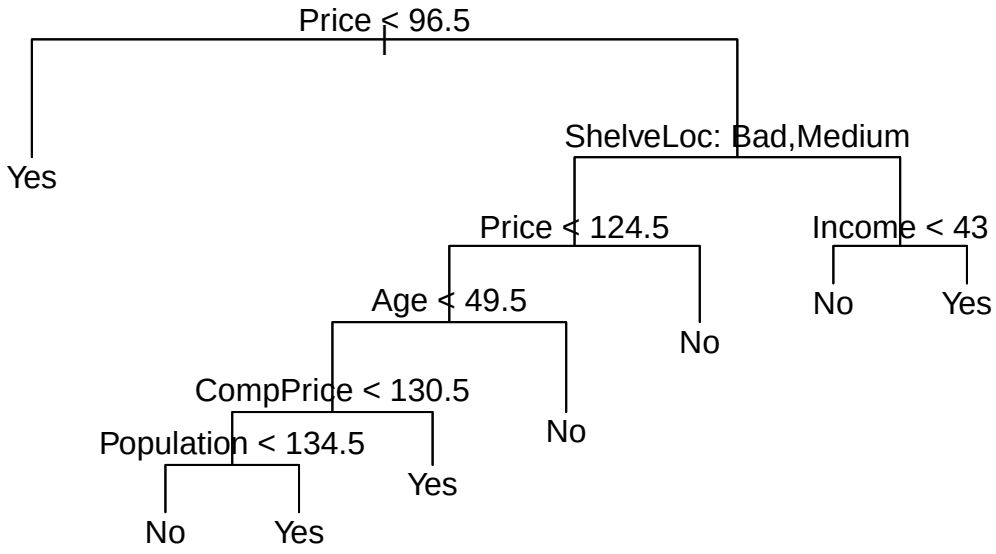


Question 8: What value corresponds to the lowest cross-validation error rate?

- 75 with number of nodes is 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)
```

```
##          High.test
## tree.pred No Yes
##       No  89  21
##       Yes 28  62
```

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

```
(28+21)/(89+21+28+62)
```

```
## [1] 0.245
```

- 24.5%, this error rate is higher before we prune the tree. However, we are not giving up too much error rate to have better interpretation of the data. It depends on how much you want to give up to have a better interpretation.

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)
```

```
##
## Regression tree:
## tree(formula = medv ~ ., data = Boston, subset = train)
## Variables actually used in tree construction:
## [1] "rm"    "lstat" "crim"  "age"
## Number of terminal nodes:  7
## Residual mean deviance:  10.38 = 2555 / 246
## Distribution of residuals:
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -10.1800  -1.7770  -0.1775   0.0000   1.9230  16.5800
```

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

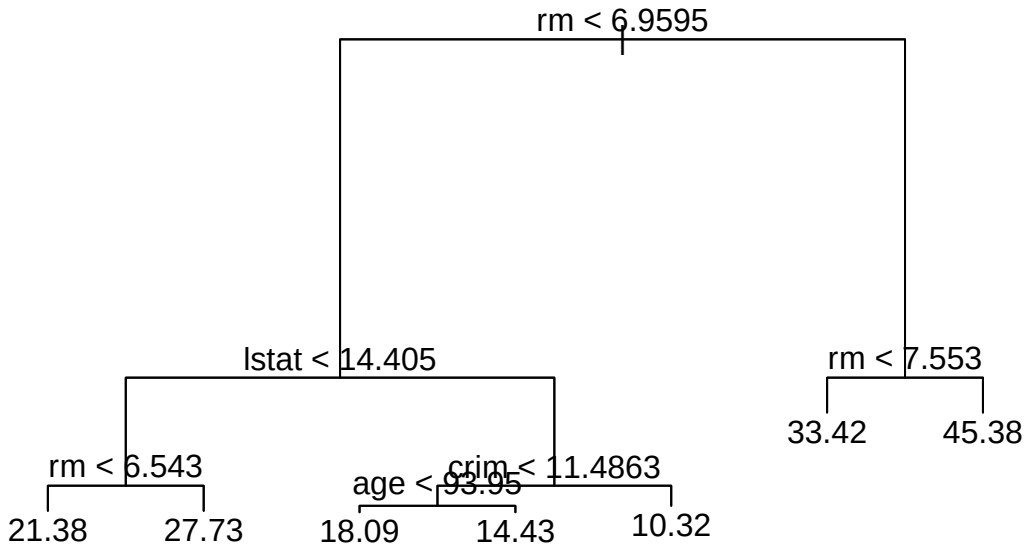
- rm, lstat, crim and age.

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

- 7 nodes are used.

Plot the tree

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



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

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?