Classification and Regression Tree, CART

Jason

2015 年 7 月 30 日

```
library(ISLR); library(MASS); library(tree)
```

Regression Tree

```
#Data in MASS package
data(Boston)
str(Boston)
```

```
'data.frame':   506 obs. of  14 variables:
 $ crim   : num  0.00632 0.02731 0.02729 0.03237 0.06905 ...
 $ zn     : num  18 0 0 0 0 0 12.5 12.5 12.5 12.5 ...
 $ indus  : num  2.31 7.07 7.07 2.18 2.18 2.18 7.87 7.87 7.87 7.87 ...
 $ chas   : int  0 0 0 0 0 0 0 0 0 0 ...
 $ nox    : num  0.538 0.469 0.469 0.458 0.458 0.458 0.524 0.524 0.524 0.524 ...
 $ rm     : num  6.58 6.42 7.18 7 7.15 ...
 $ age    : num  65.2 78.9 61.1 45.8 54.2 58.7 66.6 96.1 100 85.9 ...
 $ dis    : num  4.09 4.97 4.97 6.06 6.06 ...
 $ rad    : int  1 2 2 3 3 3 5 5 5 5 ...
 $ tax    : num  296 242 242 222 222 222 311 311 311 311 ...
 $ ptratio: num  15.3 17.8 17.8 18.7 18.7 18.7 15.2 15.2 15.2 15.2 ...
 $ black  : num  397 397 393 395 397 ...
 $ lstat  : num  4.98 9.14 4.03 2.94 5.33 ...
 $ medv   : num  24 21.6 34.7 33.4 36.2 28.7 22.9 27.1 16.5 18.9 ...
```

```
set.seed(1)
train <- sample(1:nrow(Boston), nrow(Boston)/2)
Bostree <- tree(medv ~ ., data=Boston, subset=train)
Bostree
```

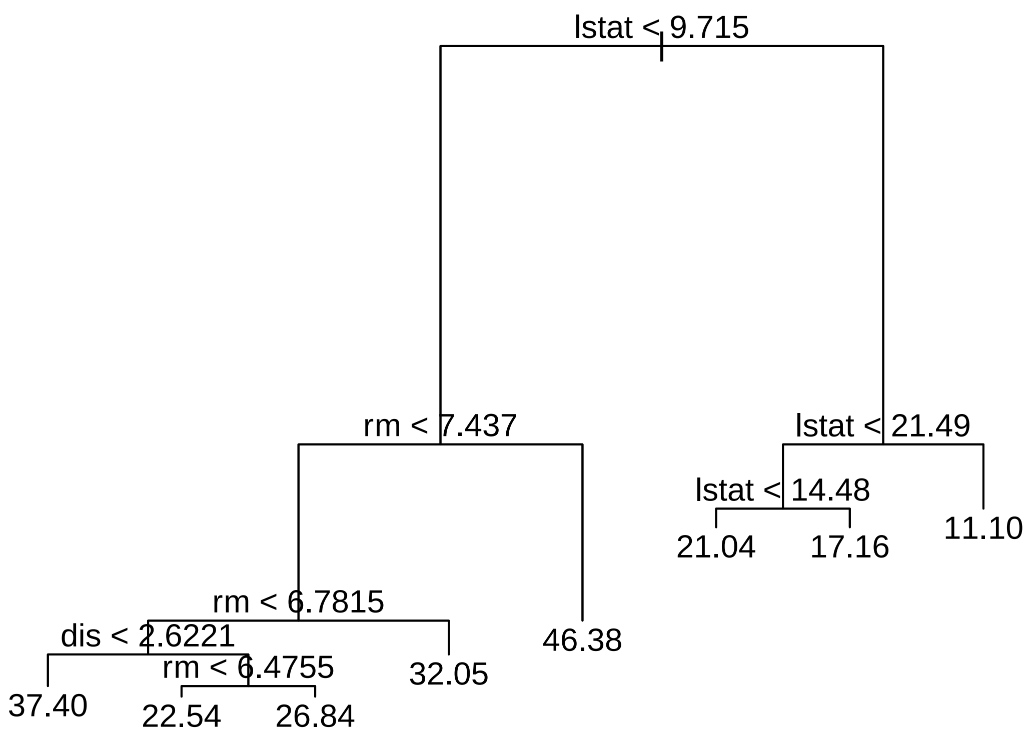
```
node), split, n, deviance, yval
      * denotes terminal node

 1) root 253 20890.0 22.67
   2) lstat < 9.715 103  7765.0 30.13
     4) rm < 7.437 89  3310.0 27.58
       8) rm < 6.7815 61  1995.0 25.52
        16) dis < 2.6221 5   615.8 37.40 *
        17) dis > 2.6221 56   610.3 24.46
          34) rm < 6.4755 31   136.4 22.54 *
          35) rm > 6.4755 25   218.3 26.84 *
       9) rm > 6.7815 28   496.6 32.05 *
     5) rm > 7.437 14   177.8 46.38 *
   3) lstat > 9.715 150  3465.0 17.55
```

```
     6) lstat < 21.49 120  1594.0 19.16
      12) lstat < 14.48 62   398.5 21.04 *
      13) lstat > 14.48 58   743.3 17.16 *
     7) lstat > 21.49 30   311.9 11.10 *
```

```r
summary(Bostree)
```

```
Regression tree:
tree(formula = medv ~ ., data = Boston, subset = train)
Variables actually used in tree construction:
[1] "lstat" "rm"    "dis"
Number of terminal nodes:  8
Residual mean deviance:  12.65 = 3099 / 245
Distribution of residuals:
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
-14.10000  -2.04200  -0.05357   0.00000   1.96000  12.60000
```

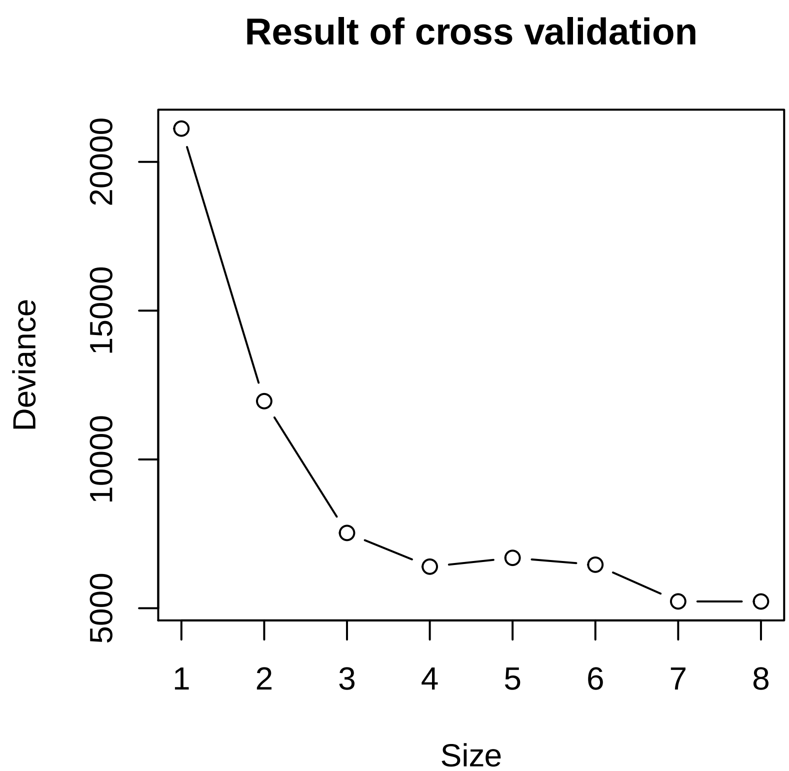
```r
plot(Bostree)
text(Bostree, pretty=0)
```



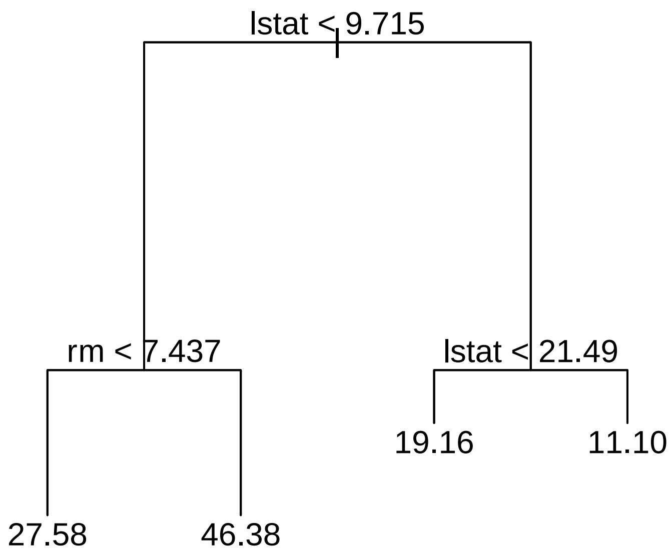
```r
str(cv.tree)
```

```
function (object, rand, FUN = prune.tree, K = 10, ...)
```

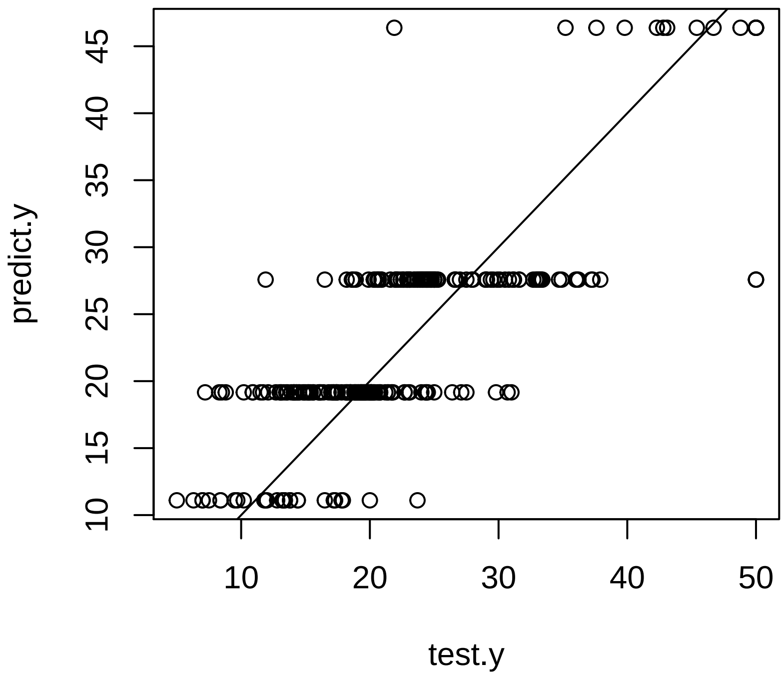
```r
cv.result <- cv.tree(Bostree)
plot(cv.result$size, cv.result$dev, type="b",
     main="Result of cross validation",
     xlab="Size", ylab="Deviance")
```




```r
fit <- prune.tree(Bostree, best=4)
plot(fit)
text(fit, pretty=0)
```




```
test.y <- Boston$medv[-train]
test.x <- Boston[-train, ]
predict.y <- predict(fit, newdata=test.x)
plot(test.y, predict.y)
abline(0, 1)
```

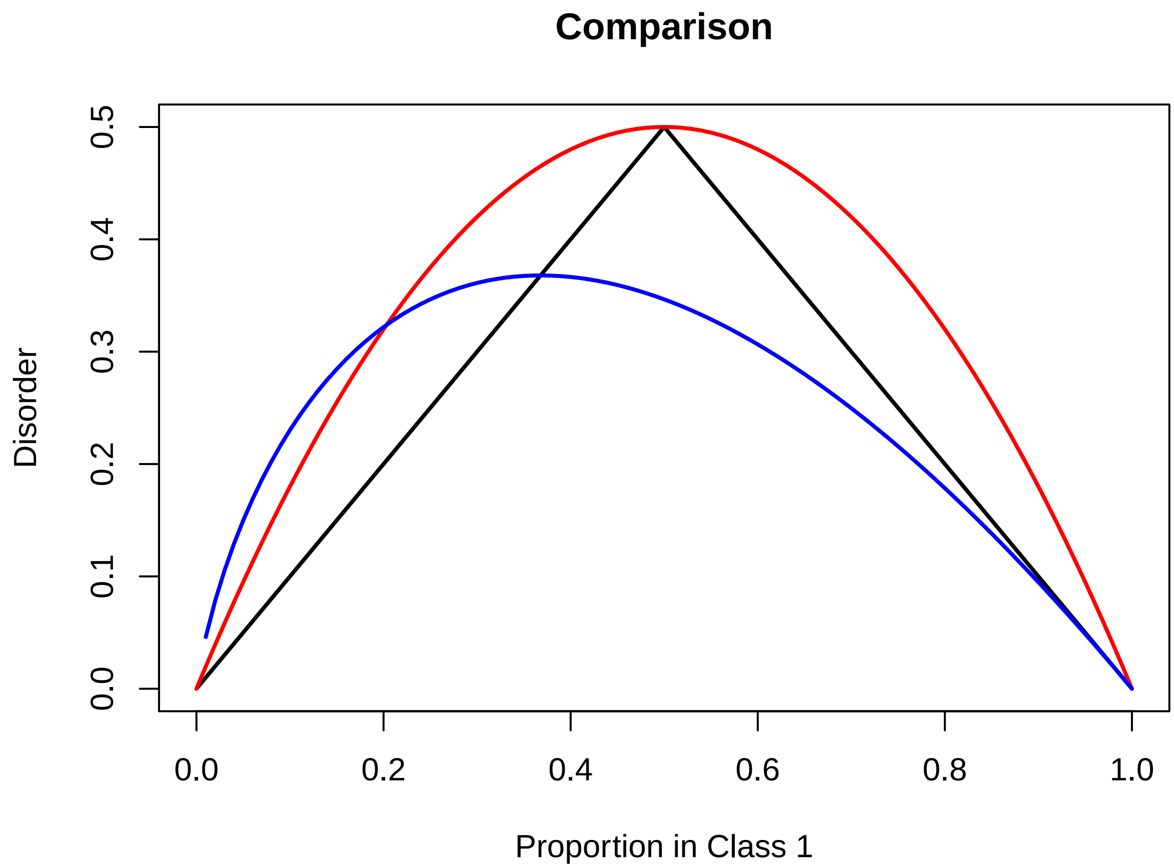



```r
mean((test.y - predict.y)^2)
```

```
[1] 32.22697
```

Classification

```r
misclass <- function(x){
  min(x, 1 - x)
}
gini <- function(x){
  2*x*(1 - x)
}
entropy <- function(x){
  -x*log(x)
}

p <- seq(0, 1, by=0.01)
plot(p, sapply(p, misclass), lwd=2, type="l",
     main="Comparison",
     xlab="Proportion in Class 1", ylab="Disorder")
lines(p, gini(p), col="red", lwd=2)
lines(p, entropy(p), col="blue", lwd=2)
```




```r
data(Carseats)
str(Carseats)
```

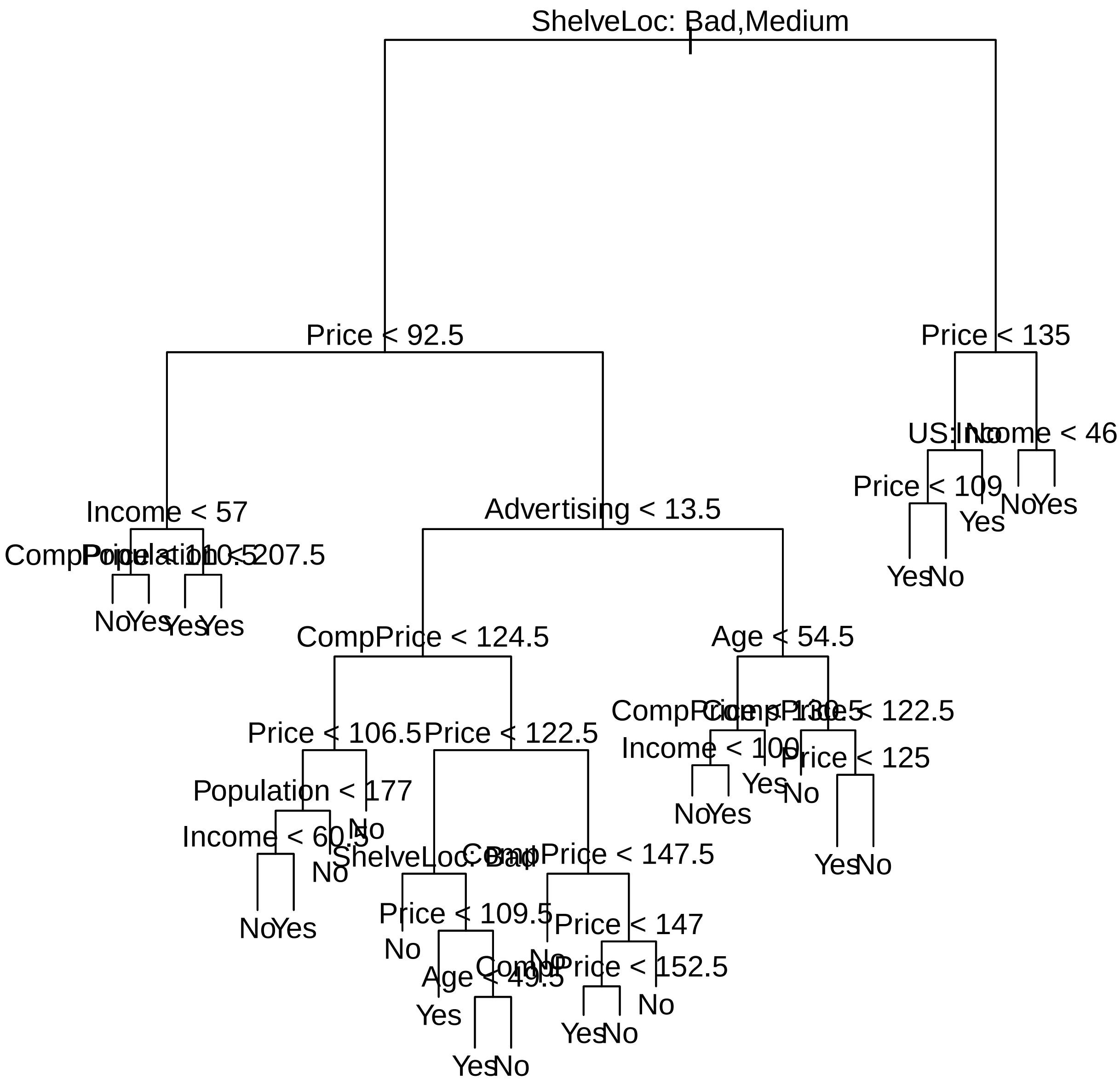
```
'data.frame':   400 obs. of  11 variables:
 $ Sales      : num  9.5 11.22 10.06 7.4 4.15 ...
 $ CompPrice  : num  138 111 113 117 141 124 115 136 132 132 ...
 $ Income     : num  73 48 35 100 64 113 105 81 110 113 ...
 $ Advertising: num  11 16 10 4 3 13 0 15 0 0 ...
 $ Population : num  276 260 269 466 340 501 45 425 108 131 ...
 $ Price      : num  120 83 80 97 128 72 108 120 124 124 ...
 $ ShelveLoc  : Factor w/ 3 levels "Bad","Good","Medium": 1 2 3 3 1 1 3 2 3 3 ...
 $ Age        : num  42 65 59 55 38 78 71 67 76 76 ...
 $ Education  : num  17 10 12 14 13 16 15 10 10 17 ...
 $ Urban      : Factor w/ 2 levels "No","Yes": 2 2 2 2 2 1 2 2 1 1 ...
 $ US         : Factor w/ 2 levels "No","Yes": 2 2 2 2 1 2 1 2 1 2 ...
```

```r
High <- ifelse(Carseats$Sales <= 8, "No", "Yes")
Carseats <- data.frame(Carseats, High)
```

```r
Cartree <- tree(High ~ . - Sales, data=Carseats)
summary(Cartree)
```

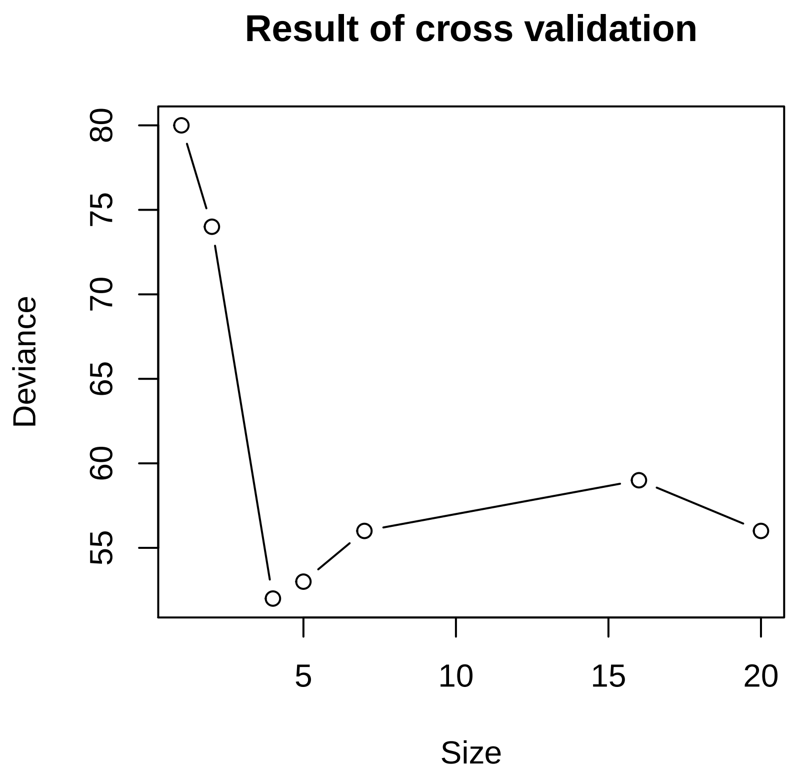
```
Classification tree:
tree(formula = High ~ . - Sales, data = Carseats)
Variables actually used in tree construction:
[1] "ShelveLoc"   "Price"       "Income"      "CompPrice"   "Population"
[6] "Advertising" "Age"         "US"
Number of terminal nodes:  27
Residual mean deviance:  0.4575 = 170.7 / 373
Misclassification error rate: 0.09 = 36 / 400
```

```r
plot(Cartree)
text(Cartree, pretty=0)
```

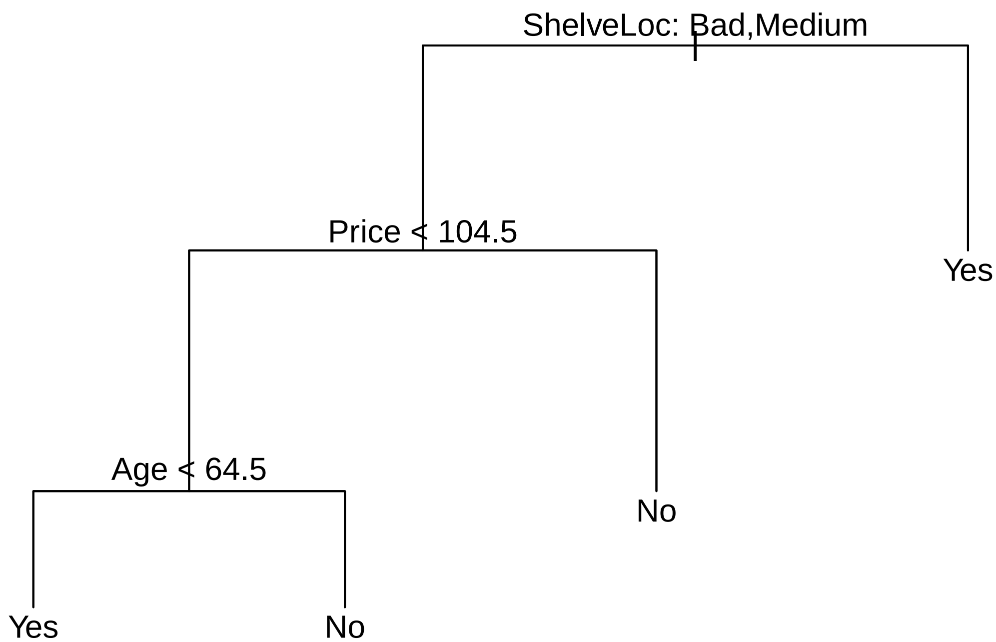


```r
set.seed(1)
train <- sample(1:nrow(Carseats), nrow(Carseats)/2)
Cartree <- tree(High ~ . - Sales, data=Carseats, subset=train)
cv.result2 <- cv.tree(Cartree, FUN=prune.misclass)
```

```r
plot(cv.result2$size, cv.result2$dev, type="b",
     main="Result of cross validation",
     xlab="Size", ylab="Deviance")
```




```r
fit2 <- prune.misclass(Cartree, best=4)
plot(fit2)
text(fit2, pretty=0)
```




```
test.y <- High[-train]
test.x <- Carseats[-train, ]
predict.y <- predict(fit2, newdata=test.x, type="class")
table(Prediction=predict.y, True=test.y)
```

```
          True
Prediction No Yes
       No  92  29
       Yes 24  55
```