Mehmet Duman

Solution to Homework #7—MTH 522

Problem 1 (Chapter 8 Exercises 7):
In the lab, we applied random forests to the Boston data using mtry=6 and using ntree=25 and ntree=500. Create a plot displaying the test error resulting from random forests on this data set for a more comprehensive range of values for mtry and ntree. You can model your plot after Figure 8.10. Describe the results obtained.

```
> set.seed(1)
> library(MASS)
> library(randomForest)
randomForest 4.6-12
Type rfNews() to see new features/changes/bug fixes.
```

We randomly divided the observations into a training and a test set

mtry : I applied random forests to the training set for three different values of the number of splitting variables m.
m = p = dim(Boston)[2] - 1 = 13
m = p/2
m = sqrt(p) ntree: A range of ntree from 1 to 500.

```
> train = sample(dim(Boston)[1], dim(Boston)[1]/2)
> X.train = Boston[train, -14]
> X.test = Boston[-train, -14]
> Y.train = Boston[train, 14]
> Y.test = Boston[-train, 14]

> p = dim(Boston)[2] - 1
> p.2 = p/2
> p.sq = sqrt(p)

> rf.boston.p = randomForest(X.train, Y.train, xtest = X.test, ytest = Y.test,
+     mtry = p, ntree = 500)

> rf.boston.p.2 = randomForest(X.train, Y.train, xtest = X.test, ytest = Y.test,
+     mtry = p.2, ntree = 500)

> rf.boston.p.sq = randomForest(X.train, Y.train, xtest = X.test, ytest = Y.test,
+     mtry = p.sq, ntree = 500)

> plot(1:500, rf.boston.p$test$mse, col = "green", type = "l", xlab = "Number of Trees",
+     ylab = "Test MSE", ylim = c(10, 19))

> lines(1:500, rf.boston.p.2$test$mse, col = "red", type = "l")
> lines(1:500, rf.boston.p.sq$test$mse, col = "blue", type = "l")
> legend("topright", c("m=p", "m=p/2", "m=sqrt(p)"), col = c("green", "red", "blue"),
+     cex = 1, lty = 1)

> dev.copy2pdf(file = "MTH522_hw7_p1.pdf", width = 8, height = 6, out.type = "pdf")
> dev.off()
```

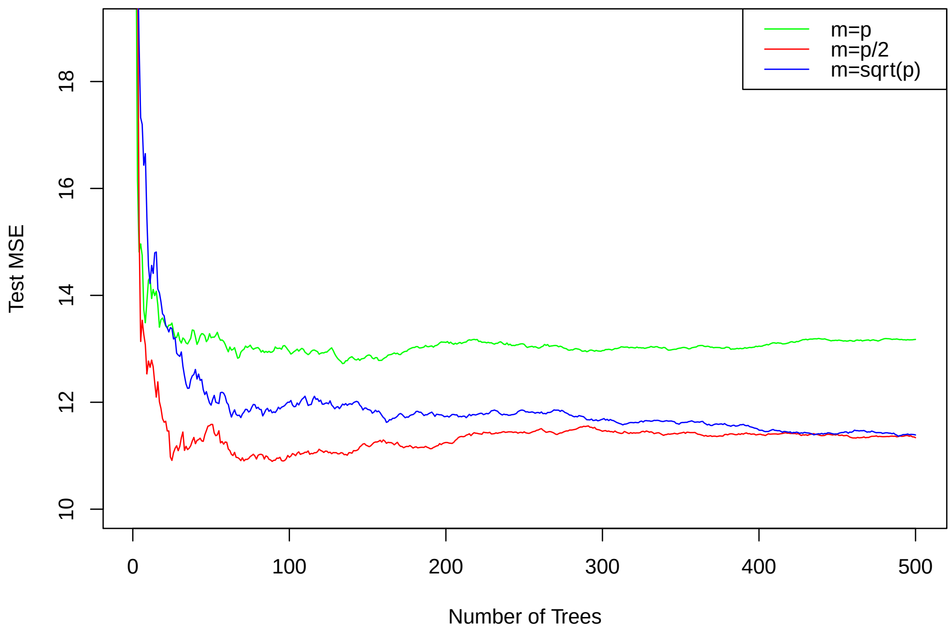


Figure 1: .

.

ntree: A range of ntree from 1 to 500. From Plot: For single treee the "Test MSE" is very high like 19,but when number of tree increases in the model the Test MSE drop down very quickly and after adding 100 trees to the model it stabilizes.
The Test MSE (m = p ,Green color) for all predictors is higher than other two of the number of predictors (m=p/2 and m=sqrt(p))

Problem 2 (Chapter 8 Exercises 8):
In the lab, a classification tree was applied to the Carseats data set after converting Sales into a qualitative response variable. Now we will seek to predict Sales using regression trees and related approaches, treating the response as a quantitative variable. (a) Split the data set into a training set and a test set.

```
> library(ISLR)
> attach(Carseats)
> set.seed(1)
> train = sample(dim(Carseats)[1], dim(Carseats)[1]/2)
> Carseats.train = Carseats[train, ]
> Carseats.test = Carseats[-train, ]
```

(b) Fit a regression tree to the training set. Plot the tree, and interpret the results. What test MSE do you obtain?

```
> install.packages("tree")
trying URL 'https://cran.cnr.berkeley.edu/bin/macosx/mavericks/contrib/3.3/tree_1.0-37.tgz'
Content type 'application/x-gzip' length 112331 bytes (109 KB)
==================================================
downloaded 109 KB


The downloaded binary packages are in
/var/folders/28/5cht8_x964n2w_4_wf_1tn3m0000gn/T//Rtmpi5oCSV/downloaded_packages




> library(tree)
> tree.carseats = tree(Sales ~ ., data = Carseats.train)
> summary(tree.carseats)

Regression tree:
tree(formula = Sales ~ ., data = Carseats.train)
Variables actually used in tree construction:
[1] "ShelveLoc"   "Price"       "Age"         "Advertising" "Income"      "CompPrice"
Number of terminal nodes:  18
Residual mean deviance:  2.36 = 429.5 / 182
Distribution of residuals:
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-4.2570 -1.0360  0.1024  0.0000  0.9301  3.9130


> plot(tree.carseats, cex = 0.2)
> text(tree.carseats, pretty = 0, cex = 0.7)
> dev.copy2pdf(file = "MTH522_hw7_p2a.pdf", width = 8, height = 6, out.type = "pdf")
> dev.off()
```

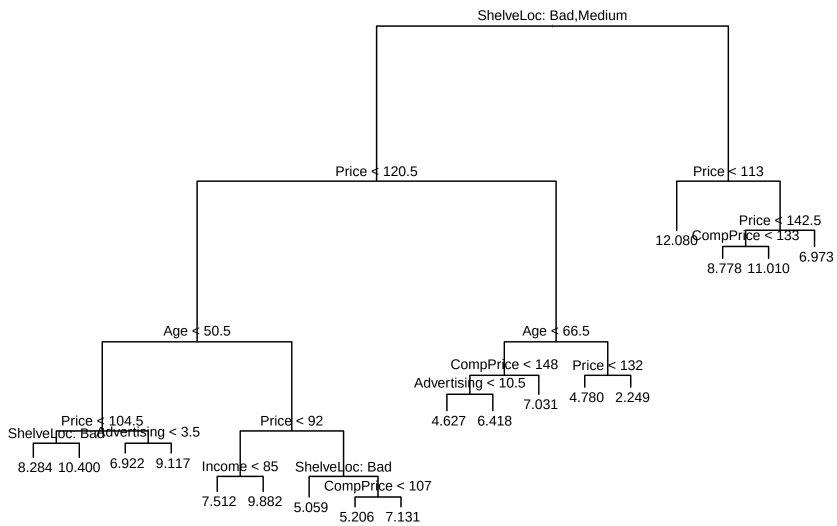


Figure 2: .

```
> pred = predict(tree.carseats, Carseats.test)
> mean((Carseats.test$Sales - pred)^2)

[1] 4.148897
```

Result, Test MSE is 4.148897

(c) Use cross-validation in order to determine the optimal level of tree complexity. Does pruning the tree improve the test MSE?

```
> cv.carseats = cv.tree(tree.carseats)
> plot(cv.carseats$size, cv.carseats$dev, type = "b")
> dev.copy2pdf(file = "MTH522_hw7_p2c.pdf", width = 8, height = 6, out.type = "pdf")
\ > dev.off()
```

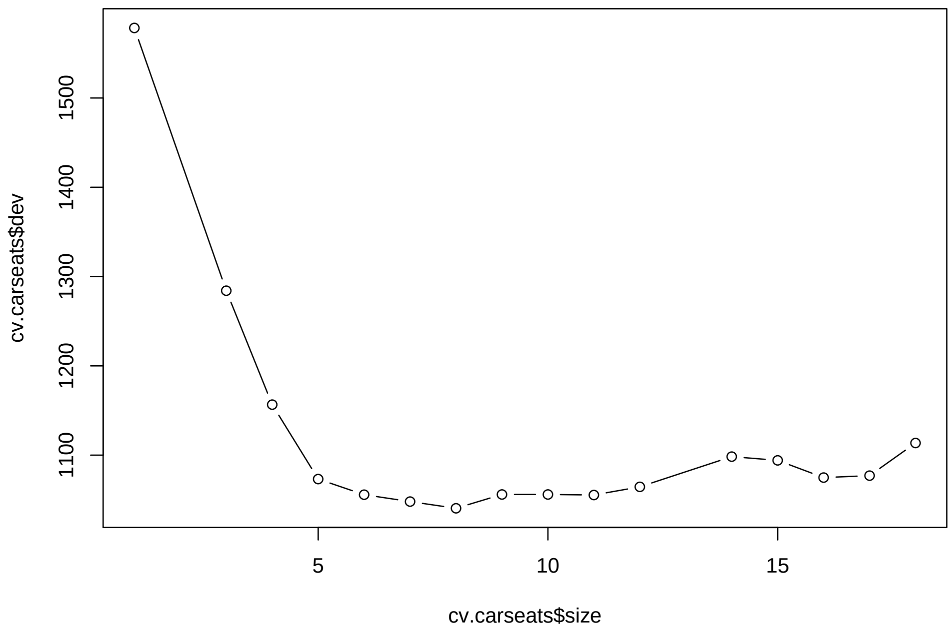


Figure 3: .

The tree of size 8 is selected by cross-validation. We now prune the tree to obtain the 8-node tree.

```
> pruned.carseats = prune.tree(tree.carseats, best = 8)
> plot(pruned.carseats)
> text(pruned.carseats, pretty = 0)
> dev.copy2pdf(file = "MTH522_hw7_p2c_2.pdf", width = 8, height = 6, out.type = "pdf")
> dev.off()
```

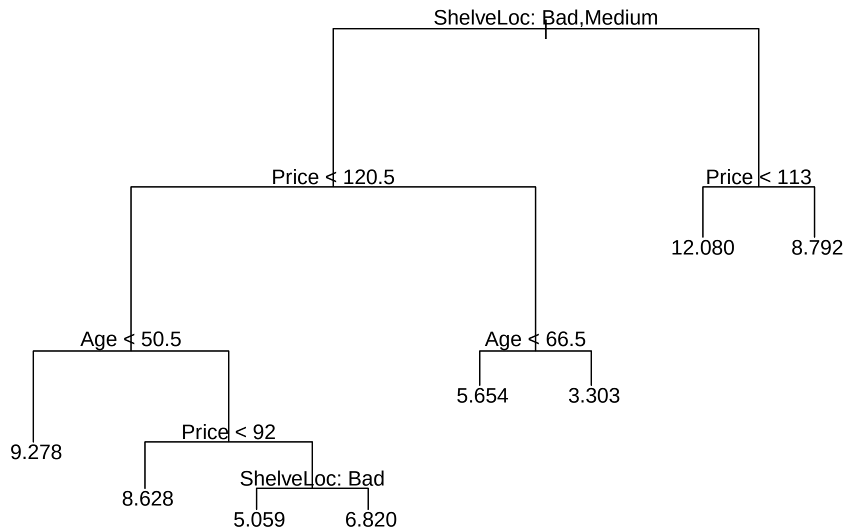


Figure 4: .

```
> pred = predict(pruned.carseats, Carseats.test)
> mean((Carseats.test$Sales - pred)^2)

[1] 5.09085
```

We may say that pruning the tree increases the Test MSE to 5.09085.

(d) Use the bagging approach in order to analyze this data. What test MSE do you obtain? Use the importance() function to determine which variables are most important.

```
> library(randomForest)
> carseats = randomForest(Sales ~ ., data = Carseats.train, mtry = 10, ntree = 500,
+      importance = T)
>  pred = predict(carseats, Carseats.test)
>  mean((Carseats.test$Sales - pred)^2)

[1] 2.583251
```

```
> importance(bag.carseats)

%IncMSE IncNodePurity
CompPrice   13.06849545    130.785188
Income       4.54737141     79.502339
Advertising 13.53207401    131.998744
Population   0.08549938     60.498175
Price       57.10673896    504.769865
ShelveLoc   43.30069615    320.415045
Age         21.91388932    184.199205
Education    2.50663690     40.334715
Urban       -3.11593207      8.526444
US           5.62400085     14.288796
```

We can say that bagging decrease the Test MSE error to 2.58.

"Price", "ShelveLoc" and "Age" are most important predictor veriables.

(e) Use random forests to analyze this data. What test MSE do you obtain? Use the importance() function to determine which variables are most important. Describe the effect of m, the number of variables considered at each split, on the error rate obtained.

```
> rf.carseats = randomForest(Sales ~ ., data = Carseats.train, mtry = 5, ntree = 500,
+      importance = T)
>  pred = predict(rf.carseats, Carseats.test)
>  mean((pred - Carseats.test$Sales )^2)

[1] 2.837792
```

```
> importance(rf.carseats)
%IncMSE IncNodePurity
CompPrice   11.7995764     123.99149
Income       4.5044593     109.90659
Advertising 13.0146889     132.11725
Population   2.0893802      84.74694
Price       46.3418812     442.97829
ShelveLoc   38.1022103     276.41698
Age         17.7985501     199.49882
Education    2.2709849      53.47066
Urban        0.1619189      11.04973
US           5.2523002      23.87314
```

We can say that random forest worsens the Test MSE error to 2.837792.

"Price", "ShelveLoc" and "Age" are still most important predictor veriables.

Problem 3 (Chapter 8 Exercises 9):
This problem involves the OJ data set which is part of the ISLR package.

```
> library(ISLR)
> attach(OJ)
> set.seed(1)
> summary(OJ)

Purchase WeekofPurchase     StoreID        PriceCH         PriceMM
CH:653   Min.   :227.0   Min.   :1.00   Min.   :1.690   Min.   :1.690
MM:417   1st Qu.:240.0   1st Qu.:2.00   1st Qu.:1.790   1st Qu.:1.990
Median :257.0   Median :3.00   Median :1.860   Median :2.090
Mean   :254.4   Mean   :3.96   Mean   :1.867   Mean   :2.085
3rd Qu.:268.0   3rd Qu.:7.00   3rd Qu.:1.990   3rd Qu.:2.180
Max.   :278.0   Max.   :7.00   Max.   :2.090   Max.   :2.290
DiscCH           DiscMM         SpecialCH        SpecialMM
Min.   :0.00000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
1st Qu.:0.00000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
Median :0.00000   Median :0.0000   Median :0.0000   Median :0.0000
Mean   :0.05186   Mean   :0.1234   Mean   :0.1477   Mean   :0.1617
3rd Qu.:0.00000   3rd Qu.:0.2300   3rd Qu.:0.0000   3rd Qu.:0.0000
Max.   :0.50000   Max.   :0.8000   Max.   :1.0000   Max.   :1.0000
LoyalCH          SalePriceMM     SalePriceCH      PriceDiff       Store7
Min.   :0.000011   Min.   :1.190   Min.   :1.390   Min.   :-0.6700   No :714
1st Qu.:0.325257   1st Qu.:1.690   1st Qu.:1.750   1st Qu.: 0.0000   Yes:356
Median :0.600000   Median :2.090   Median :1.860   Median : 0.2300
Mean   :0.565782   Mean   :1.962   Mean   :1.816   Mean   : 0.1465
3rd Qu.:0.850873   3rd Qu.:2.130   3rd Qu.:1.890   3rd Qu.: 0.3200
Max.   :0.999947   Max.   :2.290   Max.   :2.090   Max.   : 0.6400
PctDiscMM        PctDiscCH       ListPriceDiff       STORE
Min.   :0.0000   Min.   :0.00000   Min.   :0.000   Min.   :0.000
1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0.140   1st Qu.:0.000
Median :0.0000   Median :0.00000   Median :0.240   Median :2.000
Mean   :0.0593   Mean   :0.02731   Mean   :0.218   Mean   :1.631
3rd Qu.:0.1127   3rd Qu.:0.00000   3rd Qu.:0.300   3rd Qu.:3.000
Max.   :0.4020   Max.   :0.25269   Max.   :0.440   Max.   :4.000
```

(a) Create a training set containing a random sample of 800 observations, and a test set containing the remaining observations.

```
> train = sample(dim(OJ)[1], 800)
> OJ.train = OJ[train, ]
> OJ.test = OJ[-train, ]
```

(b) Fit a tree to the training data, with Purchase as the response and the other variables except for Buy as predictors. Use the summary() function to produce summary statistics about the tree, and describe the results obtained. What is the training error rate? How many terminal nodes does the tree have?

```
> library(tree)
> oj.tree = tree(Purchase ~ ., data = OJ.train)
> summary(oj.tree)

Classification tree:
tree(formula = Purchase ~ ., data = OJ.train)
Variables actually used in tree construction:
[1] "LoyalCH"       "PriceDiff"     "SpecialCH"     "ListPriceDiff"
Number of terminal nodes:  8
Residual mean deviance:  0.7305 = 578.6 / 792
Misclassification error rate: 0.165 = 132 / 800
```

The tree uses variables are "LoyalCH", "PriceDiff", "SpecialCH", "ListPriceDiff" and has 8 terminal nodes and a misclassification (training) error rate of 0.165.

(c) Type in the name of the tree object in order to get a detailed text output. Pick one of the terminal nodes, and interpret the information displayed.

```
> oj.tree


node), split, n, deviance, yval, (yprob)
* denotes terminal node

1) root 800 1064.00 CH ( 0.61750 0.38250 )
2) LoyalCH < 0.508643 350  409.30 MM ( 0.27143 0.72857 )
4) LoyalCH < 0.264232 166  122.10 MM ( 0.12048 0.87952 )
8) LoyalCH < 0.0356415 57   10.07 MM ( 0.01754 0.98246 ) *
9) LoyalCH > 0.0356415 109  100.90 MM ( 0.17431 0.82569 ) *
5) LoyalCH > 0.264232 184  248.80 MM ( 0.40761 0.59239 )
10) PriceDiff < 0.195 83   91.66 MM ( 0.24096 0.75904 )
20) SpecialCH < 0.5 70   60.89 MM ( 0.15714 0.84286 ) *
21) SpecialCH > 0.5 13   16.05 CH ( 0.69231 0.30769 ) *
11) PriceDiff > 0.195 101  139.20 CH ( 0.54455 0.45545 ) *
3) LoyalCH > 0.508643 450  318.10 CH ( 0.88667 0.11333 )
6) LoyalCH < 0.764572 172  188.90 CH ( 0.76163 0.23837 )
12) ListPriceDiff < 0.235 70   95.61 CH ( 0.57143 0.42857 ) *
13) ListPriceDiff > 0.235 102   69.76 CH ( 0.89216 0.10784 ) *
7) LoyalCH > 0.764572 278   86.14 CH ( 0.96403 0.03597 ) *
```

The asterisk (*) in the line shows that it is a terminal node. I picked terminal node label 9. The splitting value of this node is 0.0356415. There are 109 subtree below this node with deviance of 100.9. The prediction at this node is Sales = MM. About 17.431% of the observations in this node take the value of CH and remaining 82.569% take the value of MM.

(d) Create a plot of the tree, and interpret the results.

```
> plot(oj.tree, cex = 0.2)
> text(oj.tree, pretty = 0, cex = 0.7)
> dev.copy2pdf(file = "MTH522_hw7_p3d.pdf", width = 8, height = 6, out.type = "pdf")
> dev.off()
```

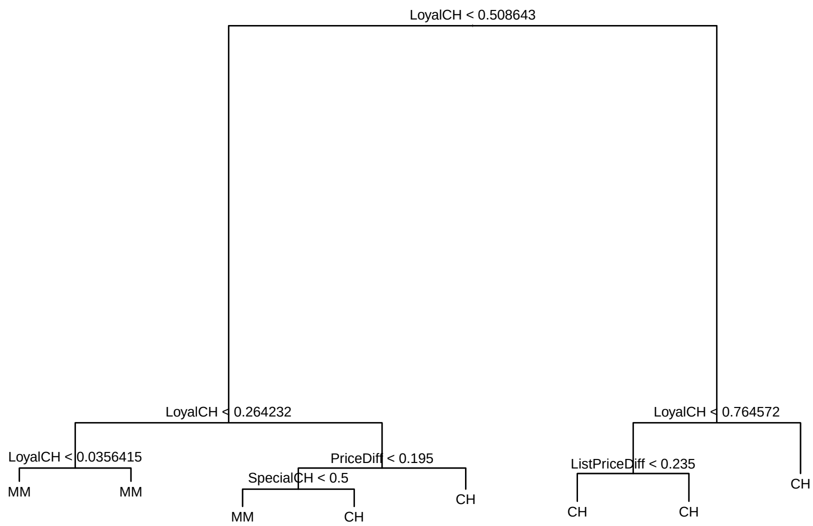


Figure 5: .

"LoyalCH" is the most important indicator of tree. The top three nodes contain "LoyalCH". If "LoyalCH" ¡ 0.264243, the tree predicts MM. If "LoyalCH" =¿ 0.764572, the tree predicts CH. Between those values of "LoyalCH", the decision depends on the value of "PriceDiff" and "ListPriceDiff"

(e) Predict the response on the test data, and produce a confusion matrix comparing the test labels to the predicted test labels. What is the test error rate?

```
> oj.pred = predict(oj.tree, OJ.test, type = "class")
> table(OJ.test$Purchase, oj.pred)


oj.pred
CH  MM
CH 147  12
MM  49  62





> 1 - (147 + 62) / (147+12+49+62)
[1] 0.2259259
```

Test error rate : 0.2259259

(f) Apply the cv.tree() function to the training set in order to determine the optimal tree size.

```
> cv.oj = cv.tree(oj.tree, FUN = prune.tree)
> cv.oj


$size
[1] 8 7 6 5 4 3 2 1

$dev
[1]  689.1001  685.8030  654.9314  653.7774  666.8890  721.2494  733.6936 1066.6499

$k
[1]      -Inf  11.20965  14.72877  17.88334  23.55203  38.37537  43.02529 337.08200

$method
[1] "deviance"

attr(,"class")
[1] "prune"         "tree.sequence"
```

(g) Produce a plot with tree size on the x-axis and cross-validated classification error rate on the y-axis.

```
> plot(cv.oj$size, cv.oj$dev, type = "b", xlab = "Tree Size", ylab = "Deviance")
> dev.copy2pdf(file = "MTH522_hw7_p3g.pdf", width = 8, height = 6, out.type = "pdf")
> dev.off()
```

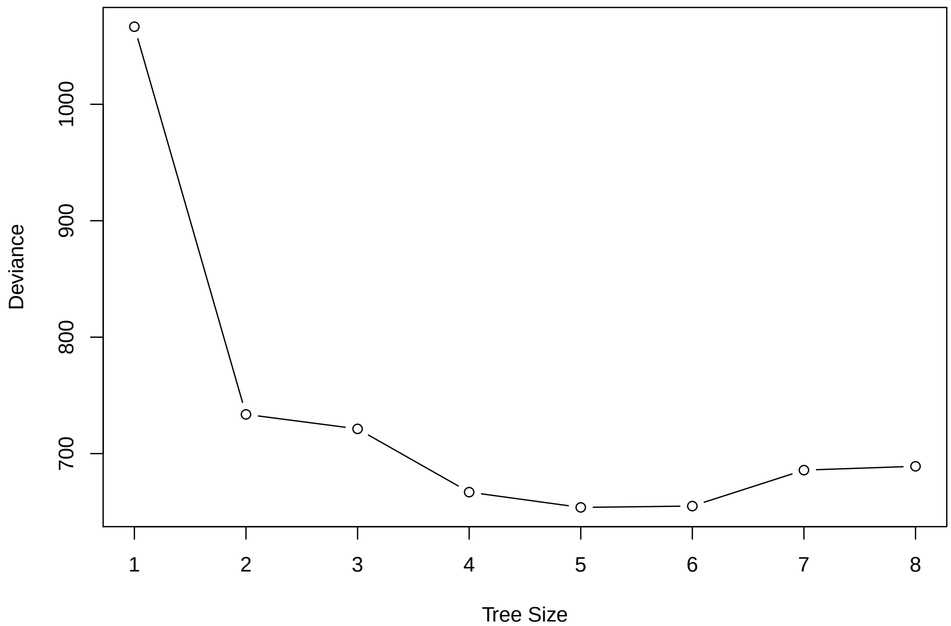


Figure 6: .

(h) Which tree size corresponds to the lowest cross-validated classi- fication error rate?

```
The 5-node tree is the smallest tree with the lowest cross-validation error rate.
```

(i) Produce a pruned tree corresponding to the optimal tree size obtained using cross-validation. If cross-validation does not lead to selection of a pruned tree, then create a pruned tree with five terminal nodes.

```
> oj.prune = prune.tree(oj.tree, best = 5)
> plot(oj.prune)
> text(oj.prune, pretty = 0)
> dev.copy2pdf(file = "MTH522_hw7_p3i.pdf", width = 8, height = 6, out.type = "pdf")
> dev.off()
```

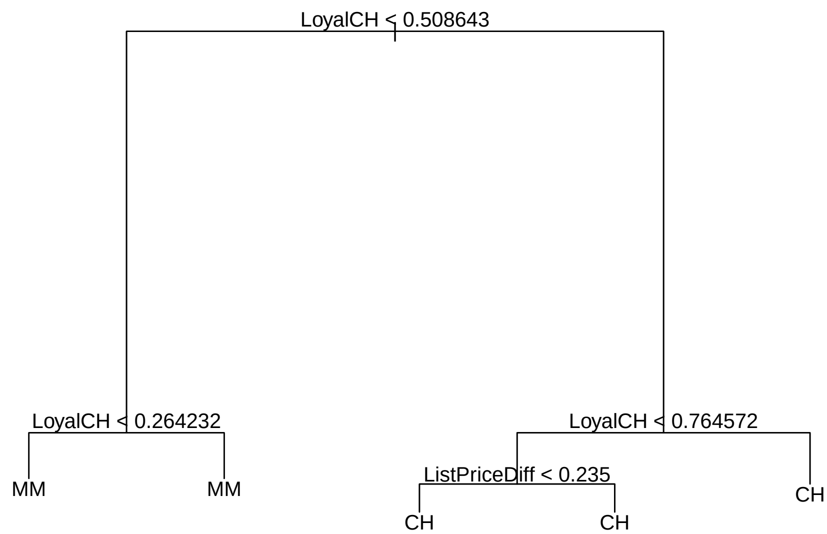


Figure 7: .

(j) Compare the training error rates between the pruned and un- pruned trees. Which is higher?

```
> summary(oj.tree)


Classification tree:
tree(formula = Purchase ~ ., data = OJ.train)
Variables actually used in tree construction:
[1] "LoyalCH"       "PriceDiff"     "SpecialCH"     "ListPriceDiff"
Number of terminal nodes:  8
Residual mean deviance:  0.7305 = 578.6 / 792
Misclassification error rate: 0.165 = 132 / 800

> summary(oj.prune)

Classification tree:
snip.tree(tree = oj.tree, nodes = 4:5)
Variables actually used in tree construction:
[1] "LoyalCH"       "ListPriceDiff"
Number of terminal nodes:  5
Residual mean deviance:  0.7829 = 622.4 / 795
Misclassification error rate: 0.1825 = 146 / 800
```

Misclassification error of prune tree (0.1825) is slightly higher than original treel tree (0.165).

(k) Compare the test error rates between the pruned and unpruned trees. Which is higher?

```
> pred.unprune = predict(oj.tree, OJ.test, type = "class")
> table(pred.unprune, OJ.test$Purchase)

pred.unprune  CH  MM
CH 147  49
MM  12  62

> 1 - (147+62) / (147+49+12+62)
[1] 0.2259259


> pred.prune = predict(oj.prune, OJ.test, type = "class")
> table(pred.prune, OJ.test$Purchase)

pred.prune  CH  MM
CH 119  30
MM  40  81

> 1 - (119 + 81) / (119+30+40+81)
[1] 0.2592593
```

Prune test error rate 0.2592593 is slightly higher than unpruned trees error rate of 0.2259259 .

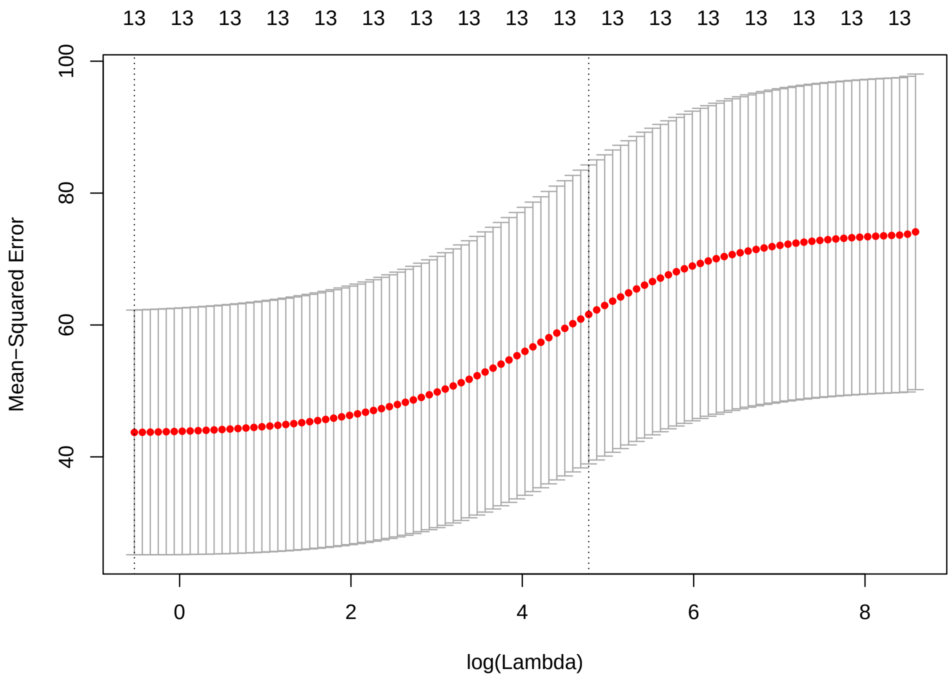
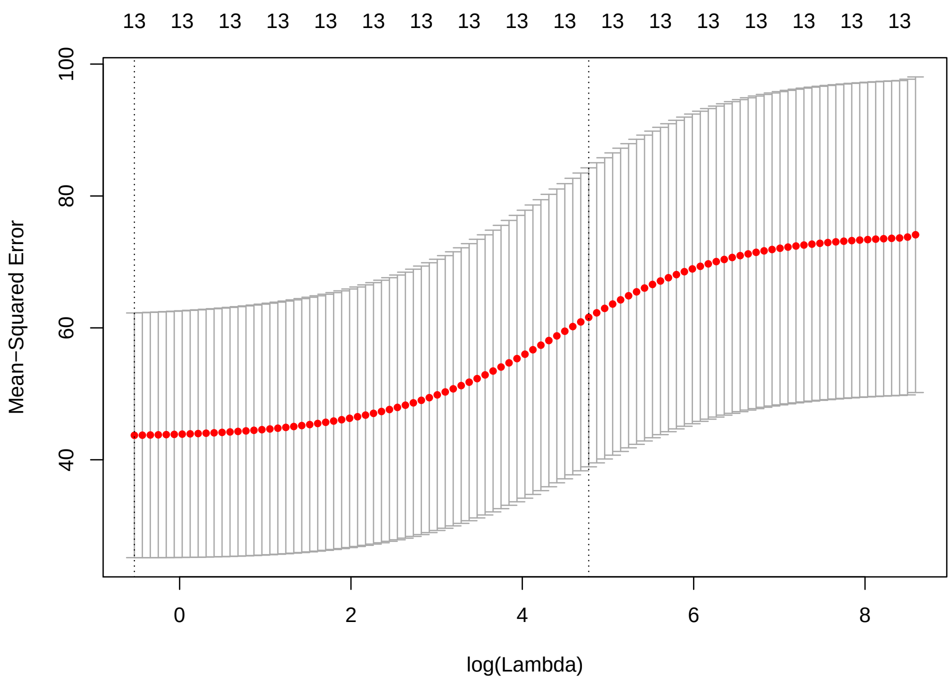

13 13 13 13 13 13 13 13 13 13 13 13 13 13 13 13 13
100
80
60
40
Mean−Squared Error
0
2
4
6
8
log(Lambda)

Figure 8: .

deeneme



13 13 13 13 13 13 13 13 13 13 13 13 13 13 13 13 13
Mean−Squared Error
100
80
60
40
0
2
4
6
8
log(Lambda)

Figure 9: .