R Lab #6 – Solutions

---

1. 8.3.1: Fitting Classification Trees

Work through pages 324-325 of the lab until you enter "tree.carseats." Make sure you understand the output. For example, "Income < 57 10 12.220 No (0.70000 0.30000)" means there are 10 observations at this node, which splits on "Income < 57," and 70% of these 10 observations (i.e., 7) have a response values of "No" while 30% of these 10 observations (i.e., 3) have a response values of "Yes." Thus the predicted response at this node is "No." The "12.220" is the deviance at this node. You should find that the next line of output (corresponding to "CompPrice < 110.5") is starred, which indicates a terminal (leaf) node. What is the number of "No" observations at this node and what is the predicted response?

a) 1, Yes
b) 2, Yes
c) 3, No
d) 4, No
e) 5, No

Solution: E

```
# Income < 57 10  12.220 No ( 0.70000 0.30000 )
.70 * 10  # 7 observations have response values of "No"
.30 * 10  # 3 observations have response values of "Yes"

# CompPrice < 110.5 5   0.000 No ( 1.00000 0.00000 ) *
1.0 * 5  # 5 observations have the response value of "No"
```

2. 8.3.1: Fitting Classification Trees

Continue working through the lab from the bottom of page 325 to the top of page 327. You should find that the pruned nine-node tree (prune.carseats) uses five predictor variables. Three of these variables are ShelveLoc, Price, and Advertising. What are the other two?

a) Income
b) CompPrice
c) Population
d) Age
e) US

Solution: B, D

```
prune.carseats <- prune.misclass(tree.carseats, best = 9)
summary(prune.carseats)
prune.carseats
par(mfrow = c(1, 1))
```

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

3. 8.3.1: Fitting Classification Trees

Continue with this section of the lab on page 327 by assessing the nine-node pruned tree on the test data set. Confirm that 77% of the test observations are correctly classified. The instructions in the lab go on to show how to fit a pruned tree with 15 terminal nodes that has 74% classification accuracy. True or false? A pruned tree with 13 terminal nodes achieves the same 77% classification accuracy as the nine-node pruned tree.

Solution: True

```
# If we decrease the value of best to 13, we obtain the same classification accuracy as best = 9
prune.carseats <- prune.misclass(tree.carseats, best = 13)
plot(prune.carseats)
text(prune.carseats, pretty = 0)
tree.pred <- predict(prune.carseats, Carseats.test, type = "class")
table(tree.pred, High.test)
mean(tree.pred == High.test) # Correct Prediction Rate = 77%
```

4. 8.3.2 Fitting Regression Trees

Work through section 8.3.2 on pages 327-328. Confirm that the full (unpruned) tree predicts a median house price of 46.4 thousand dollars ($46,400) for larger homes in suburbs in which the percentage of individuals with lower socioeconomic status is less than 9.715% (rm>=7.437 and lstat<9.715). What is the predicted median house price (in thousands of dollars) for suburbs in which the percentage of individuals with lower socioeconomic status is greater than or equal to 21.49%? (Round your answer to 1 decimal place.)

Solution: 11.1

```
set.seed(1)
train <- sample(1:nrow(Boston), nrow(Boston)/2)
names(Boston)
tree.boston <- tree(medv ~ ., data = Boston, subset = train)
summary(tree.boston)
plot(tree.boston)
text(tree.boston, pretty = 0)
```

5. 8.3.3 Bagging and Random Forests

Work through this section of the lab on pages 328 to 330. Note the warning at the bottom of page 328 that your results may not match those in the book exactly depending on the versions of R and the randomForest package that you're using. For example, I obtain test set MSEs of 13.47 for the bagged model based on 500 trees (vs. 13.16 in the book), 13.43 for the bagged model based on 25 trees (vs. 13.31 in the book), and 11.48 for the random forest based on mtry=6 (vs. 11.31 in the book). Compare the scatterplot of fitted values versus response values for the bagged model based on 500 trees with the similar plot for the regression tree fit in the previous section. Match the plots to the descriptions of their appearance.

a) Bagged model
b) Regression tree

- General positive association overall, but the points are arranged in vertical lines at eight discrete horizontal axis values. – B
- General positive association overall, with the points distributed more randomly in the horizontal direction. – A

Solution:

```
plot(yhat.bag,boston.test)
plot(yhat,boston.test)
```

6. 8.3.3 Bagging and Random Forests

Enter set.seed(1), then fit another random forest model but exclude the mtry argument so that it uses its default value. Confirm that the number of variables tried at each split is 13/3=4 (to the nearest whole number) by typing rf.boston (or whatever name you have used for the model). True or false? The test set MSE for this model is lower than the test set MSE you obtained for the random forest model with mtry=6.

Solution: False

```
set.seed(1)
rf2.boston <- randomForest(medv ~., data = Boston, subset = train, importance = T)
yhat.rf2 <- predict(rf2.boston, newdata = Boston[-train,])
mean((yhat.rf2 - boston.test)^2) # test MSE = 11.72
```

7. 8.3.4 Boosting

Work through this section on pages 330-331 and confirm the results in the book. Create a partial dependence plot for the third most important variable, dis (a weighted mean of distances to five Boston employment centres). Do median house prices increase or decrease with dis?

a) Increase
b) Decrease

Solution: B

plot(boost.boston, i = "dis")

8. 8.3.4 Boosting

Enter set.seed(1) and then fit a boosted model with n.trees=5000, interaction.depth=4, shrinkage=0.01. What is the test MSE? (Round your answer to 1 decimal place.)

Solution: 10.2

```
set.seed(1)
boost.boston2 <- gbm(medv ~., data = Boston[train,], distribution = "gaussian", n.trees = 5000,
                     interaction.depth = 4, shrinkage = 0.01)
yhat.boost2 <- predict(boost.boston2, newdata = Boston[-train,], n.trees = 5000)
mean((yhat.boost2 - boston.test)^2)  # test MSE = 10.2
```

9. 8.3.4 Boosting

Enter set.seed(1) and then fit a boosted model with n.trees=5000, interaction.depth=3, shrinkage=0.01. What is the test MSE? (Round your answer to 1 decimal place.)

Solution: 10.5

```
set.seed(1)
boost.boston3 <- gbm(medv ~., data = Boston[train,], distribution = "gaussian", n.trees = 5000,
                     interaction.depth = 3, shrinkage = 0.01)
yhat.boost3 <- predict(boost.boston3, newdata = Boston[-train,], n.trees = 5000)
mean((yhat.boost3 - boston.test)^2)  # test MSE = 10.5
```