Regression Tree Exercises

Shin, Joonwoo

Exercise 1 Getting Ready

Description: We will use "ISLR2" package and "tree" package for consequent exercises. Open your R program and install "ISLR2" and "tree" packages. Alternatively, you may also use 'rpart' for building trees.

Solution:

```r
install.packages("ISLR2")
install.package("tree")
library(ISLR2)
library(tree)

#rpart
install.packages("rpart")
library(rpart)
```

Exercise 2 Fitting Regression Trees

Description : Fit a regression tree to the Boston data set. First create a training set and fit the tree to the training data.

Solution :

```r
set.seed(1)
train_set <- sample(1:nrow(Boston),nrow(Boston)/2)
tree.boston <- tree(medv ~ . ,Boston, subset = train_set)
summary(tree.boston)
```

Exercise 3 Pruning Regression Trees

Description : Use cv.tree to prune the tree obtained in Exercise 2.

Exercise 4 Recursive Partition and Regression Tree

Description In this exercise we shall use rpart package. Use kyphosis data and fit kyphosis to age, number and start variable. Plot the result.

solution :

```r
fit <- rpart(Kyphosis ~ Age + Number + Start, data = kyphosis)
par(mfrow = c(1,2), xpd = NA) # otherwise on some devices the text is
    clipped
plot(fit)
text(fit, use.n = TRUE)
```

Exercise 5 Simple Linear Model versus Regression Trees

Description : Generate a random data consist of two dimensional vectors. Choose one component as a label and one component as a covariate. Fit a linear regression model and regression tree to the generated data. Compare summary statistics and plot the results on the XY plane.

Exercise 6 Fitting Regression Trees

Description : This exercise is from ”An Introduction to Statistical Learning: with applications in R” second edition, section 8.4 exercise 8. Use Carseats data included in ISLR2 package to answer following questions (a) Split the data set into traing set and a test set (b) Fit a regression tree to the training data to predict Sales. Plot the tree and interpret the result. What test MSE do you obtain? (c) Use Cross Validation to determine optimal level of tree complexity. Does Pruning the tree improve the test MSE?

solution :

```
# Exercise 8 of ISLR2, Section 8.4

#Exercise 8
#(a) Splitt the Carseat data into two : training set and test set
set.seed(1)
# Total number of observations
data(Carseats)
n <- nrow(Carseats)

# Proportion of data to use for training (we will use 50%)
train_size <- floor(0.5 * n)

# Randomly sample row indices for the training set
train_indices <- sample(seq_len(n), size = train_size)

# Split the data
train_data <- Carseats[train_indices, ]
test_data  <- Carseats[-train_indices, ]

# Check sizes
cat("Training set size:", nrow(train_data), "\n")
cat("Test set size:", nrow(test_data), "\n")

#b. Fit a regression tree to the training data to predict Sales. Plot the tree and interpret the result. What test MSE do you obtain?
tree.carseats <- tree(Sales ~ ., data=train_data)
summary(tree.carseats)
par(mar=c(1,1,1,1))
plot(tree.carseats)
text(tree.carseats, pretty=0)

#Computing MSE
yhat <- predict(tree.carseats, newdata=test_data)
carseats.test <- test_data["Sales"]
y <-carseats.test[[1]]
plot(yhat, y)
abline(0,1)
mean((yhat - y)^2)

```

```
#(c) use c.v to determine 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")
prune.carseats <- prune.tree(tree.carseats, best=5)
plot(prune.carseats)
text(prune.carseats, pretty=0)

yhat_prune <- predict(prune.carseats, newdata=test_data)
mean((yhat_prune-y)^2)
```

Exercise 7 More regression trees

Description See Exercise 9 of the same book as above. Use OJ data in ISLR2 package to answer following questions.

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

(b) Fit a tree to the training data, with purchase as the response and the other variables as predictors. Use 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?

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

References

[1] Islr2 package description. https://cran.r-project.org/web/packages/ISLR2/index.html.

[2] rpart package description. https://cran.r-project.org/web/packages/rpart/rpart.pdf.

[3] tree package description. https://cran.r-project.org/web/packages/tree/index.html.

[4] Trevor Hastie Robert Tibshirani Gareth James, Daniela Witten. *Introduction to Statistical Learning with applications in R*. Springer.