Homework: Ch 05

STAT 4510/7510

Due Tuesday, March 8, 11:59 pm

Instructions: Please list your name and student number clearly. In order to receive credit for a problem, your solution must show sufficient detail so that the grader can determine how you obtained your answer.

Submit a single pdf file for your final outcome. All `R` code should be included, as well as all output produced. Upload your work to the Canvas course site.

Chapter 5 Lab Exercise

I strongly suggest you to read the textbook 5.3 Lab: Cross-Validation and the Bootstrap, found on pages 190 - 197. The data set files `Auto.csv`, `Portfolio.csv`, and the `R` code used in this Lab can be found in Canvas. Just run each line of the code and see what happens.

- You don't have to submit your work for this, but to solve the following homework problems, you may want to do this first.

Problem 1

In this question, we will use the `Default` data set used in Chapter 4. You can find the data set file under Module 4 in Canvas. We will use logistic regression to predict the probability of `default` using `income` and `balance`.

We will estimate the test error of this model using the validation set approach. Read the data set as follows.

```
Default <- read.csv("Default.csv", stringsAsFactors = TRUE)
Default <- Default[,-1] # Remove the first index column
summary(Default)
```

```
##  default    student       balance           income
##  No :9667   No :7056   Min.   :   0.0   Min.   :  772
##  Yes: 333   Yes:2944   1st Qu.: 481.7   1st Qu.:21340
##                        Median : 823.6   Median :34553
##                        Mean   : 835.4   Mean   :33517
##                        3rd Qu.:1166.3   3rd Qu.:43808
##                        Max.   :2654.3   Max.   :73554
```

(a) Using the entire observations in the data set, fit a logistic regression model that uses `income` and `balance` to predict `default`. Show the summary of the fitted model.

(b) Obtain a predicted probability of default status for each individual in the data set by using `predict()` function, and classify the individual to the default category if the predicted probability is greater than 0.5. Provide the confusion matrix, and compute the error rate.

(c) Now we will use the validation set approach to estimate the test error rate of this model. Create a vector `train` that includes the indices of training set (70%) by using `sample()` function. The data corresponding to the remaining indices will be used as a validation set (30%). Before doing this, set the seed number as follows.

```
set.seed(1)
```

(d) Using the training data, fit a logistic regression model that uses `income` and `balance` to predict `default`. Show the summary of the fitted model.

(e) Using the fitted model obtained in (d), predict the class labels for the validation set. Find the error rate and compare with the training error rate what you obtained from (b).

Problem 2

In this question, we will use the `Auto` data set again to implement the LOOCV without using the `cv.glm()` function. Read the data set as follows.

```
Auto <- read.csv("Auto.csv", na.strings ="?") # With the option, R recognizes ? as NA.
Auto <- na.omit(Auto) # Remove data rows including NA.
Auto$origin <- as.factor(Auto$origin) # Coerce the type of origin into factor
```

(a) We will fit a simple linear regression model with `mpg` as the response variable and `horsepower` as the predictor variable. Run the following code and compare the outcome with what we obtain from `cv.glm` function provided by `boot` library (See the page 17 of Chapter 5 lecture slides).

```
# LOOCV
res <- numeric(length = 392)
for (i in 1:392) {
  lmfit.loocv <- lm(mpg ~ horsepower, data = Auto[-i, ])
  yhat <- predict(lmfit.loocv, data.frame(horsepower = Auto$horsepower[i]))
  res[i] <- Auto[i,]$mpg - yhat
}
mean(res^2)
```

(b) From the code in (a), explain in your own words about the option `data = Auto[-i, ]` in the `lm` function.

(c) From the code in (a), explain in your own words about the input `data.frame(horsepower = Auto$horsepower[i])` in the `predict` function.