

# Homework 6

Alex Soupir

March 16, 2020

*Packages:* ISLR

*Collaborators:*

Answer all questions specified on the problem and include a discussion on how your results answered/addressed the question.

Submit your **.rmd** file with the knitted **PDF** (or knitted Word Document saved as a PDF). If you are having trouble with .rmd, let us know and we will help you, but both the .rmd and the PDF are required.

This file can be used as a skeleton document for your code/write up. Please follow the instructions found under Content for Formatting and Guidelines. No code should be in your PDF write-up unless stated otherwise.

There is no need for plotting in this assignment.

You do not need to include the above statements.

Please do the following problems from the text book ISLR or written otherwise.

1. Question 5.4.3 pg 198

We now review  $k$ -fold cross-validation.

- (a) Explain how  $k$ -fold cross-validation is implemented.

**$k$ -fold cross-validation is implemented by splitting the data into  $k$  sets, then removing one of the sets and training the model and testing on set. This is repeated for all  $k$  sets. Error is then calculated by averaging the MSE from all of the models.**

- (b) What are the advantages and disadvantages of  $k$ -fold cross-validation relative to:

- i. The validation set approach?

**Validation set is easy to implement, but can cause a bias based on which samples are included in the training and testing set, but the variation is much greater than that of the  $k$ -fold cross-validation.**

- ii. LOOCV?

**Leave one out can use a lot of computational power as mentioned in lecture. As shown on one of the slides, LOOCV may have slightly lower MSE than  $k$ -fold cross-validation but  $k$ -fold variation is low.**

2. Question 5.4.5 pg 198 (use `set.seed(702)` to make results replicable)

In Chapter 4, we used logistic regression to predict the probability of `default` using `income` and `balance` on the `Default` data set. We will now estimate the test error of this logistic regression model using the validation set approach. Do not forget to set a random seed before beginning your analysis.

- (a) Fit a logistic regression model that uses `income` and `balance` to predict `default`.

- (b) Using the validation set approach, estimate the test error of this model. In order to do this, you must perform the following steps:
- Split the sample set into a training set and a validation set.
  - Fit a multiple logistic regression model using only the training observations.
  - Obtain a prediction of default status for each individual in the validation set by computing the posterior probability of default for that individual, and classifying the individual to the `default` category if the posterior probability is greater than 0.5.
  - Compute the validation set error, which is the fraction of the observations in the validation set that are misclassified.

```
## [1] 0.9720812
```

- (c) Repeat the process in (b) three times, using three different splits of the observations into a training set and validation set. Comment on the results obtained.

**Maybe easier to use a function to run this 3 times?**

```
## [1] 0.9720812
## [1] 0.9699746
## [1] 0.9769585
```

- (d) Now consider a logistic regression model that predicts the probability of `default` using `income`, `balance`, and a dummy variable for `student`. Estimate the test error for this model using the validation set approach. Comment on whether or not including a dummy variable for `student` leads to a reduction in the test error rate.

```
## [1] 0.9720812
```

**Adding the variable for `student` produces similar (same) error when setting the seed to the same one as used in (b). This suggests that the dummy variable doesn't reduce the error test rate when included.**

### 3. Question 5.4.7 pg 200

In Sections 5.3.2 and 5.3.3, we saw that the `cv.glm()` functions can be used in order to compute the LOOCV test error estimate. Alternatively, one could compute those quantities using just the `glm()` and `predict.glm()` functions, and a for loop. You will now take this approach in order to compute the LOOCV error for a simple logistic regression model on the `Weekly` data set. Recall that in the context of classification problems, the LOOCV error is given in (5.4).

- Fit a logistic regression model that predicts `Direction` using `Lag1` and `Lag2`.
- Fit a logistic regression model that predicts `Direction` using `Lag1` and `Lag2` *using all but the first observation*.
- Use the model from (b) to predict the direction of the first observation. You can do this by predicting that the first observation will go up if  $P(\text{Direction}=\text{"Up"}|\text{Lag1}, \text{Lag2}) > 0.5$ . Was this observation correctly classified?

```
##      1
## "Up"
## Actual Direction:
## [1] Down
## Levels: Down Up
```

**The observation was incorrectly classified.**

- (d) Write a for loop from  $i = 1$  to  $i = n$ , where  $n$  is the number of observations in the data set, that performs each of the following steps:
  - i. Fit a logistic regression model using all but the  $i$ th observation to predict `Direction` using `Lag1` and `Lag2`.
  - ii. Compute the posterior probability of the market moving up for the  $i$ th observation.
  - iii. Use the posterior probability for the  $i$ th observation in order to predict whether or not the market moves up.
  - iv. Determine whether or not an error was made in predicting the direction for the  $i$ th observation. If an error was made, then indicate this as a 1, and otherwise indicate it as a 0.
- (e) Take the average of the  $n$  numbers obtained in (d)iv in order to obtain the LOOCV estimate for the test error. Comment on the results.

```
## [1] 0.4499541
```

The error is rather high. Looking at the values of individual errors and comparing them with the `Direction` variable in the data, the model doesn't seem to be predicting the same value for all observations, meaning it isn't predicting "Up" or "Down" for each observation. Interestingly though, if it were to predict Up for all of the observations, the error (the observations that are incorrect over the total number of observations) would be  $484/1089$ , which is 0.444 and is lower than that of the model's prediction error.

4. Write your own code (similar to Q #3. above) to estimate test error using 6-fold cross validation for fitting linear regression with  $mpg \sim horsepower + horsepower^2$  from the Auto data in the ISLR library. You should show the code in your final PDF.

Going off of the slide for 5.1.3 k-fold Cross Validation from lecture, I will split the data into  $k$  (6) different sets. Then, I'll remove one set, train the model on the remaining data, and test on the set that was removed, calculating the MSE. I'll do this for all  $k$  data sets, storing the MSE for each as a term in a vector. After all  $k$  sets are tested, I will take the average of all of the MSE terms.

```
set.seed(702)
data("Auto")
error = double()
sample.ind = sample(6, nrow(Auto), replace=true)
for(i in seq(max(sample.ind))){
  train.4 = Auto[c(sample.ind != i),]
  valid.4 = Auto[c(sample.ind == i),]

  model.4 = glm(mpg ~ horsepower + horsepower^2, data=train.4)
  preds.4 = round(predict(model.4, valid.4), digits=0)
  error[i] = mean((preds.4 - valid.4$mpg)^2)
}
mean(error)
```

```
## [1] 24.67371
```

5. Last homework you started analyzing the dataset you chose from [this website](#). Now continue the analysis and perform Logistic Regression, KNN, LDA, QDA, MclustDA, MclustDA with EDDA if appropriate. If it is not possible to perform any of the methods mentioned above please justify why.

## References

- ISLR Book
-