

STAT 4700/5700 Homework 2 Due 3/14/2023

1. Chapter 4 (Exercises 10)

This question should be answered using the **Weekly** data set, which is part of the **ISLR** package. This data is similar in nature to the **Smarket** data from this chapter's lab, except that it contains 1,089 weekly returns for 21 years, from the beginning of 1990 to the end of 2010.

- (a) Produce some numerical and graphical summaries of the **Weekly** data. Do there appear to be any patterns?
- (b) Use the full data set to perform a logistic regression with **Direction** as the response and the five lag variables plus **Volume** as predictors. Use the summary function to print the results. Do any of the predictors appear to be statistically significant? If so, which ones?
- (c) Compute the confusion matrix and overall fraction of correct predictions. Explain what the confusion matrix is telling you about the types of mistakes made by logistic regression.
- (d) Now fit the logistic regression model using a training data period from 1990 to 2008, with **Lag2** as the only predictor. Compute the confusion matrix and the overall fraction of correct predictions for the held out data (that is, the data from 2009 and 2010).
- (e) Repeat (d) using LDA.
- (f) Repeat (d) using QDA.
- (g) Repeat (d) using KNN with $K = 1$.
- (h) Which of these methods appears to provide the best results on this data?
- (i) Experiment with different combinations of predictors, including possible transformations and interactions, for each of the methods. Report the variables, method, and associated confusion matrix that appears to provide the best results on the held out data. Note that you should also experiment with values for K in the KNN classifier.

2. Chapter 5 (Exercies 7)

In Sections 5.3.2 and 5.3.3, we saw that the `cv.glm()` function 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).

- (a) Fit a logistic regression model that predicts `Direction` using `Lag1` and `Lag2`.
- (b) Fit a logistic regression model that predicts `Direction` using `Lag1` and `Lag2` *using all but the first observation*.
- (c) 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?
- (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.

3. Chapter 5 (Exercies 9)

We will now consider the **Boston** housing data set, from the **MASS** library.

- (a) Based on this data set, provide an estimate for the population mean of **medv**. Call this estimate $\hat{\mu}$.
- (b) Provide an estimate of the standard error of $\hat{\mu}$. Interpret this result.

Hint: We can compute the standard error of the sample mean by dividing the sample standard deviation by the square root of the number of observations.

- (c) Now estimate the standard error of $\hat{\mu}$ using the bootstrap. How does this compare to your answer from (b)?
- (d) Based on your bootstrap estimate from (c), provide a 95 % confidence interval for the mean of **medv**. Compare it to the results obtained using `t.test(Boston$medv)`.

Hint: You can approximate a 95 % confidence interval using the formula $[\hat{\mu} - 2SE(\hat{\mu}), \hat{\mu} + 2SE(\hat{\mu})]$.

- (e) Based on this data set, provide an estimate, $\hat{\mu}_{med}$, for the median value of **medv** in the population.
- (f) We now would like to estimate the standard error of $\hat{\mu}_{med}$. Unfortunately, there is no simple formula for computing the standard error of the median. Instead, estimate the standard error of the median using the bootstrap. Comment on your findings.
- (g) Based on this data set, provide an estimate for the tenth percentile of **medv** in Boston suburbs. Call this quantity $\hat{\mu}_{0.1}$. (You can use the `quantile()` function.)
- (h) Use the bootstrap to estimate the standard error of $\hat{\mu}_{0.1}$. Comment on your findings.