Chapter – 4

ISLR 4.7

Q5. We now examine the differences between LDA and QDA.

a. **If the Bayes decision boundary is linear, do we expect LDA or QDA to perform better on the training set? On the test set?**

   We assume that QDA will perform better on the training set if the Bayes decision boundary is linear because its greater flexibility might result in a better match. Since QDA might overfit the linearity on the Bayes decision boundary, we anticipate LDA to perform better on the test set than QDA.

b. **If the Bayes decision boundary is non-linear, do we expect LDA or QDA to perform better on the training set? On the test set?**

   We assume that QDA will perform better on both the training and test sets if the Bayes decision boundary is non-linear.

c. **In general, as the sample size nn increases, do we expect the test prediction accuracy of QDA relative to LDA to improve, decline, or be unchanged? Why?**

   If the training set is very large, QDA is advised so that the classifier's variance is not a major worry (QDA is more flexible than LDA, hence it has more variance).

d. **True or False: Even if the Bayes decision boundary for a given problem is linear, we will probably achieve a superior test error rate using QDA rather than LDA because QDA is flexible enough to model a linear decision boundary. Justify your answer.**

   False. A more flexible approach, such as QDA, may produce overfitting when there are fewer sample points, which could result in a worse test error rate.

Q6. Suppose we collect data for a group of students in a statistics class with variables X1 =hours studied, X2=undergrad GPA, and Y = receive an A. We fit a logistic regression and produce estimated coefficient, ˆβ0 = −6, ˆβ1 =0.05, ˆβ2 = 1.

a. **Estimate the probability that a student who studies for 40 hours and has an undergrad GPA of 3.5 gets an A in the class.**

It suffices to plug in the beta values in the equation for predicted probability

$$\hat{p}(X) = \frac{e^{-6+0.05X_1+X_2}}{\left(1+e^{-6+0.05X_1+X_2}\right)}$$

where $X_1 = 40$ & $X_2 = 3.5$

$$= \frac{e^{-6+0.05(40)+3.5}}{1+e^{-6+0.05(40)+3.5}}$$

$$= 0.3775$$

b. **How many hours would the student in part (a) need to study to have a 50% chance of getting an A in the class?**

the equation for predicted probability

$$\frac{e^{-6+0.05X_1+3.5}}{\left(1+e^{-6+0.05X_1+3.5}\right)} = 0.5,$$

which is equal to

$$e^{-6+0.05X_1+3.5} = 1$$

taking log on both sides

$$X_1 = \frac{2.5}{0.05} = 50$$

Q8. Suppose that we take a data set, divide it into equally sized training and test sets, and then try out two different classification procedures. First, we use logistic regression and get an error rate of 20% on the training data and 30% on the test data. Next, we use 1-nearest neighbors (i.e., K = 1) and get an average error rate (averaged over both test and training data sets) of 18%. Based on these results, which method should we prefer to use for classification of new observations? Why?

for KNN with K=1, the training error rate is 0 as all the training data will be classified correctly. Hence for the KNN with average error rate of 18%, the test error rate is 36% which makes the logistic regression with test error rate 30% is better for classification.

Chapter – 5

ISLR 5.4

Q3. We now review k-fold cross-validation.

a. Explain how k-fold cross-validation is implemented.

By randomly dividing the set of n observations into k non-overlapping groups, k-fold cross-validation is carried out. Each of these groups serves as a training set, with the others serving as validation sets. By averaging the k resulting MSE estimations, the test error is estimated.

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

i) The validation set approach?

The validation set strategy is conceptually straightforward and simple to apply because it only requires splitting the training data into two sets. There are two negatives, though: 1. Depending on which observations are included in the training and validation sets, the estimate of the test error rate can vary greatly. 2. The test error rate for the model fit on the full data set may tend to be underestimated by the validation set error rate.

ii) LOOCV?

With k = n, LOOCV is a specific instance of k-fold cross-validation. LOOCV is the method that requires the greatest computing because the model needs to be fit n times. Additionally, compared to k-fold CV, LOOCV has reduced bias but higher variance.

Q4. Suppose that we use some statistical learning method to make a prediction for the response Y for a particular value of the predictor X. Carefully describe how we might estimate the standard deviation of our prediction.

We may use the bootstrap method to estimate the standard deviation of our forecast if we were to generate a prediction for the response Y using some statistical learning method for a certain value of the predictor X. The bootstrap method involves repeating collecting observations from the original data set (with replacement) B times for some large value of B, fitting a new model each time, and then calculating the RMSE of the estimates for all B models.