DAT 500S Module 4

Textbook Quiz:

Exercise 4.7 Q5

Group Assignment:

Exercise 4.7 Q11 (e)

Exercise 5.4 Q5

Individual Assignment:

Exercise 4.7 Q10 (f)

Exercise 5.4 Q8

Exercise 4.7 Q5

- (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?

On the training set, QDA will perform better than LDA because QDA is a more non-linear method.
Since the Bayes decision boundary is linear, it is more accurately approximated by LDA than by QDA on the test set.

- (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?

If the Bayes decision boundary is non-linear, we expect QDA to perform better both on the training and test sets.

Exercise 4.7 Q5

- (c) In general, as the sample size n increases, do we expect the test prediction accuracy of QDA relative to LDA to improve, decline, or be unchanged? Why?

Improve. The answer lies in the bias-variance trade-off.

Roughly speaking, QDA is recommended if the training set is very large, so that the variance of the classifier is not a major concern (from ISLR).

Exercise 4.7 Q5

- (d) 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.

False. Since the Bayes decision boundary is linear, it is more accurately approximated by LDA than by QDA. The QDA decision boundary is inferior, because it suffers from higher variance without a corresponding decrease in bias.

Exercise 4.7 Q11

- (e) Perform QDA on the training data in order to predict mpg01 using the variables that seemed most associated with mpg01 in (b). What is the test error of the model obtained?

For example, use log(displacement), log(weight), and year as predictor, because the others are either strongly correlated with these variables or do not seem to associate with mpg01.

```
-> library(MASS)
-> qda.fit = qda (mpg01~log(weight) + log(displacement) + year, data= data.training)
-> qda.pred = predict (qda.fit, data.test)
```

Exercise 4.7 Q11

```
 -> qda.class = qda.pred$class

 -> table.qda = table(qda.class,data.test$mpg01)
-> 1 - mean(qda.class == data.test$mpg01)

# replace “lda” to “qda”
```

Exercise 5.4 Q5

(a) Fit a logistic regression model

-> library(ISLR)

-> default = data(Default)

-> glm.fit = glm (default~income+balance, data = Default, family = binomial)

(b) i. Split the sample set into a training set and a validation set.

ii. Fit a multiple logistic regression model using only the training observations.

```
-> set.seed(42)
-> train = sample(nrow(Default),nrow(Default)*0.7, replace = F)
-> train.data = Default[train,]
-> test.data = Default[-train,]
-> glm.fit = glm (default~income+balance, data = train.data, family = binomial)
-> summary(glm.fit)
```

(b) iii. Obtain a prediction of default status for each individual in the validation set by computing the posterior probability of default for that individual

iv. Compute the validation set error

```
-> glm.probs = predict(glm.fit, test.data, type = "response")
-> glm.pred = ifelse(glm.probs>0.5, "Yes", "No")
-> validation_error = mean(glm.pred!=test.data$default)
```

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

- Set different seed for three times and calculate their validation errors
- Comment: Variability exists in validation rate, because we include different observations in training and test datasets. Thus, the coefficient estimates can also vary.

(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.

```
-> class(Default$student)
-> glm.fit = glm (default~income+balance+student, data = train.data,
family = binomial)
```

Comment: The error rate does not decrease much or even increase for this validation set approach compared to the previous glm model. However, the validation errors of these models can happen by chance. You need to try this with different splits before coming to a conclusion.