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

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

**Answer:** QDA would likely fit the training set better due to its higher flexibility finding a slightly better fit, however LDA would perform better on the test set because QDA likely overfit the data.

1. 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?

**Answer:** We would expect QDA to better fit both training and test set due to its higher flexibility in fitting a non-linear boundary.

1. 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?

**Answer:** We would expect the accuracy to improve because QDA is more flexible, and has more variance inherently, but as n increases, the variance of the classifier has a smaller effect.

1. 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.

**Answer:** False, if the observations, n, is small, QDA will overfit the training set and perform much worse than LDA.

6. 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 coefficients, B0 = -6, B1 = 0.05, B2 = 1.

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

**Answer:**

1. How many hours would the student in part a need to study to have a 50% chance of getting an A?

**Answer:**

50 hours is needed for a predicted probability of 50% of getting an A for the student

7. Suppose that we wish to predict whether a given stock will issue a dividend this year based on X, last year’s percent profit. We examine a large number of companies and discover that the mean value of X for companies that issued a dividend was 10, and the mean for those that didn’t was 0. Both sets of companies' variance was 36. 80% of companies issued a dividend. Assume X follows a normal distribution, what is the probability that a company will issue a dividend given its percentage profit was 4 last year.

**Answer:** P(Y=dividend|X=4) = ?

Probability of a dividend given the last years profit percent of 4% is 70.29%

8. Suppose that we take a data set, divided into equal parts test and training, and try out logistic regression and k-nearest neighbor (K=1). For logistic regression the error rates were 20% and 30% for training and test respectively. For KNN, they were both 18%. Which method should we prefer to use for new observations? Why?

**Answer:** Because K=1, the training error is 0%, which implies the test error is 36%. Therefore, we’d choose to use logistic regression because there is a smaller jump between train and test error rates as well as a lower test error rate overall.

9. About odds

1. On average, what fraction of people with an odds of 0.37 of defaulting on their credit card payment will in fact default?

**Answer:** , implies p(x) = 0.27

1. Suppose that an individual has a 16% chance of defaulting on her credit card payment, what are the odds she will default?

**Answer:** odds = 0.16/0.84 = 0.19