Assignment 3
INFO H-515
An Introduction to Statistical Learning (with python application) Chapter 4

Exercise-5:

(a) If the Bayes decision boundary is linear, we expect LDA to perform better on the training set compared to QDA. LDA assumes that the class-conditional densities are multivariate Gaussian with a common covariance matrix for all classes. In the case of a linear decision boundary, this assumption aligns well with the true data distribution. Therefore, LDA, which is a linear classifier, is likely to perform well on the training set.

For the test set, LDA might also perform better than QDA, especially if the test data follows a similar linear distribution as the training data. LDA, when provided with sufficient training data, tends to generalize well even when the underlying distribution is not perfectly Gaussian.

(b) If the Bayes decision boundary is non-linear, we expect QDA to perform better on the training set. QDA does not make the assumption of a common covariance matrix for all classes, and it is more flexible in capturing non-linear decision boundaries. Therefore, in the presence of a non-linear decision boundary, QDA is likely to fit the training data better.

For the test set, the performance of QDA will depend on the complexity of the non-linear decision boundary. If QDA overfits the training data, it may not generalize well to the test data. However, if the non-linearity in the data is well-captured by QDA, it can perform better on the test set as well.

(c) In general, as the sample size (n) increases, we expect the test prediction accuracy of QDA relative to LDA to improve. QDA typically benefits from larger sample sizes because it has more parameters to estimate (due to separate covariance matrices for each class), and a larger sample size provides more data for parameter estimation. With more data, QDA can better estimate the class-conditional distributions and, as a result, generalize more accurately to the test data. LDA, which assumes a common covariance matrix, is less affected by sample size changes, and its performance may remain relatively stable.

(d) False. If the Bayes decision boundary is linear, and the true underlying distribution is indeed linear, LDA is expected to perform at least as well as QDA, and it may even outperform QDA. QDA introduces more flexibility in modeling the data by estimating separate covariance matrices for each class, which can lead to overfitting when the true decision boundary is linear. In such cases, LDA's assumption of a common covariance matrix can act as a regularization, preventing overfitting and providing better test error rates. Therefore, it is not necessarily true that QDA will achieve a superior test error rate when the true decision boundary is linear; LDA may perform equally well or better in this scenario.

Exercise-6:

(a)Logistic regression use the following function to estimate the probability for a given Y:

$P(Y = 1) = 1 / (1 + \exp(-t))$

where $t = (\beta0 + \beta1 * X1 + \beta2 * X2)$

Substituting the values into $t = -6 + 0.05*40 + 1*3.5 = -0.5$

$P(Y = 1) = 1 / (1 + \exp(0.5))$

$P(Y = 1) = 0.37754.$

So, probability of getting A is 0.3775

(b)Now using the same formula and substituing the values with Probability equal to 0.5 and taking hours as “h”

$0.5 = 1 / (1 + \exp(-(\beta0 + \beta1 * X1 + \beta2 * X2)))$

$0.5 * (1 + \exp(-(\beta0 + \beta1 * X1 + \beta2 * X2))) = 1$

$1 + \exp(-(\beta0 + \beta1 * X1 + \beta2 * X2)) = 2$

$\exp(-(\beta0 + \beta1 * X1 + \beta2 * X2)) = 1$

$\exp(-(-6 + 0.05*h + 1*3.5)) = 1$

$6-0.05h-3.5 = \log(1)$

$0.05h = 2.5$

$h = 50$

So, student must study at least 50 hours to get A with chance of 50% or more