Unisha Aryal

Assignment 3

INFO H - 515

5. 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 expect QDA to have better performance on the training dataset due to its enhanced adaptability, resulting in a more precise fit. However, if the Bayes decision boundary is linear, we predict that LDA will outperform QDA on the test set because QDA might be susceptible to overfitting.

(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 anticipate QDA to excel on both the training and test sets, as its enhanced flexibility enables a more accurate approximation, resulting in superior performance on both datasets.

(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?*

We expect that as the sample size 'n' increases, the relative test prediction accuracy of QDA compared to LDA will improve. In general, with larger sample sizes, a more flexible method is expected to yield a more accurate fit as the variance is offset by the larger sample size.

(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. When dealing with a reduced number of sample points, the increased variability from using a highly flexible approach like QDA would likely lead to overfitting, which, in turn, would result in a higher test error rate compared to LDA.

6. Suppose we collect data for a group of students in a statistics class with variables X_1 hours studied, X_2 = undergrad GPA, and Y = receive an A. We fit a logistic regression and produce estimated coefficient, $\hat{\beta}_0 = -6$, $\hat{\beta}_1 = 0.05$, $\hat{\beta}_2 = 1$.

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

Logistic regression uses the following function to estimate the probability for a given Y:

$$Y = \frac{1}{1 + e^{-t}}$$

Where,

$$t = \beta_0 + \beta_1 * X_1 + \beta_2 * X_2$$

a) Substituting these values in above equation, we get

t = -6 + 0.05*40 + 1*3.5 = -0.5

Now, substituting value of t in first equation to get probability,

$$Y = \frac{1}{1 + e^{0.5}}$$

Y = 0.3775

So, probability of getting A is 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?*

$-6+0.05X_1+3.5=\log(0.51-0.5) \Rightarrow 0.05X_1-2.5=0 \Rightarrow X_1=50$

Here probability is 0.5. Hence,

$e^{(-6+0.05\times h+1\times 3.5)}$= p / (1−p) = 1
$-2.5+0.05\times h=0$

and hence value of h is **50**.

So, students must study 50 hours to get A with chance of 50% or more.