**Problem 1: Concept Review**

1. When fitting a logistic regression model, we are trying to find the best estimates for the β coefficients, which tell us how different factors affect the likelihood of a certain outcome. We know the outcomes and the factors that might influence the data points. The goal is to find the β values that make it most likely for the observed outcomes to occur. We use maximum likelihood estimation which is finding the most probable explanation for our data.
2. Using 0.5 as a threshold will give the smallest overall misclassification rate because it mimics the Bayes rule which gives us the lowest possible error.
3. Misclassification rate: (25+32) / 218 = 0.26146

False negatives are worse in this case - saying someone is healthy when they are actually sick. Change the posterior probability threshold so that it is easier to classify someone as sick.

1. Given X1 = 40, X2 = 3.5 and B̂0 = -6, B̂1 = 0.05, B̂2 = 1

(i) The expression of the model is 0.05(40) + 1(3.5) – 6 = -0.5

p̂(k) = ek / (1 + ek) when k = -0.5

p̂(-0.5) = e-0.5 / (1 + e-0.5) = 0.3775407

So, a student who studies 40 hours and has an undergraduate GPA of 3.5 has a probability of approximately 37.75% of getting an A in the class.

(ii) ½ = ek / (1 + ek)

1 + ek = 2ek

ln(1) = ln(ek)

ln(1) = k

0 = k

0.05X1 + 1(3.5) – 6 = 0

X1 = 50

If the student wanted a 50% chance of getting an A in class, they’d have to study for 50 hours.

1. LDA will be better on the test set because if the true decision boundary is linear then the bias of the LDA is going to be 0. We already know that the variance of LDA is smaller than QDA and so overall, out of sample error = Bias2 + Variance will be lower for LDA than for QDA.
2. QDA will be better on the training set because it will lead to a model with smaller bias and higher variance. This means we could have a more flexible model, which may yield to a closer fit.
3. The data I created can’t converge as the data is perfectly linearly separable, so there is no unique solution for the coefficient estimates since there can be infinite ways to separate the data.

B0 = -256.701, B1 = 27.678, B2 = 8.162

A computer screen shot of a black screen

Description automatically generated

1. LDA misclassification rate was 0. QDA misclassification rate was 0. Since LDA and QDA are not affected by the convergence issues that logistic regression faces in this situation, we are able to get meaningful results.

**Problem 2: Email Spam**

1. Spam proportion: 1813 / 4601 = 39.4%

Non-spam proportion: 2788 / 4601 = 60.6%

1. Training spam proportion: 927 / 2300 = 40.3%

Training non-spam proportion: 1373 / 2300 = 59.7%

Testing spam proportion: 886 / 2301 = 38.5%

Testing non-spam proportion: 1415 / 2301 = 61.5%

50-50 split for the training and test set, the proportions are relatively the same from part a

1. First 10 predicted probabilities:

0.6830159, 0.9962195, 0.9999999, 0.7700588, 0.6334144, 0.6216763, 0.9259383, 0.5188221, 0.6779338, 0.9998643

1. The model predicted the spam trend correctly 93.35% of the time.

Misclassification rate = 0.06649283

False negative rate: 0.06180556

False positive rate: 0.07433217



1. Reporting meaningful email as spam is a more critical mistake. To accommodate this, we could increase the 0.5 threshold for classifying an email as spam so it’s harder to classify an email as spam.

**Problem 3:**

1. Summary of the logistic regression

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Coefficient Estimate | Standard Error | z-value | P-value |
| Intercept | 0.26686 | 0.08593 | 3.106 | 0.0019 |
| Lag1 | -0.04127 | 0.02641 | -1.563 | 0.1181 |
| Lag2 | 0.05844 | 0.02686 | 2.175 | 0.0296 |
| Lag3 | -0.01606 | 0.02666 | -0.602 | 0.5469 |
| Lag4 | -0.02779 | 0.02646 | -1.050 | 0.2937 |
| Lag5 | -0.01447 | 0.02638 | -0.549 | 0.5833 |
| Volume | -0.02274 | 0.03690 | -0.616 | 0.5377 |

Lag2 is statistically significant.

1. To determine the % of correct predictions: (54+557) / (54+48+430+557) = 0.5611

This says that the model predicted the weekly market trend correctly 56.11% of the time. Separating in how the model correctly predicts the Up and Down trends. The model correctly predicted the Up weekly trends (557) / (48+557) = 0.9207 which is 92.07% correct. The Down weekly trends were predicted at a lower rate, (54) / (430 + 54) = 0.115 which is 11.5% correctly predicted.

A black background with white text

Description automatically generated

1. When splitting the Weekly dataset into training and test data, the model correctly predicted weekly trends at a rate of 62.5%, which is an improvement from the model that used the whole dataset. This model predicted upward trends as 91.8% and downwards as 20.93% correct. This model was able to improve significantly on correctly predicting downwards trends. The overall fraction of correct predictions is 62.5%

A black background with white text

Description automatically generated

1. Using LDA, the model correctly predicted weekly trends at a rate of 62.5%. The correct predictions were 65/104

A black background with white text

Description automatically generated

1. Using QDA, the model correctly predicted weekly trends at a rate of 58.65%. The correct predictions were 61/104

A black background with white text

Description automatically generated

1. Using Naïve Bayes, the model correctly predicted weekly trends at a rate of 58.65%. The correct predictions were 61/104

A black background with white text

Description automatically generated