DS 303 Homework 8
Due: Oct. 23, 2023 on Canvas by 11:59 pm (CT)

Instructions: Homework is to be submitted on Canvas by the deadline stated above. Please clearly print your name and student ID number on your HW.

Show your work (including calculations) to receive full credit. Please work hard to make your submission as readable as you possibly can - **this means no raw `R` output or code** (unless it is asked for specifically or needed for clarity).

Code should be submitted with your homework as a separate file (for example, a .R file, text file, Word file, or .Rmd are all acceptable). You should mark sections of the code that correspond to different homework problems using comments (e.g. `##### Problem 1 #####`).

Problem 1: Concept Review

(a) Explain in plain language how we obtain estimates for β when fitting a logistic regression model.

(b) When using logistic regression, what threshold will give us the smallest overall misclassification rate? Explain briefly why.

(c) Suppose we are trying to build a classifier where Y can take on two classes: 'sick' or 'healthy'. In this context, we consider a positive result to be testing sick (you have the virus) and a negative result to test as healthy (you don't have the virus). After fitting the model with LDA in `R`, we compare how our classifier performs with the actual outcomes of the individuals, as shown below:

```
#rows are predicted, columns are true outcomes
#so the number of actually sick people is 65

lda.pred sick healthy
   sick     40   32
   healthy  25   121
```

What is the misclassification rate for the LDA classifier above? In the context of this problem, which is more troubling: a false positive or a false negative? Depending on your answer, how could you go about decreasing the false positive or false negative rate? Comment on how this will likely affect overall the misclassification rate (consider which threshold will have the lowest overall misclassification rate).

(d) 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 model and produce estimated regression coefficients, $\hat{\beta}_0 = -6$, $\hat{\beta}_1 = 0.05$, and $\hat{\beta}_2 = 1$.

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

ii. How many hours would the student in part (i) need to study to have a 50% chance of getting an A in the class?

(e) If the true decision boundary between two groups is linear and the constant variance assumption holds, do you expect LDA or QDA to perform better on the testing set? Explain using concepts from bias/variance tradeoff.

(f) Same setup as (e), do you expect LDA or QDA to perform better on the training set? Justify your answer.

(g) Create a data set that consists of two predictors (X_1, X_2) and a binary response variable Y. Let $n = 16$ and $Y = 0$ for 8 observations and $Y = 1$ for the remaining 8 observations. Create this data set in such a way that logistic regression cannot converge when applied to this data set. Explain why logistic regression cannot converge on this data set. Report your estimates for β_0, β_1, and β_2 obtained from fitting a logistic regression model on this data set. You may copy/paste your output.

(h) Apply LDA/QDA to the dataset you created in part (h). Are you able to get meaningful results? Report the misclassification rate for LDA and QDA.

Problem 2: Email Spam

We will use a well-known dataset to practice classification. You can find it here: `https://archive.ics.uci.edu/ml/datasets/Spambase`. Read the attribute information and download the dataset onto your computer. To load this data into R, use the follow code:

```
spam = read.csv('.../spambase.data',header=FALSE)
```

The last column of the `spam` data set, called `V58`, denotes whether the e-mail was considered spam (1) or not (0).

(a) What proportion of emails are classified as spam and what proportion of emails are non-spam?

(b) Carefully split the data into training and testing sets. Check to see that the proportions of spam vs. non-spam in your training and testing sets are similar to what you observed in part (a). Report those proportions here.

(c) Fit a logistic regression model here and apply it to the test set. Use the `predict()` function to predict the probability that an email in our data set will be spam or not. Print the first ten predicted probabilities here.

(d) We can convert these probabilities into labels. If the predicted probability is greater than 0.5, then we predict the email is spam ($\hat{Y}_i = 1$), otherwise it is not spam ($\hat{Y}_i = 0$). Create a confusion matrix based on your results. What's the overall misclassification rate? Break this down and report the false negative rate and false positive rate.

(e) What type of mistake do we think is more critical here: reporting a meaningful email as spam or a spam email as meaningful? How can we adjust our classifier to accommodate this?

Problem 3: Weekly data set

We will use the `Weekly` data set from the `ISLR2` library. It contains 1,089 weekly returns for 21 years, from the beginning of 1990 to 2010.

(a) Use the full data set to perform a logistic regression with `Direction` as the response and the five lag variables plus `Volume` as predictors. Use the summary function to report your results.

(b) Set a threshold that minimizes the overall misclassification rate. Compute the confusion matrix and overall correct classification rate. Explain what the confusion matrix is telling you about the types of mistakes made by logistic regression.

(c) Now fit the logistic regression model using a training data period from 1990 to 2008, with `Lag2` as the only predictor. Set a threshold that minimizes the overall misclassification rate. Compute the confusion matrix and overall correct classification rate on the test set (that is, data from 2009 and 2010).

(d) Repeat (d) using LDA.

(e) Repeat (d) using QDA.

(f) Repeat (d) using Naive Bayes.

End of assignment