DS 303 Homework 7
Due: Oct. 25, 2021 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. 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).

b. Consider the dataset:

x	y
-2	red
5	blue
-1	red
10	blue
5	blue

We use logistic regression to fit a model to this data: that is, Y is binary variable that is either red or blue. Our model is estimating:

$$P(Y_i = \text{red}|x_i) = \frac{1}{1 + \exp(\beta_0 + \beta_1 x_i)} \quad \text{and} \quad P(Y_i = \text{blue}|x_i) = \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)}$$

for all $i = 1, 2, 3, 4, 5$. What value(s) of β_0 and β_1 would maximize the likelihood (and therefore be the estimates we would get from fitting this model)? Recall that our likelihood looks like:

$$l(\beta_0, \beta_1, X) = P(Y_1 = \text{red}|\beta_0, \beta_1, x_1) \times P(Y_2 = \text{blue}|\beta_0, \beta_1, x_2) \times \ldots \times P(Y_5 = \text{blue}|\beta_0, \beta_1, x_5).$$

Hint: What is $P(Y_i = \text{blue}|x_i > 4)$? Now what is the $P(Y_2 = \text{blue}|x_2 = 5)$? What values of β_0 and β_1 will get us close to this probability?

c. Suppose you just took on a new consulting client. He tells you he has a large dataset (say $100,000$ observations) and he wants to use this to classify whether or not to invest in a stock based on a set of $p = 10,000$ predictors. He claims KNN will work really well in this case because it is non-parametric and therefore makes no assumptions on the data. Present an argument to your client on why KNN might fail when p is large relative to the sample size.

d. For each of the following classification problems, state whether you would advise a client to use LDA, logistic regression, or KNN and explain why:

   i. We want to predict gender based on height and weight. The training set consists of heights and weights for 82 men and 63 women.

   ii. We want to predict gender based on annual income and weekly working hours. The training set consists of 770 mean and 820 women.

   iii. We want to predict gender based on a set of predictors where the decision boundary is complicated and highly non-linear. The training set consists of 960 men and 1040 women.

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 question as part (e), but what if we compare the performance of LDA and QDA on the training set? Which will perform better?

g. 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.

h. 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. Using logistic regression, obtain the predicted probabilities for data set and report them here. You may copy/paste your output.

i. 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: Practicing data simulations

Let us simulate data where we know the true $P(Y = 1|X)$. Suppose Y can only take on 0 or 1. We have 3 predictors of interest. Fill in the following code to simulate classification data.

a.
```
set.seed(1)
x1 = rnorm(1000)          # create 3 predictors
x2 = rnorm(1000)
x3 = rnorm(1000)

#true population parameters
B0 = 1
B1 = 2
B2 = 3
B3 = 2

# construct the true probability of Y =1 using the logistic function.
pr = ??

# randomly generate our response y based on these probabilities
y = rbinom(1000,1,pr)

df = data.frame(y=y,x1=x1,x2=x2, x3=x3)
```

b. On the simulated data, fit a logistic regression model with Y as the response and X_1, X_2, X_3 as the predictors. Compute the confusion matrix and the misclassification rate.

c. On the simulated data, apply LDA. Compute the confusion matrix and the misclassification rate.

d. On the simulated data, apply Naive Bayes. Compute the confusion matrix and the misclassification rate.

e. How do the 3 methods compare?

Problem 3: Weekly Data

This question should be answered using the `Weekly` data set, which is part of the `ISLR2` package. This data is similar in nature to the `Smarket` data we saw in class, except that it contains 1,098 weekly returns for 21 years, from the beginning of 1990 to the end of 2010.

a. Produce some numerical and graphical summaries of the `Weekly` data. Do there appear to be any patterns?

b. Fit the logistic regression model using a training data period from 1990 to 2008, with `Lag2` as the only predictor. Compute the confusion matrix and the overall fraction of `correct` predictions for the test data period (that is, data from 2009 and 2010).

c. Repeat (b) using LDA.

d. Repeat (b) using QDA.

e. Repeat (b) using KNN with K chosen using cross-validation.

f. Repeat (b) using Naive Bayes.

g. Which of these methods appear to provide the best results on this data?

h. Experiment with different combinations of predictors for each of the methods. Report the final model and associated confusion matrix that appears to provide the best results on the test set.

Problem 4: 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?

f. Carry out LDA, QDA, Naive Bayes and KNN on the training set. You should experiment with values for K in the KNN classifier using cross-validation. Remember to standardize your predictors for KNN. For each classifier, report the confusion matrix and overall test error rates for each of the classifiers.

g. Which classifier would you recommend for this data? Justify your answer.