DS 301: Homework 9
Due: April 20, 2022 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` code or raw `R` output** (unless it is asked for specifically or needed for clarity).

Code should be submitted with your homework as a separate file (for example, an R file, text 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. Suppose you have a training set and a testing set, both of which have a sample size of 1000. Assume our outcome Y is binary and can take on $Y = 0$ or $Y = 1$.
We obtain the following estimates for the training set:

$\hat{\mu}_0 = 3.4, \quad \hat{\mu}_1 = 5.1, \quad \hat{\sigma}^2 = 4.5, \quad \hat{\pi}_0 = 0.32, \quad \hat{\pi}_1 = 0.68$

and for the testing set:

$\hat{\mu}_0 = 3.2, \quad \hat{\mu}_1 = 5.5, \quad \hat{\sigma}^2 = 4.1, \quad \hat{\pi}_0 = 0.35, \quad \hat{\pi}_1 = 0.65$

1. Based on the information above, construct the LDA classifier. Explain when test observations will be assigned to $Y = 0$ and when they will be assigned to $Y = 1$. **Show your work for full credit.**
2. What threshold would give us the smallest possible test misclassification rate? Explain why.

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 covariance 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. LDA/QDA use Bayes Theorem (also known as Bayes Rule) to try to estimate the probabilities $P(Y = k|X)$. Bayes Theorem involves estimating 3 quantities: $P(X|Y = k)$, $P(Y = k)$, and $P(X)$. Explain in plain language (no statistics terminology) why we do not need to worry about estimating $P(X)$.

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(1000,0,0.9)           # create 3 predictors
x2 = rnorm(1000,1,1)
x3 = rnorm(1000,0,2)

#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 K-NN (obtain the optimal K using cross-validation). Remember to standardize your predictors for K-NN. Report the K you obtained. 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. You may find it helpful to reference your code from HW 8.

a. Fit LDA using a training data period from 1990 to 2008, with `Lag2` as the only predictor. Report your estimates for the class specific means and prior probabilities.

b. For the first observation in your training set, what is the predicted probability of Up? What is the predicted probability of Down? How would you classify this observation? Does this match what you observed?

c. Repeat (b) for the first observation in your test set.

d. Compute the confusion matrix and the overall fraction of **correct** predictions for the test set (that is, data from 2009 and 2010).

e. What are the two assumptions that LDA makes? Do you think those two assumptions holds? Justify your answer using graphical or numerical summaries.

f. Fit QDA on the training set with `Lag2` as the only predictor. Compute the confusion matrix and the overall fraction of **correct** predictions for the test data period.

g. Repeat (f) using KNN with K chosen using cross-validation.

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