Homework 3

#for nicer table output  
library(knitr)

Warning: package 'knitr' was built under R version 4.1.3

# Conceptual Problems

## Book Problems

Complete the following problems from the Introduction to Statistical Learning with R book (I’m not sure if the problems are in the same order in the python book so use the R book to identify which problems to do).

Section 4.8

* Book Problem 4
  1. 10% of the data will be used as the distribution is uniform and our range accounts for 10% of the total range of [0, 1]. (Not asked for, but we can show this via simulation.)

set.seed(10)  
unif\_values <- runif(100000)  
mean(unif\_values < 0.65 & unif\_values > 0.55)

[1] 0.10049

b) Since we have 0.1 probability of X1 being in the range and, separately/independently, a 0.1 probability of X2 being in the range, we have 0.1\*0.1 = 0.01 probability of a value falling in the range. So 1% of our observations would fall in the range.

unif\_values\_2 <- runif(100000)  
mean((unif\_values< 0.65 & unif\_values > 0.55) & (unif\_values\_2 > 0.3 & unif\_values\_2 < 0.4))

[1] 0.01098

c) Here only $0.1^{100} = $10^{-100} of the observations fall in the region! Almost no chance of a value falling in the interval of choice.  
  
d) We can see that, if we want to use the same 'size' region, as $p$ grows we will have fewer and fewer observations. That means we either need to have way more observations or we need to make our 'near' range larger as $p$ grows.  
  
e) Let $L$ be the length.   
  
 i) For $p = 1$, 10% of the observations fall in any area of length L = 0.1 in [0,1].   
 ii) For $p = 2$, we want $L^2 = 0.1$. This means $L=(0.1)^{1/2} = $ 0.3162278  
 iii) Generically, for $p$ we want $L^p = 0.1$ or $L = (0.1)^{1/p}$. For a few values of $p$ we can see the lengths below:

p <- 1:30  
L <- (0.1)^(1/p)  
data.frame(p, L) |>  
 knitr::kable()

| p | L |
| --- | --- |
| 1 | 0.1000000 |
| 2 | 0.3162278 |
| 3 | 0.4641589 |
| 4 | 0.5623413 |
| 5 | 0.6309573 |
| 6 | 0.6812921 |
| 7 | 0.7196857 |
| 8 | 0.7498942 |
| 9 | 0.7742637 |
| 10 | 0.7943282 |
| 11 | 0.8111308 |
| 12 | 0.8254042 |
| 13 | 0.8376776 |
| 14 | 0.8483429 |
| 15 | 0.8576959 |
| 16 | 0.8659643 |
| 17 | 0.8733262 |
| 18 | 0.8799225 |
| 19 | 0.8858668 |
| 20 | 0.8912509 |
| 21 | 0.8961505 |
| 22 | 0.9006280 |
| 23 | 0.9047357 |
| 24 | 0.9085176 |
| 25 | 0.9120108 |
| 26 | 0.9152473 |
| 27 | 0.9182543 |
| 28 | 0.9210553 |
| 29 | 0.9236709 |
| 30 | 0.9261187 |

* Book Problem 5:
  1. If the Bayes decision boundary is truly linear, then we would expect LDA to do better on the test set as QDA would be finding relationships that don’t truly exist in the data and, therefore, should not exist in the test set. As QDA is able to better fit data, we would expect QDA to perform better on the training set.
  2. If the Bayes decision boundary is truly non-linear, we expect QDA to perform better on both the training and test set. QDA generally fits more closely to the training data due to being a more complicated of a model. With a non-linear decision boundary, we’d also expect QDA to translate better to the test set.
  3. We would generally assume that QDA’s performance would improve relative to LDA as we’d have more data to fit the more complicated model better. If the true boundary is linear (the equal covariance assumption being reasonable), then LDA may continue to outpeform QDA. QDA should get closer and closer to the LDA fit in that situation as well.
  4. False - the performance of QDA may approach that of LDA as the sample size increases but QDA is generally going to be fitting too complicated of a relationship and, thus, won’t perform as well on the test set.
* Book Problem 6 Suppose we collect data for a group of students in a statistics class with variables X1 = hours studied, X2 = undergrad GPA, and Y = receive an A. We fit a logistic regression and produce estimated coefficients given by intercept = -6, slope on hours studied = 0.05, and slope on undergraduate GPA of 1.
  1. Our (esitmated) log-odds of success are equal to . Our probability of success is $ = 0.3775407
  2. To have a 50% probability we would also have a log-odds of 0. Setting our log-odds equation to 0 with GPA equal to 3.5 we get which implies we need to have $hours = $50.
* Book Problem 9
  1. On average, what fraction of people with an odds of 0.37 of defaulting on their credit card payment will in fact default?
  2. Suppose that an individual has a 16 % chance of defaulting on her credit card payment. What are the odds that she will default?

Book Problem 14 In this problem, you will develop a model to predict whether a given car gets high or low gas mileage based on the Auto data set.

1. Create a binary variable, mpg01, that contains a 1 if mpg contains a value above its median, and a 0 if mpg contains a value below its median. You can compute the median using the median() function. Note you may find it helpful to use the data.frame() function to create a single data set containing both mpg01 and the other Auto variables.
2. Explore the data graphically in order to investigate the association between mpg01 and the other features. Which of the other features seem most likely to be useful in predicting mpg01? Scatterplots and boxplots may be useful tools to answer this question. Describe your findings.
3. Split the data into a training set and a test set.
4. Perform LDA on the training data in order to predict mpg01 using the variables that seemed most associated with mpg01 in (b). What is the test error of the model obtained?
5. Perform QDA on the training data in order to predict mpg01 using the variables that seemed most associated with mpg01 in (b). What is the test error of the model obtained?
6. Perform logistic regression on the training data in order to predict mpg01 using the variables that seemed most associated with mpg01 in (b). What is the test error of the model obtained?
7. Perform naive Bayes on the training data in order to predict mpg01 using the variables that seemed most associated with mpg01 in (b). What is the test error of the model obtained?
8. Perform KNN on the training data, with several values of K, in order to predict mpg01. Use only the variables that seemed most associated with mpg01 in (b). What test errors do you obtain? Which value of K seems to perform the best on this data set?

* Book Problem 5
* Book Problem 6
* Book Problem 9

# Implementation Problems

## Book Problems

Section 4.8

* Book Problem 14
  + For this problem you may want to use the caret package in R! [The reference for the package is here](https://topepo.github.io/caret/).