Homework 2

Sean Eva

February 2022

1 Part I - Theoretical Problems

3. Given that each class has its own covariance matrix $p = 1$. That is to say that there is only one feature $x \sim N(\mu_k, \sigma_k^2)$. $f_k(x) = \frac{1}{\sqrt{2\pi}\sigma_k} e^{\frac{-(x-\mu_k)^2}{2\sigma_k^2}}$. Then $P(Y = k|X = x) = \frac{P(X=x|Y=k)P(Y=k)}{\sum_{k=1}^{K} P(X=x|Y=k)P(Y=k)} = \frac{\pi_k f_k(x)}{\pi_k f_k(x)}$. Bayes classifier assigns the class for which $P(Y = k|X = x)$ is the largest and therefore $\pi_k f_k(x)$ is largest. This implies then that $\log(\mu_k f_k(x))$ is largest. Then we have that $\sigma_k(x) = \log(\pi_k f_k(x)) = \log(\pi_k) + \log(\frac{1}{\sqrt{2\pi}\sigma_k}) - \frac{(x-\mu_k)^2}{\sigma_k^2}$. This then implies that σ_k is quadratic on x and that the Bayes classifier is quadratic.

4. (a) We can see that given the observation range that we are using $[x - 0.05, x + 0.05]$. Therefore, we know that since the data is uniformly distributed on $[0, 1]$ we get that $if X \in [0, 0.05)$ then we would use the training observations in the range of $[0, 0.1]$. Similarly if $X \in (0.95, 1]$ we would use the training observations in the range of $[0.9, 1]$. Therefore in any of these scenarios we can see that hte fraction of observations used to make each prediction will be 10%

   (b) If we assume that x_1 and x_2 are independent the fraction of available observations we will use to make the prediction is $10\% \times 10\% = 1\%$.

   (c) As we saw in the previous results of available observations we will use to make the prediction is $(10\%)^p$ where p is the number of features. So when p tends to ∞. We have that $\lim_{p\to\infty}(10\%)^p = 0$.

   (d) So we see that our fraction interval that we are able to make because very very small (0 in fact) as our number of features is very large. Therefore, the training observations do not get a large enough range to pull data from to find any data actually near the test observation. This means that the testing will only actually be base on the one point in the long run and will inaccurately fit to the rest of the data.

   (e) For $p = 1$, it is the same as part a, 0.1. For $p = 2$ the length would be $0.1^{\frac{1}{2}} = 0.316$. As we extend to $p = 100$ we get that $0.1^{\frac{1}{100}} = 0.997$. This length in whatever dimension is always supposed to include

approximately 10% of the data as the tester as required. So as we go to higher p we see that the length actually increases closer and close to 1 and this is because the data is much more spread out and so we need to cover a theoretically larger area in order to still get the required amount.

5. (a) When the Bayes decision boundary is linear, one can expect QDA to perform better on the training set because it has a higher degree of flexibility which may yield a closer fit. On the test set we expect LDA to perform better than QDA because QDA has a tendency to over fit the linearity from the Bayes Decision boundary.

   (b) If the Bayes decision boundary is non-linear, then we expect QDA to perform better on both the training and test sets because of the higher degree of flexibility.

   (c) Roughly, QDA is recommended because of the high degree of flexibility if the training set is very large, so that the variance of the classifier is not a major concern.

   (d) False. With a few sample points, the variance from using a more flexible method such as QDA may lead to over fitting. Which in terms may lead to test-error.

6. (a) Since we are given all values needed we get that $t = -6 + 0.05 * 40 + 1 * 33.5 = -0.5$. Then we get that $Y = \frac{1}{1+e^{0.5}} = 0.3775$. Therefore, the probability of getting an A is 0.3775.

   (b) In order to get $Y > 0.5$ we need that $0.5 < \frac{1}{1+e^t}$ which evaluates to $t > 1$. In order for $t > 1$ then we must have that $-6+0.05*h+1*3.5 > 1$ which implies that $h > 70$. Therefore, the number of hours a student must study to get an A with a 50% chance or more is at least 70 hours.

7. We are given that $X = 4, \sigma^2 = 36, \mu_k = 10, \mu_l = 0$. Then we get that $p_1(4) = \frac{0.8e^{\frac{-(4-10)^2}{72}}}{0.8e^{\frac{-(4-10)^2}{72}} + 0.2e^{\frac{-(4-0)^2}{72}}} =$

8. We should use the 1-nearest neighbors approach for classification. This is because the average error rate in this case is 18%, whereas the average error rate in the other approach is $\frac{20+30}{2} = 25\%$. The error rate in the fist method is more than nearest neighbors method, hence we should use the the second method for classification to get more appropriate results.

9. (a) We know that we have to use odds, $odds = \frac{p}{1-p} = 0.37$. Therefore,

$$\begin{aligned} p &= 0.37(1-p) \\ &= 0.37 - 0.37p \\ p + 0.37p &= 0.37 \\ 1.37p &= 0.37 \\ p &= \frac{0.37}{1.37} \\ p &= 0.27 \end{aligned}$$

   (b) We get that the odds are $\frac{0.16}{1-0.16} = \frac{0.16}{0.84} = 0.19$. Therefore the odds that she will default is 0.19.

12. (a) The log odds of our model is

$$\log(\frac{e^{\hat{\beta}_0+\hat{\beta}_1 x}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 x}})$$

   (b) The log odds of our friends model is

$$\log(\frac{e^{\hat{\alpha}_{orange0}+\hat{\alpha}_{orange1}x}}{e^{\hat{\alpha}_{orange0}+\hat{\alpha}_{orange1}x}+e^{\hat{\alpha}_{apple0}+\hat{\alpha}_{apple1}x}})$$

   (c) I'm not too sure what to do in this part but I am going to guess that I am supposed to set the two models equal to each other and attempt to solve for the αs in the friends model. Obviously since there are 4 unknowns it would be very difficult to solve for all of them so we are supposed to develop a relationship between all of them.

   (d) I'm also not too sure what to do with this part but I think it is like part (c) in which I set them equal to each other and attempt to solve for the βs in our model but again since there are two unknowns we can develop a relationship directly between the two so it would be much easier to develop this relationship.

   (e) Since I didn't do part (d), I will simply write notes for this section as well. This here probably refers to using the log odds of the model using the coefficients from part (d) that would lead to the odds answer.
   Sorry

2 Part II - Programming

14. (a) In code

   (b) It appears that displacement, horsepower, weight, year, and acceleration seem to be the best predictors for mpg01.

(c) In code

(d) The test error of the model is about $1 - 0.8969072 = 0.1030928$.

(e) The test error of the model is about $1 - 0.8969072 = 0.1030928$.

(f) The test error of the model is about $1 - 0.9278351 = 0.0721649$.

(g) The test error of the model is about $1 - 0.9072165 = 0.0927835$.

(h) It appears that $k = 7$ would be the best for this data set.

16. From the correlation heat plot we can see that features that have a high rate of correlation are istat, indus, nox, rad, and tax. If we put the data into boxplots based on crim we can see that rad, age, indus, nox, and tax covariates separate the data well in boxplots. When we plot the data in scatter pltos for the strong predictors that we found in the boxplots we have that the red points are above the median and the blue points are below the median. When we compare the LDA, QDA, and Logistic regression models we see that the logistic regression models perform the best while the LDA and moreso the QDA models do worse. Finally when we check the KNN model we see that KNN model with $k = 1$ performs the best with accuracy of 0.913.