MiniProject 2: Report

Raheel Ahmed

2023-02-21

Netid: rsa170130

Question 1

a. Exploratory analysis of data

From examining the data, we notice that there are 7 features: cancervol, weight, age, benpros, vesinv, capspen, and gleason with one response, psa. We also know that there are 97 observations in our dataset, which actually appears to be a small number to try and develop a reasonable model with. For our two qualitative measures, vesinv and gleason, we can see via summary() that there are more people without seminal vesicle invasion than with and that there is a reasonable mix of people with varying gleason scores, though 7, the middle score, does appear most frequently. If we look at the histogram of ages in the dataset, we can see that mostly people from 50 to 70+ are affected with the disease. Also, examination of the correlations between all the variables shows that age correlates with hyperplasia which is indicative of prostatic abnormality and cancer volume is proportional to the outgrowth of the cancer itself, which in turn is correlated to a higher psa level.

b. Is PSA appropriate as a response variable?

PSA is not appropriate as a response variable. Looking at the histogram of psa prior to the transformation shows us that it is skewed heavily to the left, not fitting the traits of a normal distribution at all. Upon using a natural log transformation, we see that the histogram is much more like a normal distribution's.

c. Fit a simple linear regression model to predict PSA for each feature.

I ran lm or glm for each feature in relation to the psa. I found that of the 7 features, only 4 – cancervol, vesinv, capspen, and gleason – were linearly related to the psa. The plots showing their relationship, and confirming the results of the lm(), are shown below:

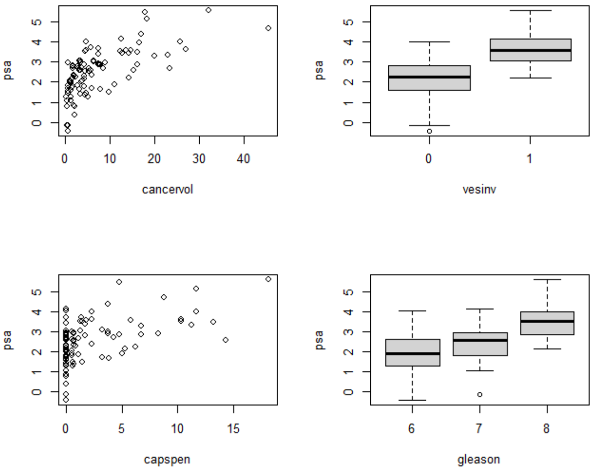

Figure 1: Linearity-plots

d. Fit a multiple regression model to predict PSA for each feature. For which predictors can we reject the null hypothesis?

The multiple regression model found a similar set of features that were linearly related to the psa. This time, it was cancervol, benpros, vesinv=1, and gleason=8, which are the features that we can reject the null hypothesis for. Again, weight, age, and other values of gleason were found to not be statistically significant in terms of a linear relationship to psa. However, this model finds benpros to be linearly related to psa and capspen to not be so (by quite a significant margin).

e. Build a "reasonably good" multiple regression model for these data.

I decided to use all the features found to be linearly related to psa, but I excluded capspen as it did not seem statistically significant when other predictors were involved and I excluded benpros as it did not have statistical significance to the psa response until other predictors were accounted for.

f. What is the equation of the final model?

The equation my final model returned was $psa = 1.60467 + 0.05875 * cancervol + 0.6259312 * vesinv1 + 0.3543990 * gleason7 + 0.7863444 * gleason8$, but since vesinv and gleason are qualitative measures I have derived the following equations for all possible cases of vesinv and gleason.

$$psa = \begin{cases} 1.60467 + 0.05875 * cancervol, & \text{with} \quad vesinv = 0, gleason = 6 \\ 1.95907 + 0.05875 * cancervol, & \text{with} \quad vesinv = 0, gleason = 7 \\ 2.39102 + 0.05875 * cancervol, & \text{with} \quad vesinv = 0, gleason = 8 \\ 2.230602 + 0.05875 * cancervol, & \text{with} \quad vesinv = 1, gleason = 6 \\ 2.58500 + 0.05875 * cancervol, & \text{with} \quad vesinv = 1, gleason = 7 \\ 3.01695 + 0.05875 * cancervol, & \text{with} \quad vesinv = 1, gleason = 8 \end{cases}$$

g. Use the final model to predict the PSA level for a patient whose quantitative predictors are at the sample means of the variables and qualitative predictors (if any) are at the most frequent category.

We find via mean() that the average cancervol is approximately 6.9987, the most common vesinv value is 0, and the most common gleason score is 7. We can use predict() with our final model and this new entry as our parameters to find that the predicted psa value for such a person is 2.370213. To confirm, we can try $psa = 1.95907 + 0.05875 * (6.9987) = 2.370244$, which gets us approximately the correct answer.

Question 2

a. Perform an exploratory analysis of data by examining appropriate plots and comment on how helpful these predictors may be in predicting response.

Again, we have a situation with not many training and test data observations, so accuracy measurements and model reliability should be evaluated with this in mind. Upon looking at plots with GPA and GMAT plotted against Group individually, we can see that GPA tends to correlate fairly well with Group in that midrange GPA values were most commonly found in students on the borderline, whereas high and low GPA values were associated with acceptance and non-acceptance, respectively. GMAT scores, however, were less significantly correlated; students with medium range values of GMAT scores were assigned to all 3 groups in noticeable quantity. Only those with very high or low scores were clearly associated with acceptance and non-acceptance, respectively.

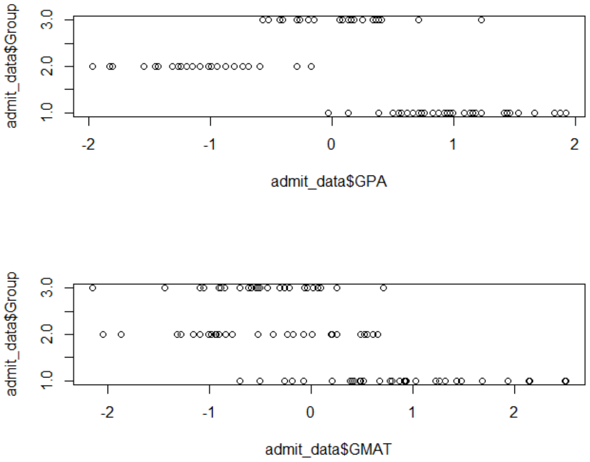

Figure 2: GPA and GMAT vs. Group

I decided to place my decision boundary graphs here for comparison.

b. Perform an LDA. Superimpose the decision boundary on an appropriate display of the training data. Does the boundary seem sensible? Compute Confusion matrix and misclassification rate for both test and training data. What do you observe?

The boundary seems sensible, with separation between all three classes clearly present. The training confusion matrix has (24,23,20) across its diagonal with only 3 instances present elsewhere. It's misclassification rate is $\sim$ 4.29%. The test confusion matrix has (4,3,5) across its main diagonal with 3 instances present elsewhere. It's misclassification rate is 20%.

c. Perform a QDA.

The boundary for this one seems sensible as well; you can see curved lines separating all three classes in the training data, but even better as the curvature in the decision boundaries brings the bounds closer to those outliers that are commonly misclassified. The training confusion matrix has (25,23,20) across its main diagonal and only 2 instances present elsewhere. It's misclassification rate is $\sim$ 2.86%. The test confusion matrix has (5,3,5) across its main diagonal and 2 instances present elsewhere. It's misclassification rate is $\sim$ $13.\overline{3}$%.

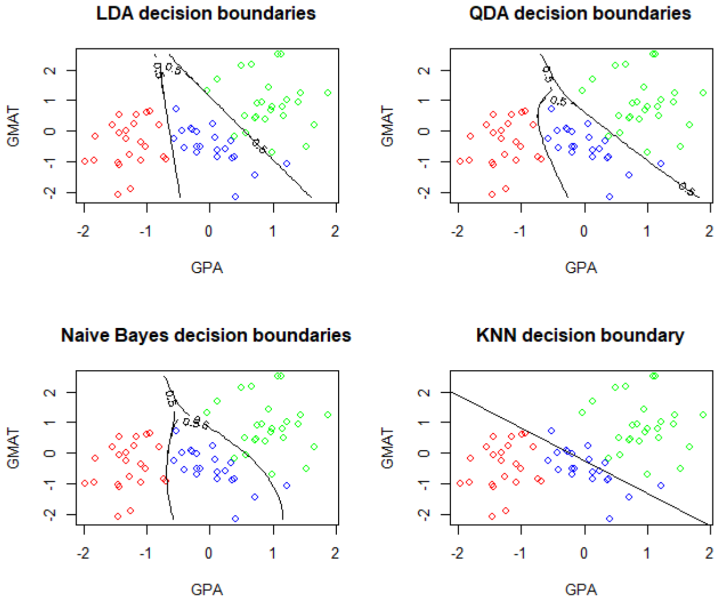

Figure 3: All decision boundaries

d. Create a naive Bayes model.

The boundary for this one seems sensible. Like the QDA, it has curvature that isolates the classes while remaining indeterminate towards the top where few class instances are present. These curves extend in a concave fashion and appear less strict and inflexible in comparison to the QDA curves. The training confusion matrix has (24,23,20) across its main diagonal and 2 instances present elsewhere. Its misclassification rate is $\sim$ 4.29%. The test confusion matrix has (4,2,5) across its main diagonal and 4 instances present elsewhere. Its misclassification rate is $\sim$ 26.$\overline{6}$%.

e. Create a KNN classifier with optimal K found via test error rates.

For this one, the classifier we used, KNN with $K = 55$, did not return probabilities for each individual class, so I estimated a 0.33$\overline{3}$% value for the decision boundary (where each class had approximately equal chance of occurring), and simply found points with probabilities close to that point (within 0.05 of that value). It drew a single line and that separated the accepted and non-accepted populations quite well, actually. Unfortunately, it did not provide much information on the borderline population. The misclassification rate for training and testing sets were 10% and 13.$\overline{3}$% respectively. The training confusion matrix had (25, 23, 15) on its main diagonal and 7 instances elsewhere. The testing confusion matrix had (4,5,4) on its main diagonal, and 2 instances elsewhere.

f. Compare results. Which one would you recommend?

QDA and Naive Bayes should more flexibility in their decisions compared to LDA. I feel as though QDA was a really solid option in this situation, but with a lot more data Naive Bayes could serve as a competitive model as well. LDA had a very reliable output and was easy to interpret, but its outliers were more noticeably away from the decision boundaries. As for KNN, with more data and more probability information (perhaps from using a separate package), one could most likely make a very reliable model as well, one where the decision boundaries wrap relatively closely around each group. In this case, however, I would recommend QDA.

Question 3

Consider a classification problem with K classes. The Bayes decision rule assigns an observation to the class for which P (Y = k|x), or equivalently, its log odds relative to class K as baseline, i.e., P (Y = k|x)/P (Y = K|x), is maximum.

Note, that we assume that $X \sim \mathcal{N}_K(\mu, \sigma^2)$ with each class having a specified mean, u_k, specified proportion, π_k, and shared covariance matrix, Σ.

a1. Assuming that the model parameters are known, verify that the log odds can be expressed as follows for LDA.

$$\log\left(\frac{P(Y=k \mid X=x)}{P(Y=K \mid X=x)}\right) = \tag{1}$$

$$\log\left(\frac{\pi_k f_k(x)}{\pi_K f_K(x)}\right) = \tag{2}$$

$$\log\left(\frac{\pi_k \exp(-\frac{1}{2}(x-\mu_k)^T\Sigma^{-1}(x-\mu_k))}{\pi_K \exp(-\frac{1}{2}(x-\mu_K)^T\Sigma^{-1}(x-\mu_K))}\right) = \tag{3}$$

$$\log\left(\frac{\pi_k}{\pi_K}\right) + \log\left(\frac{\exp(-\frac{1}{2}(x-\mu_k)^T\Sigma^{-1}(x-\mu_k))}{\exp(-\frac{1}{2}(x-\mu_K)^T\Sigma^{-1}(x-\mu_K))}\right) = \tag{4}$$

$$\log\left(\frac{\pi_k}{\pi_K}\right) + \log\left(\exp\left(-\frac{1}{2}(x-\mu_k)^T\Sigma^{-1}(x-\mu_k)\right)\right) - \log\left(\exp\left(-\frac{1}{2}(x-\mu_K)^T\Sigma^{-1}(x-\mu_K)\right)\right) = \tag{5}$$

$$\log\left(\frac{\pi_k}{\pi_K}\right) + \left[-\frac{1}{2}(x-\mu_k)^T\Sigma^{-1}(x-\mu_k)\right] - \left[-\frac{1}{2}(x-\mu_K)^T\Sigma^{-1}(x-\mu_K)\right] = \tag{6}$$

$$\log\left(\frac{\pi_k}{\pi_K}\right) - \left[\frac{1}{2}(x-\mu_k)^T\Sigma^{-1}(x-\mu_k)\right] + \left[\frac{1}{2}(x-\mu_K)^T\Sigma^{-1}(x-\mu_K)\right] = \tag{7}$$

$$\log\left(\frac{\pi_k}{\pi_K}\right) - \frac{1}{2}(\mu_k+\mu_K)^T\Sigma^{-1}(\mu_k-\mu_K) + x^T\Sigma^{-1}(\mu_k-\mu_K) = \tag{8}$$

$$a_k + \sum_{j=1}^{p} b_{kj}x_j \tag{9}$$

a2. Assuming that the model parameters are known, verify that the log odds can be expressed as follows for Naive Bayes.

$$\log\left(\frac{P(Y=k \mid X=x)}{P(Y=K \mid X=x)}\right) = \tag{10}$$

$$\log\left(\frac{\pi_k f_k(x)}{\pi_K f_K(x)}\right) = \tag{11}$$

$$\log\left(\frac{\pi_k \prod_{j=1}^{p} f_{kj}(x_j)}{\pi_K \prod_{j=1}^{p} f_{Kj}(x_j)}\right) = \tag{12}$$

$$\log\left(\frac{\pi_k}{\pi_K}\right) + \log\left(\frac{\prod_{j=1}^{p} f_{kj}(x_j)}{\prod_{j=1}^{p} f_{Kj}(x_j)}\right) = \tag{13}$$

$$\log\left(\frac{\pi_k}{\pi_K}\right) + \sum_{j=1}^{p} \log\left(\frac{f_{kj}(x_j)}{f_{Kj}(x_j)}\right) = \tag{14}$$

$$a_k + \sum_{j=1}^{p} g_{kj}(x_j) \tag{15}$$

b. Under what assumptions LDA is a special case of naive Bayes? Justify your answer.

LDA, being a classifier with a linear boundary, inherently is a special case of naive Bayes where $g_{kj}(x_j) = b_{kj}x_j$. The assumptions that have to be made, however, is that all features are normally distributed, one covariance matrix is common to all classes, and the features are independent of one another.

c. Under what assumptions naive Bayes is a special case of LDA? Justify your answer.

From the previous portions of this question, we can compare LDA to naive Bayes and see that for Naive Bayes to be a special case of LDA, we should model elements of $g_{kj}(x_j)$ so that $g_{kj}(x_j)$ is equal to $b_{kj}x_j$. To do this, we can model $f_{kj}x_j$ with a normal distribution, $\mathcal{N}(\mu_{kj}, \sigma_j^2)$, which will result in $g_{kj}(x_j) = \frac{\mu_{kj} - \mu_{Kj}}{\sigma_j^2}$. In this case, $g_{kj}(x_j) = b_{kj}x_j$ and Σ is a diagonal covariance matrix of σ's corresponding to each 1-dimensional from 1 to p. We can see this covariance matrix is the same for regardless of class and the distributions are normal; therefore, assuming sample measurements are independent, we can say that naive Bayes is a special case of LDA.