Homework 4

Andrew Boschee

2/14/2020

No collaborators. Outside Resources: Rdocumentation.com, Elements of Statisitcal Learning

Question 4.7.3, pg 168: This problem relates to the QDA model, in which the observations within each class are drawn from a normal distribution with a class-specific mean vector and a class-specific covariance matrix. We conside the simple case where p = 1; i.e., there is only one feature.

Suppose that we have K classes, and that if an observation belongs to the kth class then X comes from a one-dimensional normal distribution. X ~ N(

$$\mu_k, \sigma_k^2$$

). Recall that the density function for one-dimensional normal distribution is given in (4.11). Prove that in this case, the Bayes' classifier is not linear. Argue that it is in fact quadratic.

*Hint: For this problem, you should follow the arguments laid out in Section 4.4.2, but without making the assumption that

$$\sigma_1^2 = ... = \sigma_k^2$$

.

Results:

$$p_k(x) = \frac{\pi_k \frac{1}{\sqrt{2\pi}\sigma_k} \exp(-\frac{1}{2\sigma_k^2}(x-\mu_k)^2)}{\sum \pi_l \frac{1}{\sqrt{2\pi}\sigma_k} \exp(-\frac{1}{2\sigma_k^2}(x-\mu_l)^2)}$$

$$Constant(c) = \frac{\frac{1}{\sqrt{2\pi}}}{\sum \pi_l \frac{1}{\sqrt{2\pi}\sigma_k} \exp(-\frac{1}{2\sigma_k^2}(x-\mu_l)^2)}$$

$$p_k(x) = c\frac{\pi_k}{\sigma_k} \exp(-\frac{1}{2\sigma_k^2}(x-\mu_k)^2)$$

$$log(p_k(x)) = log(c) + log(\pi_k) - log(\sigma_k) + (-\frac{1}{2\sigma_k^2}(x-\mu_k)^2)$$

$$log(p_k(x)) = (-\frac{1}{2\sigma_k^2}(x^2 + \mu_k^2 - 2x\mu_k)) + log(\pi_k) - log(\sigma_k) + log(C')$$

We can see x to the power of two in the equation making the function quadratic.

Question 4.7.5, pg 169: We now examine the differences between LDA and QDA.

Part A. If the Bayes decision boundary is linear, do we expect LDA or QDA to perform better on the training set? On the test set?

In this situation, the QDA will most likely perform very well on the training set but potentially worse on the test set in comparison to the LDA due to overfitting by QDA.

Part B. If the Bayes decision boundary is non-linear, do we expect LDA or QDA to perform better on the training set? On the test set?

Result: In this non-linear event, I would expect for the QDA to peform better in both data sets.

Part C. In general, as the sample size (n) increases, do we expect the test prediction accuracy of QDA relative to LDA to improve, decline, or be unchanged? Why?

Result: QDA is recommended when the training size is very large to reduce concern regarding variance. As a result, as the training size increases, the QDA test prediction accuracy should improve relative to the LDA prediction accuracy.

Part D. 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.

Result: False. QDA decision boundary has a higher variance without a decrease in bias. LDA will perform better in this instance.

Question 4.7.10, pg 171: This question should be answered using the *Weekly* data set, which is part of the *ISLR* package. This data set is similar in nature to the *Smarket* data from this chapter's lab, except that it contains 1,089 weekly returns for 21 years, from the beginning of 1980 to the end of 2020.

Part E. Repeat (d) using LDA.

Results: Repeating process from homework 3, I split the weekly data into a training set and test set. Using Lag2 as predictor and direction as response variable on the training set, I then tested the model making predictions on the test set.

```
## Call:
## lda(Direction ~ Lag2, data = weeklyTrain)
##
## Prior probabilities of groups:
##      Down        Up
## 0.4477157 0.5522843
##
## Group means:
##             Lag2
## Down -0.03568254
## Up    0.26036581
##
## Coefficients of linear discriminants:
##            LD1
## Lag2 0.4414162
```

Table 1: LDA Confusion Matrix

	Down	Up
Down	9	34
Up	5	56

Table 2: LDA Accuracy

0.625

Part F: Repeat (d) using QDA

Can see that the QDA model is very optimistic and make predictions of the market always going up on the test set. Even though the accuracy is not much lower than LDA, it clearly has issues with false positves.

```
##
##         Down Up
##   Down     0 43
##   Up       0 61
```

Table 3: QDA Accuracy

0.5865385

Part G: Repeat (d) using KNN with K = 1

KNN did not perform well with k = 1. The confusion matrix shows that there were many false positves and false negatives in this model. With k = 1, I can assume there is extreme overfitting to the training set causing mediocre performance on the test set. I would expect better results adjusting k.

Table 4: KNN = 1 Confusion Matrix

	Down	Up
Down	21	22
Up	29	32

Table 5: KNN = 1 Accuracy

0.5096154

Part H. Which of these methods appears to provide the best results on this data?

To satisfy my curiosity, I brought back GLM model from prior assignment to see how it compares with LDA, QDA, and KNN. Both LDA and GLM had accuracy of .625 and identical confusion matrices. I am not surprised that KNN performed the worst but I am tempted to see how much better it can perform adjusting k.

Table 6: GLM Accuracy

	Down	Up
Down	9	34
Up	5	56

Table 7: GLM Summary

0.625

Misclassification Rate	Sensitivity	Specificity
0.375	0.9180328	0.2093023

Part I: Experiment with different combinations of predictors including possible transformation and interactions, for each of the methods. Report the variables, method, and associated confusion matrix that appears to provide the best results on the held out data. Note that you should also experiment with values for K in the KNN classifier.

K-Nearest Neighbors - k = 3, 10, 15

Table 9: KNN = 3 Confusion Matrix

	Down	Up
Down	16	27
Up	20	41

Table 10: KNN = 3 Accuracy

0.5480769

While k = 15 did give nearly ten percent increase in performance, I was expecting slightly better performance. Appears that K Nearest Neighbors is not a good modeling method when it comes to analyzing the stock market.

Table 11: KNN Accuracy Comparison

k = 3	k = 10	k = 15
0.5480769	0.5673077	0.5865385

GLM, QDA, LDA Interaction Comparison

GLM, QDA, and LDA are fairly similar regarding accuracy. GLM and LDA seem a little more consistent from the few times that I have ran the model. QDA seems to give more false positives on a regular basis.

Table 12: Polynomial Model Accuracy Comparison

GLM	LDA	QDA
0.625	0.6153846	0.625

Table 13: Polynomial LDA Confusion Matrix

	Down	Up
Down	8	35
Up	4	57

Table 14: Polynomial QDA Confusion Matrix

	Down	Up
Down	7	36
Up	3	58

Question 4.7.11, pg 172: 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.

Part D: 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?

LDA performed pretty well with only about a nine percent error rate on the test set. Similar to the prior models in the last homework, the Auto dataset has data that is easily classified and can make reasonable predictions.

Table 15: LDA Error Rate

0.1122449

Part E: 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?

Table 16: QDA Error Rate

0.1122449

KNN Model Tuning

Part G: 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?

Surprisingly, KNN again performs worse than LDA and QDA. Going all the way up to k =50, the model does not see much improvement.

Table 17: Accuracy Comparison

k=1	k=5	k=10	k=30	k=50
0.8265306	0.877551	0.8877551	0.8877551	0.877551

Classification Methods Summary

Question 5: Read the paper "Statistical Classification Methods in Consumer Credit Scoring: A Review" posted on D2L. Write a one page (no more, no less) summary.

I believe that consumer credit scoring is a very suitable application for statistical modeling. With many factors easily available to analyze for an applicant/consumer it is possible to make many assumptions and combine that with the available data. Immediately, with over 100,000 applicants and 100 variables, the paper shows how complex the models can become. Not only are there a large amount of possible variables, but the proportion of credit offerings goes from a very low 17% to a high 84%. This made me think about the variability in industries that we are considering and maybe the timeframe regarding the economical situations at the given time of applicants being accepted or declined. Not only is the wide range of industries making the analysis complicated, but not all of these independent variables are necessarily relevant to all applicants and I'm sure that much of the data is missing due to applicants not knowing the answer or their unwillingness to give the information. This brings another question on how to handle those who are unwilling to provide certain information and whether that will impact how much influence the variable missing has on the end result. This is a situation where it is not just the model that is needed, but human judgement and knowledge of the industry related to the model. On that topic, the selection of variables through human judgement and statistical models such as stepwise procedures is also discussed. It would be interesting to see how many independent variables get used in these models on average. The first modeling method that comes to mind right away since this is a fairly simple accept/reject outcome is logistic regression. For simplicity, K Nearest Neighbors always runs through my mind but I feel like decision trees would be the best fit when considering it from a business perspective. Recursive partitioning would be helpful and easily interpretable for those who are less familiar with some of the more complicated modeling methods. With technological advances and more efficient modeling methods, neural networks are definitely an option that may provide a slight improvement in accuracy, but I do not see that as a big enough benefit to give away a modeling method that is somewhat easily interpretable for all others. There are many other ways to apply these models for less critical decisions such as marketing and non-lifechanging impacts. Before the conclusion, they brought up the possibility of cluster analysis for market segmentation and that was one method that definitely came to my mind when considering how it would be helpful to get a higher level view of the overall demographics of applicants.

Wine Quality Dataset

Question 6: Explore this website (https://archive.ics.uci.edu/ml/datasets.html) that contains open data sets that are used in machine learning. Find one data set with a classification problem and write a description of the dataset and problem. I don't expect you to do the analysis for this homework, but feel free to if you want!

I have selected the 'Wine Quality' dataset from the website. I have meant to work with this dataset for a while after frequently seeing it on Kaggle.com. I think this could be analyzed in a few different ways. While I believe the most interesting part would be to do regression on the quality score of the wine, it may be interesting to see if it is as easy as I suspect to classify the different types of wine.

With two datasets for different types of wine, I should be able to easily combine them and compare the independent variables for exporatory data analysis and the challenge of classification.

```
##   fixed.acidity volatile.acidity citric.acid residual.sugar chlorides
## 1           7.4             0.70        0.00            1.9     0.076
## 2           7.8             0.88        0.00            2.6     0.098
## 3           7.8             0.76        0.04            2.3     0.092
## 4          11.2             0.28        0.56            1.9     0.075
## 5           7.4             0.70        0.00            1.9     0.076
## 6           7.4             0.66        0.00            1.8     0.075
```

```
##   free.sulfur.dioxide total.sulfur.dioxide density   pH sulphates alcohol
## 1                  11                   34  0.9978 3.51      0.56     9.4
## 2                  25                   67  0.9968 3.20      0.68     9.8
## 3                  15                   54  0.9970 3.26      0.65     9.8
## 4                  17                   60  0.9980 3.16      0.58     9.8
## 5                  11                   34  0.9978 3.51      0.56     9.4
## 6                  13                   40  0.9978 3.51      0.56     9.4
##   quality
## 1       5
## 2       5
## 3       5
## 4       6
## 5       5
## 6       5
```

```
##   fixed.acidity volatile.acidity citric.acid residual.sugar chlorides
## 1           7.0             0.27        0.36           20.7     0.045
## 2           6.3             0.30        0.34            1.6     0.049
## 3           8.1             0.28        0.40            6.9     0.050
## 4           7.2             0.23        0.32            8.5     0.058
## 5           7.2             0.23        0.32            8.5     0.058
## 6           8.1             0.28        0.40            6.9     0.050
##   free.sulfur.dioxide total.sulfur.dioxide density   pH sulphates alcohol
## 1                  45                  170  1.0010 3.00      0.45     8.8
## 2                  14                  132  0.9940 3.30      0.49     9.5
## 3                  30                   97  0.9951 3.26      0.44    10.1
## 4                  47                  186  0.9956 3.19      0.40     9.9
## 5                  47                  186  0.9956 3.19      0.40     9.9
## 6                  30                   97  0.9951 3.26      0.44    10.1
##   quality
## 1       6
## 2       6
## 3       6
## 4       6
## 5       6
## 6       6
```