Homework 4

Rylie Fleckenstein

2/14/21

Exercises (ISLR)

1. 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 consider 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(μ_k, σ_k^2). Recall that the density function for the 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 σ^2 1 = ... = σ^2

We can assume that each class has its own covariance matrix in order to fall in line with the quadratic assumption. We can also state that $p = 1$ meaning there is only 1 feature. Therefore, we know X ~ N(μ_k, σ_k^2).

$$f_k(x) = \frac{1}{\sqrt{2\pi\sigma_k}} exp(-\frac{1}{2\sigma_k^2}(x-\mu_k)^2)$$

$$P(Y=k|X=x) = \frac{P(x=x|y=k)P(y=k)}{\sum_{k=1}^{K} P(x=x|y=k)P(y=k)} = \frac{\pi_k f_k(x)}{\sum \mu_k f_k(x)}$$

We know that $\mu_k f_k(x)$ is the largest so using Bayes' Classifier

$$\delta_k(x) = log[\mu_k f_k(x)] = log\mu_k + log[\frac{1}{\sqrt{2\mu\sigma_k}}] - \frac{1}{2\sigma_k^2}(x-\mu_k)^2$$

Therefore, the Bayes' classifier is quadratic.

2. Question 4.7.5 pg 169

We now examine the differences between LDA and QDA.

(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?

If the Bayes decision boundary is linear we can expect the QDA to perform better on the training set because it is a more flexible approach which would allow the model to fit the data more closely. On the test set, however, the LDA model would score much better than the QDA because the QDA model tends to overfit the training data when the Bayes decision boundary is linear.

(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?

If the Bayes decision boundary is non-linear the QDA should perform better on the training set and the test set. LDA does not perform well on non-linear data sets.

(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?

As the sample size n increases we expect QDA to perform better than LDA because the model is more flexible which means it can account for more variance in the data. an LDA approach would have a higher level of bias which is good for smaller data sets but not so good as n increases.

(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.

This is false. The reasoning for this is because the QDA model will likely overfit the training data which leads to negative returns in the test error rate. The goal of a machine learning model is to determine the coefficients that properly generalize the data so that it can perform well on data it has not seen. When we overfit to the training set the model does not know how to generalize the patterns it has learned and therefore does not know what to do when it sees data that is different from the training data.

3. Continue from Homework 3 Question 4.7.10(e-i) pg 171

(e) (d) Now fit the LDA model using a training data period from 1990 to 2008, with Lag2 as the only predictor. Compute the confusion matrix and the overall fraction of correct predictions for the held out data (that is, the data from 2009 and 2010).

```
## $`Confusion Matrix`
##       pred
## valCol Down Up
##   Down    9 34
##   Up      5 56
##
## $`Overall Percentage Correct`
## [1] 62.5
```

(f) Repeat (d) using QDA.

```
## $`Confusion Matrix`
##       pred
## valCol Down Up
##   Down    0 43
##   Up      0 61
##
## $`Overall Percentage Correct`
## [1] 58.65385
```

(g) Repeat (d) using KNN with K = 1.

```
## $`Confusion Matrix`
##       pred
## valCol Down Up
##   Down   21 30
##   Up     22 31
##
## $`Overall Percentage Correct`
## [1] 50
```

(h) Which of these methods appears to provide the best results on this data?

If we take a look at the overall percentage of correctly predicted outcomes we can see the LDA performed the best with 62.5% accuracy. The next best in overall prediction accuracy is the QDA model, however, the model was unable to predict "Down" occurrences in the market and therefore has a sensitivity of 0. This makes the second best model the KNN classifier with k=1 coming in with an overall prediction accuracy of 50% but a sensitivity of 41% and a specificity of 58%.

(i) Experiment with different combinations of predictors, including possible transformations 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.

After experimenting with different combinations of predictors, different interaction terms, and different value of k in the knn classifier, I was not able to create a model that performed better than the original LDA model using lag2 as the only predictor. The next closest model was a knn model that used lag2 as the predictor and used a value of 13 for k. The overall percentage of correctly predicted outcomes was 59% which is close to the LDA model with 62.5% but not quite as good.

```
## $`Confusion Matrix`
##       pred
## valCol Down Up
##   Down   20 19
##   Up     23 42
##
## $`Overall Percentage Correct`
## [1] 59.61538
```

4. Continue from Homework 3 Question 4.7.11(d,e,g) pg 172

(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?

After performing LDA on the training data and making predictions on the test data, we achieved a test error of 11.864.

```
## $`Test Error`
## [1] 11.86441
```

(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?

After performing QDA on the training data and then making predictions on the test data set, we achieve a test error of 14.406.

```
## $`Test Error`
## [1] 14.40678
```

(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?

Using only the variables that seemed most associated with mpg01 from part (b) I was able to produce a matrix for use in the knn algorithm. I then looped through different values for k in the following sequence, c(1, 3, 5, 7, 9, 11, 13) and was able to obtain the below ordered test errors. From studying the figure 2.17 on page 42 of ISLR we can see the nearest neighbors algorithm starts to overfit the data and the test error starts to increase again around k = 13 so it would be counterproductive to train our model with anything higher. As we can see, the best error rates were obtained using k = 1 and k = 11 for this data set.

```
## $`Test Error`
## [1] 15.25424
##
## $`Test Error`
## [1] 16.94915
##
## $`Test Error`
## [1] 16.10169
##
## $`Test Error`
## [1] 16.10169
##
## $`Test Error`
## [1] 16.10169
##
## $`Test Error`
## [1] 15.25424
##
## $`Test Error`
## [1] 16.10169
```

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.

The paper "Statistical Classification Methods in Consumer Credit Scoring: A Review" written by D.J> Hand and W.E. Henley covers the how companies have approached determining which potential customers would be considered a good risk and which would be considered a bad risk. Within the paper they use credit "to refer to an amount of money that is loaned to a consumer by a financial institution and which by be repaid, with interest, in instalments". When financial institutions loan money there is always risk involved. There is always the chance that the borrower with default on the loan or that the borrower will fall behind on payments, so it is important for these institutions to try to gain some insight into the risks they are taking to improve their business plans and maximize their profits or minimize their losses. A typical method for discrimination is the use of credit scoring. Most statistical models used are called scorecards or classifiers and they predict the probability of defaulting. Common statistical methods include linear regression, logistic regression, linear discriminant analysis, and decision trees. It is important to note that there are more subtle types of analysis that can be performed that looks beyond simply defaulting on loans or not. One prediction that could also be useful would be how many payments a borrower is likely to be behind by a certain period. The reason these subtleties are important is because financial institutions are not necessarily concerned with who will default and who will not. They are more concerned with maximizing their profits which sometimes means taking risks. According to the paper, low risk borrowers who make quick payments are not going to return the profits of someone who is higher risk that might not make all payments on time. The reasoning for this is because higher risk borrowers are often given higher interest rates. When it comes to determining which features are to be used in the statistical models there are some hurtles that need to be jumped through. Some important and potentially relevant information is not legally allowed to be collected such as gender. Another important factor to keep in mind when collecting information from customers is that most will only withstand so much questioning before they chose to go elsewhere. One strategy that is implemented to deal with this is using a screening method that splits the clients into three categories, good risk, bad risk, unknown and then only completes further screening on the unknown. When it comes to the data itself, data collected for credit scoring usually has a substantial amount of missing values. Some of the missing values are said to be structurally missing, in the sense that the question was a follow up to a previous question and if the previous question was not applicable, then by logical reasoning the sequential answer will missing. Also, some missing values are just randomly missing. There is, however, information to be gained from the missing values and the paper talks about several approaches to dealing with them. A few of the outlined approaches are creating a new attribute for the missing values, dropping incomplete vectors, or substituting values in for the missing values. When comes to making the final classification there have been several methods used in the credit industry, some of which are still being used and some of which have been proven to be useful but not necessary. A list of the different approaches outlined in the paper is the following: logistic regression, mathematical programming methods, recursive partitioning, expert systems, neural networks, smoothing nonparametric methods, and time varying models. It has been found that each method can be applied to the problem and each method produces similar results they just get to those results in different ways. It is for that reasoning, that most of the "black box" methods are not industry standard. It is more common to use methods that are more explainable so that the decisions being made can be clearly justified and explained by the models which helps with legal and business regulations. The paper then goes on to talk about how the future advances in the improvement of these models will be in improving the model algorithms, gathering better characteristics, pulling information from the subset of customers who were rejected for the loans in the first place. A common practice that was discussed in the paper that is used for this issue is called reject inference. Reject inference is the approach that tries to draw information from the people who were classified as bad risks and try to infer their true class. The financial institution is going to know how the people they gave loans to with perform in due time. They will, however, never know how those individuals they never gave loans to would have performed if they were given a chance. Important information for tuning a model can be learned from these people and one approach to gaining insight is to use data collected by other financial institutions that loaned the bad risk borrowers money to see how they performed once they were given a chance.

6. Explore this website that contains open data sets that are used in machine learning. Select a data set that has classification as a Default Task and describe, in your own words, the task, including a description of the data set. Look for data sets that are amenable to the analyses we have learned thus far. Pay attention to the characteristics of the data with selecting an analysis method. I do not expect you to do the analysis for this homework, but feel free to if you want!

The data set that I decided to go with was the Breast Cancer data set. This data set has a default task of classification and from further investigation it was concluded it is in fact a binary classification data set. The main goal is to predict if there will be recurrence events or if their will be no recurrence events in reference to breast cancer in breast cancer patients. The data is made up of 9 variables and a class variable. They are the following: Age: age range of the patient at time of diagnosis, menopause: whether the patient is pre or postmenopausal at time of diagnosis, tumor size: diameter of the tumor, inv-nodes: number of lymph nodes with visible breast cancer, Node caps: (yes/no) the replacement of lymph nodes by the cancer, degree of malignancy: range from 1-3, breast: which breast the cancer occurs in, breast quadrant: area of the breast where cancer is located, irradiation: (yes/no) have they recieved radiation therapy or not and then the Class attribute is the final attribute. The class distribution is 201 instances of no-recurrence events and 85 instances of reccurence events. A good place to start with this data set would be plotting the relationships between the variables based upon the two classes. We can gain valuable information and insight into which features might have the largest effect on the class outcome through visualizations. Then, I would start with performing a logistic regression on the data set. Since this is a relativiely small data set I would implement some form of cross validation into the model building process. Also, from just looking at the data set I would want to make sure there aren't variables that present collinearity and also would look for potential signs that variables might have non-linear relationships with the response and would benefit from transformations.

Sources:

Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani. (2013). An introduction to statistical learning : with applications in R. New York :Springer,

D. J. Hand & W. E. Henley, 1997. "Statistical Classification Methods in Consumer Credit Scoring: a Review," Journal of the Royal Statistical Society Series A, Royal Statistical Society, vol. 160(3), pages 523-541, September.