Homework 4

Alex Soupir

February 14, 2020

Packages: ISLR, MASS, class

Collaborators:

Please do the following problems from the text book ISLR or written otherwise.

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, $X\ N(\mu_k, \sigma_k^2)$. 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 $\sigma_1^2 = ... = \sigma_K^2$.

If σ isn't the same for all K then for (4.12) σ must maintain K:

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

Taking the log of both sides will remove the *exp* on the top:

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

Multiply the bottom to the other side:

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

Expand last quadratic term:

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

indicating that there is a quadratic term present for μ_K^2.

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 trianing set? On the test set?

The QDA is much more flexible so it will probably fit the training set better than the LDA. However, due to the flexibility chances are it is going to overfit the training set making the LDA fit the testing set better.

(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?

For a non-linear boundary, the QDA will probably perform much better on both the training and the testing. LDA would likely miss a great deal.

(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?

I think this was mentioned in lecture, that if the sample size is small, LDA generalizes much better and QDA doesn't. So if n is increasing, QDA should have more data to be trained on and therefore improve in prediction accuracy to the LDA.

(d) True or False: Even if they Bayes decision layer 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 linea decision bounday. Justify your answer.

False. LIke stated for part (c), the QDA is more flexible but will overfit if there aren't a lot of data points for a linear boundary layer, causing it to have poor test error rates.

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

This question should be answered using the `Weekly` data set, which is part of the `ISLR` package. This data is similar to the `Smarket` data from this chapter's lab, except that it contains 1,089 weekly returs for 21 years, from the beginning of 1990 to the end of 2010.

```
## Confusion Matrix of data from 2009 and 2010 using Logistic Regression:

##
## preds2.d Down Up
##     Down    9  5
##     Up     34 56

## Overall fraction of correct predictions:

## [1] 0.625
```

(e) Repeat (d) using LDA. [(d) Now fit the logistic regression 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 to 2010.)]

```
## Confusion Matrix of data from 2009 and 2010 using LDA:

##
## preds3.e_class Down Up
##           Down    9  5
##           Up     34 56

## Overall fraction of correct predictions:

## [1] 0.625
```

(f) Repeat (d) using QDA.

```
## Confusion Matrix of data from 2009 and 2010 using QDA:
```

```
##
## preds3.f_class Down Up
##           Down    0  0
##           Up     43 61
```

```
## Overall fraction of correct predictions:
```

```
## [1] 0.5865385
```

I don't know if this is the right ourcome from using QDA, but if not I for the life of me cannot figure out why it is only predicting up, unless this is part of QDA not getting enough data to make good predictions. Looking at the posterior, most of the probabilities fall close to 0.5 but all having slightly over 0.5 for "Up".

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

```
## Confusion matrix of data from 2009 and 2010 using KNN (K=1):
```

```
##
## preds3.g Down Up
##     Down   21 29
##     Up     22 32
```

```
## Overall fraction of correct predictions:
```

```
## [1] 0.5096154
```

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

The logistic regression with a 0.5 threshold and the linear discriminat analysis both predicted 62.5% of the correct responses. As previously mentioned, the QDL was reporting some really weird responses with everything being "Up" and following the Lab in the book, I think I have everything correct or if not I am not sure what is not correct. Even with reporting everything as "Up" it was predicting the correct response on the 2009 and 2010 data 58.7% of the time. K-nearest neighbor performed the worst with $K = 1$, predicting only 51.0% of the responses correctly.

(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 experiemnt with values for K in the KNN classifier.

It looks like the highest for the logistic regression is what was being done for class. Tested several interactions, individual terms, and combinations of covariates and *Lag2* alone performed better than all of the other combinations. Same seems to be the case for the linear discriminant analysis. For QDL, *Lag1*, *Lag2*, and *Lag3* alone produced the same result of 58.7% accurate predictions, but they were producing all "Up" repsonses which may work in the long-run hope of the stock, as in "long term returns" but probably isn't a good strategy for week to week trading. I also tried some interaction terms and combinations of the covariates. Finally I ran KNN with a range of K on normalized data to see if that would be beneficial. Using all of the data (all Lags, and Volume) the best predicion accuracy was acheived with $K = 31$ at 58.7% correct.

The best model was either the logistic regression or the LDA. The logistic regression with just Lag2 cannot be moved from 0.5 without the overall fraction of correct prediction decreasing.

Issue I had made: I thought the KNN was doing a really great job with a 93% overall fraction being correct but I had accidentally included "Today" which obviously would be a high accuracy because it is related to the direction.. So I had to start over with it and thats where I ended up with a slightly higher KNN but lower still than the logistic regression or the LDA.

```
## Confusion Matrix of data from 2009 and 2010 using Logistic Regression:
```

```
##
## preds2.d Down Up
##     Down    9  5
##     Up     34 56
```

```
## Overall fraction of correct predictions:
```

```
## [1] 0.625
```

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

```
## Columns that were determined to be most associated with mpg01 in (b):
```

```
## [1] "cylinders"    "displacement" "horsepower"   "weight"       "mpg01"
```

(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?

```
## Confusion matrix of hold out data from LDA:
```

```
##
## preds4.d  0  1
##        0 37  0
##        1  7 36
```

```
## Testing error:
```

```
## [1] 0.0875
```

(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?

```
## Confusion matrix of hold out data from QDA:
```

```
##
## preds4.e  0  1
##        0 38  1
##        1  6 35
```

```
## Testing error:
```

```
## [1] 0.0875
```

(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 values of K seem to perform the best on this data set?

```
## K with the highest fraction of correct predictions:
```

```
## [1] 6
```

```
## Fraction of correct predictions:
```

```
## [1] 0.9375
```

```
## K with the lowest testing error:
```

```
## [1] 6
```

```
## Testing error:
```

```
## [1] 0.0625
```

```
## Confusion matrix of best K:
```

```
##
## preds4.g  0  1
##        0 40  1
##        1  4 35
```

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.

Summary:

The authors are focusing on the importance of credit scoring to determine whether or not applicants will default on the lending which they are given, and even state that this is important to both the lending organization (that is expected) but also to the applicant so they are not pulled into a loan that they can not feasibly pay off. Typical methods of analysis are the discribinant analysis, linear regression, logistic regression, as well as decision trees. These are used to analyze the different attributes taht are collected about an applicant during the application process. It is also stated that having too long of an application process may deter the applicant from a particular organization causing them to search for credit somewhere else, or remove those whom do not truly need the credit; this may remove those with higher risk of repayment but also decrease the number of individuals that are able to be used for training a model.

Data that is collected on individuals can be wide ranging in both amount and completeness. Some studies have used several methods of data selection and pooling. For example, one study used 2500 different attributes collected on individuals but later subset those to the best 10 variables. Continuous data is sometimes pooled into ranges, converting the continuous variable to a factor level variable. Another issue stated is the completeness of the data that is collected. Some responses are only conditional on the previous question, and some may not be *required* but may still pull weight in creating a model. There are ways to get around the issue of incomplete data, but doing so can introduce bias. The selection of attributes can be done with expert knowledge, step-wise selection, or selection based on those variables with large difference between those that default and those that do not.

Like with any model creation, the ability for the model to generalize to different data is important. Authors state that there are a large amount of data available and so using a testing split or different set of data can be used to judge the model’s performance. Since the other data that is used for testing has known risk already associated with each applicant, a confusion matrics can be created. Something like a ROC-AUC is sometimes used as a measure as well. Another metric may just be the error of the model as by the confusion matrix. It is also mentioned that a resulting risk of 0 does not always mean that the applicants score card is bad, so to compare back to the score card from the model needs other methods.

The way of scoring a model is not quantitative rather than how classifications used to be made with subjective judgements. This quantitative scoring has resulted in more accurate predictions than the subjective judgement from experts. The quantitative objective scoring though, at the time of this article, apparently had drawn some concern from the applicants even though it was more accurate (stated possibly due to not being as personal of a process as having experts make the decision, i.e. computer vs person). One of the first models that did this was used in 1941, but there are many different methods that have been used to solve this problem, or come close to solving, since. The best method is highly dependent on the what data is being used and the situation.

Models are assessed in different ways and the combination of these is what would make the model ‘good’. One metric that is probably most important is the accuracy of the model as a poor predicting model is most likely not of good use. However, things like time to run the model are also important. Another aspect mentioned is the interpretability of the model, and why it decided to make the classification that it had. For the interpretability, methods like linear models and decision trees are nice, but where the data has bad structure, neural networks can provide decent accuracy but not easy to interpret. To increase the accuracy of the predictions, the authors state that improvements may come from new attributes, and even new modeling methods of the data. New metrics may carry legal issues though, such as the use of sex and race in the decision making process. Also metrics like spending patterns that are used in credit card fraud detection.

6. Explore this website that contains open datasets that are used in machine learning. Find one dataset 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!

The classification data set that I had found and think is interesting is titled "Mamographic Mass Data". The data set has 6 columns in the `.data` file. The first column is a BI_RADS assessment with values between 1 and 5 with 1 being benign mass and 5 being a high probability of it being malignant, as reviewed by 2 doctors. This is probably a metric that isn't to be used for prediction since its likely to be highly correlated with the response. The second column is age, followed by shape (round, oval, lobular, irregular). The last 2 predictors are margin (circouscribed, microlobated, obscured, ill-defined, spiculated) and density (high, iso, low, fat-containing). These are to predict whether the mass is benign or malignant. There is an issue with some missing values throughout the data frame though, which may either need the missing observations removed (might amount to a lot smaller of a data set) but has a total of 961 observations.

References:

- stackoverflow
- datasharkie
- ISLR book for the lab portion