# Homework 4

Prakash Paudyal

Please do the following problems from the text book ISLR.

### $\mathbf{1.}\, \boldsymbol{Question 4.7.3\ pg\ 168}$

3. This problem relates to the QDA model, in which the observations within each class are drawn from a normal distribution with a classspecific 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(??k, ??2k). 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 ??2 1 = . . . = ??2K.

**ANS:** From the Bayes 'theorem Equation (4.10) and Gaussian function in equation(4.11) if $\sigma$ varies from 1 to $k$ then we can get equation (4.12) by substituting Gaussian function equation (4.11) in equation (4.10) $p_k(x) = \frac{\pi_k \frac{1}{\sqrt{2\pi}\sigma_k} \exp(-\frac{1}{2\sigma_k^2}(x-\mu_k)^2)}{\sum \pi_l \frac{1}{\sqrt{2\pi}\sigma_l} \exp(-\frac{1}{2\sigma_l^2}(x-\mu_l)^2)}$

Taking the log both side above equation becomes $\log(p_k(x)) = \frac{\log(\pi_k)+\log(\frac{1}{\sqrt{2\pi}\sigma_k})+-\frac{1}{2\sigma_k^2}(x-\mu_k)^2}{\log(\sum \pi_l \frac{1}{\sqrt{2\pi}\sigma_l} \exp(-\frac{1}{2\sigma_l^2}(x-\mu_l)^2))}$

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

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

As we can see this expression is quadratic function of x. Hence Bayes'classifier in above case is not linear.

### $\mathbf{2.}\, \boldsymbol{Question 4.7.5 pg 169}$

5. 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?

**Ans** If the Bayes decision boundary is linear, QDAA will perform better than LDA on training set beceuse QDA is more felexibile model and give closer fit. In test set LDA wil perform better because it avoids overfitting of test data.

(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?

**Ans:** For non linear Bayes decision boundary, QDA will perform better.

(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?

**Ans:** If data size is large, test prediction accuracy of QDA wil improve, because as the sample size increase the more flexibile method will fit better as the variance is offset by the larger smaple size.

(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.

**Ans:** False, because QDA will overfit and gives higher test error rate than LDA.

### $\mathbf{3}.\boldsymbol{ContinuefromHomework\#3Question4.7.10(e-i)pg171}$

10. This question should be answered using the Weekly data set, which is part of the ISLR package. This data 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 1990 to the end of 2010.

(e) Repeat (d) using LDA.

LDA FIT and confusion matrix

```
## Call:
## lda(Direction ~ Lag2, data = Weekly, subset = train)
##
## 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
```

```
##       Direction.test
##        Down Up
##   Down    9  5
##   Up     34 56
```

Fraction of correct predictions

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

(f) Repeat (d) using QDA.

QDA FIT and confusion matrix

```
## Call:
## qda(Direction ~ Lag2, data = Weekly, subset = train)
## 
## Prior probabilities of groups:
##      Down        Up 
## 0.4477157 0.5522843 
## 
## Group means:
##             Lag2
## Down -0.03568254
## Up    0.26036581

##       Direction.test
##        Down Up
##   Down    0  0
##   Up     43 61
```

Fraction of correct predictions

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

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

KNN FIT and confusion matrix

```
##         Direction.test
## knn.pred Down Up
##     Down   21 30
##     Up     22 31
```

Fraction of correct predictions

```
## [1] 0.5
```

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

**ANS:** Logistic regression and LDA are the best models which provide the 62.5% accuracy of predection.

(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.

## Logistic regression with Lag2+Lag4+Lag5:Lag4

```
##         Direction.test
## glm.pred Down Up
##     Down    9  4
##     Up     34 57
```

fraction of correct predictions

```
## [1] 0.6346154
```

## LDA with Lag2 interaction with Lag1

```
##       Direction.test
##        Down Up
##   Down    0  1
##   Up     43 60
```

fraction of correct predictions

```
## [1] 0.5769231
```

## QDA with sqrt(abs(Lag2))

```
##          Direction.test
## qda.class Down Up
##      Down   12 13
##      Up     31 48
```

fraction of correct predictions

```
## [1] 0.5769231
```

## KNN k =4

```
##         Direction.test
## knn.pred Down Up
##     Down   16 20
##     Up     27 41
```

fraction of correct predictions

```
## [1] 0.5480769
```

## KNN k = 15

```
##         Direction.test
## knn.pred Down Up
##     Down   20 20
##     Up     23 41
```

fraction of correct predictions

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

### 4. ContinuefromHomework#3Question4.7.11(d, e, g)pg172

11.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.

(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?

```
## Call:
## lda(mpg01 ~ displacement + horsepower + weight + year, data = train1)
##
## Prior probabilities of groups:
##         0         1
## 0.5068027 0.4931973
##
## Group means:
##   displacement horsepower   weight     year
## 0     276.0604  130.69799 3623.631 74.30872
## 1     116.0034   78.77241 2324.159 77.68966
##
## Coefficients of linear discriminants:
##                       LD1
## displacement -0.007095375
## horsepower    0.011334815
## weight       -0.001277128
## year          0.132173702
```

```
##          Actual
## Predected  0  1
##         0 37  1
##         1 10 50
```

test error

```
## [1] 0.1122449
```

(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?

```
## Call:
## qda(mpg01 ~ displacement + horsepower + weight + year, data = train1)
##
## Prior probabilities of groups:
##         0         1
## 0.5068027 0.4931973
##
## Group means:
```

```
##   displacement horsepower   weight     year
## 0     276.0604  130.69799 3623.631 74.30872
## 1     116.0034   78.77241 2324.159 77.68966
```

```
##          Actual
## Predected  0  1
##         0 40  3
##         1  7 48
```

test error

```
## [1] 0.1020408
```

(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?

I used function to calculate missclassification error for different values of k(1to100). Test error are given below. At k=43 it gaves minimum error value 0.09183673.

```
## $K
## [1] 100
## 
## $misclass
##   [1] 0.15306122 0.13265306 0.13265306 0.13265306 0.14285714 0.13265306
##   [7] 0.14285714 0.13265306 0.14285714 0.14285714 0.15306122 0.14285714
##  [13] 0.15306122 0.16326531 0.14285714 0.14285714 0.14285714 0.16326531
##  [19] 0.12244898 0.12244898 0.12244898 0.11224490 0.12244898 0.12244898
##  [25] 0.11224490 0.11224490 0.11224490 0.11224490 0.11224490 0.11224490
##  [31] 0.11224490 0.11224490 0.11224490 0.10204082 0.11224490 0.10204082
##  [37] 0.11224490 0.11224490 0.11224490 0.10204082 0.10204082 0.12244898
##  [43] 0.09183673 0.11224490 0.11224490 0.11224490 0.11224490 0.11224490
##  [49] 0.11224490 0.11224490 0.11224490 0.10204082 0.11224490 0.11224490
##  [55] 0.10204082 0.10204082 0.10204082 0.09183673 0.10204082 0.11224490
##  [61] 0.11224490 0.12244898 0.10204082 0.09183673 0.10204082 0.09183673
##  [67] 0.09183673 0.11224490 0.11224490 0.11224490 0.11224490 0.11224490
##  [73] 0.10204082 0.10204082 0.10204082 0.11224490 0.11224490 0.10204082
##  [79] 0.10204082 0.10204082 0.11224490 0.11224490 0.11224490 0.11224490
##  [85] 0.11224490 0.11224490 0.11224490 0.11224490 0.11224490 0.11224490
##  [91] 0.11224490 0.11224490 0.11224490 0.11224490 0.11224490 0.11224490
##  [97] 0.11224490 0.11224490 0.11224490 0.11224490
## 
## $Kmin
## [1] 43
```

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.

Title: Statistical Classification Methods in Consumer Credit Scoring: A Review

Authors: D. J. Hand and W. E. Henley

This paper describes about the various problems that need to be addressed in order to effectively determine whether a financial institution should or should not loan money to a particular consumer. Since there has been a drastic increase in consumer credit in the last 3 decades, organizations like banks, building societies, and retailers are in dire need of a proper mechanism that can use predictor variables from application forms as well as the consumer's previous credit history and provide information about whether the customer will be able to pay back the amount or not.

This paper discusses about three main complications that consumer credit scoring applications need to handle. First, the population of potential borrowers always keeps on evolving because of the multiple steps of selection that happens in between of the process of customers receiving an application form from the bank to the bank finally deciding on the individuals whom they are going to offer a further loan. The second complication arises due to population drift, which is the tendency of populations to evolve with time so that the distributions change. This phenomenon is usually caused by economic pressures and a changing competitive environment and leads to a poor performance of the metrics in use. This demands for a constant adjustment of classification thresholds and dynamic update of the classification metrics. The third issue that the authors mention is that that only those applicants who are accepted for credit will be followed up to find out whether they are good or bad risks, and this leads to the design sample being a biased sample from the overall population. Moreover, even though there is a widespread use of statistical techniques in the consumer credit industry, the need for confidentiality in the organization has resulted in very less literature being published.

The datasets that are available for developing credit scoring metrics are usually very large with multivariate categorical values. Some of the common preprocessing steps that need to be performed on such datasets include coding categorical values into numerical form, dealing with missing values, and deciding on the number of variables to used so that the metric does not over-fit or under-fit the training data. Important features or characteristics are usually filtered out by using expert knowledge, applying stepwise statistical approaches, or by using a measure of the difference between the distributions of the good and bad risks on that characteristic. In addition to these, we also need to be aware of the legislative restrictions on the information which may or may not be used in constructing a credit classification rule. Even though discriminant analysis and linear regression are two of the most widely used metrics, other statistical techniques like logistic regression, probit analysis, nonparametric smoothing methods, mathematical programming, Markov chains models, recursive partitioning, expert systems, genetic algorithms, neural networks and conditional independence model have also been used in the industry. The data structure that is being used, features that have chosen, and the extent to which it is possible to separate the classes by using those features and the objective of the classification are some of the important factors that can help in deciding the metric that should be selected. Measure like accuracy, overall misclassification rate, cost-weighted misclassification rate, bad risk rate among those accepted, and profitability can be used to assess the performance of different credit scoring metrics.

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!

   **ANS:**

**Dataset: Activity recognition with healthy older people using a battery-less wearable sensor dataset**

This dataset consists of sequential motion data from 14 healthy older people that are 66 to 86 years old. The data was collected by using a battery-less, wearable sensor that was placed on top of their clothing for the recognition of activities in clinical environments. Participants were allocated in two clinical room settings (S1 and S2). The setting of S1 (Room1) used 4 RFID reader antennas around the room (one on ceiling level, and 3 on wall level) for the collection of data, whereas the room setting S2 (Room2) used 3 RFID reader antennas (two at ceiling level and one at wall level) for the collection of motion data. The 14 subjects were asked to perform broadly scripted activities like walking to the chair, sitting on the chair, getting off the chair, walking to the bed, lying on the bed, getting of the bed, and walking to the door. The possible class labels that will be assigned for every sensor observation are: sitting on bed, sitting on chair, lying on bed, and ambulating (where ambulating includes standing, walking around the room).

The dataset has 75128 instances and 9 attributes: time in seconds, acceleration reading in G for frontal axis, acceleration reading in G for vertical axis, acceleration reading in G for lateral axis, Id of the antenna reading sensor, received signal strength indicator, phase, frequency, and label of activity (1: sitting on bed, 2: sit on chair, 3: lying, 4: ambulating). Hence given the values of the other 8 attributes, we need to identify the label of activity. Since there are 4 different possible class labels, we will need to perform a multiclass classification in this dataset.