STAT 602 - Homework 4 - Resubmission

Snigdha Peddi in collaboration with John Herbert

Reusable Functions

- The misclassification function created in homework 3 (*misclass.fun.SP*).

Exercises

Question 1 (ISLR 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,$X \approx N(\mu_k, \sigma_k^2)$. 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 $\sigma_1^2 = ... = \sigma_K^2$.

Answer

Starting from the normal or Gaussian and one dimensional equation wher p=1 (4.11)

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

Plugging this function into the Bayes' theorem (4.10), we get the following:

$$p_k(x) = \frac{\pi_k \frac{1}{\sqrt{2\pi}\sigma_k} exp\left(-\frac{1}{2\sigma_k^2}(x-\mu_k)^2\right)}{\sum_{l=1}^{K} \pi_l \frac{1}{\sqrt{2\pi}\sigma_k} exp\left(-\frac{1}{2\sigma_k^2}(x-\mu_k)^2\right)}$$

Simplifying above equation we get,

$$p_k(x) = \frac{\pi_k exp\left(-\frac{1}{2\sigma_k^2}(x-\mu_k)^2\right)}{\sum_{l=1}^{K} \pi_l exp\left(-\frac{1}{2\sigma_k^2}(x-\mu_k)^2\right)}$$

Applying natural log to above equation,

$$= \ln(\pi_k) + \ln((-\frac{1}{2\sigma_k^2}(x-\mu_k)^2)) - \ln \sum_{l=1}^{K} \pi_l exp\left(-\frac{1}{2\sigma_k^2}(x-\mu_k)^2\right)$$

Simplifying and considering that x belongs to kth class in denominator, we get:

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

Simplyfying and rearranging above equation we get,

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

$$\delta_k(x) = ln(\pi_k) - \frac{1}{2\sigma_k^2}x^2 - \frac{\mu_k^2}{2\sigma_k^2} + \frac{\mu_k x}{\sigma_k^2}$$

From the above equation it is clear that the probability of x beloging to kth class is quadratic with an x^2 term, therefore we can say that the Bayes' classifier is not linear, but quadratic for QDA classifier.

REFERENCES

- Chapter 4,Linear Discriminant Analysis,*An Introduction to Statistical Learning with Applications in R* by Gareth James.

Question 2 (ISLR 4.7.5 pg 169):

We now examine the differences between LDA and QDA.

Question 2(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?

Answer

If the Bayes decision boundary is linear, we expect the LDA to perform better on the training set and test set assuming a normal distribution and similar covariance for all classes. QDA might be more flexible andlead to higher error rate.

Question 2(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?

Answer

If the Bayes Decision boundary is non-linear, I would expect QDA to preform better on both the training and test data sets over LDA. A non-linear decision boundary indicates a different variance among the classes and QDA is more flexible to behave like the Bayes decision boundary and separate the classes properly.

Question 2(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?

Answer

With increase in number of observations I would expect QDA to improve the test prediction as it provides more flexibility and QDA works best with large number of observations.

Question 2(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.

Answer

False. If the Bayes decision boundary is linear, LDA is most likely to produce a superior test error rate as QDA is more flexible and might not classify properly.

REFERENCES

- Chapter 4,Linear Discriminant Analysis,*An Introduction to Statistical Learning with Applications in R* by Gareth James.
- Stat 602, Lecture,*Classification Part2 LDA and QDA* by Dr. Saunders.

Question 3 (4.7.10 pg 171):

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.

Answer

I reran a Logistic Regression Model with Lag2 as predictor variable and Direction as Response variable using the weekly data from Years 1990 to 2008(train data). The weekly data for Years 2009 and 2010 is used as a test data.(repeated from Homework 3-ISLR:4.7.10.d)

```
weekly.glm <- glm(data=weekly.train,Direction~Lag2,family='binomial')
```

Summary of the model shows that Lag2 variable is statistically significant with a p value of 0.04298.Predictions were made on test data using the logistic regression model.The misclassification function is used to calculate the matrices like misclassification rate,sensitivity and specificity and were presented in the table below.This model shows a prediction accuracy of 44.2% (Accuracy=1-misclassification rate).

Table 1: Results of Logistic Regression

	Metric
Misclassification Rate	55.769
Sensitivity	0.082
Specificity	0.953

Question 3(e)

Repeat (d) using LDA.

Answer

Linear Discriminant analysis (LDA) was performed on weekly data using *Direction* as the target and *Lag2* as the predictor.

```
weekly.lda <- lda(Direction~Lag2,data=weekly.train)
```

Table 2: Results of LDA

	Metrics
Misclassification Rate	37.500
Sensitivity	0.918
Specificity	0.209

Predictions were made and and misclassification rate,sensitivity and specificity were reported in above table.

Question 3(f)

Repeat (d) using QDA.

Answer

Quadretic Discriminant analysis (QDA) was performed on weekly data using *Direction* as the target and *Lag2* as the predictor.

```
weekly.qda <- qda(Direction~Lag2,data=weekly.train)
```

Predictions were made and and misclassification rate,sensitivity and specificity were reported in below table.

Table 3: Results of QDA

	Metrics
Misclassification Rate	41.346
Sensitivity	1.000
Specificity	0.000

Question 3(g)

Repeat (d) using KNN with K = 1.

Answer

Same predictor,Lag2 and response variable(Direction) aree used for K Nearest Neighbor analysis.Individual data frames for test and train predictor variable and a factor of training response variable were created and used to predict the test response variable with K=1. Misclassification fucntion is used to report misclassification rate, sensitivity and specificity of the model.

Table 4: Results of KNN

	Metrics
Misclassification Rate	50.000
Sensitivity	0.508
Specificity	0.488

Question 3(h)

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

Answer

All the metrics from different models of Weekly data were presented in the table.

Table 5: Comparitive Model Results for Weekly Data

	GLM	LDA	QDA	KNN
Misclassification Rate	55.769	37.500	41.346	50.000
Sensitivity	0.082	0.918	1.000	0.508
Specificity	0.953	0.209	0.000	0.488

The misclassification rate of the LDA model is less at 37.5% compared to the other models. However, the specificity of the model is low predicting more upward market though it has a good sensitivity. The next best model is QDA with a misclassification rate of 41.3% and only predicted upward movement with a sensitivity of 1 and specificity of 0.KNN model has a 50% misclassification rate with about 50% of sensitivity and specificity. Logistic regression model has a 55.8% error rate and a lower sensitivity. Considering all the model metrics I think LDA is better model with lower misclassification rate and this can be improved along with the sensitivity and specificity by considering more predictors,logarithmic terms,interaction terms etc.

Question 3(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.

Answer

New polynomial variables and interaction terms were added to the dataset and used for the analysis.Logarithmic terms were not used as the all the variables have negative p values.A stepwise regression is used to select the final variables for the analysis. The pent value of 0.4 is used so that the variable below 0.4 p value will enter and prem is set to 0.6 to remove variables with p values more than that value.

```
weekly.glm.I <- glm(Direction~.+Lag1*Lag2*Lag3*Lag4*Lag5,data=weekly.train.I,family='binomial')
```

The predictors the stepwise regression chose with AIC as its qualifier are: * Lag1 * Lag2 * Lag4 * Lag3 Squared * Lag1 Squared * Lag5 Squared * Lag1:Lag3 Interaction * Lag2:Lag3:Lag4 Interaction * Lag2:Lag3:Lag5 Interaction * Lag1:Lag3:Lag4 Interaction

The Logistic Regression model is fit with the predictors selected by stepwise regression method.

```
weekly.glm2 <- glm(Direction~Lag1+Lag2+Lag4+Lag3_2+Lag1_2+Lag5_2
+Lag1:Lag3+Lag2:Lag3:Lag4+Lag2:Lag3:Lag5+Lag1:Lag3:Lag4,data=weekly.train.I)
```

Table 6: Results of Logistic Regression

	Metric
Misclassification Rate	46.154
Sensitivity	0.754
Specificity	0.233

The LDA is fit with the predictors selected by stepwise regression method.

```
weekly.lda2 <- lda(Direction~Lag1+Lag2+Lag4+Lag3_2+Lag1_2+Lag5_2
+Lag1:Lag3+Lag2:Lag3:Lag4+Lag2:Lag3:Lag5+Lag1:Lag3:Lag4,data=weekly.train.I)
```

Table 7: Results of LDA

	Metrics
Misclassification Rate	43.269
Sensitivity	0.869
Specificity	0.140

The QDA is fit with the predictors selected by stepwise regression method.

```
weekly.qda2 <- qda(Direction~Lag1+Lag2+Lag4+Lag3_2+Lag1_2+Lag5_2
+Lag1:Lag3+Lag2:Lag3:Lag4+Lag2:Lag3:Lag5+Lag1:Lag3:Lag4,data=weekly.train.I)
```

Table 8: Results of QDA

	Metrics
Misclassification Rate	49.038
Sensitivity	0.721
Specificity	0.209

Using the original predictor variable Lag2 KNN model is fit with K ranging from 1-100.All the merics were reported.

Table 9: Metrics for KNN Model Where K=40

Metrics	Values
Misclassification	47.208
Sensitivity	0.577
Specificity	0.467

From the above plot shows the all metrics at corresponding K values.I have picked k=40 as it has minimum misclassification rate with a balance sensitivity and specificity.However, regardless of the k value, the misclassification does not seem to vary much and the specificity does not go above 0.5, except at K = 1.

Below table summarize the results from all the models

Table 10: Comparitive Model Results for Weekly Data

	GLM	LDA	QDA	KNN
Misclassification Rate	46.154	43.269	49.038	47.208
Sensitivity	0.754	0.869	0.721	0.577
Specificity	0.233	0.140	0.209	0.467

KNN model was analyzed for previously selected feature,Lag2 with k values 1 to 100 and we found an optimum value of K(40) where the misclassification rate and sensitivity improved and decreased specificity. Comparing the model metric summaries with and without additional variables it is clear that only the misclassification rate of logistic regression model has improved and the error rate of LDA and QDA model

increased. However, for all three model sensitivity and specificity followed same pattern and no model have performed well in terms of sensitivity and specificity. If only misclassification rate is considered the LDA model with one predictor variable (Lag2) is better compared to all other models.

Question 4 (ISLR 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.

Answer

A new response variable mpg01 is created based on the median values of predictor variable *mpg*.A 70:30 split of the data is created and used as training and test set respectively.Predictor variables Cylinders,displacement,weight and horsepower were selected based on the investigation done in 4.7.11.b and will be used for further analysis.

```
##
## Size of Auto Training Data: 274
```

```
## Size of Auto Test Data: 118
```

Question 4(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?

Answer

Using the same predictors and response variable as before a LDA model is fit

```
auto.lda <- lda(mpg01~cylinders+displacement+horsepower+weight,data=train1)
```

Table 11: Results of LDA

	Metrics
TPR	0.969
TNR	0.811
Test Error	0.102

Question 4(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?

Using the same predictors and response variable as before a QDA model is fit.

```
auto.qda <- qda(mpg01~cylinders+displacement+horsepower+weight,data=train1)
```

Table 12: Results of QDA

	Metrics
TPR	0.908
TNR	0.868
Test Error	0.110

Question 4(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?

Answer

Using the same predictors and response variable as before a knn model is fit.The graph shows the misclassification, sensitivity, and specificity for k values from 1 to 100 were investigated and k value of 2 is picked for the analysis because of lower misclassification rate and a reasonable sensitivity and specificity.

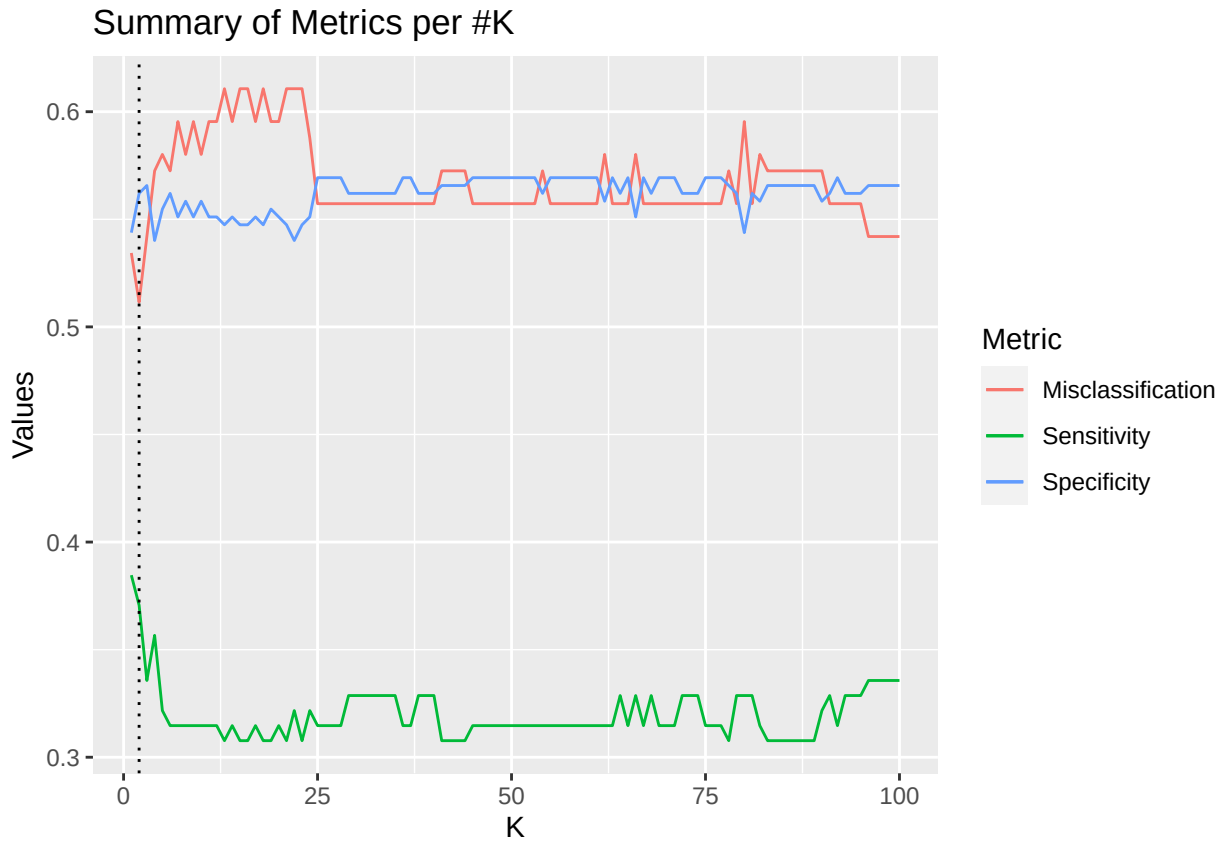

Table 13: Metrics for KNN Model Where K=2

Metrics	knn
TPR	0.371
TNR	0.562
Test Error	0.511

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.

Answer

The below summary was also included submitted as a Word document to show that it is exactly one page.

The Statistical Classification Methods in Consumer Credit Scoring: A review, paper provides an overview on traditional and formal statistical methods used for classifying the applicants into good and bad risk in terms of credit scoring (a process of determining how likely the applicants will default) along with other associated problems in the credit industry. It lists the types of characteristics to be considered, amount of data being used for developing the methods, types of assessments and different types of methods available for analysis. It also talks about reject inference, legal aspects and other potential problems that could benefit from using similar methods.

The papers define terms like credit scoring, application scoring, behavioral or performance scoring, fixed loans, rolling or revolving loans, population drift, reject inference, characteristics and attributes of scoring systems etc., that are most commonly used in the credit industry. The credit score cards (characteristics that determine the approval or rejection of loan) are usually constructed by considering the characteristics of the applicants whose credit is already approved resulting in the use of biased sample and has implications on accuracy. To overcome this problem the information from the reject applicants can be used and this process is called reject inference. After detailed exploration, Hand and Hanley (1993,1994) have concluded that it cannot work unless additional assumptions were made and only provide little information when used as is. However, this can be improved if we include the reject applicants in the accepted pool.

The data used for constructing the score cards are often very large with over 100000 records and more than 100 variables. The proportion of applications that are accepted depends on the financial product at hand, target population, risks the company is willing to take etc. Few of the characteristics considered include Time at present address, Home status, Applicants annual income, credit card, age, type of occupation, purpose of loan, marital status, time with employer, time with bank etc. The data used is often categorical (typically, continuous variables that are grouped) with few to many levels within the categories. This data usually has lot of missing values and strategies like coding a missing value as additional attribute, dropping incomplete vector etc., are used to cope with this problem. Though a large number of characteristics are required for a score card overfitting problems do not occur as the data is usually large. Three approaches are commonly used to select the characteristics for constructing a score card. They include using expert knowledge and a feeling for data and characteristics, stepwise statistical procedures and selecting individual characteristic by using a measure of difference between the distribution of the good and bad risk on that characteristic. However, typically all three methods are used together.

The paper also talks about the two classes of assessment methods. The first one being the separability measures of the good and bad risk score distributions like divergence statistic where the sample t-statistic between two classes is measured and the information statistic where information value is measured. The second method discussed was counting methods which are based on 2 x2 table of predicted-by-true-classes. Here, a threshold is imposed and all the applications falling below are considered bad risk and the above the threshold are considered good risk candidates for credit. Lorentz diagram is a good example of this method.

Different methods discussed for identifying good and bad risks include traditional Judgmental methods and formal statistical scoring methods. Traditional methods were based on subjective human assessment. Nowadays, the scoring methods are often considered over the traditional method as these methods can handle large data transformations, provide accurate classifications, highly consistent, objective, efficient and have superior predictive power. The Statistical scoring methods include the linear regression, discriminant analysis, logistic regression, decision trees, mathematical programming methods, expert systems, Neural networks, time varying models and smoothing nonparametric methods. The paper explains these methods with examples and citations on these methods by different authors. There is no best method but the choice of these methods depends on type of the product at hand and how flexible you would like your model to be.

In conclusion, the review demonstrates that most part of the developing credit scoring methods depends on constructing improved discriminant rules. The advances in technology and scoring methods will help develop more complex and sophisticated models in the future.

Question 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!

Answer

The data set I chose form the UCI Machine Learning Repository is "Breast Cancer" data.The author of this dataset is Matjaz Zweitter and Milan Soklic from Institute if Oncology,University Medical Canter,Ljublijana,Slovenia.The total number of instances in this dataset are 286 with 9 predictor variables and 1 target variable. Out of 286 instances 201 instances are of one class and 85 instances are of other class.

The predictors of the dataset are as follows,

- Class:Nominal, It is the target variable with 2 classes, no-recurrence-events, recurrence- events.
- age:Nominal,The patients are grouped into 9 groups based on their age[10-19,20-29,30-39,40-49,50-59,60-69,70-79,80-89,90-99].
- menopause:Nominal, Patients grouped into lt40, ge40 and premeno groups.
- tumor-size:Nominal, grouped in 12 bins based on the tumor size[0-4, 5-9, 10-14, 15-19, 20-24, 25-29, 30-34, 35-39, 40-44,45-49, 50-54, 55-59].
- inv-nodes:Nominal, grouped into 13 bins based on the nodes [0-2, 3-5, 6-8, 9-11, 12-14, 15-17, 18-20, 21-23, 24-26,7-29, 30-32, 33-35, 36-39].
- node-caps:Nominal, Categorized into yes and no based on presence or absence of the nodes.
- deg-malig:Nominal, Grouped into nominal groups of 1,2 and 3.
- breast:Nominal, left, right
- breast-quad:Nominal, Categorized into left-up, left-low, right-up, right-low, central.
- Irradiat:Nominal, Yes, no.

I would follow the below steps to analyze the brest cancer dataset,

1. Imported the data files:*breast-cancer.data*,*breast-cancer.names*.
2. Data Exploration and transformation: I would do in detail investigation of the data. I would check correlation of the variables,few variables have missing values and i would do missing value imputation either by categorizing the missing values to new class or removing the record. I would use box plots to study the relation of the categories withing the nominal variables. I would transform all the nominal variables by one hot encoding or by giving dummy variables.
3. Variable selection: Correlation plot will give a basic idea of presence of multicollinearity. I would use a stepwise regression analysis to select the variables that i would use for my analysis.

4. Train-Test Split: Since the proportion of no recurrence and recurrence events are not very proportional I would separate the instances based on the class. Will do a 70:30 split or 80:20 split of both the datasets. Then combine the training datasets and test datsets of instances of both classes separately. This would maintain the proportion of the classes in test and train sets rather than randomly selecting the instances from the whole data.
5. Model Selection: I would start with logistic regression model as the dependent variable is of 2 classes.I would use the variables that i have selected previously and try out other models like Linear discriminant analysis,Quadratic discriminant analysis,KNN,Random Forest(would use all the variables).Using the metrics like misclassification rate, sensitivity and specificity i would make the model selection.
6. Testing: Once my model is selected I would use the test data to make my predictions and find the accuracy of the model. If the accuracy rate is not good for the selected model I would try testing with other models or do some hyper parameter tuning or do variable transformation(polynomial variable, interations,log transformations) and do model selection and prediction all over again until a model is picked with good accuracy rate, sensitivity and specificity.