COMPARATIVE STUDY BETWEEN LOGISTIC REGRESSION AND LINEAR DISCRIMINANT ANALYSIS

MAINACK PAUL AND JISHU ADHIKARY

June 23, 2022

1 INTRODUCTION AND MOTIVATION:

Classification is the process in which we study approaches for predicting qualitative responses. In Classification, the response variable is qualitative in nature. Classification problems occur very frequently in various fields. There are many possible techniques or classifiers that one might use to predict qualitative response. In our project, we will discuss two techniques namely Logistic Regression and Linear Discriminant Analysis. Our motivation is to find out the differences between Linear Discriminant Analysis and Logistic Regression; and how to know which technique we have to apply in which case. There are some basic assumptions in both Linear Discriminant Analysis and Logistic Regression which are somewhat different which we have discussed in the later section. So, through this project we have wanted to observe how things are working and how they are related and whether we can blindly use these methods of classification anywhere. We will also provide some conclusions with the help of a comparative study between these two methods but firstly we will discuss what these methods are actually.

2 CLASSIFICATION METHODS: AN OVERVIEW

2.1 LOGISTIC REGRESSION:

The first thing that is of importance here, is that in logistic regression, we model the probability that Y belongs to a particular category rather than modeling the response Y directly. In logistic regression(here we consider the basic univariate binomial case), we use the logistic function,

$$\rho(X) = \frac{\exp\ (\beta_0 + \beta_1 x)}{1+\exp\ (\beta_0 + \beta_1 x)}.$$

After a bit of manipulation, we can write as

$$log(\frac{\rho(X)}{1-\rho(X)}) = \beta_0 + \beta_1 X.$$

The left-hand side is called the log odds or logit. We see that the logistic regression model has a logit that is linear in X. In a linear regression model, β_1 gives the average change in Y associated with a one-unit increase in X. By contrast, in a logistic regression model, increasing X by one unit changes the log odds by β_1. Equivalently, it multiplies the odds by $\exp(\beta_1)$.

Now, we will describe logistic regression in general, that is, for K(>=2) classes and for multiple predictors p(>=2). We estimate the posterior probabilities of classes given X.

We use Bayes rule and we pick the class k (out of K classes) that has the maximum posterior probability:

$$\hat{G}(x) = \arg \underset{k}{max} Pr(G = k|X = x)$$

The decision boundary between classes k and l is determined by the equation:

$$Pr(G = k|X = x) = Pr(G = l|X = x)$$

that is, the x's at which the two posterior probabilities of k and l are equal.

If we divide both sides by $Pr(G = l|X = x)$ and take the log of this ratio, the above equation is equivalent to:

$$\log \frac{Pr(G = k|X = x)}{Pr(G = l|X = x)} = 0$$

Since we want to enforce a linear classification boundary, we assume the function above is linear (below):

$$\log \frac{Pr(G = k|X = x)}{Pr(G = l|X = x)} = a_0^{(k,l)} + \sum_{j=1}^{p} a_j^{(k,l)} x_j$$

This is the basic assumption of logistic regression.

We use the superscript (k, l) on the coefficients of the linear function because, for every pair of k and l, the decision boundary would be different, determined by the different coefficients.

ASSUMPTIONS:

Now, we will clearly specify the assumptions. If we take class K as the base class, the assumed equations are:

$\log \frac{Pr(G=1|X=x)}{Pr(G=K|X=x)} = \beta_{10} + \beta_1^T x$

$\log \frac{Pr(G=2|X=x)}{Pr(G=K|X=x)} = \beta_{20} + \beta_2^T x$

$\vdots$

$\log \frac{Pr(G=K-1|X=x)}{Pr(G=K|X=x)} = \beta_{(K-1)0} + \beta_{K-1}^T x$

We can choose any class as the base class. The coefficient estimates will differ between the two fitted models due to the differing choice of baseline, but the fitted values (predictions), the log odds between any pair of classes, and the other key model outputs will remain the same. Nonetheless, interpretation of the coefficients in a multinomial logistic regression model must be done with care, since it is tied to the choice of baseline.

Once we have specified the parameters for these (K-1) log ratios, then for any pair of classes (k, l), we can derive the log ratios without introducing new parameters:

$$\log \frac{Pr(G=k|X=x)}{Pr(G=l|X=x)} = \beta_{k0} + \beta_{l0} + (\beta_k - \beta_l)^T x$$

Here, the total number of parameters are $(K-1)*(p+1)$.

For convenience, we will denote the entire parameter set by ϑ and arrange them in this way:

$$\theta = \{\beta_{10}, \beta_1^T, \beta_{20}, \beta_2^T, ..., , \beta_{(K-1)0}, \beta_{K-1}^T\}.$$

The log ratios of posterior probabilities are called log-odds or logit transformations.

Under these assumptions, the posterior probabilities are given by the following two equations:

$$Pr(G=k|X=x) = \frac{\exp\ (\beta_{k0} + \beta_k^T x)}{1 + \sum_{l=1}^{K-1} \exp\ (\beta_{l0} + \beta_l^T x)} \text{ for } k = 1, ..., K-1.$$

$$Pr(G=K|X=x) = \frac{1}{1 + \sum_{l=1}^{K-1} \exp\ (\beta_{l0} + \beta_l^T x)}$$

For $Pr(G=k|X=x)$ given above, obviously

$$\sum_{k=1}^{K} Pr(G=k|X=x) = 1$$

These must sum up to 1: A simple calculation shows that the assumptions are satisfied.

FITTING LOGISTIC REGRESSIONS:

Here, we will find the parameters that maximize the conditional likelihood of class labels G given X. We are not interested in the distribution of X, instead, our focus is on the conditional probabilities of the class labels given X.

Given point x_i , the posterior probability for the class to be k is denoted by:

$$p_k(x_i;\theta) = Pr(G=k|X=x_i;\theta)$$

Given the first input x_1, the posterior probability of its class, denoted as g_1, is computed by:

$$Pr(G = g_1|X = x_1).$$

Since samples in the data set are assumed independent, the posterior probability for the N sample points each having class $g_i, i = 1, 2, \cdots, N$, given their inputs $x_1, x_2, \cdots, x_N$ is:

$$\prod_{i=1}^{N} Pr(G = g_i|X = x_i)$$

In other words, the joint conditional likelihood is the product of the conditional probabilities of the classes given every data point.

The conditional log-likelihood of the class labels in the data set becomes a summation:

$$l(\theta) = \sum_{i=1}^{N} logPr(G = g_i|X = x_i)$$

$$= \sum_{i=1}^{N} logp_{g_i}(x_i; \theta)$$

BINARY CLASSIFICATION:

Logistic regression models are usually fit by maximum likelihood, using the conditional likelihood of G given X. Since $Pr(G|X)$ completely specifies the conditional distribution, the multinomial distribution is appropriate. The log-likelihood for N observations is

$$l(\beta) = \sum_{i=1}^{N} logp_{g_i}(x_i; \beta)$$

where $p_k(xi; \theta) = Pr(G = k|X = xi; \theta)$.

We discuss in detail the two-class case, since the algorithms simplify considerably. It is convenient to code the two-class g_i via a 0/1 response y_i , where $y_i = 1$ when $g_i = 1$, and $y_i = 0$ when $g_i = 2$. Let $p_1(x; \theta) = p(x; \theta)$, and $p_2(x; \theta) = 1 - p(x; \theta)$. The log-likelihood can be written

$$l(\beta) = \sum_{i=1}^{N} logp_{g_i}(x_i; \beta)$$

$$= \sum_{i=1}^{N} (y_i \log\, p(x_i; \beta) + (1 - y_i) \log\, (1 - p(x_i; \beta)))$$

Here$\beta = \beta_{10}, \beta_1$, and we assume that the vector of inputs x_i includes the constant term 1 to accommodate the intercept.

To maximize the log-likelihood, we set its derivatives to zero. These score equations are

$$\frac{\partial l(\beta)}{\partial \beta_j} = \sum_{i=1}^{N} x_i(y_i - p(x_i; \beta)) = 0$$

which are $(p + 1)$ equations nonlinear in β. To solve these (p+1) score equations, we generally use the Newton–Raphson algorithm.

REMARK:

Logistic regression involves directly modeling Pr(Y = k|X = x) using the logistic function. We model the conditional distribution of the response Y , given the predictor(s) X. We now consider an alternative and less direct approach to estimating these probabilities. In this new approach, we model the distribution of the predictors X separately in each of the response classes (i.e. for each value of Y). We then use Bayes' theorem to flip these around into estimates for Pr(Y = k|X = x). But still why do we need another method, when we have logistic regression? There are several reasons:

i. If the distribution of the predictors X is approximately normal in each of the classes and the sample size is small, then the approaches in this section may be more accurate than logistic regression.

ii. The methods in this section can be naturally extended to the case of more than two response classes.

2.2 LINEAR DISCRIMINANT ANALYSIS (LDA):

Here, in this project we have done LDA using Fisher's argument. The idea was to transform the multivariate observations x to univariate observations y such that the y's derived from population π_1 and $\pi_2$were separated as much as possible. Fisher suggested taking linear combinations of x to create y's because they are simple enough functions of the x to be handled easily. Fisher's approach does not assume that the populations are normal. The only assumption is that the population covariance matrices are equal.

A fixed linear combination of the x's takes the values y_{11} ,... , $y_{1,n1}$ for the observations from the first population and the values y_{21}, ... , $y_{2,n2}$ for the observations from the second population. The separation of these two sets of univariate y's is assessed in terms of the difference between $\overline{y_1}$ and $\overline{y_2}$ expressed in standard deviation units.

An Allocation Rule Based on Fisher's Discriminant Function:

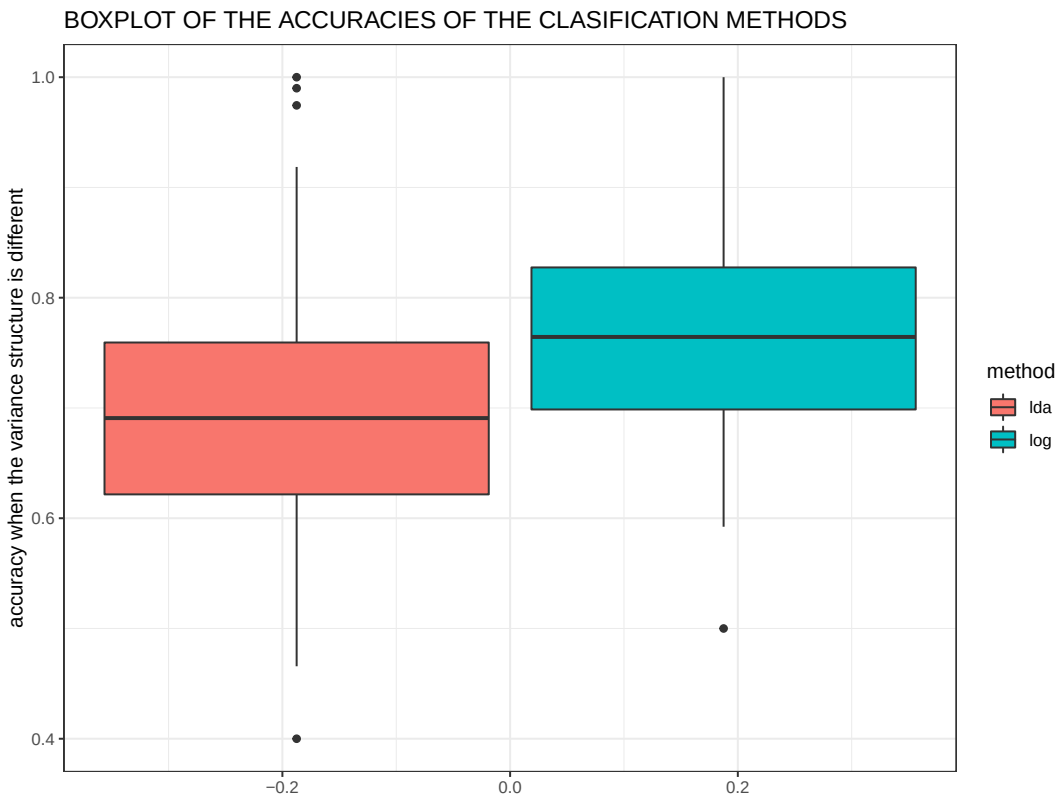
Allocate x_0 to π_1 if

$$\hat{y_0} = (\overline{x_1} - \overline{x_2})' S_*^{-1} x_0 \geq \hat{rn} = \frac{1}{2}(\overline{x_1} - \overline{x_2})' S_*^{-1} (\overline{x_1} + \overline{x_2})$$

$$where \hat{a}' S_* \hat{a} = \frac{\sum_{j=1}^{n_1} (y_{1j} - \overline{y_1})^2 + \sum_{j=1}^{n_2} (y_{1j} - \overline{y_2})^2}{n_{1+} n_2 - 2} \; with y_{1j} = \hat{a}' x_{1j} \; and \; y_{2j} = \hat{a}' x_{1j}$$

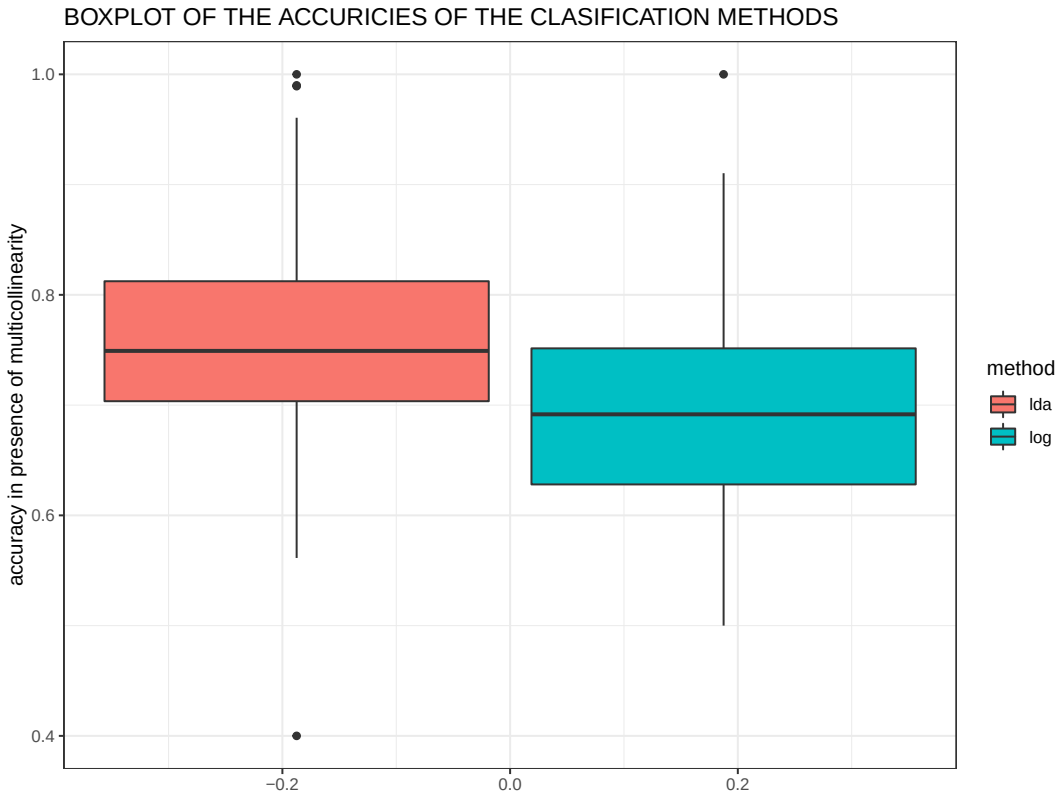
3 COMPARISON OF THE METHODS:

We now compare the empirical (practical) performance of logistic regression and LDA . We generated data from five different scenarios, each of which involves a binary (two-class) classification problem.

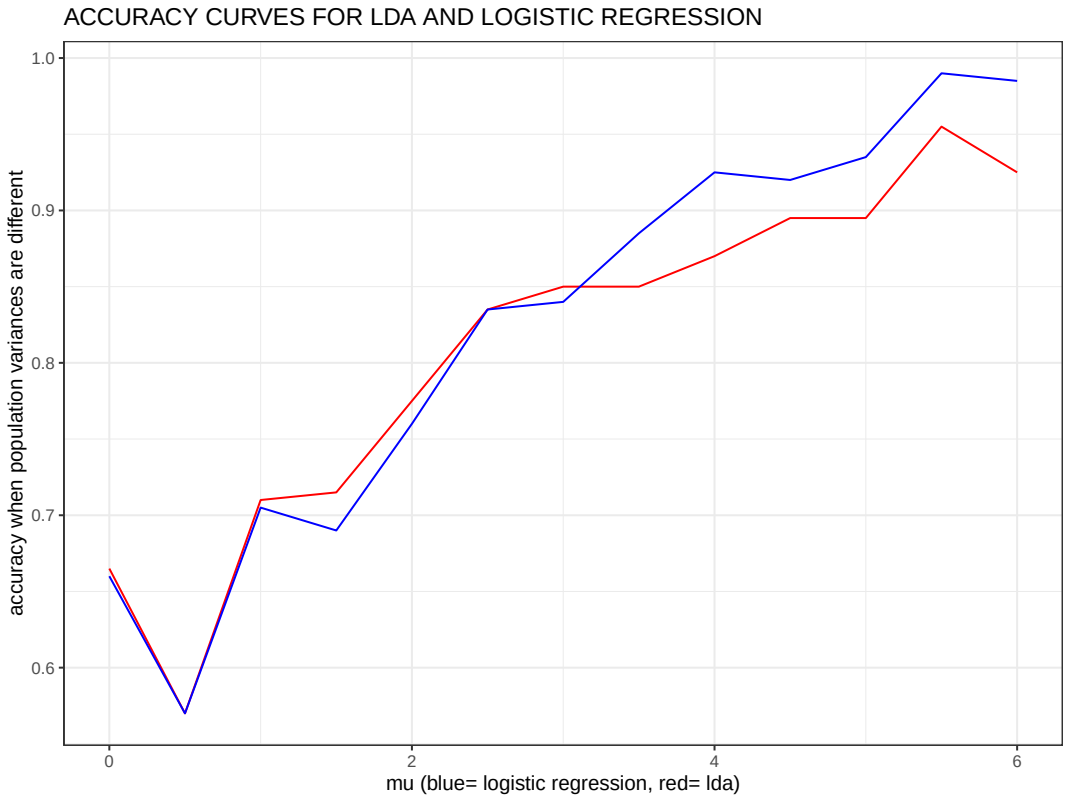


Boxplot of test set accuracies from the two methods on the simulated data.Here population 1 (π_1) $\equiv N(0, 0; 5, 5, \rho = 0)$ and population 2 (π_2) $\equiv N(6, 6; 2, 2, \rho =$

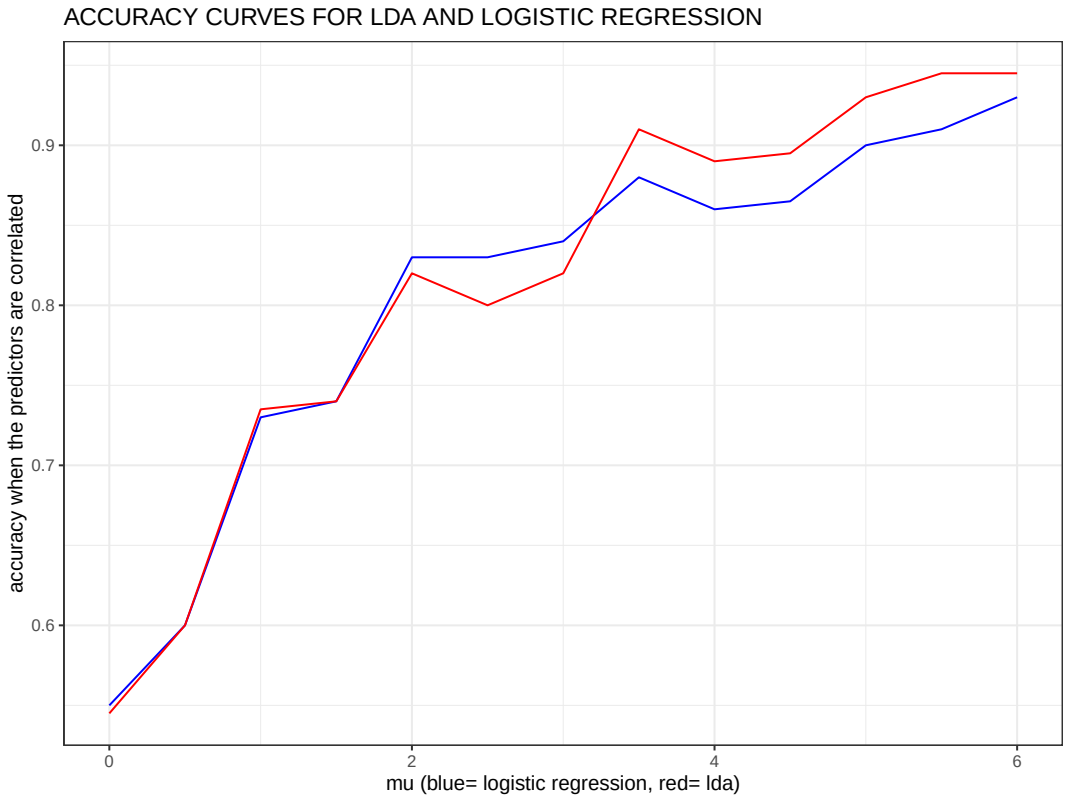
0) .In this setup the coveriance structure of the populations are different so the assumptions of LDA are not satisfied .So logistic regression works better.



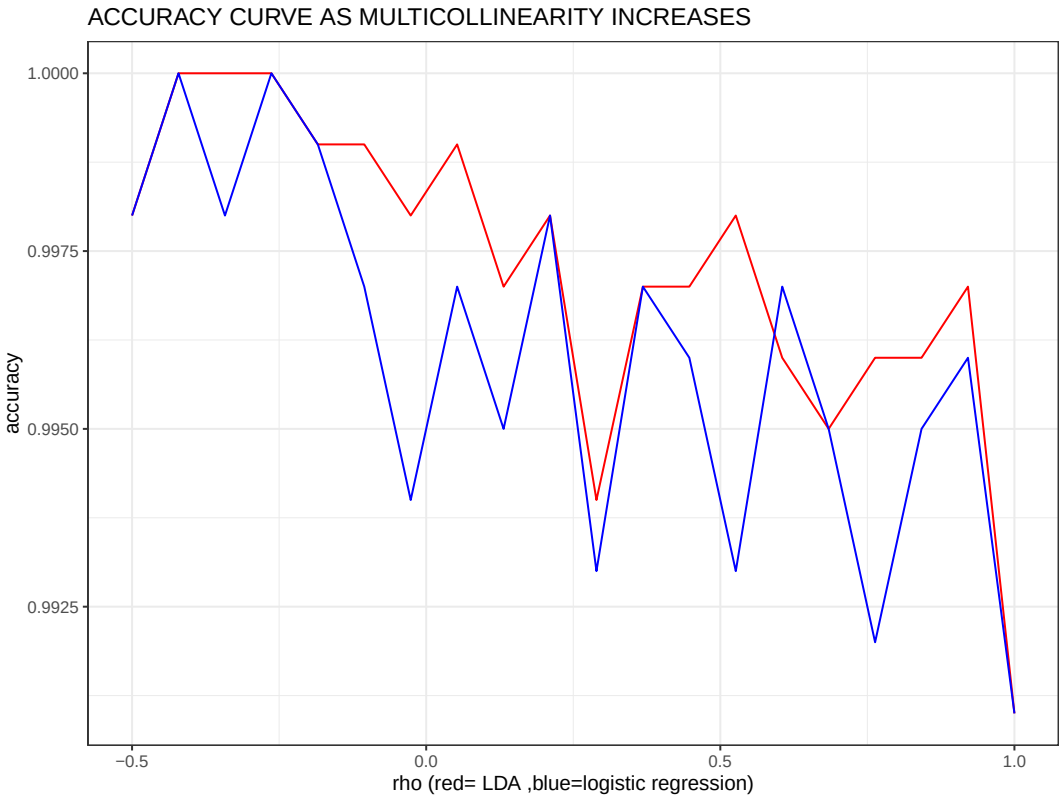
Boxplot of the test set accuracy of LDA and logistic in the previous set up.Only difference is here the correlation between the variables $\rho = -0.5$ .In this set up LDA performs better.



Accuracy curve of LDA (red) and logistic regression (blue) as μincreases.Here $(\pi_1) \equiv N(0, 0; 1, 1, \rho = 0)$ and $(\pi_2) \equiv N(\mu, \mu; 5, 5, \rho = 0)$.



Accuracy curve of LDA (red) and logistic regression (blue) as μincreases.Here $(\pi_1) \equiv N(0, 0; 1, 1, \rho = 0)$ and $(\pi_2) \equiv N(\mu, \mu; 1, 1, \rho = 0.8)$.



Accuracy curve of LDA (red) and logistic regression (blue) as ρ decreases.Here $(\pi_1) \equiv N(0, 0; 5, 5, \rho)$ and $(\pi_2) \equiv N(6, 6; 5, 5, \rho)$.

4 CASE STUDY: STOCK MARKET DATASET & ANACONDA DATASET

The main motive of studying these two cases together is that, here, we want to show that Logistic Regression performs very poorly in presence of Multicollinearity while Linear Discriminant Analysis method works quite well in presence of Multicollinearity.

STOCK MARKET DATASET:

This data set consists of percentage returns for the S&P 500 stock index over 1,250 days, from the beginning of 2001 until the end of 2005. For each date, we have recorded the percentage returns for each of the five previous trading days, Lag1 through Lag5. We have also recorded Volume (the number of shares traded on the previous day, in billions), Today (the percentage return on the date in

question) and Direction (whether the market was Up or Down on this date). Our goal is to predict Direction (a qualitative response) using the covariates.

After computing, it is seen that the correlations between the lag variables and today's returns are close to zero . In other words, there appears to be little correlation between today's returns and previous days' returns and hence multicollinearity is more or less absent in this dataset.

Next, we have fitted a logistic regression model in order to predict Direction using Lag1 through Lag5, Today and Volume.

Logistic Regression:

In order to better assess the accuracy of the logistic regression model in this setting, we can fit the model using part of the data, and then examine how well it predicts the held out data. This will yield a more realistic error rate, in the sense that in practice we will be interested in our model's performance not on the data that we used to fit the model, but rather on days in the future for which the market's movements are unknown. To implement this strategy, we will first create a vector corresponding to the observations from 2001 through 2004. We will then use this vector to create a held out data set of observations from 2005. We first fit a logistic regression model using only the subset of the observations that correspond to dates before 2005. We then obtain predicted probabilities of the stock market going up for each of the days in our test set, that is, for the days in 2005. Finally, we compute the predictions for 2005 and compare them to the actual movements of the market over that time period. We find a confusion matrix in order to determine how many observations were correctly or incorrectly classified.

```
> table(Predicted,Actual)
         Actual
Predicted Down  Up
     Down  110   0
     Up      1 141
```

From the output above, we can see that the test error rate for logistic regression came out to be 0.4% .

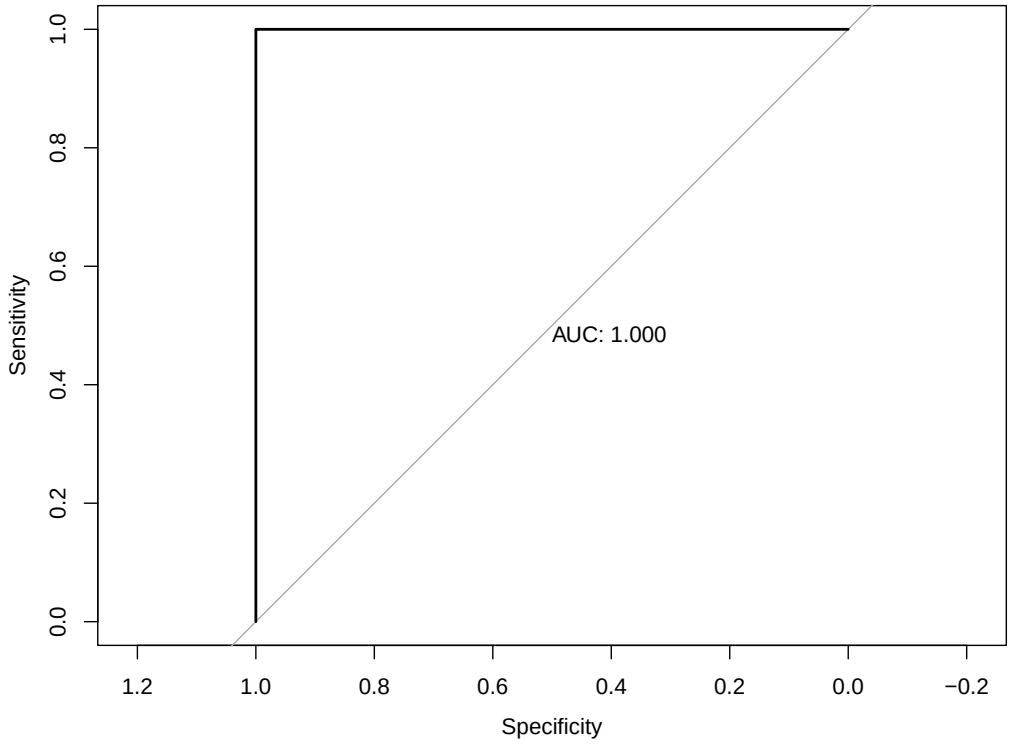
Linear Discriminant Analysis:

Similarly, here also, in order to better assess the accuracy of the linear discriminant analysis in this setting, we first fit using only the subset of the observations that correspond to dates before 2005 and then obtain predicted probabilities of the stock market for the days in 2005.

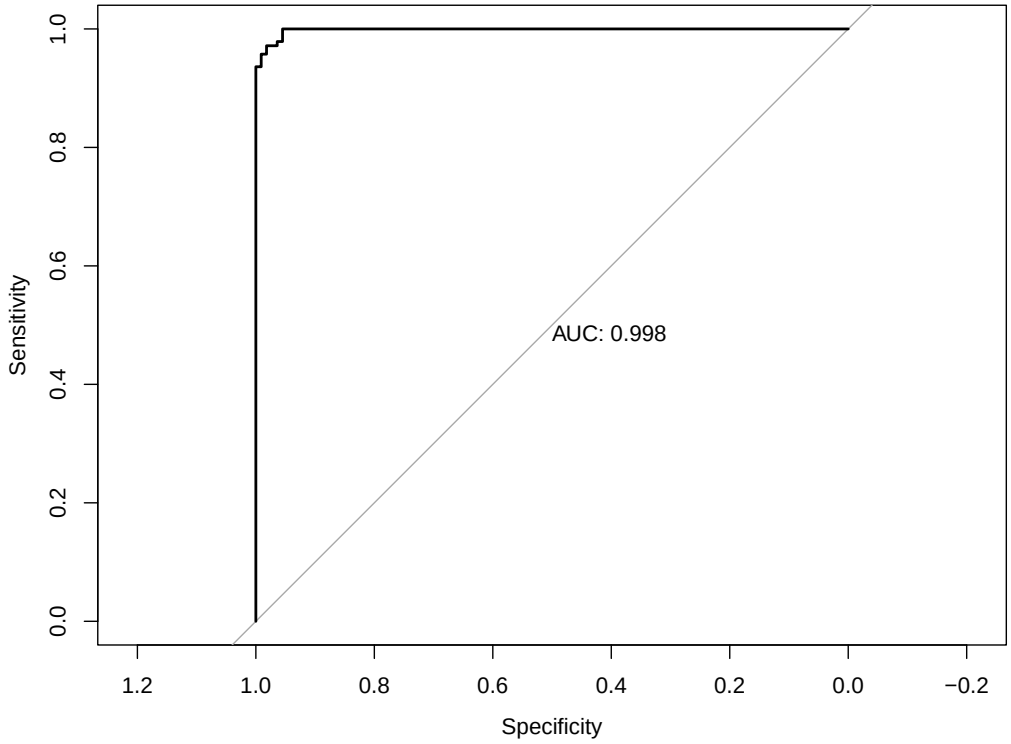
```
> table(Predicted, Actual)
         Actual
Predicted Down  Up
     Down  110   7
     Up      1 134
```

From the output above, we can see that the test error rate for came out to be 3.1% .

Also, we have plotted the ROC(receiver operating characteristic) curve which is produced by calculating and plotting the true positive rate(Sensitivity) against the false positive rate(1-Specificity).Then, we have calculated the AUC(area under ROC Curve). The higher the AUC score, the better a classifier performs for the given task.



Above figure shows the ROC curve for logistic regression.



Above figure shows the ROC curve for LDA.

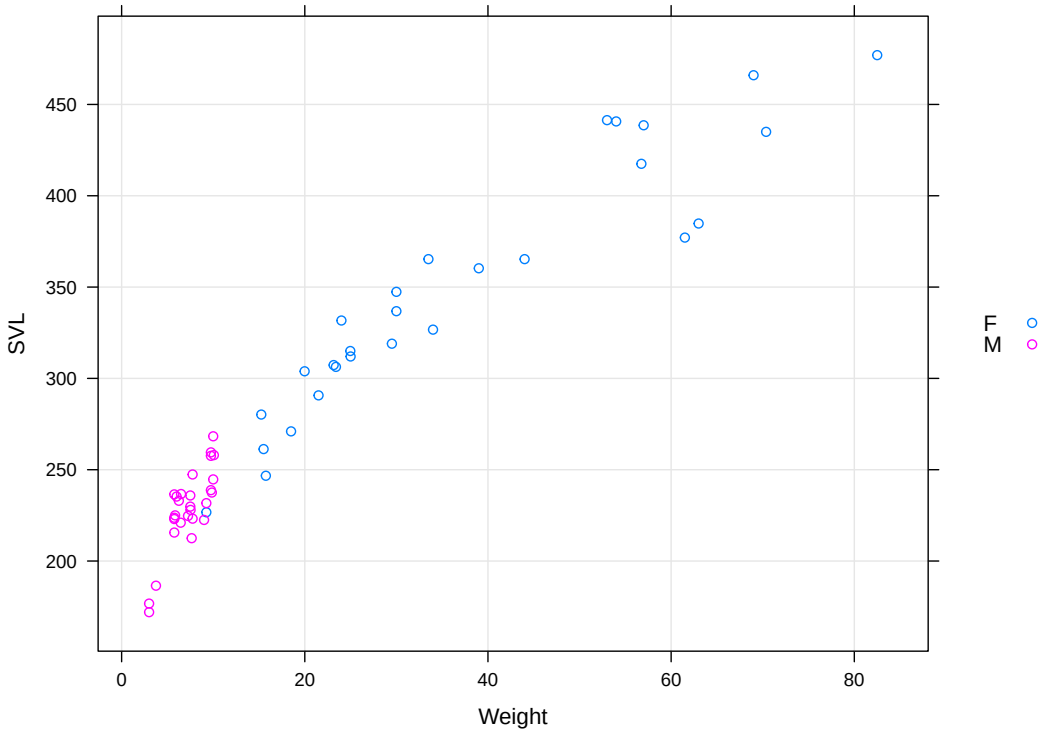
Hence, we can say that in this case, Logistic Regression outshines Linear Discriminant Analysis as it has larger AUC value.

ANACONDA DATASET:

Anaconda Data has 3 columns,one is for snount-vent length(SVL), one is for weight and last one is gender. Here, our goal is to predict gender of anacondas based on their SVL and weight.Here gender is categorial but other two covariates are continuous.

Linear Discriminant Analysis:

First we have done a basic xyplot of SVL vs weight for visualisation of the data.

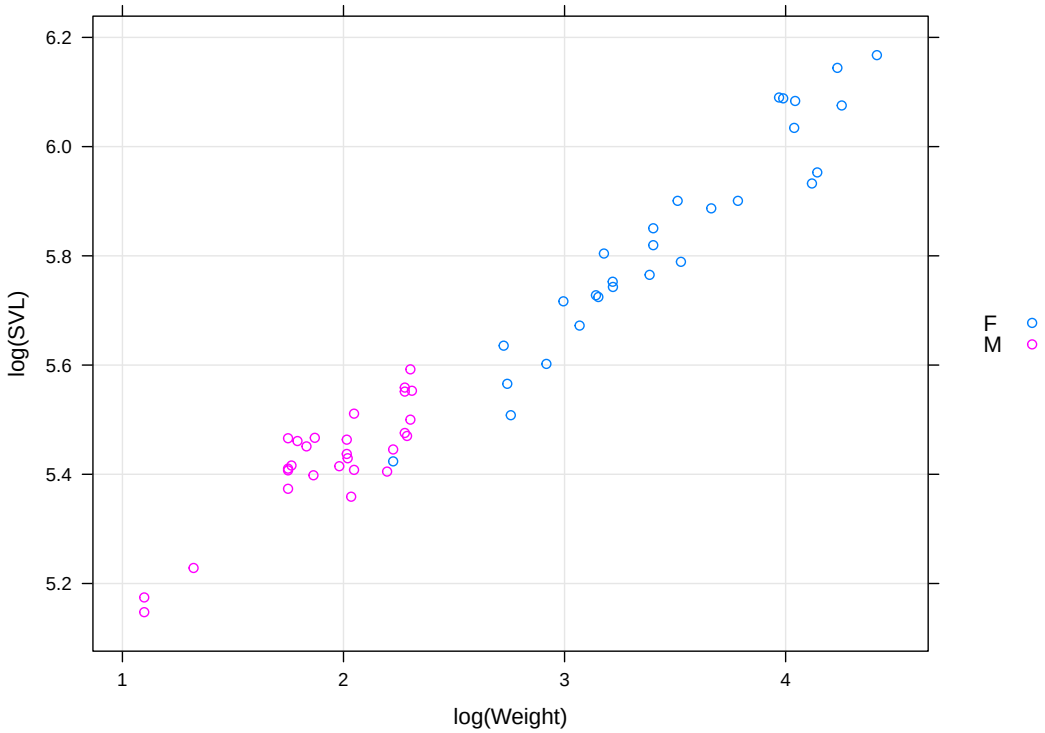


Then, we perform LDA.

```
> table(Actual, Predicted)
      Predicted
Actual  F  M
     F 23  5
     M  0 28
```

From the output above, we can see that misclassification members are somewhat large in numbers. It is due to the fact that in LDA, we assume the variance structure are same for two populations but here if we see the above xyplot, its quite clear that their dispersions are no way similar.

Thus, we perform a log transformation to make the variance structure same of the two populations and again we have done a basic xyplot of log(SVL) vs log(weight).



Then, again, we perform LDA.

```
> table(Actual, Predicted)
      Predicted
Actual  F  M
     F 27  1
     M  0 28
```

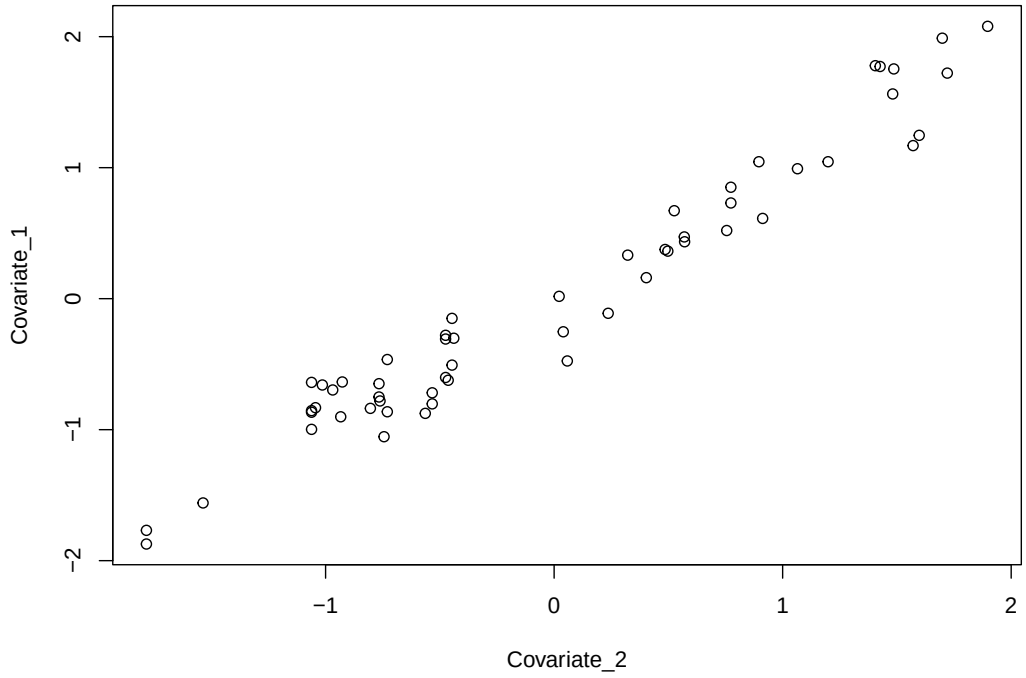
It's really a great improvement over misclassification error as we can see from the above table above that only 1 misclassification is there. The reason for this is that after log transfomation, variance structure of the two populations looks similar upto some extent which we can observe in the above plot.

Logistic Regression:

We now fit a logistic regression but in output, we are getting warning messages which are not something we can ignore, and specially it says algorithm did not converge.

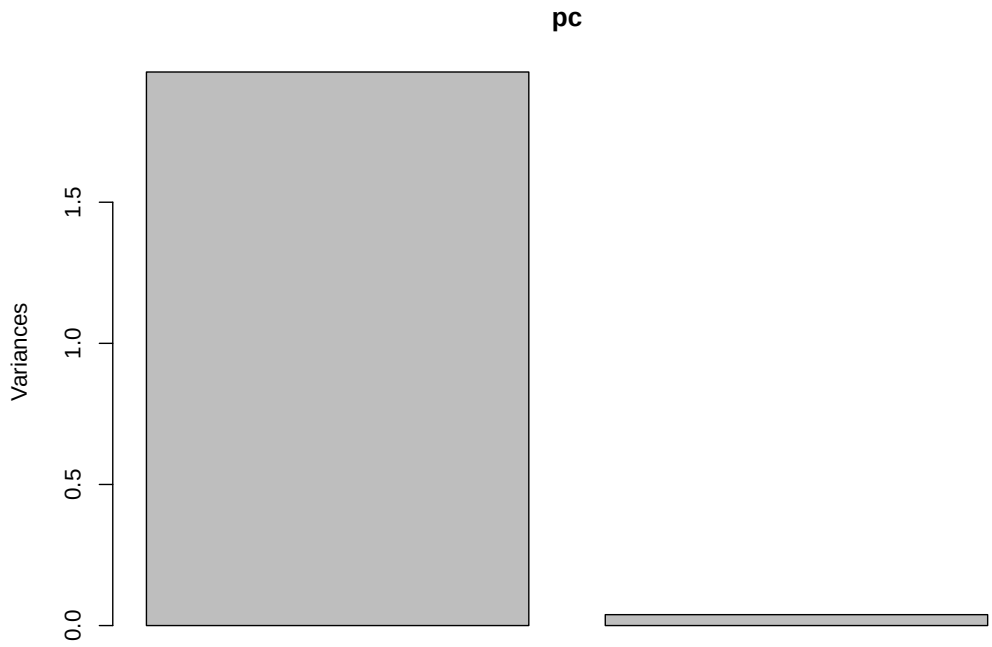
```
Warning messages:
1: glm.fit: algorithm did not converge
2: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

Now, if we see observe the dataset carefully, we can see that the two covariates are highly correlated which can be easily by a simple plot.



Hence, we can suspect that multicollinearity can be a possible reason.

We apply Principle Component Analysis (PCA) to get rid of Multicollinearity and thus, we get the principal components.



The above plot is known as Scree Plot. From it, we can see that the maximum variability is captured by the first principal component only and thus we will be keeping it only in our further analysis.

We again then fit a logistic regresssion and calculated the confusion matrix.

```
> table(Actual,predicted)
      predicted
Actual  F  M
     F 26  2
     M  1 27
```

From the above output, we can say that the error rate is 5.3%.

Hence, we see that, after we have removed multicollinearity, logistic regression works fine.

Hence, we can conclude that for this case, in presence of multicollinearity, LDA works better than logistic regression.

5 ANOTHER CASE STUDY: IRIS DATASET

The main motive of studying these case is that, here, we want to show that Logistic Regression performs very poorly in Multiclass(>2) Classification problems compared to Linear Discriminant Analysis method.

IRIS DATASET

Iris data set gives the measurements in centimeters of the variables sepal length and width; and petal length and width, respectively. The species are Iris setosa, versicolor, and virginica.

Linear Discriminant Analysis:

We fit the LDA and get the following output.

```
> table(Actual, Predicted)
            Predicted
Actual       setosa versicolor virginica
  setosa         50          0         0
  versicolor      0         48         2
  virginica       0          1        49
```

From the output above, we can see that the error rate is 2% and it works more or less well for multiclass classification problems.

Logistic Regression:

We fit the Logistic Regression and get the following output.

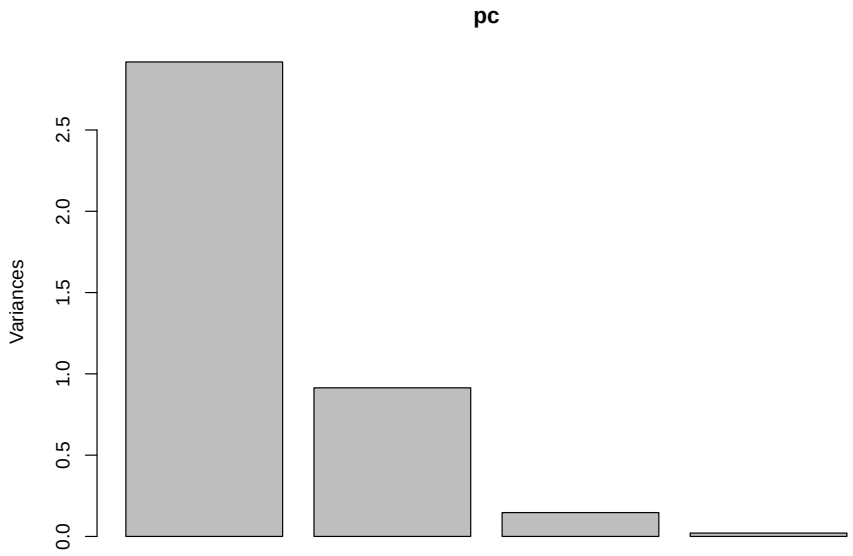
```
> fit=multinom(Species~Sepal.Length+Sepal.Width+Petal.Length+Petal.Width,data=iris)
# weights:  18 (10 variable)
initial  value 164.791843
iter  10 value 16.177348
iter  20 value 7.111438
iter  30 value 6.182999
iter  40 value 5.984028
iter  50 value 5.961278
iter  60 value 5.954900
iter  70 value 5.951851
iter  80 value 5.950343
iter  90 value 5.949904
iter 100 value 5.949867
final  value 5.949867
stopped after 100 iterations
```

Hence, from the output above, we can see that the numerical method did not converge and we donot get a proper estimate of the coefficients and thus Logistic Regression fails.

But here also, the problem of multicollinearity is present. So, we should find out the reason that why logistic regression is not working properly. Is it due to the multicollinearity problem or more due to the multiclass classification? Hence, we perform PCA to remove the multicollinearity and then we will be

again performing the logistic regression in absence of multicollinearity and then, finally we will be able to decide what is actually causing the problem.

```
> cor(iris[,-5])
             Sepal.Length Sepal.Width Petal.Length Petal.Width
Sepal.Length    1.0000000  -0.1175698    0.8717538   0.8179411
Sepal.Width    -0.1175698   1.0000000   -0.4284401  -0.3661259
Petal.Length    0.8717538  -0.4284401    1.0000000   0.9628654
Petal.Width     0.8179411  -0.3661259    0.9628654   1.0000000
```



The above plot is known as Scree Plot. From it, we can see that the maximum variability is captured by the first two principal components and thus we will be keeping these two only in our further analysis. Hence, the output obtained after this is as follows:

```
> f=multinom(Species~pc1+pc2,data=iris)
# weights:  12 (6 variable)
initial  value 164.791843
iter  10 value 26.854470
iter  20 value 18.719880
iter  30 value 18.603163
iter  40 value 18.600719
iter  50 value 18.599063
iter  60 value 18.597375
iter  70 value 18.596743
iter  80 value 18.596221
iter  90 value 18.595639
iter 100 value 18.595038
final  value 18.595038
stopped after 100 iterations
```

We see that still the numerical method did not converge and we donot get a proper estimate of the coefficients and thus Logistic Regression fails.

Hence, multicollinearity is not the only problem in this case and due to the presence of multiclass, Logistic regression is not performing well..

ACKNOLEDGEMENT

This is to acknowledge all those without whom this project would not have been reality. Firstly, I would wish to thank our professor Dr. Deepayan Sarkar who gave his support, dedicated his time towards it.

BIBLIOGRAPHY:

The contents of this project is mainly our work.We have been guided by our professor greatly.Here are some books we got help from:

1.An Introduction To Statistical Learning: With Applications In R

2.Applied Multivariate analysis ,Johnson and Wichern

THANK YOU