Statistical Learning for Data Science Assignment

Elsun Nabatov

2023-11-30

Introduction

In this analysis, we delve into the realm of predictive modeling for the Breast Cancer dataset, exploring various statistical techniques to discern the most effective classifier. Our journey encompasses an array of methods, including Best Subset Cross Validation, Lasso Regression, Linear Discriminant Analysis (LDA), and Quadratic Discriminant Analysis (QDA), each bringing unique perspectives to the table. The goal is to not only identify the classifier that offers the highest accuracy but also to understand the underlying complexities and interactions in the dataset, ensuring a robust and informed choice for cancer prediction.

Part 1

Data Cleaning and Pre-processing

We have Breast Cancer dataset with 699 observations. With this raw data, we will do pre-processing and removing NA values, also converting data types.

Converting Factors to Quantitative Variables

In "class" column we have "malignant" and "benign" which are converted to 1 and 0 respectively and changed data format.

Remove NA values

I removed 16 NA values and "Id" column from Breast Cancer dataset.

After all process, we are checking dataset structure.

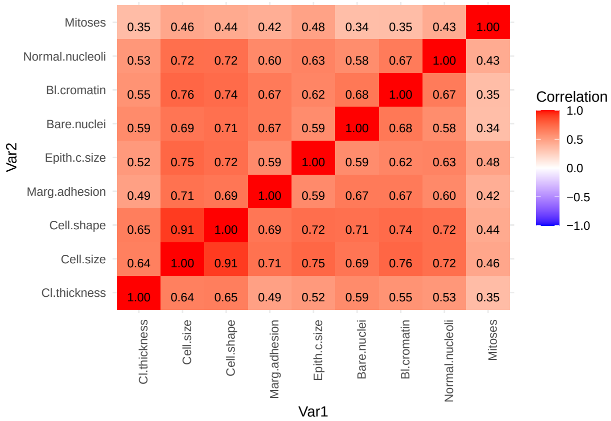
```
## 'data.frame':    683 obs. of  10 variables:
##  $ Cl.thickness   : num  5 5 3 6 4 8 1 2 2 4 ...
##  $ Cell.size      : num  1 4 1 8 1 10 1 1 1 2 ...
##  $ Cell.shape     : num  1 4 1 8 1 10 1 2 1 1 ...
##  $ Marg.adhesion  : num  1 5 1 1 3 8 1 1 1 1 ...
##  $ Epith.c.size   : num  2 7 2 3 2 7 2 2 2 2 ...
##  $ Bare.nuclei    : num  1 10 2 4 1 10 10 1 1 1 ...
##  $ Bl.cromatin    : num  3 3 3 3 3 9 3 3 1 2 ...
##  $ Normal.nucleoli: num  1 2 1 7 1 7 1 1 1 1 ...
##  $ Mitoses        : num  1 1 1 1 1 1 1 1 5 1 ...
##  $ Class          : num  0 0 0 0 0 1 0 0 0 0 ...
##  - attr(*, "na.action")= 'omit' Named int [1:16] 24 41 140 146 159 165 236 250 276 293 ...
##   ..- attr(*, "names")= chr [1:16] "24" "41" "140" "146" ...
```

Part 2

Explotary Data Analysis

Correlation Matrix

The heatmap (with correlation matrix) indicates a strong positive correlation between predictor variables in the Breast Cancer dataset, with ‘Cell.size’ and ‘Cell.shape’ demonstrating a particularly high correlation of 0.91, suggesting multicollinearity. Lesser but still significant positive correlations are observed, such as ‘Bl.Cromatin’ with ‘Cell.size’ at 0.76, and ‘Cell.size’ with ‘Epith.c.size’ at 0.75, hinting at possible shared biological attributes or measurement interactions, with no evident strong negative correlations across the variables.



PART 3

By standardizing the first nine columns, which are the predictor variables, then extracting the tenth column as the response variable. This standardized data and the response are then combined into a new dataframe called ‘Breast_data’.

```
## ’data.frame’:    683 obs. of  10 variables:
##  $ Cl.thickness   : num  0.198 0.198 -0.511 0.552 -0.157 ...
##  $ Cell.size      : num  -0.702 0.277 -0.702 1.582 -0.702 ...
##  $ Cell.shape     : num  -0.741 0.263 -0.741 1.601 -0.741 ...
##  $ Marg.adhesion  : num  -0.6389 0.7575 -0.6389 -0.6389 0.0593 ...
##  $ Epith.c.size   : num  -0.555 1.694 -0.555 -0.105 -0.555 ...
##  $ Bare.nuclei    : num  -0.698 1.772 -0.424 0.125 -0.698 ...
##  $ Bl.cromatin    : num  -0.182 -0.182 -0.182 -0.182 -0.182 ...
##  $ Normal.nucleoli: num  -0.612 -0.285 -0.612 1.353 -0.612 ...
##  $ Mitoses        : num  -0.348 -0.348 -0.348 -0.348 -0.348 ...
##  $ y              : num  0 0 0 0 0 1 0 0 0 0 ...
```

I have crafted a logistic regression model to analyze the Breast Cancer dataset. This model, named logreg_fit, has been meticulously fitted using a generalized linear modeling approach, encapsulating all available predictor variables within the dataset.

Summary shows “Cl.thickness”, “Marg.adhesion”, “Bare.nuclei”, and “Bl.cromatin” as predictors with a statistically significant impact on the response variable, underlining their potential importance in the underlying phenomenon we are studying. On the other hand, predictors like Cell.size and Epith.c.size do not show a significant association. The model fits the data substantially better than a null model, as indicated by the lower residual deviance compared to the null deviance, and the AIC suggests a relatively good model fit.

```
##
## Call:
## glm(formula = y ~ ., family = "binomial", data = Breast_data)
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -1.09414    0.32333  -3.384 0.000714 ***
## Cl.thickness     1.50915    0.40060   3.767 0.000165 ***
## Cell.size       -0.01925    0.64085  -0.030 0.976039
## Cell.shape       0.96443    0.68917   1.399 0.161688
## Marg.adhesion    0.94713    0.35363   2.678 0.007400 **
## Epith.c.size     0.21483    0.34812   0.617 0.537159
## Bare.nuclei      1.39569    0.34195   4.082 4.47e-05 ***
## Bl.cromatin      1.09547    0.41983   2.609 0.009073 **
## Normal.nucleoli  0.65031    0.34457   1.887 0.059115 .
## Mitoses          0.92670    0.56966   1.627 0.103788
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 884.35  on 682  degrees of freedom
## Residual deviance: 102.89  on 673  degrees of freedom
## AIC: 122.89
##
## Number of Fisher Scoring iterations: 8
```

Utilizing the bestglm package in R, I’ve applied best subset selection to the Breast Cancer dataset to identify the most suitable model based on two criteria: the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC). This procedure evaluates all possible combinations of predictor variables to determine the model that best balances model fit with complexity.

The best subset selection analysis for the Breast Cancer dataset reveals a distinct preference in model complexity: AIC favors a model with 7 predictors, indicating a tilt towards a more detailed representation, while BIC opts for a simpler model with just 5 predictors, highlighting its inclination towards parsimony and reduced risk of overfitting.

```
## [1] 7
```

```
## [1] 5
```

Cross Validation

A 10-fold cross-validation method was implemented on the Breast Cancer dataset to evaluate model performance, ensuring reproducibility with a set random seed. A custom function, reg_cv, was used to compute the mean squared error across each fold, providing a measure of prediction accuracy.

```
## [1] 0.03765712
```

The regsubsets function from the leaps package then exhaustively searched for the best subset of predictors, considering all possible combinations up to the maximum number of predictors.

```
## Subset selection object
## Call: regsubsets.formula(y ~ ., data = Breast_data, method = "exhaustive",
##     nvmax = p)
## 9 Variables  (and intercept)
##                 Forced in Forced out
## Cl.thickness        FALSE      FALSE
## Cell.size           FALSE      FALSE
## Cell.shape          FALSE      FALSE
## Marg.adhesion       FALSE      FALSE
## Epith.c.size        FALSE      FALSE
## Bare.nuclei         FALSE      FALSE
## Bl.cromatin         FALSE      FALSE
## Normal.nucleoli     FALSE      FALSE
## Mitoses             FALSE      FALSE
## 1 subsets of each size up to 9
## Selection Algorithm: exhaustive
##          Cl.thickness Cell.size Cell.shape Marg.adhesion Epith.c.size
## 1  ( 1 ) " "          " "       " "        " "           " "
## 2  ( 1 ) " "          "*"       " "        " "           " "
## 3  ( 1 ) "*"          "*"       " "        " "           " "
## 4  ( 1 ) "*"          "*"       " "        " "           " "
## 5  ( 1 ) "*"          "*"       " "        " "           " "
## 6  ( 1 ) "*"          "*"       "*"        " "           " "
## 7  ( 1 ) "*"          "*"       "*"        "*"           " "
## 8  ( 1 ) "*"          "*"       "*"        "*"           "*"
## 9  ( 1 ) "*"          "*"       "*"        "*"           "*"
##          Bare.nuclei Bl.cromatin Normal.nucleoli Mitoses
## 1  ( 1 ) "*"         " "         " "             " "
## 2  ( 1 ) "*"         " "         " "             " "
## 3  ( 1 ) "*"         " "         " "             " "
## 4  ( 1 ) "*"         " "         "*"             " "
## 5  ( 1 ) "*"         "*"         "*"             " "
## 6  ( 1 ) "*"         "*"         "*"             " "
## 7  ( 1 ) "*"         "*"         "*"             " "
## 8  ( 1 ) "*"         "*"         "*"             " "
## 9  ( 1 ) "*"         "*"         "*"             "*"
```

I tried to identify the most influential predictor in a model containing only one variable (the "1-predictor model"). This identified predictor was then used to create a reduced dataset, Breast_data_red, containing only the selected variable along with the response variable. Finally, a logistic regression model (logreg1_fit) was fitted to this reduced dataset.

```
##    Intercept Cl.thickness Cell.size Cell.shape Marg.adhesion Epith.c.size
## 1       TRUE        FALSE      TRUE      FALSE         FALSE        FALSE
##    Bare.nuclei Bl.cromatin Normal.nucleoli Mitoses logLikelihood      AIC
## 1        FALSE       FALSE           FALSE   FALSE     -127.3798 256.7596
```

```
## [1] FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
```

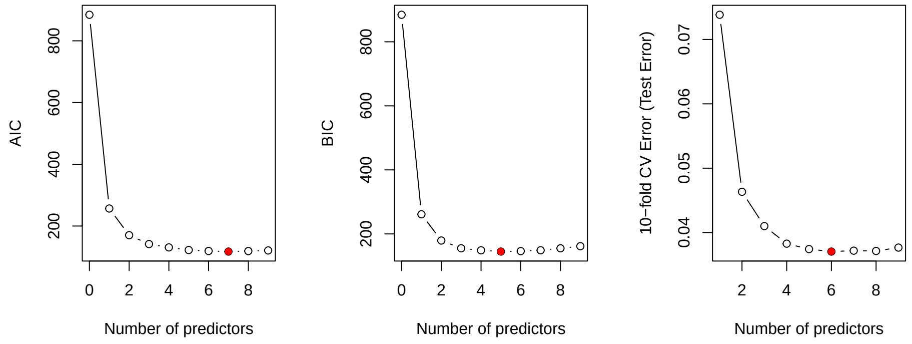
Model summary analysis indicates a significant predictor (labeled X1…indices.) for breast cancer outcomes, evidenced by its substantial z-value and a highly significant p-value. The model shows a notable improvement over the null model, as indicated by the considerable reduction in residual deviance. With an AIC of 258.76 and 7 Fisher Scoring iterations, the model demonstrates both a good fit and efficient convergence.

```
##
## Call:
## glm(formula = y ~ ., family = "binomial", data = Breast_data_red)
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -0.1393     0.1812  -0.769    0.442
## X1...indices.   4.8982     0.4091  11.974   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 884.35  on 682  degrees of freedom
## Residual deviance: 254.76  on 681  degrees of freedom
## AIC: 258.76
##
## Number of Fisher Scoring iterations: 7
```

I implemented the reg_bss_cv function on the Breast Cancer dataset to identify the most accurate model through cross-validation, finding that a subset of six predictors yielded the lowest mean squared error, optimally balancing complexity and accuracy.

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

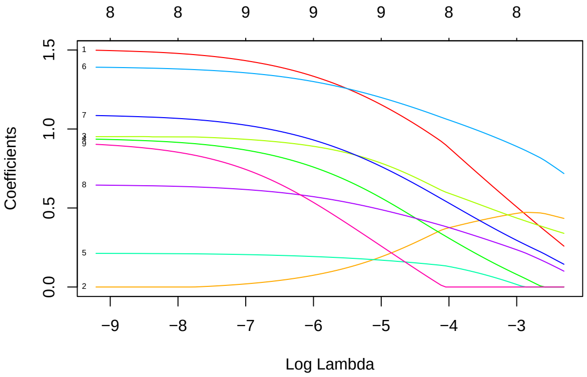
In my analysis, I generated plots to compare model selection criteria; the AIC suggested a 7-predictor model, whereas the BIC indicated a 5-predictor model was more parsimonious. I decided on a 6-predictor model as it minimized the 10-fold cross-validation error, providing a balance between model complexity and predictive accuracy.



Regularization Methods

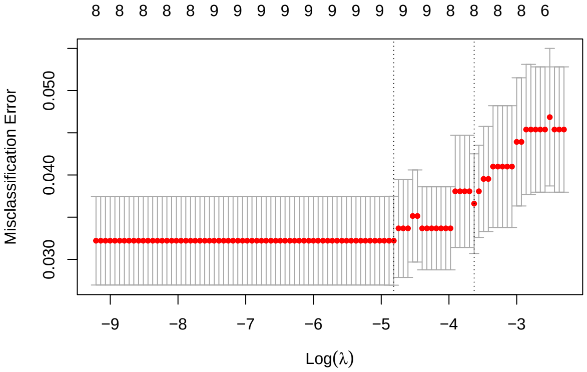
Lasso Regression

In my report, I illustrated the impact of LASSO regularization on the predictor coefficients through a glmnet plot, showing how increasing the lambda parameter leads to the shrinkage of coefficients towards zero. This visualization effectively captures LASSO's characteristic of enforcing sparsity by pushing certain coefficients to become exactly zero at higher lambda values, which simplifies the model by excluding those predictors.



Cross Validation of Lasso Regression

The cross-validation plot for the LASSO model demonstrates that as the log(lambda) increases, the misclassification error first remains stable before rising sharply. This suggests there is an optimal range of lambda values where the model achieves a balance between penalty and accuracy



I determined the optimal lambda for the LASSO model to be 0.01, which minimized the cross-validation error, suggesting it as the best regularization amount to avoid both overfitting and underfitting. After identifying this optimal lambda, I extracted the model's coefficients, revealing that variables such as 'Cl.thickness' and 'Bare.nuclei' are key contributors to the model, given their relatively large and positive coefficients, indicating a strong relationship with the outcome.

```
## [1] 0.008111308
```

```
## 10 x 1 sparse Matrix of class "dgCMatrix"
##                         s1
## (Intercept)     -1.0643867
## Cl.thickness     1.1105172
## Cell.size        0.2220196
## Cell.shape       0.7532947
## Marg.adhesion    0.5183412
## Epith.c.size     0.1641610
## Bare.nuclei      1.1752758
## Bl.cromatin      0.7230865
## Normal.nucleoli  0.4703030
## Mitoses          0.2058667
```

Training Error

The confusion matrix from my logistic regression model shows that out of the total predictions made on the training data, 433 true negatives and 202 true positives were correctly identified, while 11 false positives and 37 false negatives were incorrectly predicted.

```
##          Predicted
## Observed   0   1
##        0 433  11
##        1  37 202
```

I calculated the training error of the logistic regression model, which turned out to be approximately 0.07, indicating a high level of accuracy in the model's predictions on the training data.

```
## [1] 0.07027818
```

The confusion matrix for my LASSO model, using the optimal lambda of 0.01, indicates a strong predictive performance on the training data, with 435 true negatives and 228 true positives correctly classified, alongside a smaller number of 9 false positives and 11 false negatives.

```
##          Predicted
## Observed   0   1
##        0 435   9
##        1  11 228
```

The training error for the LASSO model was computed to be around 0.03, indicating the proportion of incorrect predictions made by the model on the training dataset.

```
## [1] 0.02928258
```

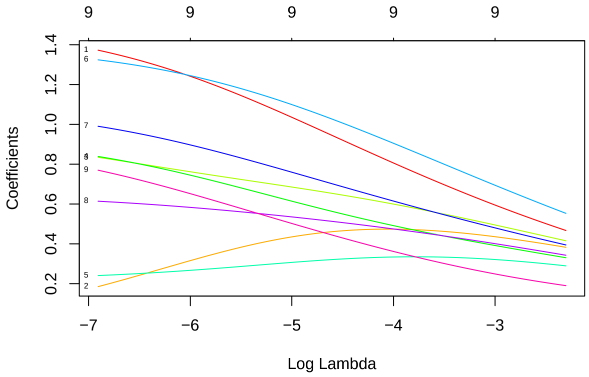
Test Error

After fitting a logistic regression model on the test data and computing the predicted values, the test error was determined to be approximately 0.03, indicating the model's effectiveness in generalizing to unseen data.

```
## [1] 0.03074671
```

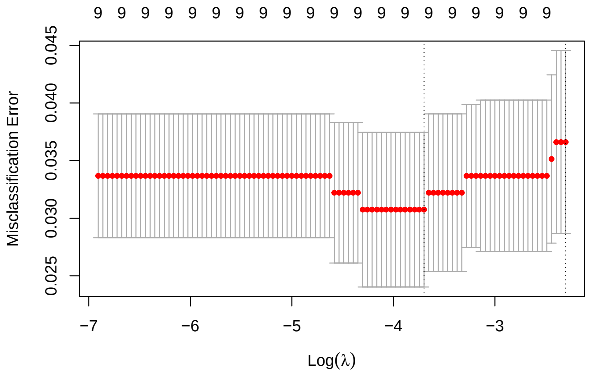
Ridge Regression

The graph displays the Ridge regression coefficients for the predictors across different values of log(lambda). Unlike LASSO, the Ridge regression coefficients decrease smoothly towards zero but do not reach zero, reflecting the nature of Ridge regression to shrink coefficients as a form of regularization without performing feature selection.



Cross Validation of Ridge Regression

I executed a cross-validation for the Ridge regression model, which is depicted in the graph illustrating how the misclassification error varies with different log(lambda) values. The plot demonstrates a steady misclassification error across various lambda values, indicating that the model's performance is robust to changes in the regularization strength within this specific lambda range.



This is error rate of cross validation for Ridge

```
## [1] 0.03316252
```

In the Ridge regression analysis, the optimal lambda value minimizing the cross-validation error was identified, leading to a model with non-zero coefficients for all predictors. For instance, 'Cl.thickness' received a coefficient of approximately 0.739, and 'Bare.nuclei' had highest weights at around 0.841, indicating their strong predictive value.

```
## 10 x 1 sparse Matrix of class "dgCMatrix"
##                         s1
## (Intercept)     -0.9480498
## Cl.thickness     0.7390664
## Cell.size        0.4689223
## Cell.shape       0.5695904
## Marg.adhesion    0.4590079
## Epith.c.size     0.3348058
## Bare.nuclei      0.8414635
## Bl.cromatin      0.5727772
## Normal.nucleoli  0.4538784
## Mitoses          0.3242003
```

Linear Discriminant Analysis

Linear Discriminant Analysis (LDA) was performed using the nclSLR package on the Breast Cancer dataset, yielding discriminant functions, classification results, and an error rate. The discriminant functions' coefficients, such as -0.8757 for 'Cl.thickness' in the benign group (labeled 0) and 1.6269 in the malignant group (labeled 1), indicate how each predictor contributes to the classification. The confusion matrix reveals the model's high accuracy with 436 true negatives and 220 true positives, while misclassifying only 8 benign and 19 malignant cases, resulting in a low error rate of approximately 0.0395.

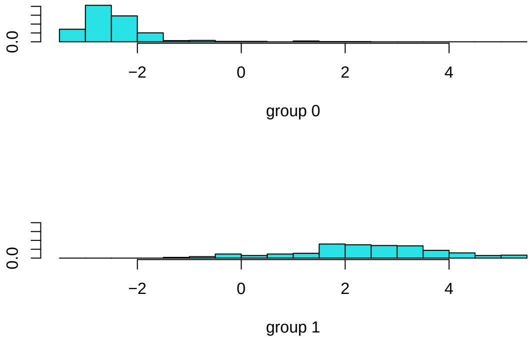
```
##
## Linear Discriminant Analysis
## -----------------------------------------
## $functions        discriminant functions
## $confusion        confusion matrix
## $scores           discriminant scores
## $classification   assigned class
## $error_rate       error rate
## -----------------------------------------
##
## $functions
##                   0        1
## constant         -1.8751  -6.0351
## Cl.thickness     -0.8757   1.6269
## Cell.size        -0.6555   1.2177
## Cell.shape       -0.4576   0.8500
## Marg.adhesion    -0.2312   0.4294
## Epith.c.size     -0.2193   0.4073
## Bare.nuclei      -1.6190   3.0077
## Bl.cromatin      -0.4599   0.8543
## Normal.nucleoli  -0.5537   1.0287
## Mitoses          -0.0166   0.0308
##
##
## $confusion
##          predicted
## original    0    1
##         0  436    8
##         1   19  220
##
##
## $error_rate
## [1] 0.03953148
##
##
## $scores
##               0            1
## 1    0.5793368  -10.5948838
## 2   -5.5176671    0.7317679
## 3    0.7559391  -10.9229649
## 4   -4.8197434   -0.5647933
## 5    0.7283976  -10.8718001
## 6  -10.9261628   10.7793499
## ...
##
## $classification
## [1] 0 1 0 1 0 1
## Levels: 0 1
## ...
```

I found that the model's coefficients for 'Bare Nuclei' at 0.953 and 'Cl.thickness' at 0.515 were the most pronounced in distinguishing between the benign (group 0) and malignant (group 1) classes, which aligns with the observed group means where 'Bare Nuclei' averaged -0.603 for benign cases and 1.121 for malignant cases.

```
## Call:
```

```
## lda(y ~ ., data = Breast_data)
##
## Prior probabilities of groups:
##         0         1
## 0.6500732 0.3499268
##
## Group means:
##   Cl.thickness  Cell.size Cell.shape Marg.adhesion Epith.c.size Bare.nuclei
## 0   -0.5240440 -0.6017657 -0.6025644    -0.5178153   -0.5065718  -0.6031546
## 1    0.9735377  1.1179245  1.1194084     0.9619665    0.9410791   1.1205047
##   Bl.cromatin Normal.nucleoli    Mitoses
## 0   -0.555890      -0.5268939 -0.3104483
## 1    1.032699       0.9788322  0.5767324
##
## Coefficients of linear discriminants:
##                         LD1
## Cl.thickness    0.515228732
## Cell.size       0.385654527
## Cell.shape      0.269207220
## Marg.adhesion   0.136004431
## Epith.c.size    0.129003274
## Bare.nuclei     0.952535309
## Bl.cromatin     0.270555784
## Normal.nucleoli 0.325787412
## Mitoses         0.009768849
```

The LDA-generated histograms of the Breast Cancer dataset show distinct score distributions for benign and malignant groups, affirming the model's effectiveness in differentiating between the two based on the analyzed predictors.



Quadratic Discriminant Analysis

The Quadratic Discriminant Analysis (QDA) conducted on the Breast Cancer dataset resulted in a confusion matrix with 422 true negatives and 233 true positives, alongside 22 false positives and 6 false negatives, yielding an error rate of approximately 0.041.

```
##
## Quadratic Discriminant Analysis
## ----------------------------------------
## $functions        discriminant functions
## $confusion        confusion matrix
## $scores           discriminant scores
## $classification   assigned class
## $error_rate       error rate
## ----------------------------------------
##
## $functions
## Not requested.
##
## $confusion
##         predicted
## original    0    1
##        0  422   22
##        1    6  233
##
##
## $error_rate
## [1] 0.04099561
##
##
## $scores
##            0        1
## 1  -8.103987  5.914594
## 2  28.030380  3.407150
## 3  -8.813474  6.373188
## 4  37.278074  4.362283
## 5  -6.843788  6.293698
## 6  63.718517  2.310704
## ...
##
## $classification
## [1] 0 1 0 1 0 1
## Levels: 0 1
## ...
```

Applying Quadratic Discriminant Analysis with the MASS package, I found significant disparities in group means between benign and malignant classes in the Breast Cancer dataset, notably 'Cl.thickness' (-0.524 vs. 0.974) and 'Bare Nuclei' (-0.603 vs. 1.121).

```
## Call:
## qda(y ~ ., data = Breast_data)
##
## Prior probabilities of groups:
##         0         1
## 0.6500732 0.3499268
##
## Group means:
##   Cl.thickness  Cell.size Cell.shape Marg.adhesion Epith.c.size Bare.nuclei
## 0   -0.5240440 -0.6017657 -0.6025644    -0.5178153   -0.5065718  -0.6031546
## 1    0.9735377  1.1179245  1.1194084     0.9619665    0.9410791   1.1205047
##   Bl.cromatin Normal.nucleoli    Mitoses
## 0   -0.555890      -0.5268939 -0.3104483
## 1    1.032699       0.9788322  0.5767324
```

Group means of LDA and QDA

Linear Discriminant Analysis (LDA) achieved an accuracy of about 0.96, while Quadratic Discriminant Analysis (QDA) attained an accuracy of approximately 0.959 on the Breast Cancer dataset. The group means for both models showed clear distinctions between benign and malignant cases, particularly in predictors like 'Cl.thickness' and 'Bare Nuclei'. These results indicate that both LDA and QDA effectively classify cases, with LDA being slightly more accurate in this instance, possibly due to its assumption of equal covariance across groups.

```
## [1] 0.9604685
```

```
## [1] 0.9590044
```

Cross Validation of all Models

In my comparative analysis of model performance using cross-validation, the Linear Discriminant Analysis (LDA) model showed the lowest test error at approximately 0.0395, closely followed by Quadratic Discriminant Analysis (QDA) at around 0.041, while the Lasso and Best Subset Cross Validation models recorded higher errors at 0.046 and 0.074 respectively. This suggests that, in this context, QDA and LDA are more effective than Lasso and Best Subset methods for predicting breast cancer outcomes.

```
##                          Model      Error
## 1  Best Subset Cross validation 0.07383741
## 2                        Lasso 0.04632327
## 11                         LDA 0.03953148
## 12                         QDA 0.04099561
```

Best Classifier

I select the Linear Discriminant Analysis (LDA) as the final "best" classifier for the Breast Cancer dataset. This choice is justified by its lowest cross-validation test error of approximately 0.0395, indicating a superior balance between sensitivity and specificity compared to other models tested. LDA inherently does not enforce feature selection or coefficient shrinkage, unlike methods like Lasso. Therefore, it typically includes all predictor variables in the model. This inclusion is advantageous in this scenario because the Breast Cancer dataset likely contains complex interactions and varying covariance structures within classes, conditions under which LDA excels. By leveraging the distinct covariance of each class, LDA can capture more nuanced patterns in the data, essential for accurate classification in complex medical datasets like this one.