MAS8404 - PROJECT REPORT

SUSHANTH SHIVPURA RAMESH

S.NO: 210441057

INTRODUCTION:

The Breast Cancer data set records the characteristics of the tissue sample collected from 699 women in Wisconsin using an invasive method. The collected tissue sample are assed with nine cytological characteristics i.e. Cl.thickness, Cell.size, Cell.shape, Marg.adhesion, Epith.c.size, Bare.nuclei, Bl.cromatin, Normal.nucleoli and Mitoses . These characteristics of tissues were measured on a scale of 0 to 10. The smaller the number it indicates that the cell is healthy. Furthermore, examination establishes whether the sample is benign (non-cancerous cell) and malignant (cancer cell).

	Id	Cl.thickness	Cell.size	Cell.shape	Marg.adhesion	Epith.c.size	Bare.nuclei	Bl.cromatin	Normal.nucleoli	Mitoses	Class
1	1000025	5	1	1	1	2	1	3	1	1	benign
2	1002945	5	4	4	5	7	10	3	2	1	benign
3	1015425	3	1	1	1	2	2	3	1	1	benign
4	1016277	6	8	8	1	3	4	3	7	1	benign
5	1017023	4	1	1	3	2	1	3	1	1	benign
6	1017122	8	10	10	8	7	10	9	7	1	malignant
7	1018099	1	1	1	1	2	10	3	1	1	benign

First few rows of the Breast cancer data set

The dimensions of the data set are 699 rows X 11 columns. Our aim of this project is to build classifiers for the Class (response variable) – benign or malignant based on the nine cytological characteristics (predictor variable).

DATA PRE-PROCESSING:

- The data set contains missing observation, which are marked as *NA*. We remove all the rows where there are missing values. We use "na.omit(BreastCancer)" function in R to remove all the *NA* values . The dimension of the dataset is reduced to 683 rows X 11 columns, by removing 16 rows/samples with *NA* values.

24	1057013	8	4	5	1	2	*NA*	7	3	1	malignant

Example row containing missing value

- For the purpose of numerical analysis, the response variable Class (benign and malignant) is converted/encode to 0 and 1 i.e., benign = 0 and malignant = 1.

- The cytological characteristics/ predictor values in the original data set are not numeric. To perform numeric analysis on the data set we convert the values to numeric using "as.numeric()" function .

```
'data.frame':   683 obs. of  11 variables:
 $ Id             : chr  "1000025" "1002945" "1015425"
 $ 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 ...
```

Predictor and response variables in numerical form after conversion

EXPLORATORY DATA ANALYSIS:

The Breast Cancer data set contains more of the benign samples (444) which is about 65 % of the dataset and malignant tissue samples (239) which is 35% of the data set . This can be calculated using the “table()” function as below .

```
> table(Bre_orig$Class)

  0   1
444 239
```

By using “ggpairs()” from “library(GGally) ” we can produce a scattered plot and correlation matrix between the variables as shown below .

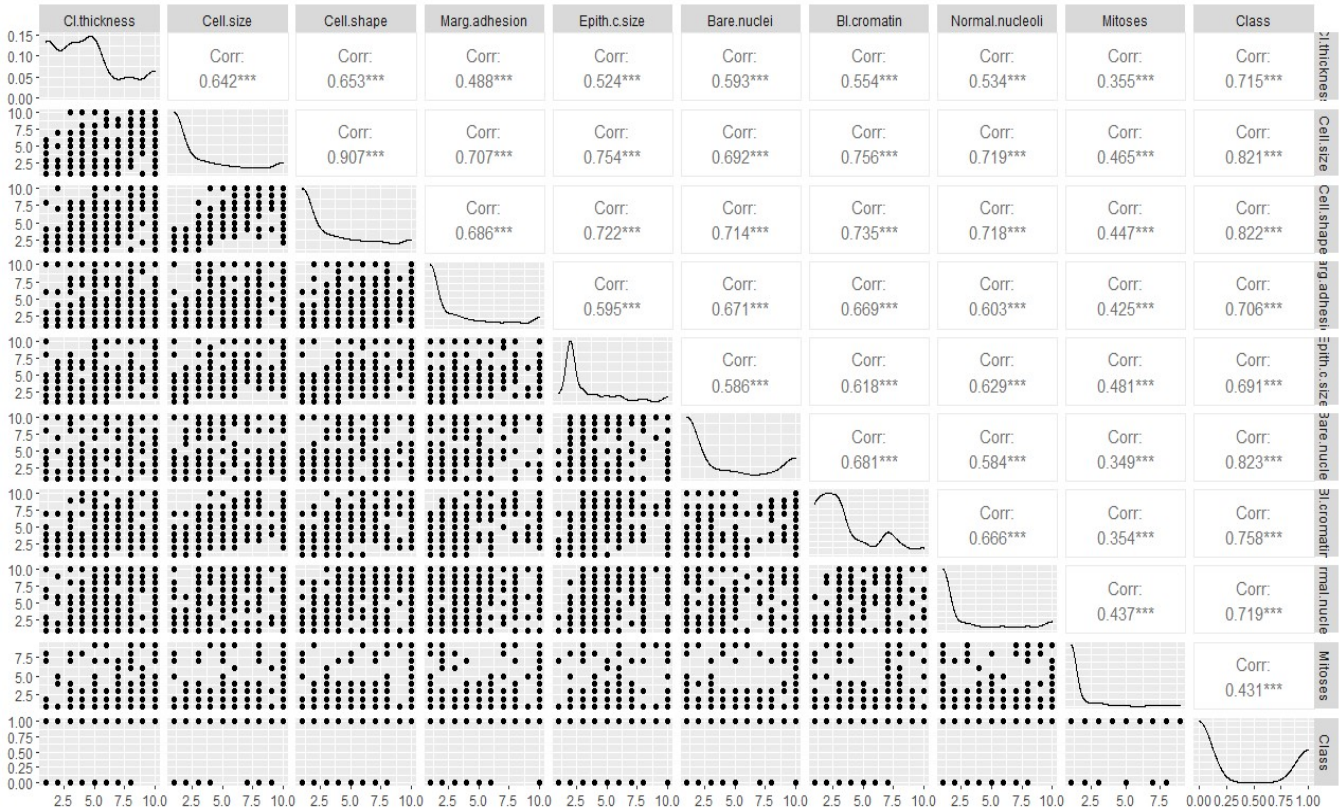

Fig 1) Scattered plot and correlation matrix between the variables

From the above matrix shows that the response variable Class is highly corelated with the predictor variables Bare.nuclei (0.823), Cell.size(0.821), Cell.shape(0.822). It can also clear that the a few predictor variables are by themselves corelated (multicollinearity) such as cell.size and cell.shape (0.907), Bl.cromatin and cell.size(0.756).This suggests that we can drop some predictor variables in the regression models. From the scattered plot we don't find any obvious pattern visually.

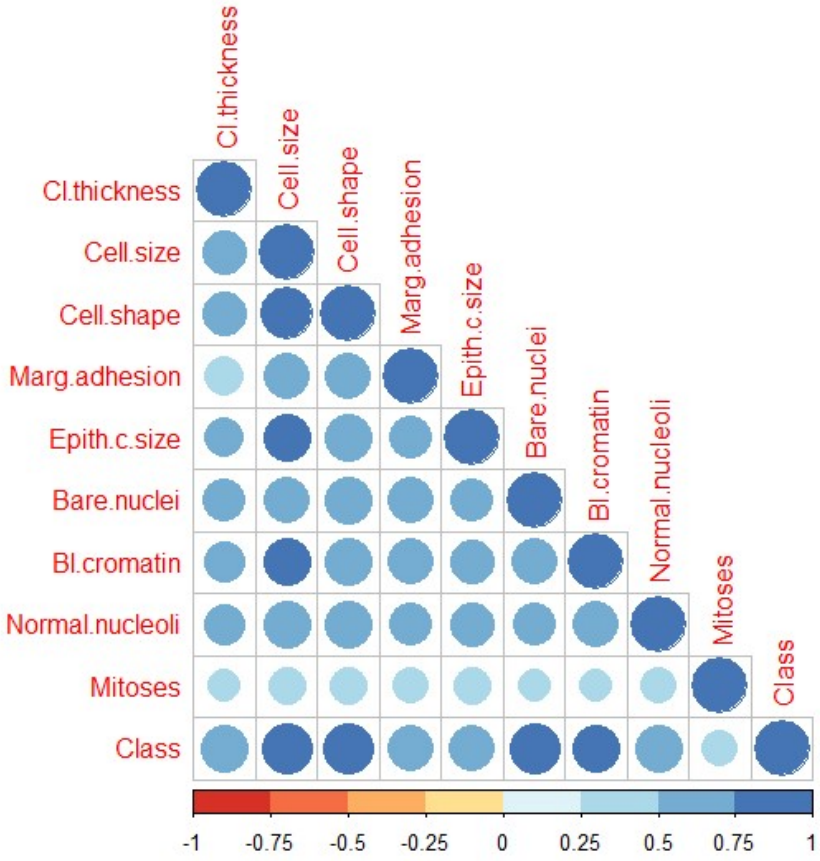

Fig2) Visual exploratory on correlation matrix

The above plot is produced using "corrplot()" function which provides a visual exploratory tool on correlation matrix. The above matrix helps in visually identifying which of the variables are highly correlated. The darker and larger the circle towards blue the variables are highly correlated. As seen in the scattered plot and correlation matrix we can see the Class (response variable) is highly correlated with Bl.cromatin , Cell.size , Cell.shape, Bare.nuclei.

BUILDING CLASSIFIERS

Our goal is to build classifiers for the class(response) – benign and malignant of the the tissue samples using the 9 cytological variables (predictors). The following are different methods with which we can build classifiers.

1) LOGISTIC REGRESSION:

We can fit a logistic regression on the breast cancer data using "glm function" and set the argument family to binomial so as to perform logistic regression. The summary of the model fit gives the following output.

```
> summary(logreg_Bre)

Call:
glm(formula = y ~ ., family = "binomial", data = Bre_data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.4855  -0.1152  -0.0619   0.0222   2.4702

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)     -10.110096   1.173774  -8.613  < 2e-16 ***
Cl.thickness      0.535256   0.141938   3.771 0.000163 ***
Cell.size        -0.005943   0.209158  -0.028 0.977332
Cell.shape        0.322136   0.230644   1.397 0.162510
Marg.adhesion     0.330694   0.123462   2.679 0.007395 **
Epith.c.size      0.096797   0.156568   0.618 0.536415
Bare.nuclei       0.383015   0.093865   4.080 4.49e-05 ***
Bl.cromatin       0.447401   0.171392   2.610 0.009044 **
Normal.nucleoli   0.213074   0.112894   1.887 0.059109 .
Mitoses           0.538551   0.325615   1.654 0.098138 .
---
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: 102.90  on 673  degrees of freedom
AIC: 122.9

Number of Fisher Scoring iterations: 8
```

The maximum likelihood estimates of the regression coefficients are

$\hat{\beta}0 = -10.110$, $\hat{\beta}1 = 0.535$, $\hat{\beta}2 = -0.006$, $\hat{\beta}3 = 0.322$, $\hat{\beta}4 = 0.330$, $\hat{\beta}5 = 0.096$, $\hat{\beta}6 = 0.383$, $\hat{\beta}7 = 0.383$,

$\hat{\beta}8 = 0.44$, $\hat{\beta}9 = 0.21$, $\hat{\beta}10 = 0.538$

We can see that there are variables with large p-values, which contribute very little to the model. As seen in the EDA there are predictor variables that are highly correlated, which suggests that we do not require all the predictor variables in the model. Including all the variables can also decrease the predictive performance as it inflates the variance of the parameter estimators.

The test error is calculated using cross validation

TEST ERROR = 0.03367496

```
> cv_lr
[1] 0.03367496
```

We can improve the model by using some of the regularisation or dimension-reduction techniques as follows.

2) BEST SUBSET SELECTION METHOD IN LOGISTIC REGRESSION

We can use " bestglm fuction" on the BreastCancer data set for best subset selection using the AIC and BIC .We can identify the best fitting models for AIC and BIC which is 7 predictor values for AIC and 5 Values for BIC.

```
> (best_Bre_AIC=bss_fit_Bre_AIC$ModelReport$Bestk) # identifying best fitting models for AIC
[1] 7
> (best_Bre_BIC=bss_fit_Bre_BIC$ModelReport$Bestk)  # identifying best fitting models for BIC
[1] 5
```

To choose a single best model we can plot how the criteria vary with the number of predictors as shown below:

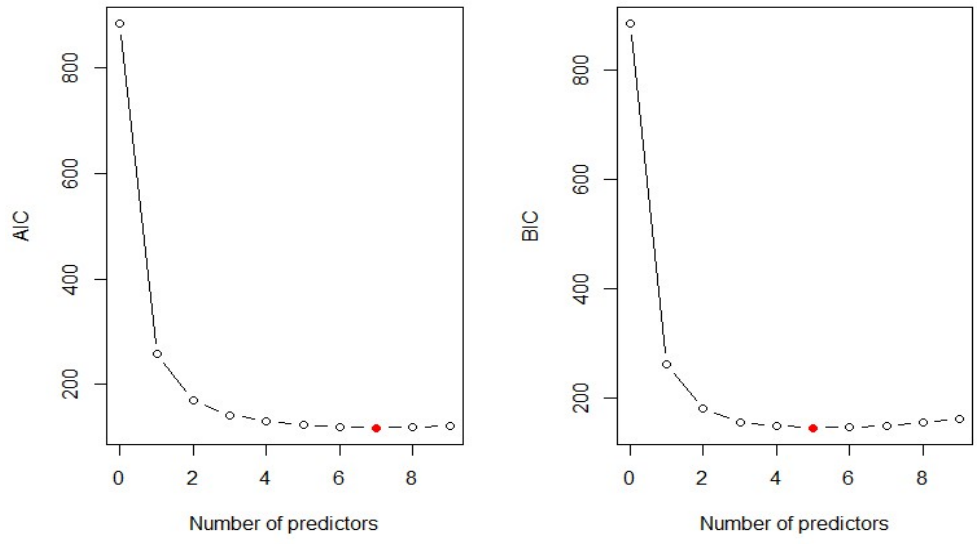

Fig 3) Best Subset selection for BreastCancer data

From the above plot we pick the model with 5 predictor variables i.e. BIC as a good compromise.

```
> bss_fit_Bre_BIC$Subsets
   Intercept Cl.thickness Cell.size Cell.shape Marg.adhesion Epith.c.size Bare.nuclei Bl.cromatin Normal.nucleoli Mitoses logLikelihood      BIC
0       TRUE        FALSE     FALSE      FALSE         FALSE        FALSE       FALSE       FALSE           FALSE   FALSE    -442.17509 884.3502
1       TRUE        FALSE     FALSE      FALSE         FALSE        FALSE       FALSE       FALSE           FALSE   FALSE    -127.37980 261.2861
2       TRUE        FALSE      TRUE      FALSE         FALSE        FALSE        TRUE       FALSE           FALSE   FALSE     -83.15598 179.3649
3       TRUE         TRUE      TRUE      FALSE         FALSE        FALSE        TRUE       FALSE           FALSE   FALSE     -67.77778 155.1351
4       TRUE         TRUE     FALSE       TRUE         FALSE        FALSE        TRUE        TRUE           FALSE   FALSE     -61.37155 148.8491
5*      TRUE         TRUE     FALSE      FALSE          TRUE        FALSE        TRUE        TRUE            TRUE   FALSE     -56.13177 144.8960
6       TRUE         TRUE     FALSE       TRUE          TRUE        FALSE        TRUE        TRUE            TRUE   FALSE     -53.57186 146.3027
7       TRUE         TRUE     FALSE       TRUE          TRUE        FALSE        TRUE        TRUE            TRUE    TRUE     -51.63998 148.9654
8       TRUE         TRUE     FALSE       TRUE          TRUE         TRUE        TRUE        TRUE            TRUE    TRUE     -51.45031 155.1126
9       TRUE         TRUE      TRUE       TRUE          TRUE         TRUE        TRUE        TRUE            TRUE    TRUE     -51.44991 161.6383
> |
```

The predictors included in best fit model BIC as above, which is denoted by a star on the first column Infront of 5

We can extract the variables form the best fitting BIC model with 5 predictor variables from the output of the besglm function and obtain the regression coefficients from the model. The summary of the fit is as below:

```
> summary(logreg_Bre1_fit)

Call:
glm(formula = y ~ ., family = "binomial", data = Bre_data_red)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.8603  -0.1228  -0.0534   0.0227   2.1903

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)     -10.13060    1.09454  -9.256  < 2e-16 ***
Cl.thickness      0.74129    0.13189   5.621 1.90e-08 ***
Marg.adhesion     0.39515    0.11592   3.409 0.000652 ***
Bare.nuclei       0.44733    0.08797   5.085 3.68e-07 ***
Bl.cromatin       0.55287    0.15019   3.681 0.000232 ***
Normal.nucleoli   0.33419    0.09781   3.417 0.000634 ***
---
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: 112.26  on 677  degrees of freedom
AIC: 124.26

Number of Fisher Scoring iterations: 8
```

The regression coefficients are

ˆβ0 = -10.130, ˆβ1 = 0 .741, ˆβ2 = 0.395, ˆβ3 = 0.447, ˆβ4 = 0 .552, ˆβ5 = 0.334.

We can observe from the above model the 4 predictor variables i.e. Cell.size, Cell.shape, Epith.c.size, and Mitoses are dropped out from the model .

The test error for this model can be calculated using cross validation as follows:

TEST ERROR = 0.03221083

```
> sub_error
[1] 0.03221083
```

We can continue building our classifier and check the performance by adding penalty

3) LOGISTIC REGRESSION WITH LASSO PENALTY

By applying Logistic Regression with LASSO Penalty on the full breast cancer data set using the **optimum value for the lambda 0.01072267 from** the cross-validation, we get the following Regression coefficients:

ˆβ0 = -8.290, ˆβ1 = 0 .466, ˆβ2 = 0.074, ˆβ3 = 0.274, ˆβ4 = 0 .236, ˆβ5 = 0.048, ˆβ6 = 0 .361, ˆβ7 = 0.328, ˆβ8 = 0 .189, ˆβ9 = 0.128

```
> (lambda_lasso_min=lasso_cv_fit$lambda.min)
[1] 0.01072267
> which_lasso
[1] 70
```

```
> coef(lasso_fit, s = lambda_lasso_min)
10 x 1 sparse Matrix of class "dgCMatrix"
                      s1
(Intercept)     -8.29080949
Cl.thickness     0.46662313
Cell.size        0.07406160
Cell.shape       0.27426482
Marg.adhesion    0.23672310
Epith.c.size     0.04874647
Bare.nuclei      0.36146492
Bl.cromatin      0.32806258
Normal.nucleoli  0.18923253
Mitoses          0.12827339
```

From the following model we observe that none of the predictor variables are dropped from the model. The test error with LASSO Penalty using cross validation is calculated using the “lasso_cv_fit$cvm() function ” with the fold being constant to make the comparison fare .

TEST ERROR = 0.03367496

```
> lasso_cv_fit_test$cvm[which_lasso_1]
[1] 0.03367496
```

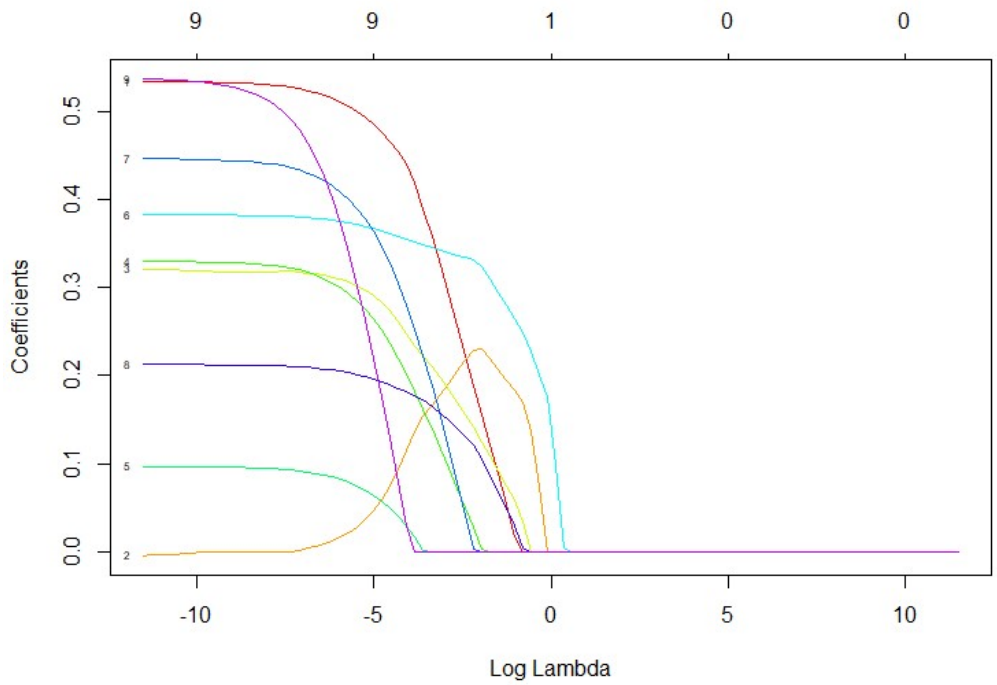

Fig 4) Effect of varying tuning parameters on the logistic regression with LASSO penalty on the data set

From the above graph we can see that the predicator variable are converging after 0 and the numbers at the top of the graph represent the variables that tend to zero as the lamda increases.

4) BAYES CLASSIFIER FOR LINEAR DISCRIMINANT ANALYSIS (LDA)

The probability of a sample being malignant is 34.99 % and benign is 65.0% from the probability of the groups. The group means of the variables are as shown below. The average mean of cell thickness variable if its benign is 2.96 and its 7.18 if its malignant. The same can be interpreted for the remaining variables as well by looking into the group means.

```
> LDA_Bre_fit
Call:
lda(y ~ ., data = Bre_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 Bl.cromatin Normal.nucleoli  Mitoses
0     2.963964  1.306306   1.414414      1.346847     2.108108    1.346847    2.083333        1.261261 1.065315
1     7.188285  6.577406   6.560669      5.585774     5.326360    7.627615    5.974895        5.857741 2.543933

Coefficients of linear discriminants:
                     LD1
Cl.thickness    0.182556105
Cell.size       0.125687035
Cell.shape      0.090054130
Marg.adhesion   0.047213478
Epith.c.size    0.057570551
Bare.nuclei     0.261447573
Bl.cromatin     0.110626289
Normal.nucleoli 0.106511431
Mitoses         0.008591172
```

The Cross-validation test error for LDA calculated using the function

TEST ERROR = 0.03806735

```
> lda_cv(Bre_data[,1:9],Bre_data[,10],fold_index)
[1] 0.03806735
```

5) BAYES CLASSIFIER FOR QUADRATIC DISCRIMINANT ANALYSIS (QDA)

As with LDA the group means for QDA is as shown below.

```
> QDA_Bre_fit
Call:
qda(y ~ ., data = Bre_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 Bl.cromatin Normal.nucleoli  Mitoses
0     2.963964  1.306306   1.414414      1.346847     2.108108    1.346847    2.083333        1.261261 1.065315
1     7.188285  6.577406   6.560669      5.585774     5.326360    7.627615    5.974895        5.857741 2.543933
```

The Cross-Validation test error for QDA calculated using the function

TEST ERROR = 0.04685212

```
> qda_cv(Bre_data[,1:9],Bre_data[,10],fold_index)
[1] 0.04685212
```

CONCLUSION:

MODEL	CV TEST ERROR
LOGISTIC REGRESSION	**0.03367496**
BEST SUBSET SELECTION	**0.03221083**
LASSO PENALTY	**0.03367496**
LDA	**0.03806735**
QDA	**0.04685212**

From the above table we can see that Logistic regression with Best subset selection has the least Test error with the least number of predictor variables. The predictor variables that are included **Cl.thickness, Marg.adhesion, Bare.nuclei, Bl.cromatin, and Normal.nucleoli** and the regression coefficients are **$\hat{\beta}0 = -10.130$, $\hat{\beta}1 = 0.741$, $\hat{\beta}2 = 0.395$, $\hat{\beta}3 = 0.447$, $\hat{\beta}4 = 0.552$, $\hat{\beta}5 = 0.334$.** The variables that are dropped are due to multi collinearity and the rest don't have significant impact on the model .