Final Report - MAS8404 - Statistical Learning for Data Science

Kaustubh Kulkarni - 230195431

2023-12-01

Introduction

In this project, I will analyse the BreastCancer data set which concerns characteristics of breast tissue samples collected from 699 women in Wisconsin using fine needle aspiration cytology (FNAC). There are total 9 different characteristics recorded for each tissue sample scaling from 1 to 10 (1 indicating healthier). I will build classifiers for this data and my goal will be to determine the best classifier among them.

Data Wrangling

To start with, first I will load the BreastCancer dataset from 'mlbench' package.

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

Above are few rows from the dataset which shows the characteristics I will be working on : ID (Sample Code Number), Predictor Variables - Cl.thickness(Clump Thickness), Cell.size(Uniformity of Cell Size), Cell.shape(Uniformity of Cell Shape), Marg.adhesion(Marginal Adhesion), Epith.c.size(Single Epithelial Cell Size), Bare.nuclei(Bare Nuclei), Bl.cromatin(Bland Chromatin), Normal.nucleoli(Normal Nucleoli), Mitoses(Mitoses) and Response Variable - Class. The predictor variables are in the form of factors. Before beginning our analysis, I will convert the factor variables into quantitative variables.

```
## 'data.frame':    699 obs. of  11 variables:
##  $ Id             : chr  "1000025" "1002945" "1015425" "1016277" ...
##  $ 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          : Factor w/ 2 levels "benign","malignant": 1 1 1 1 1 2 1 1 1 1 ...
```

Data Cleaning

Next, the data has some NA values which has to be removed before doing further analysis. Hence, I will be identifying the rows with NA values using is.na() function.

```
##          Id Cl.thickness Cell.size Cell.shape Marg.adhesion Epith.c.size
## 24  1057013            8         4          5             1            2
## 41  1096800            6         6          6             9            6
## 140 1183246            1         1          1             1            1
## 146 1184840            1         1          3             1            2
## 159 1193683            1         1          2             1            3
## 165 1197510            5         1          1             1            2
## 236 1241232            3         1          4             1            2
## 250  169356            3         1          1             1            2
## 276  432809            3         1          3             1            2
## 293  563649            8         8          8             1            2
## 295  606140            1         1          1             1            2
## 298   61634            5         4          3             1            2
## 316  704168            4         6          5             6            7
## 322  733639            3         1          1             1            2
## 412 1238464            1         1          1             1            1
## 618 1057067            1         1          1             1            1
##     Bare.nuclei Bl.cromatin Normal.nucleoli Mitoses     Class
## 24           NA           7               3       1 malignant
## 41           NA           7               8       1    benign
## 140          NA           2               1       1    benign
## 146          NA           2               1       1    benign
## 159          NA           1               1       1    benign
## 165          NA           3               1       1    benign
## 236          NA           3               1       1    benign
## 250          NA           3               1       1    benign
## 276          NA           2               1       1    benign
## 293          NA           6              10       1 malignant
## 295          NA           2               1       1    benign
## 298          NA           2               3       1    benign
## 316          NA           4               9       1    benign
## 322          NA           3               1       1    benign
## 412          NA           2               1       1    benign
## 618          NA           1               1       1    benign
```

Next, I will remove these rows from our data using na.omit() function. As we can see below, after cleaning the rows have been reduced to 683 from 699 i.e. all the 16 rows are successfully removed from the data.

```
## 'data.frame':    683 obs. of  11 variables:
##  $ Id             : chr  "1000025" "1002945" "1015425" "1016277" ...
##  $ 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          : Factor w/ 2 levels "benign","malignant": 1 1 1 1 1 2 1 1 1 1 ...
##  - 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" ...
```

Exploratory Data Analysis

Summary - Predictors vs Response Variable

Now, let's understand the data by taking a look at the numerical summary of each column.

```
##       Id             Cl.thickness      Cell.size        Cell.shape
##  Length:683         Min.   : 1.000   Min.   : 1.000   Min.   : 1.000
##  Class :character   1st Qu.: 2.000   1st Qu.: 1.000   1st Qu.: 1.000
##  Mode  :character   Median : 4.000   Median : 1.000   Median : 1.000
##                     Mean   : 4.442   Mean   : 3.151   Mean   : 3.215
##                     3rd Qu.: 6.000   3rd Qu.: 5.000   3rd Qu.: 5.000
##                     Max.   :10.000   Max.   :10.000   Max.   :10.000
##  Marg.adhesion    Epith.c.size     Bare.nuclei      Bl.cromatin
##  Min.   : 1.00   Min.   : 1.000   Min.   : 1.000   Min.   : 1.000
##  1st Qu.: 1.00   1st Qu.: 2.000   1st Qu.: 1.000   1st Qu.: 2.000
##  Median : 1.00   Median : 2.000   Median : 1.000   Median : 3.000
##  Mean   : 2.83   Mean   : 3.234   Mean   : 3.545   Mean   : 3.445
##  3rd Qu.: 4.00   3rd Qu.: 4.000   3rd Qu.: 6.000   3rd Qu.: 5.000
##  Max.   :10.00   Max.   :10.000   Max.   :10.000   Max.   :10.000
##  Normal.nucleoli    Mitoses            Class
##  Min.   : 1.00   Min.   : 1.000   benign   :444
##  1st Qu.: 1.00   1st Qu.: 1.000   malignant:239
##  Median : 1.00   Median : 1.000
##  Mean   : 2.87   Mean   : 1.603
##  3rd Qu.: 4.00   3rd Qu.: 1.000
##  Max.   :10.00   Max.   :10.000
```

Next, I will explain the relation of few predictor variables with the response variable by visualizing them.

1. Cell thickness against Class:

Fig 1.1 shows that most of the samples having cell thickness greater than 5 belong to malignant class and remaining to benign class.

2. Cell size against Class:

Fig 1.2 displays a trend of cell size and how the number of benign samples decrease and that of malignant increases as the size of cell increases.

Fig 1.1 Boxplot of Cell thickness, by Class

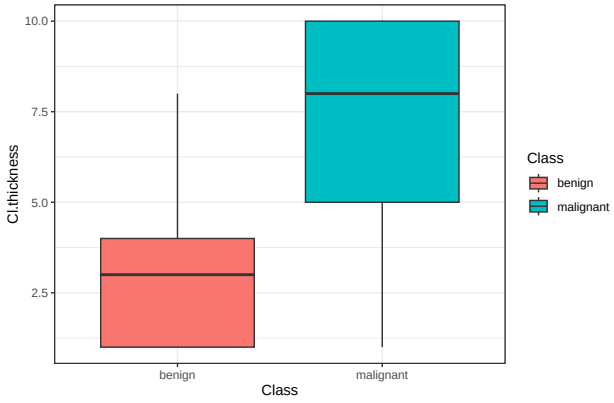
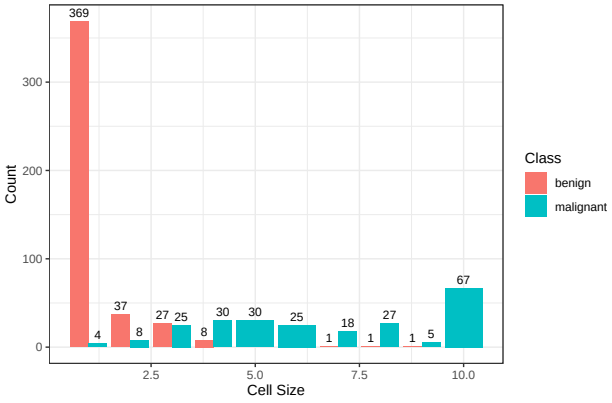


Fig 1.2 Bar plot of Cell size by Class with Counts



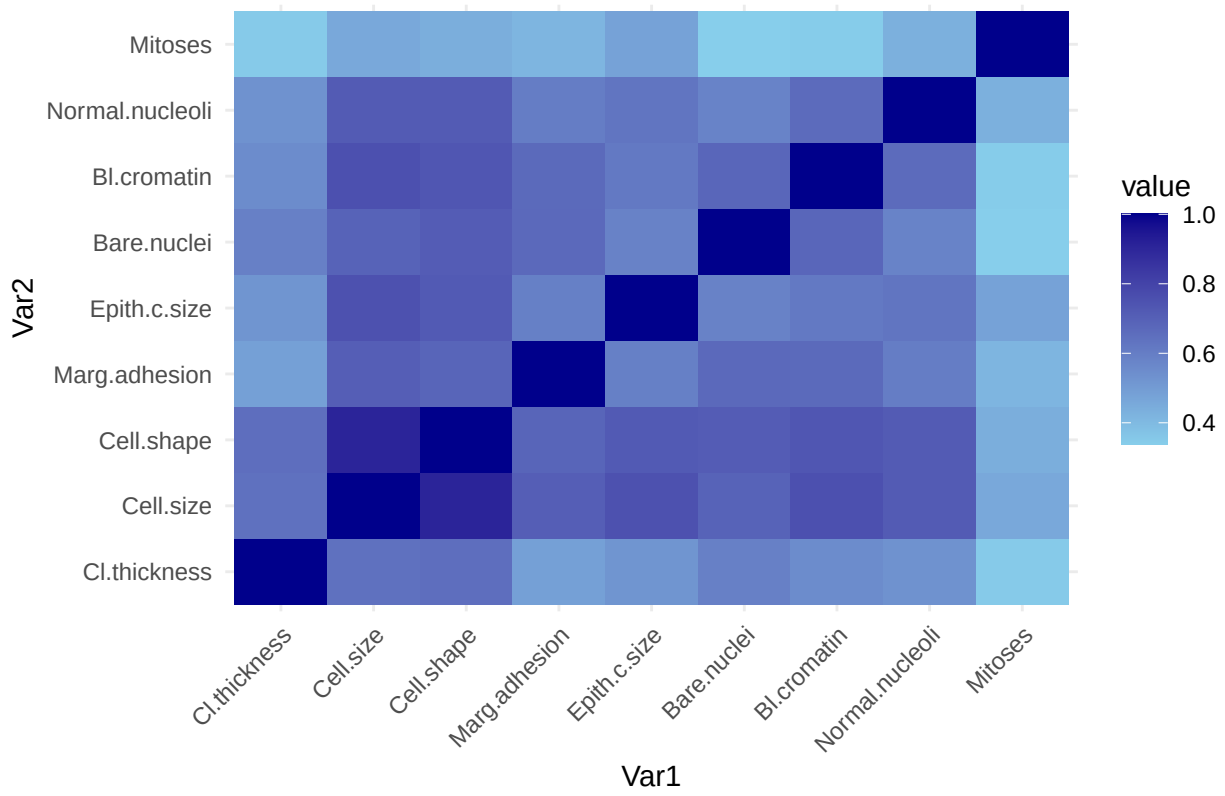
Summary - Between Predictors

Now, we will understand the relation between the predictors. First we will take a look at the correlation matrix.

```
##                Cl.thickness Cell.size Cell.shape Marg.adhesion Epith.c.size
## Cl.thickness       1.0000000 0.6424815  0.6534700     0.4878287    0.5235960
## Cell.size          0.6424815 1.0000000  0.9072282     0.7069770    0.7535440
## Cell.shape         0.6534700 0.9072282  1.0000000     0.6859481    0.7224624
## Marg.adhesion      0.4878287 0.7069770  0.6859481     1.0000000    0.5945478
## Epith.c.size       0.5235960 0.7535440  0.7224624     0.5945478    1.0000000
## Bare.nuclei        0.5930914 0.6917088  0.7138775     0.6706483    0.5857161
## Bl.cromatin        0.5537424 0.7555592  0.7353435     0.6685671    0.6181279
## Normal.nucleoli    0.5340659 0.7193460  0.7179634     0.6031211    0.6289264
## Mitoses            0.3509572 0.4607547  0.4412576     0.4188983    0.4805833
##                Bare.nuclei Bl.cromatin Normal.nucleoli   Mitoses
## Cl.thickness     0.5930914   0.5537424       0.5340659 0.3509572
## Cell.size        0.6917088   0.7555592       0.7193460 0.4607547
## Cell.shape       0.7138775   0.7353435       0.7179634 0.4412576
## Marg.adhesion    0.6706483   0.6685671       0.6031211 0.4188983
## Epith.c.size     0.5857161   0.6181279       0.6289264 0.4805833
## Bare.nuclei      1.0000000   0.6806149       0.5842802 0.3392104
## Bl.cromatin      0.6806149   1.0000000       0.6656015 0.3460109
## Normal.nucleoli  0.5842802   0.6656015       1.0000000 0.4337573
## Mitoses          0.3392104   0.3460109       0.4337573 1.0000000
```

To simplify this further, a correlation heatmap is created below showing the relations between the predictors with darker colour is high correlation. One strong insight can be taken that there is a good relation between cell shape and cell size.

Fig 1.3 Correlation Heatmap



Logistic Regression :

After having a good understanding of the data, the next step is to build the classifiers for our data. Before going further, first, the predictor variables are scaled to support the comparison. After scaling the data looks like this:

```
##   Cl.thickness  Cell.size Cell.shape Marg.adhesion Epith.c.size Bare.nuclei
## 1    0.1977598 -0.7016978 -0.7412304   -0.63889730   -0.5552016  -0.6983413
## 2    0.1977598  0.2770488  0.2625905    0.75747664    1.6939247   1.7715689
## 3   -0.5112687 -0.7016978 -0.7412304   -0.63889730   -0.5552016  -0.4239068
## 4    0.5522740  1.5820442  1.6010185   -0.63889730   -0.1053763   0.1249621
## 5   -0.1567545 -0.7016978 -0.7412304    0.05928967   -0.5552016  -0.6983413
## 6    1.2613024  2.2345419  2.2702324    1.80475710    1.6939247   1.7715689
##   Bl.cromatin Normal.nucleoli    Mitoses         y
## 1   -0.181694      -0.6124785 -0.3481446    benign
## 2   -0.181694      -0.2848960 -0.3481446    benign
## 3   -0.181694      -0.6124785 -0.3481446    benign
## 4   -0.181694       1.3530163 -0.3481446    benign
## 5   -0.181694      -0.6124785 -0.3481446    benign
## 6    2.267589       1.3530163 -0.3481446 malignant
```

In this project, 3 types of classifiers will be built :

1. Subset Selection -

For subset selection, I will be performing the best subset selection of logistic regression using "bestglm" function. In this, two types of models were computed "AIC" and "BIC".

```
## Fitting algorithm:  AIC-glm
## Best Model:
##             df deviance
## Null Model 675 103.2668
## Full Model 682 884.3502
##
##  likelihood-ratio test - GLM
##
## data:  H0: Null Model vs. H1: Best Fit AIC-glm
## X = 781.08, df = 7, p-value < 2.2e-16
```

```
## Fitting algorithm:  BIC-glm
## Best Model:
##             df deviance
## Null Model 677 112.2635
## Full Model 682 884.3502
##
##  likelihood-ratio test - GLM
##
## data:  H0: Null Model vs. H1: Best Fit BIC-glm
## X = 772.09, df = 5, p-value < 2.2e-16
```

Next the subset of models are extracted:

1. AIC Subset -

```
##    Intercept Cl.thickness Cell.size Cell.shape Marg.adhesion Epith.c.size
## 0       TRUE        FALSE     FALSE      FALSE         FALSE        FALSE
## 1       TRUE        FALSE      TRUE      FALSE         FALSE        FALSE
## 2       TRUE        FALSE      TRUE      FALSE         FALSE        FALSE
## 3       TRUE         TRUE      TRUE      FALSE         FALSE        FALSE
## 4       TRUE         TRUE     FALSE       TRUE         FALSE        FALSE
## 5       TRUE         TRUE     FALSE      FALSE          TRUE        FALSE
## 6       TRUE         TRUE     FALSE       TRUE          TRUE        FALSE
## 7*      TRUE         TRUE     FALSE       TRUE          TRUE        FALSE
## 8       TRUE         TRUE     FALSE       TRUE          TRUE         TRUE
## 9       TRUE         TRUE      TRUE       TRUE          TRUE         TRUE
##    Bare.nuclei Bl.cromatin Normal.nucleoli Mitoses logLikelihood      AIC
## 0        FALSE       FALSE           FALSE   FALSE    -442.17509 884.3502
## 1        FALSE       FALSE           FALSE   FALSE    -127.37980 256.7596
## 2         TRUE       FALSE           FALSE   FALSE     -83.15598 170.3120
## 3         TRUE       FALSE           FALSE   FALSE     -67.77778 141.5556
## 4         TRUE        TRUE           FALSE   FALSE     -61.37155 130.7431
## 5         TRUE        TRUE            TRUE   FALSE     -56.13177 122.2635
## 6         TRUE        TRUE            TRUE   FALSE     -53.57186 119.1437
## 7*        TRUE        TRUE            TRUE    TRUE     -51.63338 117.2668
## 8         TRUE        TRUE            TRUE    TRUE     -51.44455 118.8891
## 9         TRUE        TRUE            TRUE    TRUE     -51.44410 120.8882
```

2. BIC Subset -

```
##    Intercept Cl.thickness Cell.size Cell.shape Marg.adhesion Epith.c.size
## 0       TRUE        FALSE     FALSE      FALSE         FALSE        FALSE
## 1       TRUE        FALSE      TRUE      FALSE         FALSE        FALSE
## 2       TRUE        FALSE      TRUE      FALSE         FALSE        FALSE
## 3       TRUE         TRUE      TRUE      FALSE         FALSE        FALSE
## 4       TRUE         TRUE     FALSE       TRUE         FALSE        FALSE
## 5*      TRUE         TRUE     FALSE      FALSE          TRUE        FALSE
## 6       TRUE         TRUE     FALSE       TRUE          TRUE        FALSE
## 7       TRUE         TRUE     FALSE       TRUE          TRUE        FALSE
## 8       TRUE         TRUE     FALSE       TRUE          TRUE         TRUE
## 9       TRUE         TRUE      TRUE       TRUE          TRUE         TRUE
##    Bare.nuclei Bl.cromatin Normal.nucleoli Mitoses logLikelihood      BIC
## 0        FALSE       FALSE           FALSE   FALSE    -442.17509 884.3502
## 1        FALSE       FALSE           FALSE   FALSE    -127.37980 261.2861
## 2         TRUE       FALSE           FALSE   FALSE     -83.15598 179.3649
## 3         TRUE       FALSE           FALSE   FALSE     -67.77778 155.1351
## 4         TRUE        TRUE           FALSE   FALSE     -61.37155 148.8491
## 5*        TRUE        TRUE            TRUE   FALSE     -56.13177 144.8960
## 6         TRUE        TRUE            TRUE   FALSE     -53.57186 146.3027
## 7         TRUE        TRUE            TRUE    TRUE     -51.63338 148.9522
## 8         TRUE        TRUE            TRUE    TRUE     -51.44455 155.1011
## 9         TRUE        TRUE            TRUE    TRUE     -51.44410 161.6266
```

The model number with * is the best model suggested by both methods. The number of predictors of best model of each method -

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

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

To best understand this situation, graphs are plotted below as follows :

Fig 2.1.1 Number of Predictors based on AIC

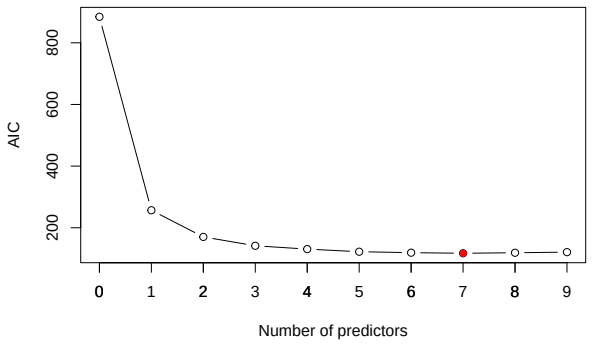
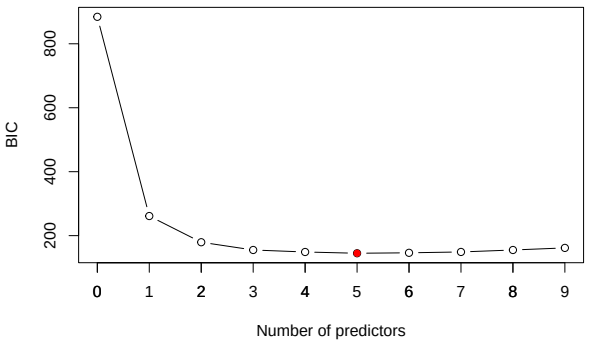


Fig 2.1.2 Number of Predictors based on BIC



From the plots above, it can be understood that model with 6 predictors(M6) would be a good compromise as it shows an optimal balance between model complexity and goodness of fit. Hence, in the next step the subset of M6 will be extracted as follows :

```
##    Intercept Cl.thickness Cell.size Cell.shape Marg.adhesion Epith.c.size
```

```
## 6       TRUE         TRUE     FALSE       TRUE          TRUE        FALSE
##    Bare.nuclei Bl.cromatin Normal.nucleoli Mitoses logLikelihood      AIC
## 6         TRUE        TRUE            TRUE   FALSE     -53.57186 119.1437
```

Storing values of each predictor variable of M6 :

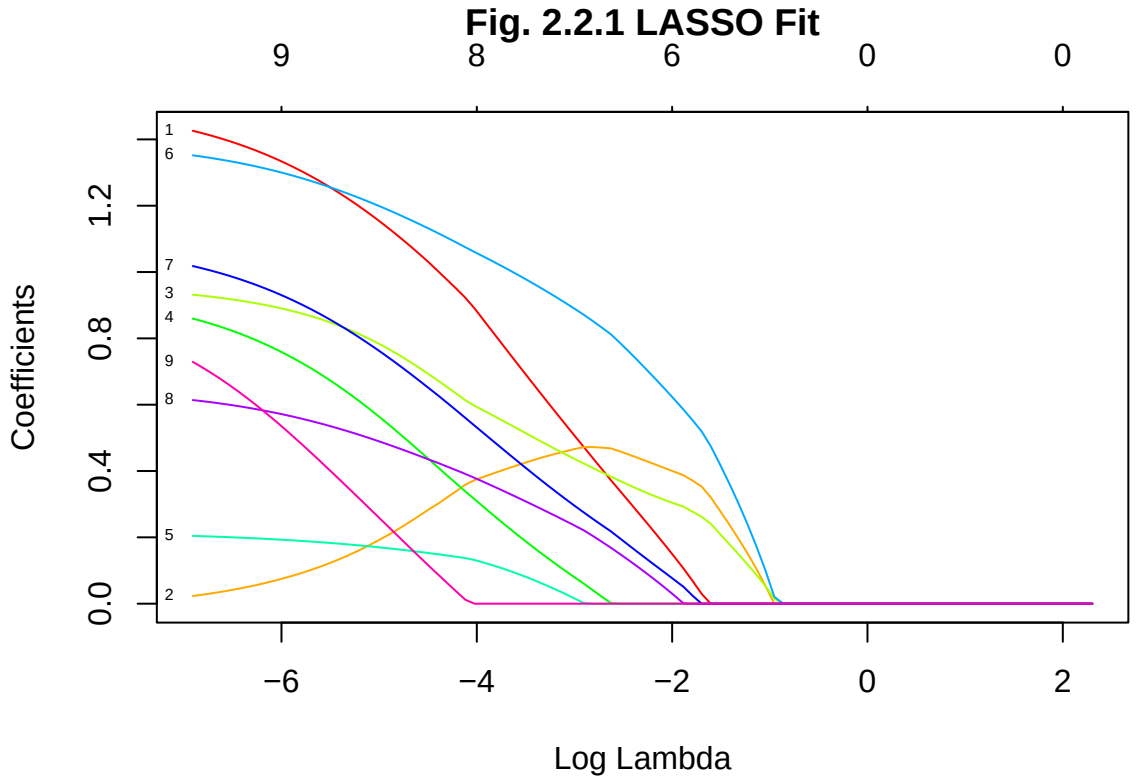
```
## [1]  TRUE FALSE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE FALSE
```

Creating dataframe with variables of M6 and response variable and passing them to a new logistic regression model. The extracted subset and their coefficients are as follows :

```
##     (Intercept)    Cl.thickness      Cell.shape   Marg.adhesion     Bare.nuclei
##      -1.2592045       1.7560138       1.0445414       0.9668875       1.3793829
##     Bl.cromatin Normal.nucleoli
##       1.1546299       0.7423195
```

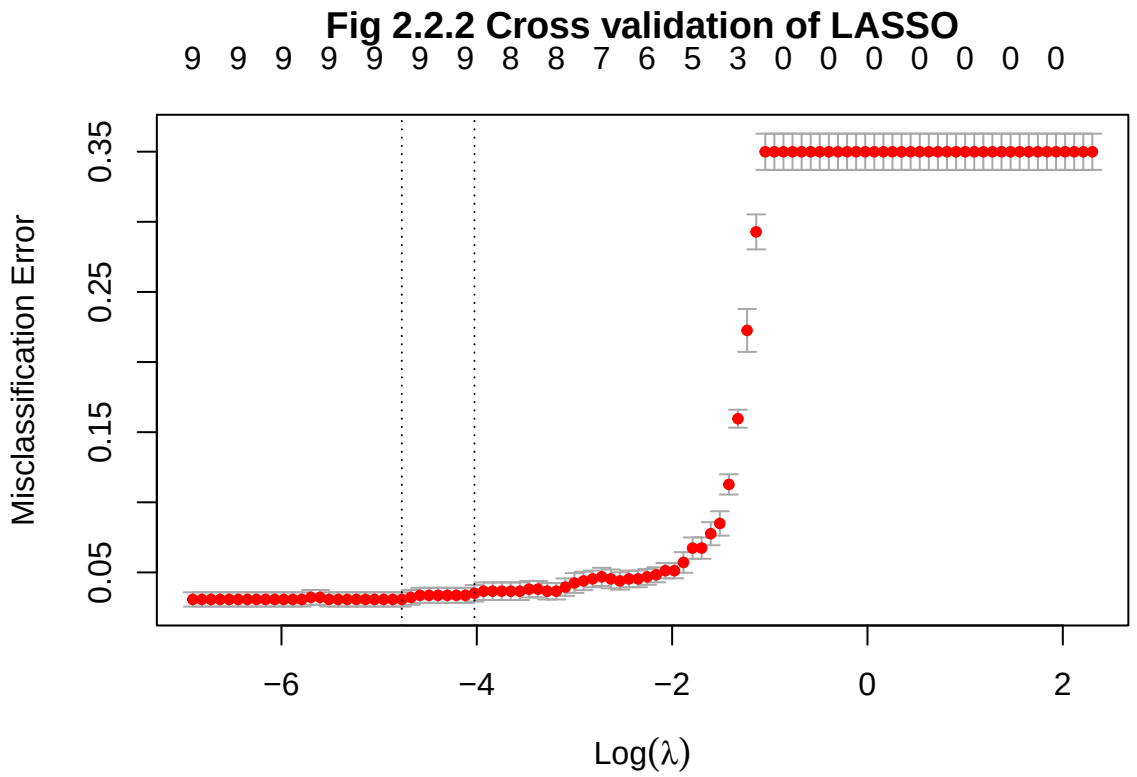
2. Regularisation Method : LASSO -

Next classifier will be built using one of the regularisation methods - LASSO using the glmnet function. The parameters are tuned into a grid and passed to lambda. The model is visualized which shows the coeffecients of each model against the negative log likelihood function.



The plot above shows the sequence in which the variables drop out and shrink towards zero. The last one to drop out is the 6th variable i.e. Bare.Nuclei. The sequence of dropping is as follows : Mitoses->Epith.c.size->Marg.adhesion->Normal.nucleoli->Bl.cromatin->Cl.thickness->Cell.size->Cell.Shape->Bare. nuclei

Next, a single value for the tuning parameter is selected using cross-validation function “cv.glmnet”



The plot is used to visualize how the test error varies with the tuning parameter. Next, the optimal value for tuning parameter is identified and corresponding parameter estimates are fetched. The optimal value is the minimum value of lambda.

```
## [1] "Minimum value for lambda  0.00849753435908644"
```

```
## [1] "Row number with least lambda value  77"
```

```
## 10 x 1 sparse Matrix of class "dgCMatrix"
##                          s1
## (Intercept)     -1.0621863
## Cl.thickness     1.0996458
## Cell.size        0.2314299
## Cell.shape       0.7447282
## Marg.adhesion    0.5067295
## Epith.c.size     0.1625488
## Bare.nuclei      1.1691865
## Bl.cromatin      0.7128376
## Normal.nucleoli  0.4652597
## Mitoses          0.1925597
```

As the coefficients shown above the regression coefficients for most of the variables have shinked towards zero.

3. Discriminant Analysis - LDA

The final type of classifier will be built for discriminant analysis and for this LDA model is created. From this model, the parameters like coefficients and the group means of predictor variables are fetched.

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

It can be summarized that :

- The prior probabilities suggest that the dataset is slightly imbalanced, with more instances of the benign class.
- Group means and coefficients suggest that features like Cl.thickness, Cell.size, Cell.shape, Bare.nuclei, etc., have significant contributions to distinguishing between the two classes.
- The group means for benign are negative and as that of malignant are positive.
- Positive coefficients across all predictors in LD1 indicate that higher values in these predictors generally contribute to the malignant class, while lower values contribute to the benign class.

Cross Validation

After building all three types of classifiers and extracting the essential parameters from each of them, now cross validation must be done to examine the error rates of each model in order to decide the "best" classifier among them. The K-fold method is used for cross validation from the "caret" package. K-fold method randomly divide the data into k folds and compute the average test error obtained by successively holding a single fold back as validation data, with the other folds serving as training data.

K-fold method is used over the validation set method because it provides a better estimate of model performance as it averages results over multiple validations, potentially reducing variability. Also K-fold cross-validation is preferred when the dataset size is limited because it uses the data more effectively for both training and validation.

1. Subset Selection

The function trainControl is used to decide the method (cross-validation) and the number of folds to be performed. Next, the train function is used to fit the model. For subset selection the method for fitting the model is "glm". Lastly, the mean test error is calculated by the formula -> (1 - mean(Accuracy)).

```
## [1] "Test error of subset selection is  0.0321611253196931"
```

2. Regularisation method - LASSO

The same method of cross validation is used here like the one used for subset selection except the method parameter in the train function is set to "glmnet". Here we get multiple accuracies as all variables are kept but shrinked to zero as mentioned above while building the LASSO classifier. Hence, the mean of errors is considered.

```
## [1] "Test error of LASSO model is  0.0354101543999242"
```

3. Discriminant Analysis - LDA

The K - fold cross validation of 10 folds is perfomed for LDA similar to that of LASSO and Subset selection except the method used in train function here is "lda". The mean test error is calculated for the model.

```
## [1] "Test error of LDA is  0.0394501278772379"
```

Conclusion - Best Classifier:

This is the final stage of the project where the best classifier is decided based on the performance metrics of all three models.

```
##                         Method Accuracy Error_rate
## 1       Best Subset Selection    96.78       3.22
## 2   Regression Method - LASSO    96.46       3.54
## 3 Discriminant Analysis - LDA    96.05       3.95
```

The metrics fetched for each classifier above are almost close to each other. But, based on the error rate and accuracy, it can be concluded that "Best Subset Selection" is the best classifier. This classifier does not include all predictor variables, it includes only 6 according to our computation. Having less predictor variables can lower the complexity and can give more significant predictions from the model.