Breast Cancer Classification

Vaishnavi Jeurkar

Contents

1 Abstract **1**

2 Data Exploration **2**

3 Fitting a logistic regression model. **5**

4 Best Subset Selection in logistic regression **7**

5 Regularized Logistic regression with Lasso penalty **9**

6 Bayes classifier for Linear Disciminant Analysis **11**

7 Cross validation and motivation **12**

8 Conclusion **13**

9 References **13**

10 Appendix **13**

10.1 Appendix A - Best Subset Selection Outputs 13

10.2 Appendix B - Plots 14

1 Abstract

This study delves into the examination of data collected from 699 women in Wisconsin who underwent a biopsy known as fine needle aspiration cytology (FNAC) to assess breast tissue. Nine characteristics, such as cell size and shape, were measured on a scale of one to ten, indicating cell health. The main objective is to determine if these characteristics alone can accurately classify tissue samples as benign or malignant. Assuming these women represent a random subset experiencing breast cancer symptoms, the project will extensively analyze this dataset. It will involve fitting a logistic regression model using best subset selection and implementing the Lasso penalty method. Additionally, Linear Discriminant Analysis will be employed. The aim is to evaluate the reliability of these characteristics in distinguishing between benign and malignant breast tissue. A successful outcome could significantly impact breast cancer diagnosis, aiding in more informed treatment decisions.

2 Data Exploration

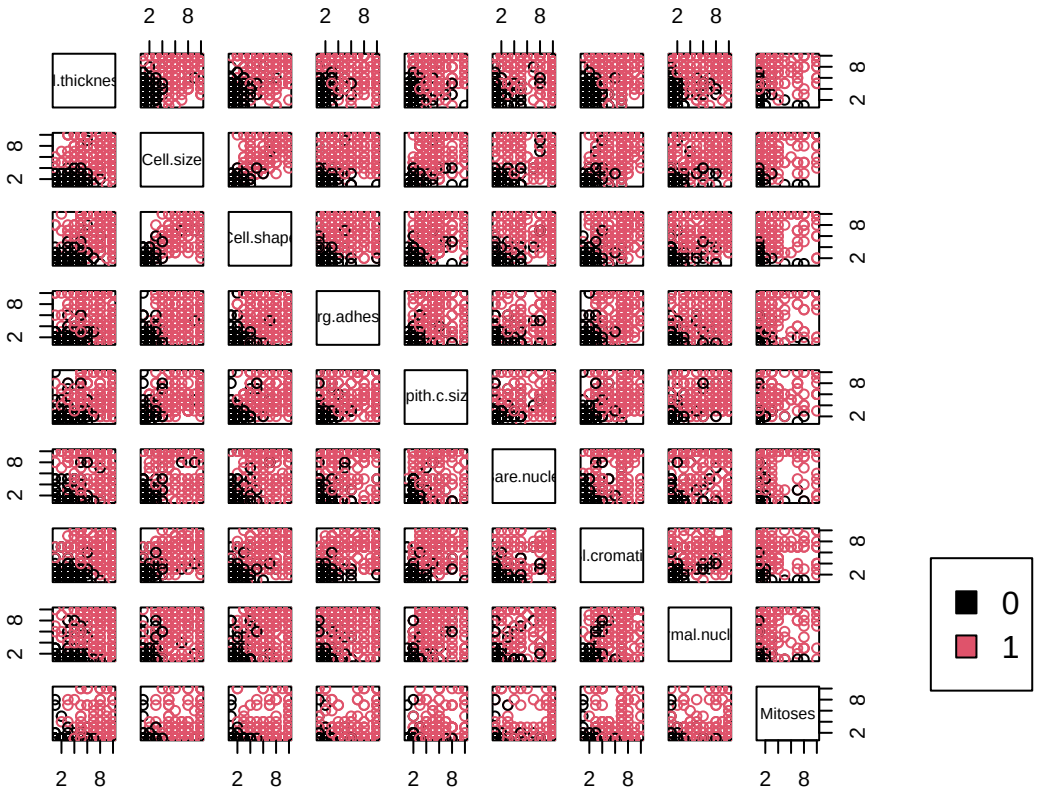
Initially, the data exploration and preparation involved converting the variables from factors to numerical representations. Following this, the class variables were transformed from categorical to numerical, with 'benign' denoted as 0 and 'malignant' as 1. Notably, the dataset contained 16 missing attributes in the 'Bare.Nuclei' column. To address this, the rows with missing attributes were removed. Consequently, the dataset was reduced to 444 observations classified as benign and 239 observations classified as malignant.

2.0.1 Data Summary

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

The summary provides insights into the range, spread, and central tendencies of the predictor variables, showcasing their variability and distribution across the dataset. Features like 'Cl.thickness' exhibits higher means and broader ranges, hinting at potentially significant variability within the dataset. Mitoses has the lowest mean and variability across the dataset.

2.0.2 Scatter plot matrix



The scatterplot matrix reveals a distinct separation between the two classes across response variables, highlighting a clear distinction. However, weaker separations are noticeable in normal.nucleoli, bare.nuclei, marg.adhesion, and epith.c.size, suggesting overlapping values between classes in these specific variables. Notably, a robust positive relationship exists between cell.size and cell.shape, indicating a strong correlation where an increase in one corresponds to an increase in the other. These findings offer valuable insights into the dataset's class separations and interrelationships among predictor variables.

2.0.3 Covariance matrix

```
##                 Cl.thickness Cell.size Cell.shape Marg.adhesion Epith.c.size
## Cl.thickness        7.956694  5.554922   5.508800      3.941776     3.283363
## Cell.size           5.554922  9.395113   8.310604      6.207468     5.134708
## Cell.shape          5.508800  8.310604   8.931615      5.872385     4.799947
## Marg.adhesion       3.941776  6.207468   5.872385      8.205717     3.786179
## Epith.c.size        3.283363  5.134708   4.799947      3.786179     4.942109
## Bare.nuclei         6.096061  7.725660   7.774099      7.000264     4.744656
## Bl.cromatin         3.826365  5.673248   5.383535      4.691541     3.366253
## Normal.nucleoli     4.598758  6.730824   6.550081      5.274024     4.268107
## Mitoses             1.715289  2.447021   2.284936      2.079140     1.851150
##                 Bare.nuclei Bl.cromatin Normal.nucleoli  Mitoses
## Cl.thickness       6.096061    3.826365        4.598758 1.715289
## Cell.size          7.725660    5.673248        6.730824 2.447021
## Cell.shape         7.774099    5.383535        6.550081 2.284936
## Marg.adhesion      7.000264    4.691541        5.274024 2.079140
```

```
## Epith.c.size       4.744656    3.366253        4.268107 1.851150
## Bare.nuclei       13.277695    6.075403        6.499229 2.141645
## Bl.cromatin        6.075403    6.001013        4.977439 1.468652
## Normal.nucleoli    6.499229    4.977439        9.318772 2.294262
## Mitoses            2.141645    1.468652        2.294262 3.002160
```

The covariance matrix provides insights into the relationships between predictor variables in the dataset. Observing the matrix, higher covariance values between variables such as 'Cell.size', 'Cell.shape', and 'Bare.nuclei' indicate stronger positive relationships among these features. This implies that as one of these variables increases, the others tend to increase as well, suggesting potential multicollinearity among them. Conversely, lower covariance values, such as those between 'Cl.thickness', 'Marg.adhesion', 'Epith.c.size', and other variables, suggest weaker relationships or less linear dependency among these particular features. Mitoses has weak positive relationship with all the variables. The diagonal elements in the matrix represent the variables' variances, highlighting the spread or variability of each predictor variable individually.

2.0.4 Total variance

The total variance of 71.03 represents the overall variability or dispersion captured within the dataset by the predictor variables considered. This is useful because it suggests that these features are quite important in understanding the data and could be helpful in understanding whether a sample is benign or malignant in a breast cancer dataset.

2.0.5 Generalized variance

Generalized variance of 53335.79 suggests the overall spread or variability across the dataset, considering all variables together. A higher generalized variance indicates that there's considerable diversity or differences among the data points when taking into account all the variables collectively. This information is valuable as it implies that the dataset covers a wide range of values or patterns across various features.

2.0.6 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
## Class              0.7147899 0.8208014  0.8218909     0.7062941    0.6909582
##                Bare.nuclei Bl.cromatin Normal.nucleoli   Mitoses     Class
## Cl.thickness     0.5930914   0.5537424       0.5340659 0.3509572 0.7147899
## Cell.size        0.6917088   0.7555592       0.7193460 0.4607547 0.8208014
## Cell.shape       0.7138775   0.7353435       0.7179634 0.4412576 0.8218909
## Marg.adhesion    0.6706483   0.6685671       0.6031211 0.4188983 0.7062941
## Epith.c.size     0.5857161   0.6181279       0.6289264 0.4805833 0.6909582
## Bare.nuclei      1.0000000   0.6806149       0.5842802 0.3392104 0.8226959
## Bl.cromatin      0.6806149   1.0000000       0.6656015 0.3460109 0.7582276
## Normal.nucleoli  0.5842802   0.6656015       1.0000000 0.4337573 0.7186772
## Mitoses          0.3392104   0.3460109       0.4337573 1.0000000 0.4234479
```

```
## Class             0.8226959   0.7582276       0.7186772 0.4234479 1.0000000
```

Correlation Between Response and Predictor Variables:

The 'Class' variable demonstrates relatively strong positive correlations with predictor variables 'Cl.thickness', 'Cell.size', 'Cell.shape', 'Marg.adhesion', 'Epith.c.size', 'Bare.nuclei', and 'Bl.cromatin'. These correlations range between 0.71 to 0.82. This suggests that as these predictor variables increase, there tends to be a higher likelihood or association with the 'Class' variable, potentially indicating their importance in predicting whether a sample is benign or malignant. The 'Mitoses' variable has a weaker correlation (0.42) with the 'Class' variable compared to other predictors, implying a relatively less strong relationship in predicting the class.

Correlation Among Predictor Variables:

Among the predictor variables themselves, there are notable strong positive correlations observed between 'Cell.size', 'Cell.shape', 'Bare.nuclei' and 'Bl.cromatin'. These features exhibit correlations ranging from 0.69 to 0.82, suggesting potential multicollinearity among these variables, indicating that changes in one of these variables might be associated with changes in others. Similarly, 'Cell.size' and 'Cell.shape' show a strong positive correlation of approximately 0.91, implying a highly correlated relationship between these two predictors. Similar strong relationships are observed between 'Cell.size' or 'Cell.shape' and 'Bl.cromatin'.

2.0.7 Standard deviation

```
##    Cl.thickness       Cell.size      Cell.shape   Marg.adhesion    Epith.c.size
##        2.820761        3.065145        2.988581        2.864562        2.223085
##     Bare.nuclei     Bl.cromatin Normal.nucleoli         Mitoses
##        3.643857        2.449697        3.052666        1.732674
```

The standard deviation values highlight the dispersion of data points within each predictor variable. Higher standard deviations, such as those observed in 'Normal.nucleoli', 'Bare.nuclei' and 'Cell.size', suggest greater variability in their values across the dataset, indicating a wider spread from their respective means. Conversely, 'Mitoses' exhibits lower variability, with data points clustered closer to its mean.

3 Fitting a logistic regression model.

The dataset underwent a division into two subsets: 80% training set and 20% test set. Both the training and test sets were scaled and a logistic regression model was fit using the glm function.

```
##
## Call:
## glm(formula = y_train ~ ., family = "binomial", data = CancerTrain_data)
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -1.00202    0.36608  -2.737 0.006197 **
## Cl.thickness     1.09545    0.44071   2.486 0.012932 *
## Cell.size        0.50367    0.75026   0.671 0.502018
## Cell.shape       0.81736    0.76142   1.073 0.283064
## Marg.adhesion    0.91998    0.41029   2.242 0.024943 *
## Epith.c.size     0.09204    0.44533   0.207 0.836265
## Bare.nuclei      1.51514    0.38957   3.889 0.000101 ***
## Bl.cromatin      1.39072    0.52504   2.649 0.008078 **
## Normal.nucleoli  0.45675    0.38858   1.175 0.239817
```

```
## Mitoses          0.89052    0.61064   1.458 0.144747
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 708.985  on 546  degrees of freedom
## Residual deviance:  80.026  on 537  degrees of freedom
## AIC: 100.03
##
## Number of Fisher Scoring iterations: 8
```

The maximum likelihood estimates of the regression coefficients are therefore

$\hat{\beta}_0 = -1.002, \hat{\beta}_1 = 1.095, \hat{\beta}_2 = 0.503, \hat{\beta}_3 = 0.817, \hat{\beta}_4 = 0.919, \hat{\beta}_5 = 0.092, \hat{\beta}_6 = 1.515, \hat{\beta}_7 = 1.390, \hat{\beta}_8 = 0.456, \hat{\beta}_9 = 0.890$

The p-value for Cl.thickness, Marg.adhesion, Bare.nuclei and Bl.cromatin is less than 0.05 If we examine the table produced by the summary function we see that a number of the variables have very large p-values meaning that, individually, they contribute very little to a model which contains all the other predictors. Inclusion of more predictors than are necessary can inflate the variance of the parameter estimators leading to a deterioration in predictive performance.

3.0.1 Cross-Validation

Cross validation on all the models is performed using Validation set approach.

```
##          Predicted
## Observed   0   1
##        0 348   7
##        1   7 185
```

From the confusion matrix we can see that benign is correctly predicted 348 times out of 348 + 7 = 355 observations. The model correctly identified malignant cancer 185 times. The training error is the proportion of misclassified observations

```
## [1] "Training error for logistic regression is: "
```

```
## [1] 0.02559415
```

i.e. around 2.56% of training error is seen.

```
##          Predicted
## Observed  0  1
##        0 87  2
##        1  4 43
```

The model trained using training data is applied on the test data and we get the corresponding confusion matrix. This model is able to predict benign cancer 87 out of 89 times and it correctly identifies malignant cancer 43 out of 47 observations.

```
## [1] "Test error for logistic regression is: "
```

```
## [1] 0.04411765
```

The test error is around 4.4% which is higher than the training error. When the test error is higher than the training error, it suggests that the model might be overfitting the training data. Factors contributing to overfitting include excessive model complexity, using too many features relative to the amount of data, or inadequate regularization techniques. In contrast, if the test error is close to the training error or even lower, it indicates that the model has successfully captured the underlying patterns and can generalize well to new data, showcasing its robustness. Regularization methods, cross-validation, or reducing model complexity are strategies commonly employed to address overfitting and mitigate the disparity between training and test errors.

4 Best Subset Selection in logistic regression

In the earlier model it is observed that some of the features do not have any significant effect on the model's output. Therefore to find the optimal model we apply different feature selection techniques.
We can apply best subset selection using AIC and BIC using the bestglm package.

By construction the implied models $M_0, M_1, \ldots, M_p$ are same in both AIC and BIC, (See Appendix A). Only difference is the final column named AIC and BIC. The model minimising AIC and BIC are starred in each case. As for logistic regression, different criteria often suggest different models are "best", and this is the case here. In order to reconcile the differences and choose a single "best" model we generate a plot to show how the criteria vary with the number of predictors. (See Appendix B Figure 1)

AIC: Tends to select larger models that might fit the data better but could be more complex.
BIC: Penalizes complexity more than AIC and often selects smaller models compared to AIC.
Here AIC has selected best model with 7 predictors and BIC suggests 5 predictor model to be the best.
In order to identify the best model we will train both models and compare their test errors.

4.0.1 Best subset selection with BIC

```
##
## Call:
## glm(formula = y_train ~ ., family = "binomial", data = Cancer_data_red_BIC)
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -1.0505     0.3242  -3.241 0.001193 **
## Cl.thickness    1.4338     0.4116   3.483 0.000495 ***
## Cell.size       1.5398     0.5093   3.023 0.002502 **
## Marg.adhesion   1.0085     0.3907   2.581 0.009847 **
## Bare.nuclei     1.6313     0.3701   4.408 1.04e-05 ***
## Bl.cromatin     1.4802     0.4845   3.055 0.002250 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 708.985  on 546  degrees of freedom
## Residual deviance:  86.171  on 541  degrees of freedom
## AIC: 98.171
##
## Number of Fisher Scoring iterations: 8
```

The maximum likelihood estimates of the regression coefficients are

$\hat{\beta}_0 = -1.05, \hat{\beta}_1 = 1.433, \hat{\beta}_2 = 1.539, \hat{\beta}_3 = 1.008, \hat{\beta}_4 = 1.631, \hat{\beta}_5 = 1.480$

The model summary clearly indicates a robust association between the predictor and response variables. Each variable exhibits positive coefficients, signifying a positive relationship. Additionally, all variables demonstrate p-values below 0.05, indicating a strong statistical significance and reinforcing the presence of a compelling positive correlation among the variables.

This model has selected Cl.thickness, Cell.size, Marg.adhesion, Bare.nuclei and Bl.cromatin variables and rest all are dropped from the model. These variables showed strong positive correlation with Class variable in the earlier correlation matrix. 4 of the variables except Cell.size had p-values less than 0.05 in earlier simple logistic regression model.

4.0.2 Test error

```
## [1] "Confusion matrix of subset selection with BIC"
```

```
##          Predicted
## Observed  0  1
##        0 87  2
##        1  3 44
```

```
## [1] "Test error for best subset selection with BIC is: "
```

```
## [1] 0.03676471
```

The test error for best subset selection with BIC is 3.67%. This error is less compared to the first regression model.

4.0.3 Best subset selection with AIC

```
##
## Call:
## glm(formula = y_train ~ ., family = "binomial", data = Cancer_data_red_AIC)
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -1.0198     0.3613  -2.822  0.00477 **
## Cl.thickness      1.1564     0.4304   2.687  0.00722 **
## Cell.shape        1.1768     0.5564   2.115  0.03441 *
## Marg.adhesion     0.9468     0.3898   2.429  0.01515 *
## Bare.nuclei       1.5384     0.3773   4.077 4.56e-05 ***
## Bl.cromatin       1.4378     0.4996   2.878  0.00400 **
## Normal.nucleoli   0.5440     0.3717   1.464  0.14330
## Mitoses           0.9230     0.6115   1.509  0.13120
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 708.985  on 546  degrees of freedom
## Residual deviance:  80.622  on 539  degrees of freedom
```

```
## AIC: 96.622
##
## Number of Fisher Scoring iterations: 8
```

The maximum likelihood estimates of the regression coefficients are

$\hat{\beta}_0 = -1.019, \hat{\beta}_1 = 1.156, \hat{\beta}_2 = 1.176, \hat{\beta}_3 = 0.946, \hat{\beta}_4 = 1.538, \hat{\beta}_5 = 1.437, \hat{\beta}_5 = 0.5444, \hat{\beta}_5 = 0.923$

The model summary indicates a association between the predictor and response variables. Each variable exhibits positive coefficients, signifying a positive relationship. All variables except Normal.nucleoli and Mitoses demonstrate p-values below 0.05, indicating a strong statistical significance and reinforcing the presence of a compelling positive correlation among the variables.

This model has additional features, Normal.nucleoli and mitoses with the features Cl.thickness, Cell.size, Marg.adhesion, Bare.nuclei and Bl.cromatin that were present in BIC and rest all are dropped from the model.

4.0.4 Test error

```
##          Predicted
## Observed  0  1
##        0 87  2
##        1  4 43
```

```
## [1] "Test error for best subset selection with AIC is: "
```

```
## [1] 0.04411765
```

The test error of best subset selection with AIC is same as the logistic regression model. This suggests that for BreastCancer data a model fitted with 5 features is better than 7 feature model.

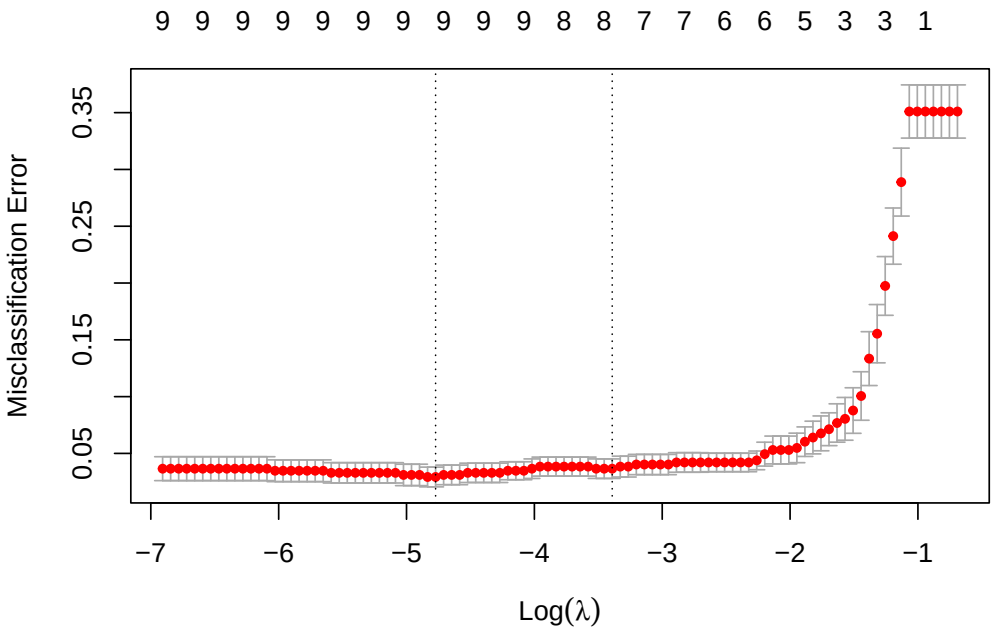
5 Regularized Logistic regression with Lasso penalty

The method operates by introducing a penalty, which is scaled by a tuning parameter, into the loss function. This modified loss function, in logistic regression, represents the negative logarithm of the likelihood function. In R, Lasso can be implemented using the 'glmnet' package.

Lasso Penalty = $\lambda \sum_{j=1}^{p} |\beta_j|$
λ represents the regularization parameter.
p denotes the number of predictors or coefficients in the model.
β_j signifies the coefficients associated with each predictor in the model.

Loss function = SSe + $\lambda \sum_{j=1}^{p} |\beta_j|$

We can use the plot function to examine how the coefficients of each variable change as the tuning parameter is increased, (See Appendix B Figure 2). In this plot we can see that as the LASSO performs variable selection, in addition to shrinkage, we see variables drop from the model as the tuning parameter increases. Each line represents the regression coefficient for a different variable. First variable to drop is mitoses followed by Epith.c.size. The last variable to drop is cell.shape.



The regression coefficients obtained by performing the LASSO with the chosen value of lambda are:

```
## 10 x 1 sparse Matrix of class "dgCMatrix"
##                          s1
## (Intercept)     -1.0040075
## Cl.thickness     0.8698421
## Cell.size        0.5183630
## Cell.shape       0.6819046
## Marg.adhesion    0.4613698
## Epith.c.size     0.1023056
## Bare.nuclei      1.2511103
## Bl.cromatin      0.8237321
## Normal.nucleoli  0.3699991
## Mitoses          0.2181063
```

At the optimal solution none of the variables drop out of the model.

5.0.1 Training error

```
##          Predicted
## Observed   0   1
##        0 348   7
##        1   7 185
```

```
## [1] "Training error for logistic regression with Lasso is: "
```

```
## [1] 0.02559415
```

The training error of regularized logistic regression with Lasso penalty is same as logistic regression model fitted without penalty

5.0.2 Test error

```
##         Predicted
## Observed  0  1
##        0 87  2
##        1  5 42
```

```
## [1] "Test error for logistic regression with Lasso is: "
```

```
## [1] 0.05147059
```

The test error (5.1%) is slightly higher for the model fitted with the LASSO penalty. Therefore of the two models, it seems that the model fitted without penalty performs better, based on this particular partition of the data into training and validation sets.

6 Bayes classifier for Linear Disciminant Analysis

All the variables have been used in the LDA model.

```
## Call:
## lda(y_train ~ ., data = data.frame(X_train))
##
## Prior probabilities of groups:
##         0         1
## 0.6489945 0.3510055
##
## Group means:
##   Cl.thickness  Cell.size Cell.shape Marg.adhesion Epith.c.size Bare.nuclei
## 0   -0.5127239 -0.6107099 -0.6119085    -0.5198588   -0.5104219  -0.6057326
## 1    0.9480052  1.1291771  1.1313933     0.9611973    0.9437489   1.1199743
##   Bl.cromatin Normal.nucleoli    Mitoses
## 0  -0.5598119      -0.5316167 -0.3250101
## 1   1.0350689       0.9829372  0.6009302
##
## Coefficients of linear discriminants:
##                         LD1
## Cl.thickness     0.44530128
## Cell.size        0.45703871
## Cell.shape       0.28054047
## Marg.adhesion    0.12117563
## Epith.c.size     0.10777290
## Bare.nuclei      0.97393852
## Bl.cromatin      0.29214794
## Normal.nucleoli  0.32313497
## Mitoses         -0.03166195
```

Above model shows, Prior probabilities of groups:
64.89% belongs to benign cancer and 35.10% belongs to malignant cancer.

Group means
It shows the class wise average (standardised) values for each predictor variables. This helps in comparing

how the average values of variables varies between two class. A large difference in average values suggests good seperation between the classes.

Coefficients of linear discriminants: The discriminant function is a linear combination of 9 variables.

$0.445 \times \text{Cl.thickness} + 0.457 \times \text{Cell.size} + 0.280 \times \text{Cell.shape} + 0.121 \times \text{Marg.adhesion} + 0.107 \times \text{Epith.c.size} + 0.973 \times \text{Bare.nuclei} + 0.292 \times \text{Bl.cromatin} + 0.323 \times \text{Normal.nucleoli} - 0.031 \times \text{Mitoses}$

6.0.1 Training error

```
##         Predicted
## Observed   0   1
##        0 349   6
##        1  12 180
```

```
## [1] "Training error for logistic regression with LDA is: "
```

```
## [1] 0.03290676
```

The training error for LDA is higher than the model fitted with Lasso penalty. There have been 12 instances where the model incorrectly classified benign cases as malignant.

6.0.2 Test error

```
##         Predicted
## Observed  0  1
##        0 87  2
##        1  7 40
```

```
## [1] "Test error for logistic regression with LDA is: "
```

```
## [1] 0.06617647
```

The test error for the linear discriminant analysis model is 6.6% which is highest among all the methods implemented on the Breast Cancer dataset.

7 Cross validation and motivation

The cross validation method used in this analysis is validation set approach. This is one of the most basic and simple techniques for evaluating a model. This simplicity is beneficial in scenarios where rapid model prototyping or quick iterations are necessary. This approach makes the comparison fair as same datasets are used for training and testing for all the models implemented. The validation set provides a single performance estimate for the model, allowing for a straightforward evaluation of its generalization capability. It gives a clearer picture of how the model might perform on unseen data. Given size of data was sufficient to perform cross validation using this method.

Comparing the performance of different models using cross validation based on the test error helps in evaluating the performance of each model on unseen data. Test error of logistic regression and best subset selection with AIC is 4.4%. Test error for logistic regression with Lasso penalty is 5.1%. linear discriminant analysis model has the highest test error i.e. 6.6% and best subset selection model with BIC least test error 3.6% among all the models.

8 Conclusion

Among the five variants of logistic regression applied to the Breast Cancer dataset to discern the nature of the cancer (benign or malignant), the model employing the best subset selection method using BIC (Bayesian Information Criterion) demonstrated superior performance. This particular model exhibited an error rate of 3.6%, signifying its accuracy in prediction.

This selected logistic regression model comprises five predictor variables: Cl.thickness, Cell.size, Marg.adhesion, Bare.nuclei, and Bl.cromatin. These variables showcase a notably strong positive correlation with the target class variable. Moreover, they exhibit statistical significance with p-values less than 0.05, further affirming their relevance in the prediction process.

Including more than five variables in the logistic regression model, particularly employing methods such as best subset selection with AIC (Akaike Information Criterion) or utilizing all variables in methods like Lasso or LDA (Linear Discriminant Analysis), results in a higher error rate. This indicates that the additional variables beyond the optimal subset or the complete set of variables do not significantly contribute to improving the predictive capability of the model.

These supplementary variables, when included in the model, do not provide substantial additional information relevant to the prediction of cancer type (benign or malignant). Consequently, their inclusion tends to introduce noise or irrelevant information, resulting in an increased error rate without a corresponding improvement in predictive accuracy. Therefore, the optimal model performance is achieved when considering a limited set of five predictor variables that demonstrate strong associations with the target class variable while maintaining statistical significance and minimizing the error rate.

9 References

1. https://www.geeksforgeeks.org/cross-validation-in-r-programming/
2. https://www.geeksforgeeks.org/convert-factor-to-numeric-and-numeric-to-factor-in-r-programming/
3. https://rpubs.com/Subhalaxmi/742119
4. https://bookdown.org/yihui/rmarkdown-cookbook/cross-ref.html

10 Appendix

10.1 Appendix A - Best Subset Selection Outputs

1. **AIC Subsets**

```
##    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      TRUE      FALSE         FALSE        FALSE
## 5       TRUE         TRUE      TRUE      FALSE          TRUE        FALSE
## 6       TRUE         TRUE     FALSE       TRUE          TRUE        FALSE
## 7*      TRUE         TRUE     FALSE       TRUE          TRUE        FALSE
## 8       TRUE         TRUE      TRUE       TRUE          TRUE        FALSE
## 9       TRUE         TRUE      TRUE       TRUE          TRUE         TRUE
##    Bare.nuclei Bl.cromatin Normal.nucleoli Mitoses logLikelihood       AIC
## 0        FALSE       FALSE           FALSE   FALSE    -354.49257 708.98514
## 1        FALSE       FALSE           FALSE   FALSE     -91.55415 185.10830
```

```
## 2         TRUE       FALSE           FALSE   FALSE     -60.82666 125.65332
## 3         TRUE       FALSE           FALSE   FALSE     -52.36195 110.72390
## 4         TRUE        TRUE           FALSE   FALSE     -46.59161 101.18323
## 5         TRUE        TRUE           FALSE   FALSE     -43.08541  96.17083
## 6         TRUE        TRUE           FALSE    TRUE     -41.44240  94.88481
## 7*        TRUE        TRUE            TRUE    TRUE     -40.31109  94.62218
## 8         TRUE        TRUE            TRUE    TRUE     -40.03435  96.06869
## 9         TRUE        TRUE            TRUE    TRUE     -40.01295  98.02591
```

2. **BIC Subsets**

```
##    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      TRUE      FALSE         FALSE        FALSE
## 5*      TRUE         TRUE      TRUE      FALSE          TRUE        FALSE
## 6       TRUE         TRUE     FALSE       TRUE          TRUE        FALSE
## 7       TRUE         TRUE     FALSE       TRUE          TRUE        FALSE
## 8       TRUE         TRUE      TRUE       TRUE          TRUE        FALSE
## 9       TRUE         TRUE      TRUE       TRUE          TRUE         TRUE
##    Bare.nuclei Bl.cromatin Normal.nucleoli Mitoses logLikelihood      BIC
## 0        FALSE       FALSE           FALSE   FALSE    -354.49257 708.9851
## 1        FALSE       FALSE           FALSE   FALSE     -91.55415 189.4128
## 2         TRUE       FALSE           FALSE   FALSE     -60.82666 134.2622
## 3         TRUE       FALSE           FALSE   FALSE     -52.36195 123.6372
## 4         TRUE        TRUE           FALSE   FALSE     -46.59161 118.4010
## 5*        TRUE        TRUE           FALSE   FALSE     -43.08541 117.6931
## 6         TRUE        TRUE           FALSE    TRUE     -41.44240 120.7115
## 7         TRUE        TRUE            TRUE    TRUE     -40.31109 124.7533
## 8         TRUE        TRUE            TRUE    TRUE     -40.03435 130.5043
## 9         TRUE        TRUE            TRUE    TRUE     -40.01295 136.7659
```

10.2 Appendix B - Plots

1. **Figure 1**

2. **Figure 2**

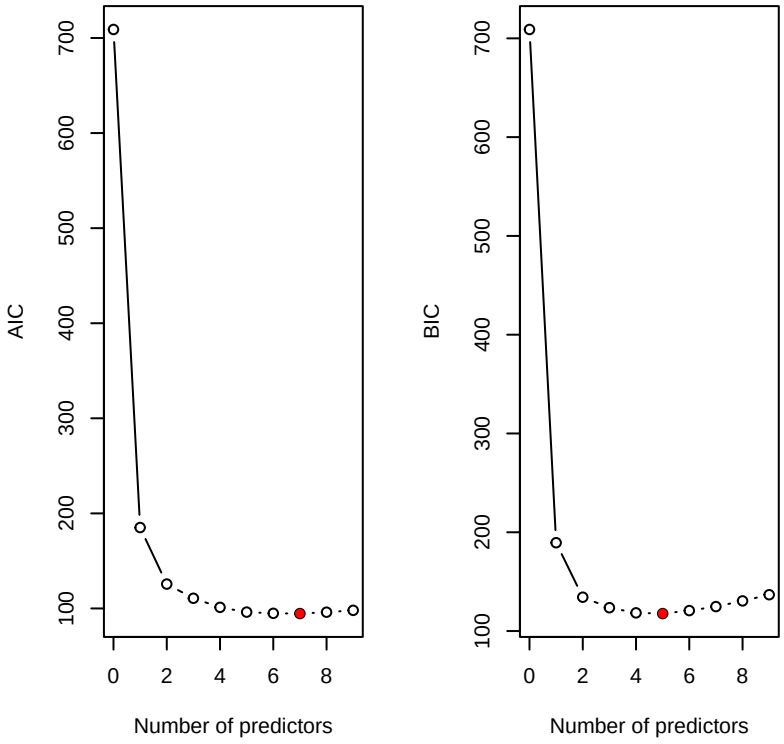


Figure 1: Best Subset selection for Cancer data

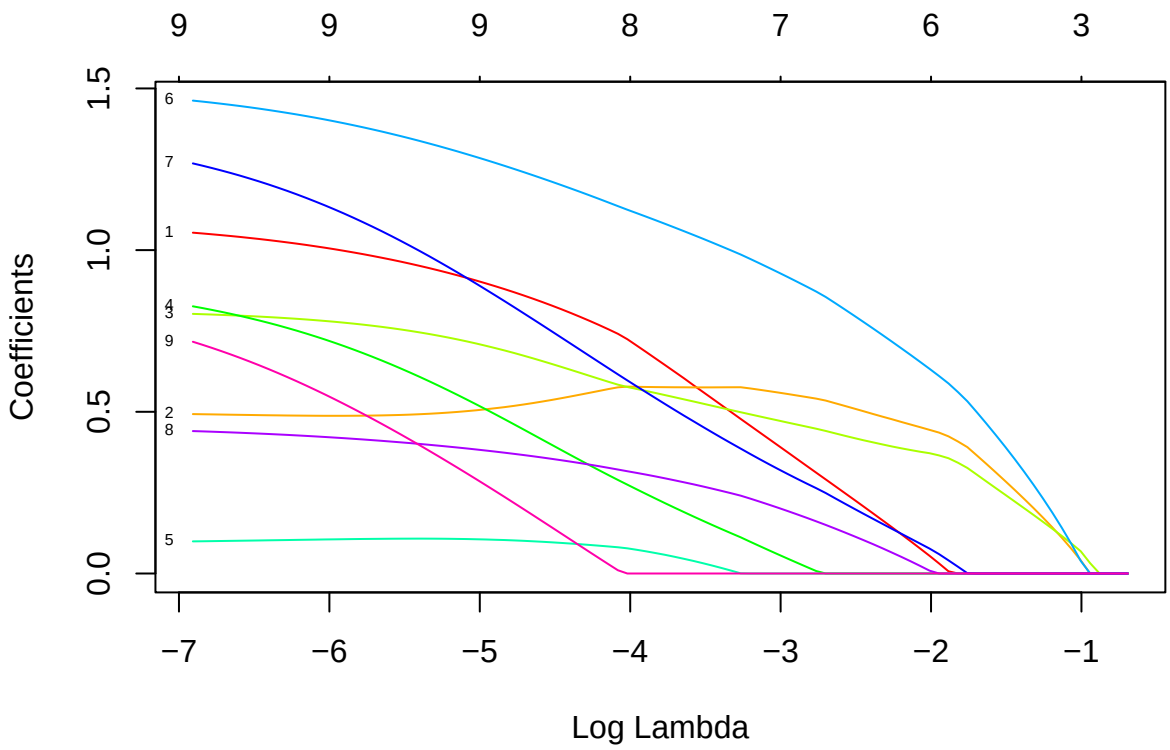


Figure 2: The effect of varying the tuning parameter in the logistic regression model with LASSO penalty for the Weekly data.