MAS8404: Breast Cancer Analysis Report

Sandra M Nino Arbelaez

2023-12-01

Exploratory Data Analysis

For the Breast Cancer data, we define a malignant sample ("no-benign") as the negative class labelled with 0, and a sample is benign as the positive class labelled with 1. In this way, we are interested in the distribution of the *Class* variable, which is the response variable, to understand how balanced is our data.

```
##
##   0   1
## 239 444
```

This distribution shows that 444 of the 683 samples were benign and 239 of the 683 samples were malignant. This indicates that our data is not completely balanced in the number of cases for the response variable.

To understand the relationships and strength of linear associations between predictors, we are going to produce the Sample Covariance Matrix and the Sample Correlation Matrix of the original data.

By doing the Sample Covariance Matrix, we can compute the total variation of the data, which is:

```
## [1] 70.706
```

Also, the generalised variance of the data is:

```
## [1] 47458.44
```

We have very large numbers for the total variation and the generalised variance, which means that our data is scattered; the observations are far away from the centre.

Even though, the Sample Covariance Matrix gives us information about the variance of the predictor variables, we are more interested in understanding the strength of the relationships between these variables. Therefore, the Sample Correlation Matrix of the original data gives useful insights into how they correlate with each other.

In Figure 1, we can see that all the predictor variables are highly correlated to each other. There is an extremely strong relationship between the Uniformity of Cell Shape and the Uniformity of Cell Size with a correlation value of 0.91. The variable Mitoses has a less strong relationship with all the other predictors with values below 0.5. Having highly correlated variables might not improve the performance of our models. We will take this into consideration for further conclusions.

Now, we want to see the relationship between some predictor variables with the response variable. This analysis will be limited to three variables due to page limit, but the findings can be generalised for the other explanatory variables.

In Figure 2, we can see that the distribution for Benign samples (encoded as 1) is positively skewed for all three variables, this means that the majority of the values are stacked in the smaller numbers, and then greater values decrease or are absent. On the opposite, the distribution for Malignant cells (encoded as 0) is left skewed for all three variables, we can see that small values for each predictor are less likely and greater values such as 10 are predominant.

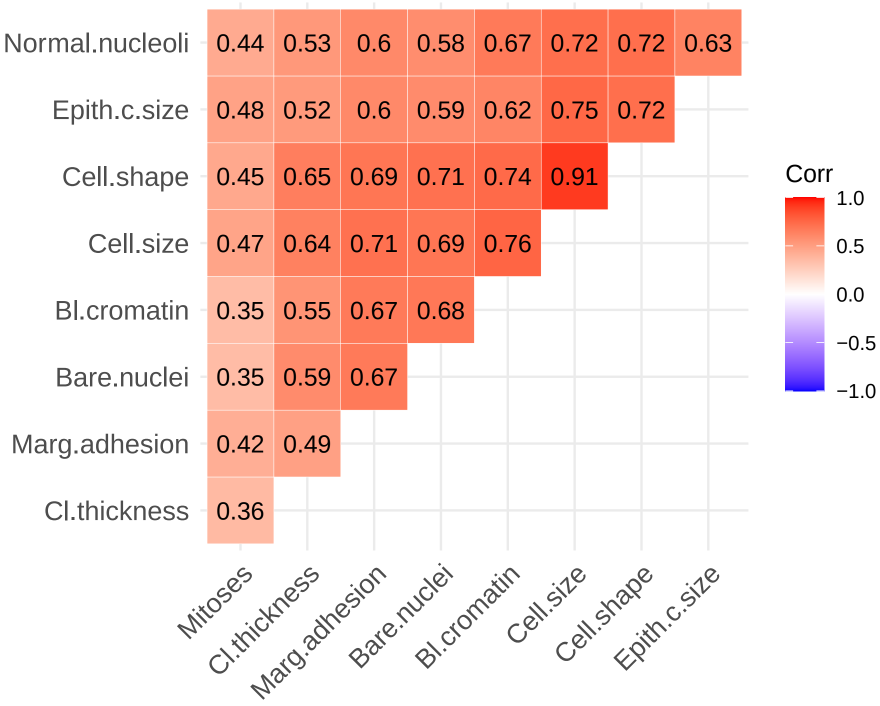

Figure 1: Sample Correlation Matrix of the original data

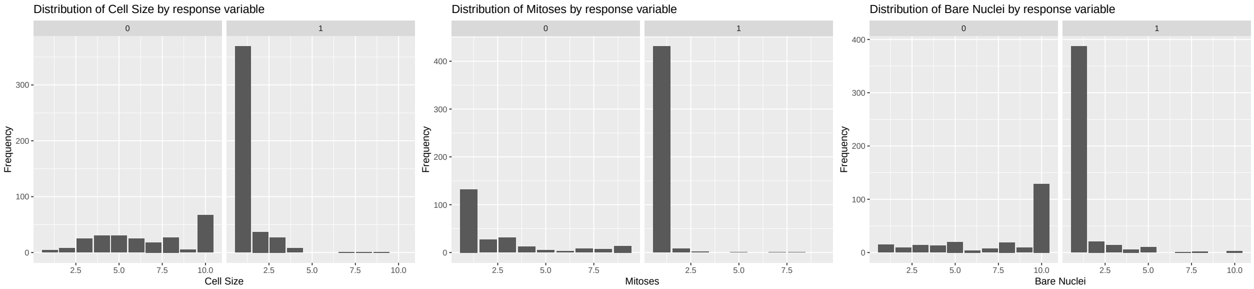

Figure 2: Distribution of predictor variables by the response variable (0 = Malignant, 1 = Beningn) for the Breast Cancer Data

In the Breast Cancer data, all nine predictor variables are ordered categorical variables, which, for breast cancer, all have 11 possible levels labelled from 0 through 10, with larger values indicating that a cell is malignant. Therefore, we can treat all the variables as quantitative variables. This will help us to build simpler models because we have only a single explanatory variable representing the effect of each one and not 11.

Logistic Regression

We want to predict a categorical variable which has two possible values: Benign and Malignant. Then, we will build a Logistic Regression classifier. This is an extension of the linear regression model and it "covers the situation where the response variable is a binary variable" (Varzani, 2014). That means it just distinguishes between two classes. It uses a sigmoid function (the cumulative distribution function of the logistic distribution) which it is S-shaped, so it transforms the input values between 0's and 1's, then there is a cut-off threshold to classify the output into one class or the other.

Before fitting our models, we will transform our data to have all the variables in the same scale. This is because our predictors can be measured on different scales.

We will fit a logistic regression model with the nine predictor variables.

```
##
## Call:
## glm(formula = y ~ ., family = "binomial", data = BreastCancer_scaled)
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)    
## (Intercept)      1.10357    0.32011   3.448 0.000566 ***
## Cl.thickness    -1.50983    0.40037  -3.771 0.000163 ***
## Cell.size        0.01822    0.64110   0.028 0.977332    
## Cell.shape      -0.96273    0.68930  -1.397 0.162510    
## Marg.adhesion   -0.94729    0.35366  -2.679 0.007395 ** 
## Epith.c.size    -0.21519    0.34806  -0.618 0.536415    
## Bare.nuclei     -1.39565    0.34203  -4.080 4.49e-05 ***
## Bl.cromatin     -1.09600    0.41986  -2.610 0.009044 ** 
## Normal.nucleoli -0.65044    0.34463  -1.887 0.059109 .  
## Mitoses         -0.88124    0.53281  -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:

$\hat{\beta}_0 = 1.104$, $\hat{\beta}_1 = -1.509$, $\hat{\beta}_2 = 0.018$, $\hat{\beta}_3 = -0.963$, $\hat{\beta}_4 = -0.947$, \ $\hat{\beta}_5 = -0.215$, $\hat{\beta}_6 = -1.396$, $\hat{\beta}_7 = -1.096$, $\hat{\beta}_8 = -0.650$, $\hat{\beta}_9 = -0.881$

By analising the coefficients, the smallest p-value is associated with Bare Nuclei. The negative coefficient for this predictor suggests that if the cell is benign, it is less likely to have Bare Nuclei. If we review in detail, all the coefficients of the predictor variables are negative, except Uniformity of Cell Size. Therefore, the negative coefficients mean the same: if the cell is benign, it is less likely to have a large value of that variable. For

Uniformity of Cell Size, the coefficient is positive which means that a large number in the Uniformity of the Cell Size, the sample is quite likely to be benign.

By reviewing the results, we can see that the t-test H_0: $\beta_i = 0$ vs H_1: $\beta_i \neq 0$ for the variables Cell Size and Single Epithelial Cell Size have very large p-values. This suggests that individually they contribute very little to a model which contains all the other eight predictors. In supervised learning models, when we include more predictors than are necessary might lead to a deterioration in predictive performance.

Consequently, we will proceed to do a subset selection in which we will not include all the predictors in our model. We will consider the best subset selection method in which we identify a "good" subset of explanatory variables. This method reduces the variance and can improve the predictive performance.

Best Subset Selection

Subset selection is a technique to identify a "good" subset of $p^* < p$ explanatory variables to fit a model on these p^* variables. The best subset selection is one of the methods for selecting a subset of the explanatory variables. This method involves fitting a logistic regression model to each possible subset of the p variables. In our case, it would be $2^9 = 512$ possible models. Then, we compare all of them to decide which one is the best. We select this method because we do not have a large number of predictor variables. However, it has a computational limitation that is, if we have a large number of p variables, the best subset selection method becomes costly in terms of computation for values of p greater than around 40.

Another subset selection method is the automated (stepwise) selection. This is more useful with large numbers of p because it is more efficient computationally compared to the best subset selection. However, we have a small number of predictor variables, just 9, so the computational cost is not a big deal. Therefore, we will use the best selection method over automated (stepwise) selection.

First of all, we will apply the best subset selection with the AIC and BIC criteria and then we will compare which is the best-fitting model containing the suggested predictors. By construction, the implied models M_0, M_1, …, M_p are the same for both criteria. However, the models minimising the AIC and BIC are starred in each case, suggesting their the best.

```
## Morgan-Tatar search since family is non-gaussian.
## Morgan-Tatar search since family is non-gaussian.
```

```
##    Intercept Cl.thickness Cell.size Cell.shape Marg.adhesion Epith.c.size
## 7*      TRUE         TRUE     FALSE       TRUE          TRUE        FALSE
##    Bare.nuclei Bl.cromatin Normal.nucleoli Mitoses logLikelihood    AIC
## 7*        TRUE        TRUE            TRUE    TRUE     -51.63998 117.28
```

```
##    Intercept Cl.thickness Cell.size Cell.shape Marg.adhesion Epith.c.size
## 5*      TRUE         TRUE     FALSE      FALSE          TRUE        FALSE
##    Bare.nuclei Bl.cromatin Normal.nucleoli Mitoses logLikelihood     BIC
## 5*        TRUE        TRUE            TRUE   FALSE     -56.13177 144.896
```

The best model based on the AIC as the information criterion is number 7 with the smallest value of 117.280. This model includes 7 predictor variables which are: Mitoses, Normal Nucleoli, Bland Chromatin, Bare Nuclei, Marginal Adhesion, Cell shape and Clump thickness.

The best model based on the BIC as the information criterion is number 5 with the smallest value of 144.896. This model includes 5 predictor variables which are: Normal Nucleoli, Bland Chromatin, Bare Nuclei, Marginal Adhesion, and Clump thickness.

As we demonstrate in the last paragraphs, different criteria suggest different models are "best". Thus, to help us to decide which is the "best" model, we will plot how both criteria vary with the different explanatory variables.

If we examine both plots in Figure 3, they suggest there is little difference between M_5 and M_7. Therefore, we might choose the best model as M_5 with few predictors.

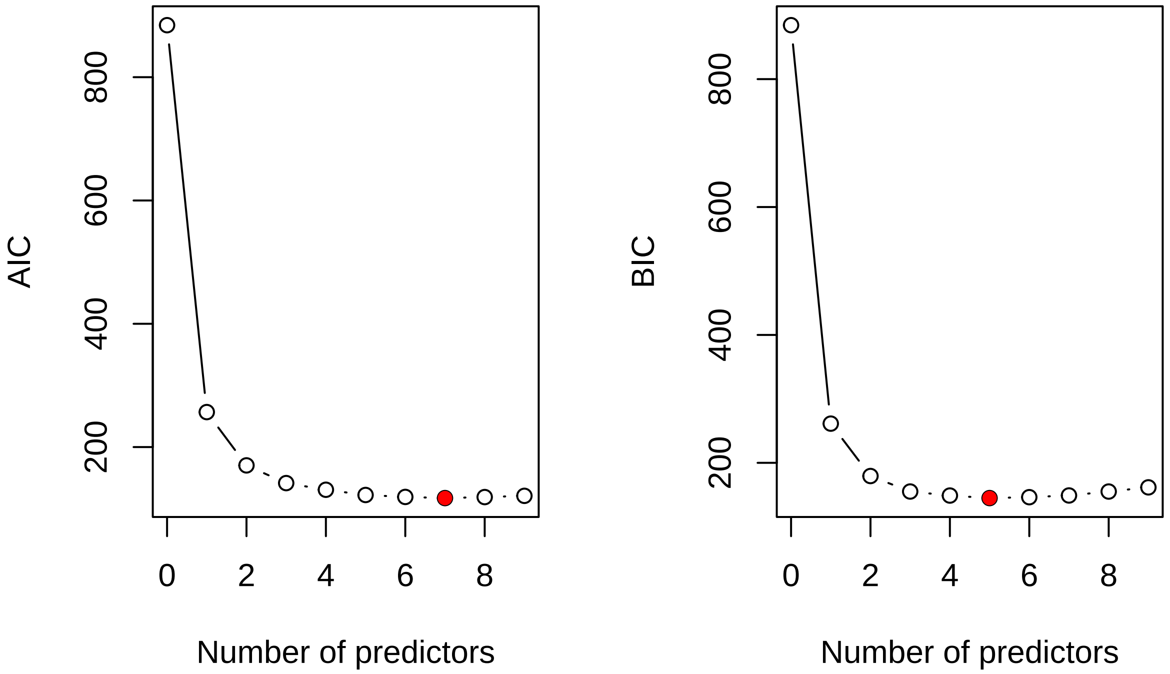

Figure 3: Best subset selection for the Breast Cancer data. Left: AIC criterion, Right: BIC criterion

Now, we are going to create a new data set with only the five selected predictors and then fit our model again.

```
##
## Call:
## glm(formula = y ~ ., family = "binomial", data = BreastData_red)
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)        1.2700     0.2825   4.495 6.97e-06 ***
## Cl.thickness      -2.0910     0.3720  -5.621 1.90e-08 ***
## Marg.adhesion     -1.1319     0.3321  -3.409 0.000652 ***
## Bare.nuclei       -1.6300     0.3206  -5.085 3.68e-07 ***
## Bl.cromatin       -1.3544     0.3679  -3.681 0.000232 ***
## Normal.nucleoli  -1.0202     0.2986  -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 new coefficients for the model with 5 predictor variables are:

$\hat{\beta}_0 = 1.270$, $\hat{\beta}_1 = -2.091$, $\hat{\beta}_2 = -1.132$, $\hat{\beta}_3 = -1.630$, $\hat{\beta}_4 = -1.354$, $\hat{\beta}_5 = -1.020$

The negative coefficients mean the same: if the cell is benign, it is less likely to have a large value of that variable.

Cross Validation - Logistic Regression

Previously, we compared the models M_0, M_1, …, M_p using AIC and BIC criteria. However, an alternative approach is to use cross-validation.

Cross-validation is an approach to assess the predictive performance of our models. We divide the data into k folds of approximately equal size. The last $k-1$ folds are used as the training data and the first fold is used as the test set. Then the process is repeated by changing the second fold as the test set, then the third and so on. In the end, to compute the overall test error is the average of the k estimates of the test error for each iteration. This procedure helps in assessing how well the classifiers generalise to unseen data without holding out a completely different portion of our original data as the test set.

In this section, we will use cross-validation to decide which model is the best for the subset selection method. Additionally, we will assess the chosen best model performance for further comparison with the next two models that will be built in this analysis. To make the comparison fair and mitigate biases, we will use the same set of folds for all three models (best subset selection, regularisation method and discriminant analysis). Moreover, in our cross-validation process, we will compute the test error and four evaluation metrics that will be explained later.

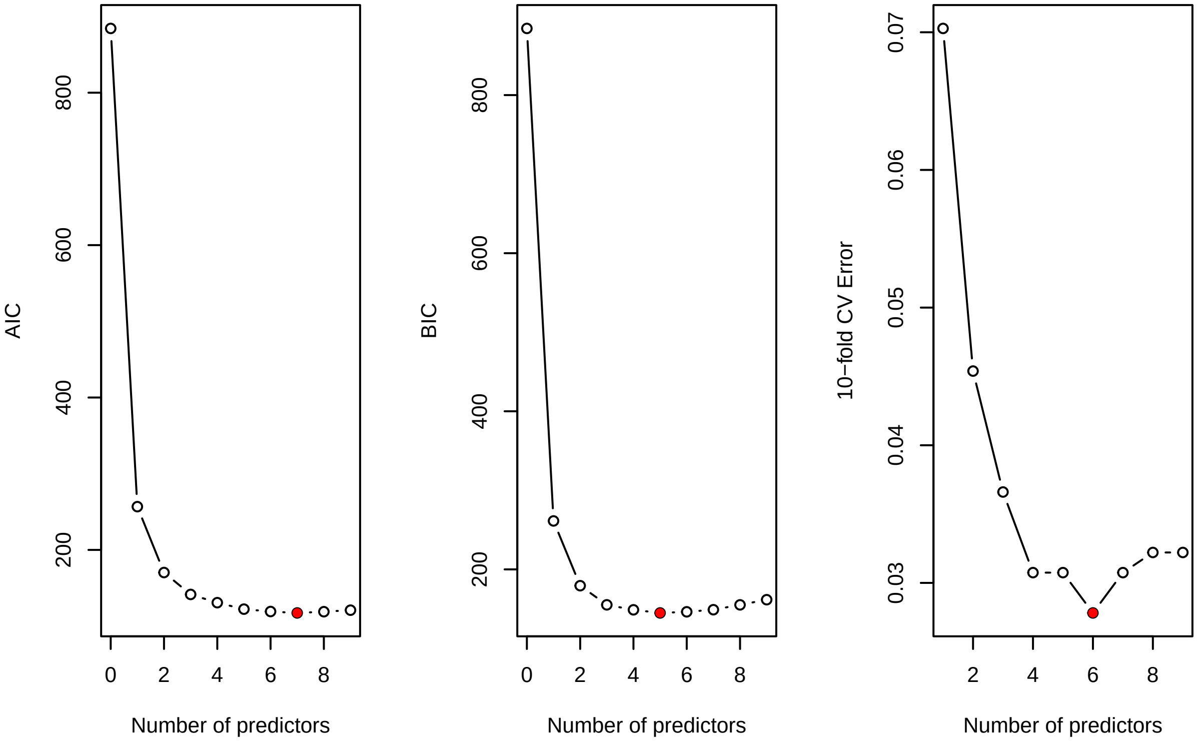

Figure 4: Best subset selection for the Breast Cancer data. Left: AIC criterion, Center: BIC criterion, Right: 10-fold Cross validation approach

In Figure 4, we can see that by doing the 10-fold cross-validation, the best model is with 6 predictors. Therefore, we are going to select this number of predictors and run again our model with this subset of variables and save the test error and evaluation metrics to compare with further models that are going to be built in this analysis.

```
##
## Call:
## glm(formula = y ~ ., family = "binomial", data = BreastCancer_red)
##
## Coefficients:
```

```
##                  Estimate Std. Error z value Pr(>|z|)    
## (Intercept)        1.2592     0.2904   4.337 1.45e-05 ***
## Cl.thickness      -1.7560     0.3868  -4.540 5.62e-06 ***
## Cell.shape        -1.0445     0.4932  -2.118  0.03419 *  
## Marg.adhesion     -0.9669     0.3312  -2.920  0.00350 ** 
## Bare.nuclei       -1.3794     0.3418  -4.035 5.45e-05 ***
## Bl.cromatin       -1.1546     0.4069  -2.837  0.00455 ** 
## Normal.nucleoli  -0.7423     0.3314  -2.240  0.02509 *  
## ---
## 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: 107.14  on 676  degrees of freedom
## AIC: 121.14
## 
## Number of Fisher Scoring iterations: 8
```

The new coefficients for the model with 6 predictor variables are:

$\hat{\beta}_0 = 1.259$, $\hat{\beta}_1 = -1.756$, $\hat{\beta}_2 = -1.045$, $\hat{\beta}_3 = -0.967$, $\hat{\beta}_4 = -1.379$, \ $\hat{\beta}_5 = -1.155$, $\hat{\beta}_6 = -0.742$

In this model with six predictors, we can see that the predictors, Cell size and Single Epithelial Cell Size, which had very large p-values in the initial model with all the nine explanatory variables were removed. Also, we can see that the AIC for this model is smaller than the one with the nine variables suggesting a performance improvement.

The average of the misclassification error by performing 10-fold cross-validation in the model with six predictors is:

```
## [1] 0.0278
```

Regularized form of Logistic Regression

We have two classic regularisation methods: ridge regression and the LASSO. The main difference is that ridge regression always includes all p explanatory variables. This is because the penalty term shrinks the coefficients towards zero, but never exactly zero. However, the LASSO sets some of the coefficients equal to zero so we only include a subset of the predictors in our fitted model. Therefore, LASSO performs subset selection AND shrinkage. For classification problems, we look for coefficients which minimise the loss function, which is the negative of the log-likelihood.

The LASSO regularisation method, drops predictor variables from the model. So as the λ value increases, explanatory variables begin to drop from the fitted model.

In Figure 5, we see that the ninth variable (Mitoses) is the first to drop out of the model, followed by the fifth variable (Single Epithelial Cell Size), then the fourth variable (Marginal Adhesion), and so on.

The penalty term is the key of the regularisation methods. Thus, λ is the tuning parameter which controls the influence of the penalty. We will use cross-validation over a grid of values to choose the appropriate value for λ which minimises the error.

Now, the optimal value of the tuning parameter that minimises the test error, this is the misclassification error, is:

```
## [1] 0.001
```

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

```
## 10 x 1 sparse Matrix of class "dgCMatrix"
```

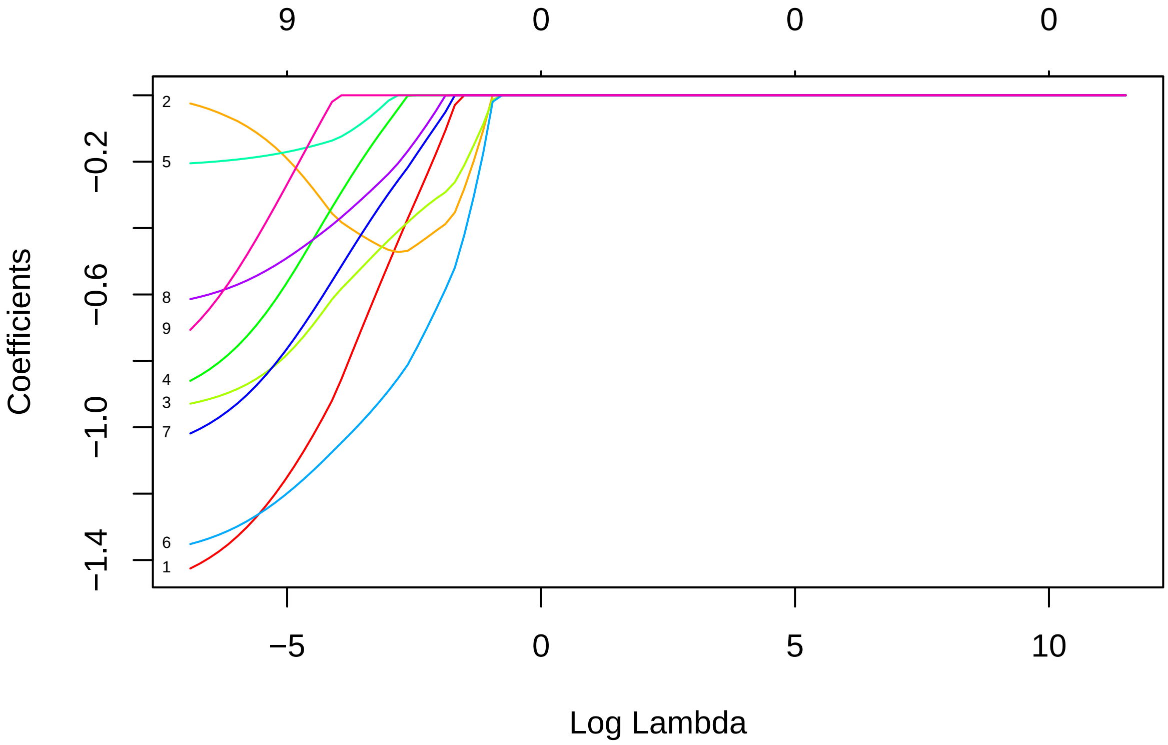

Figure 5: The effect of varying the tuning parameter in logistic regression with LASSO for the Breast Cancer data

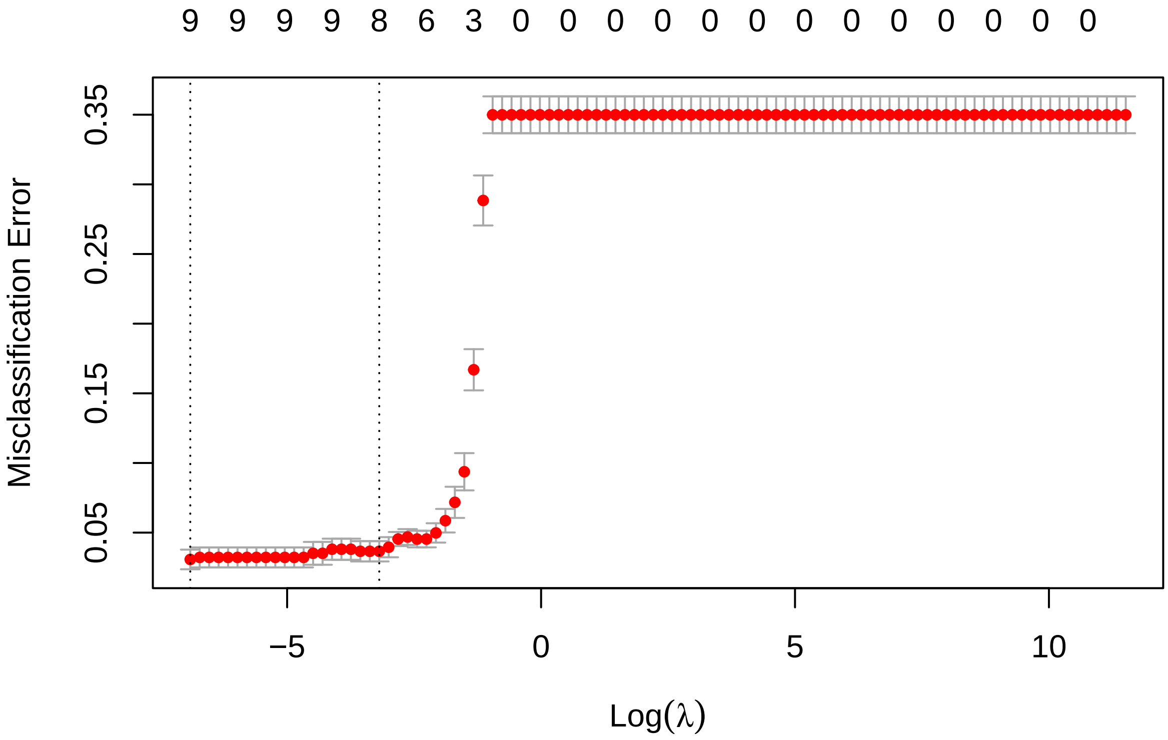

Figure 6: Cross validation scores for the Breast Cancer data using logistic regression with LASSO

```
##                          s1
## (Intercept)      1.10318970
## Cl.thickness    -1.42544234
## Cell.size       -0.02461836
## Cell.shape      -0.92904353
## Marg.adhesion   -0.86016782
## Epith.c.size    -0.20497585
## Bare.nuclei     -1.35192873
## Bl.cromatin     -1.01876948
## Normal.nucleoli -0.61398384
## Mitoses         -0.70676347
```

The choice of the tuning parameter λ is crucial in the LASSO method. A larger λ value increases the amount of regularization, leading to more coefficients being shrunk towards zero. On the opposite, a smaller λ value reduces the regularization effect, allowing more variables to have non-zero coefficients.

Having this in mind, our λ value is not large, therefore, we did not have any predictor variable equal to zero. If we compare the coefficients of the LASSO method with the ones from the best selection method with six predictor variables, it is evident that the coefficients have been shrunk towards zero. All the coefficients remain negative, suggesting the same as explained in the logistic regression section.

Cross Validation - LASSO

One of our goals is to compare the performance of our classifiers based on the test error to decide which is the best. Therefore, we are going to perform cross-validation in the same set of folds in which we performed the cross-validation for the best subset selection model with six predictor variables. As we did before, we will compute some evaluation metrics.

The average of the misclassification error by performing 10-fold cross validation is:

```
## [1] 0.0322
```

The test error with the LASSO method is bigger compared with the one obtained with the best subset selection method with a model of six predictor variables.

Discriminant Analysis Method

We are going to perform the Linear Discriminant Analysis (LDA) to the breast cancer data.

```
## Call:
## lda(y ~ ., data = BreastCancer_scaled)
##
## Prior probabilities of groups:
##         0         1
## 0.3499268 0.6500732
##
## Group means:
##   Cl.thickness  Cell.size Cell.shape Marg.adhesion Epith.c.size Bare.nuclei
## 0    0.9735377  1.1179245  1.1194084     0.9619665    0.9410791   1.1205047
## 1   -0.5240440 -0.6017657 -0.6025644    -0.5178153   -0.5065718  -0.6031546
##   Bl.cromatin Normal.nucleoli    Mitoses
## 0    1.032699       0.9788322  0.5874230
## 1   -0.555890      -0.5268939 -0.3162029
##
## Coefficients of linear discriminants:
##                         LD1
## Cl.thickness    -0.51494720
## Cell.size       -0.38524897
## Cell.shape      -0.26913404
## Marg.adhesion   -0.13524594
## Epith.c.size    -0.12798425
## Bare.nuclei     -0.95267761
## Bl.cromatin     -0.27100084
## Normal.nucleoli -0.32514387
## Mitoses         -0.01405788
```

The LDA output indicates that $\pi_1 = 0.349$ and $\pi_2 = 0.651$. This means, that 34.9% of the data corresponds to cells which are malignant and 65.1% of the data belongs to benign cells. It also provides the group means; these are the average of each predictor within each class (benign and malign), and are used by LDA as estimates of μ_k. As expected, these means suggest that all the explanatory variables tend to be negative when the cell is benign, and a tendency of each predictor to be positive when the cell is malignant.

The coefficients of linear discriminants outputs provides the linear combination of the nine explanatory variables that are used to form the LDA decision rule. For example, LD1 = -0.515×Cl.thickness - 0.385×Cell.size - 0.269×Cell.shape - 0.135×Marg.adhesion - 0.128×Epith.c.size - 0.953×Bare.nuclei - 0.271×Bl.cromatin - 0.325×Normal.nucleoli - 0.0141×Mitoses

We can also compute our discriminant functions with LDA:

```
##                          0           1
## constant        -6.03549476 -1.87522968
## Cl.thickness     1.62603717 -0.87527676
## Cell.size        1.21649199 -0.65482339
## Cell.shape       0.84983850 -0.45745811
## Marg.adhesion    0.42706307 -0.22988305
## Epith.c.size     0.40413299 -0.21754006
## Bare.nuclei      3.00824861 -1.61930500
## Bl.cromatin      0.85573325 -0.46063118
## Normal.nucleoli  1.02669946 -0.55266029
## Mitoses          0.04439024 -0.02389474
```

Therefore, our discriminant functions to define the decision boundary are:

$Q_1(\underline{x})$ = -6.035 + 1.626x_1 + 1.216x_2 + 0.849x_3 + 0.427x_4 + 0.404x_5 + 3.008x_6 + 0.855x_7 + 1.027x_8 + 0.044x_9

$Q_2(\underline{x})$ = -1.875 - 0.875x_1 - 0.655x_2 - 0.457x_3 - 0.2298x_4 - 0.217x_5 - 1.619x_6 - 0.461x_7 -0.553x_8 - 0.024x_9

As expected, the coefficients of the discriminant function for the negative class are positive, which means that with larger values of the predictors, the sample is quite likely to be malignant. Otherwise, the coefficients of the discriminant function for the positive class are negative, which means that if the cell is benign, is less likely to have larger values for the predictors.

Cross validation - LDA

Advancing to compare the three models, we will perform cross-validation for this last classifier to assess how well it generalises in new unseen data under the out-of-validation approach. We continue with the same set of folds and we will compute test error and evaluation metrics.

Now, we are going to calculate the test error using 10-fold cross validation for the LDA model:

```
## [1] 0.0395
```

Comparison of performance of the models using cross validation

To estimate the test error more accurately we used 10-fold cross-validation, rather than the validation set approach in which we split the data into training and test sets. Moreover, to make the comparison fair using cross-validation, we used the same set of folds for all three models. Due to we are performing classification models, to quantify the test error, we used the number of misclassified observations rather than the MSE.

In previous sections, we have calculated the test error for each model using the cross-validation approach. In summary, the test errors for each model are:

- Best subset selection in logistic regression (6 predictor variables): 0.0278
- Regularised form of logistic regression (LASSO): 0.0322
- Discriminant Analysis Method (LDA): 0.0395

There is very little difference between the test errors performed by the three classifiers. However, we can conclude that the Best Subset Selection with six predictor variables is the model that minimises the test error for the breast cancer data.

Evaluation Metrics - Choosing the best classifier

To select which is the best classifier, we will not rely just on selecting the model that minimises the value of the test error. That is why, in the cross-validation approach, we have additionally computed some performance metrics. In summary, accuracy is the samples correctly classified; recall computes from all the positive classes, how many were predicted correctly; precision computes from all the classes we have predicted as positive, how many are classified as positive; F1-score is the “harmonic mean of the precision and recall scores obtained for the positive class” (Kundu, 2022). In this way, we are going to choose our best model as the one that has the highest values for these performance metrics.

Considering that the positive class is that the cell is benign and the negative class is malignant, my worst error would be the false positives because we could wrongly predict a sample belonging to the positive class when it belongs to the negative class. In other words, we are predicting that a breast cancer tumour is benign when it is false, therefore, we cannot make any preventive treatment for the patient to mitigate the risk of having breast cancer. Consequently, we will focus on the precision metric which assesses how many of those predicted positive classes, were actually positive.

The performance metrics for the logistic regression with the best selection method with 6 predictor variables are:

```
## Precision: 0.98

## Accuracy: 0.972

## Recall: 0.977

## F1-score: 0.979
```

The performance metrics for the logistic regression with the regularisation method (LASSO) are:

```
## Precision: 0.973

## Accuracy: 0.968

## Recall: 0.977

## F1-score: 0.975
```

The performance metrics for the discriminant analysis method (LDA) are:

```
## Precision: 0.958

## Accuracy: 0.96

## Recall: 0.982
## F1-score: 0.97
```

When comparing these performance metrics between the three models we can see that the best-performing model is the logistic regression with the best selection method (6 predictor variables). Firstly, this method has the highest values for all the metrics compared to the other models in consideration which supports the fact that it is also the method that minimises the test error. Secondly, given the domain of the Breast Cancer data, we have said that precision is the critical metric to consider, therefore, the best selection method has the highest value for this metric. Thirdly, this method removes predictors that on its own, they contribute very little to a model which contains all the other eight predictors. Finally, according to the Exploratory Data Analysis, Cell Size and Single Epithelial Cell Size are variables with the highest correlation values with the other variables, consequently, by removing these variables we mitigate problems such as multicollinearity.

Conclusion

For the breast cancer data, we built three classifiers. One with the best selection method using a subset of six predictor variables, another with the LASSO regularisation method, and one with the Linear Discriminant Analysis. We assessed the performance of the models using the cross-validation approach and we ensured to make a fair comparison by using the common partitions for all the models in consideration. In this out-of-validation approach, we computed the test error and performance metrics such as accuracy, precision, recall and F1-score to decide which is the best model, and we found out that the best model was the one with six predictor variables using the best selection method.

References

Verzani, John (2014). *Using R for introductory statistics* (Second edition..). Boca Raton : CRC Press, Taylor & Francis Group;

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning with Applications in R.* Springer.

Chugh, Vidhi. (2023, March). *Logistic Regression in R Tutorial* https://www.datacamp.com/tutorial/logistic-regression-R

Convert Factor to Numeric and Numeric to Factor in R Programming https://www.geeksforgeeks.org/convert-factor-to-numeric-and-numeric-to-factor-in-r-programming/

Narkhede, Sarang. (2018, May 9th). *Understanding Confusion Matrix* https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62

Kundu, Rohit. (2022, Sept 13th) *Confusion Matrix: How To Use It & Interpret Results [Examples]* https://www.v7labs.com/blog/confusion-matrix-guide

Visualization of a correlation matrix using ggplot2 https://rpkgs.datanovia.com/ggcorrplot/reference/ggcorrplot.html

Chapter notes 4: Linear regression methods https://ncl.instructure.com/courses/51547/files/7367427?module_item_id=2969247

Chapter notes 5: Classification methods https://ncl.instructure.com/courses/51547/files/7367291?module_item_id=2969257

Practical 5: Subset Selection and Cross Validation https://ncl.instructure.com/courses/51547/files/7367416?module_item_id=2969249

Practical 6: Classification https://ncl.instructure.com/courses/51547/files/7367424?module_item_id=2969258