Statistical Foundations of Data Science - MAS8403

Introduction

The BreastCancer dataset is a part of the mlbench package. It concerns with characteristics of breast tissue samples collected from 699 women in Wisconsin using a fine needle aspiration cytology(FNAC). FNAC is a reliable biopsy tool that has been used for many years for diagnosing cancer in breast tissues. With the help of FNAC, nine cytological characteristics are measured on a scale of 1 to 10, which are, cell size, shape, thickness, mitosis, and a few more. A smaller value measured for a characteristic indicates a healthier cell for that characteristic. Further analysis and research on these measured characteristics and a histological examination categorized the cell into Benign or Malignant.

The classification problem refers to the prediction of a class or response variable with the help of predictor variables. The aim of the project is to build a classifier for the Class – benign or malignant – of a tissue sample based on (at least some of) the nine cytological characteristics. This means that we need to find the optimum classifier and deduce from that classifier, which of the nine cytological characteristics is statistically significant in predicting the Class - benign or malignant.

Libraries used:

- mlbench
- dplyr
- ggplot2
- GGally
- bestglm
- glmnet
- MASS
- caret

Data Preprocessing

With the help of the str() function, we find out the class of the variables. It shows that the nine cytological variables are ordinal variables encoded as factors. We convert these variables into numeric using the as.numeric() function.
There are also a few NA values that are removed from the dataset and the ID column is dropped since it is not of much significance in this analysis. After performing the munging we are left with a dataset containing 683 observations and 10 variables.

Exploratory Data Analysis:

To understand the relationship between between different variables to predict the Class- benign or malignant. We plot a pair plot to understand the correlation between the variables. For plotting pair plot we install the GGally package and use the library ggpairs.

>data(nbc)
>ggpairs(nbc, columns = 1:9, ggplot2::aes(colour=Class))

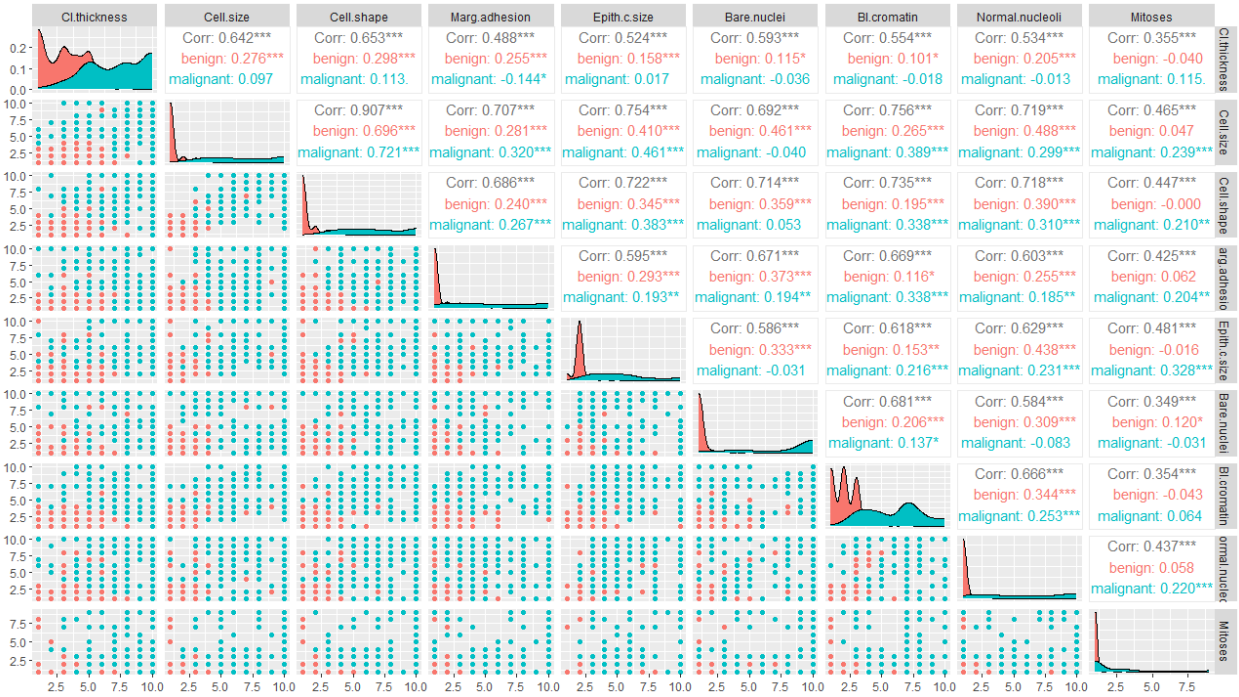

Fig 1

From Fig 1 we can deduce that cell.size and cell.shape are highly correlated, hence, we only need one of them for model prediction which would be verified when we create classifiers for the data.
Also from thes density plots in Fig 1 we can infer that cl.thickness and Bl,cromatin are able to differentiate the Class- benign or malignant fairly well compared to other variables.

Building Classifiers

For this project, we are going to build 3 classifiers. These are:

- Subset selection in logistic regression
- Ridge or Lasso regression
- Bayes classifier for LDA or QDA

Logistic Regression:

We first apply logistic regression to our data using the glm() function.

```
>logistic <- glm(Class~ ., data = nbc, family = "binomial")
>summary(logistic)
```

With the help of the summary(logistic) function, we get our coefficient values for the nine cytological characteristics.

Analyzing the coefficients we infer that since the p-value for the variables Cl.thickness , Marg.adhesion , Bare.nuclei , Bl.cromatin is less than 0.05, they are significant in prediction of the response variable. This is also confirmed by their z value is greater than 2 standard deviations.

Best subset selection in logistic regression:

The AIC is the mesaure of the goodness of fit of any estimated statistical model whilethe BIC is a type of model selection among a class of parametric models with different numbers of parameters. BIC models usually predict consistent models while AIC models generally tend to be overfit.

To apply the best subset selection in logistic regression we use the **bestglm** package. The bestglm function contains various datasets among which we use the Subset component which gives us the predictor variables which are most significant in predicting Class. We select the model where the AIC and BIC are least.

AIC = -2/N * LL + 2 * k/N :

Where N is the number of examples in the training dataset, LL is the log-likelihood of the model on the training dataset, and k is the number of parameters in the model.

The model with the lowest AIC is selected.

BIC = -2 * LL + log(N) * k :

Where log() has the base-e called the natural logarithm, LL is the log-likelihood of the model, N is the number of examples in the training dataset, and k is the number of parameters in the model.

The model with the lowest BIC is selected.

bst_fit_aic$Subsets shows us that 7th model is the best model that can be used to predict
the response variable. This means that a cluster of 7 variable model is used to predict the class of the cell ie: benign or malignant.
We reached this conclusion after seeing True values for all variables except Cell.size and Epith.c.size. and also 7th model having the lowest AIC.

bst_fit_bic$Subsets shows us that 5th model is the best model that can be used to predict the response variable. This means that a cluster of 5 variable model is used to predict the class of the cell ie: benign or malignant.
We reached this conclusion after seeing True values for all variables except Cell.size , cell.shape ,Mitosis and Epith.c.size. and also 5th model having the lowest bic score.

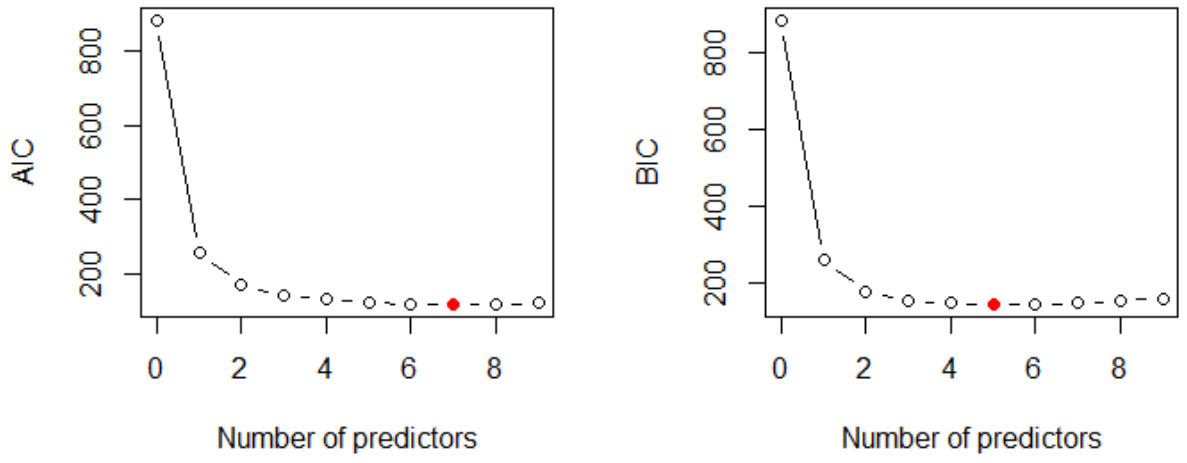

Fig2

Fig 2 Clearly summarises the above statements. For AIC, model with 7 features is considered as the best fit while for BIC model with 5 features is considered as best fit. We choose the BIC model as the best fit because as mentioned earlier, BIC produces consistent predictions and in this case there is not much significant improvement after the 5 feature model and it is also less complex.
We now fit the regression model to the new dataset containing 5 features.

```
>pstar = 5
>bst_fit_bic$Subsets[pstar+1 , ]

>indices = as.logical(bst_fit_bic$Subsets[pstar+1, 2:(p+1)])
>indices

>new_nbc = data.frame(nbc[,indices])
>head(new_nbc)

>logfit = glm(Class~., data = new_nbc, family = "binomial")
>summary(logfit)
```

Lasso Regression:

Lasso regression is a regularisation technique which performs variable selection and regularisation. In Lasso regression we introduce a small bias to the best fit line so as reduce the total variance. The value of this bias can be changed by changing the value of lambda. This bias is called the penalty which can be scaled by a tuning parameter which is equal to lambda.

We have used lasso regression over Ridge regression because in Lasso the penalty added is equal to the absolute value of the magnitude of coefficient. In this case, coefficaints of few variables may become zero and would not be considered for the best fit model.
While for Ridge regression does not result in the elimination of the variable and hence has a more complex model with a higher chance of overfitting.

We use the **glmnet** function to apply lasso regression to the data for an optimal value of lambda.

```
# Choose grid of values for the tuning parameter
>grid = 10^seq(-4, 1, length.out = 100)
# Fit a model with LASSO penalty for each value of the tuning parameter
>lasso_fit = glmnet(nbc[,1:9], nbc$Class, family="binomial", alpha=1, standardize=FALSE, lambda=grid)
# Examine the effect of the tuning parameter on the parameter estimates
```

```
>plot(lasso_fit, xvar = "lambda", col = rainbow(p), label = TRUE)
```

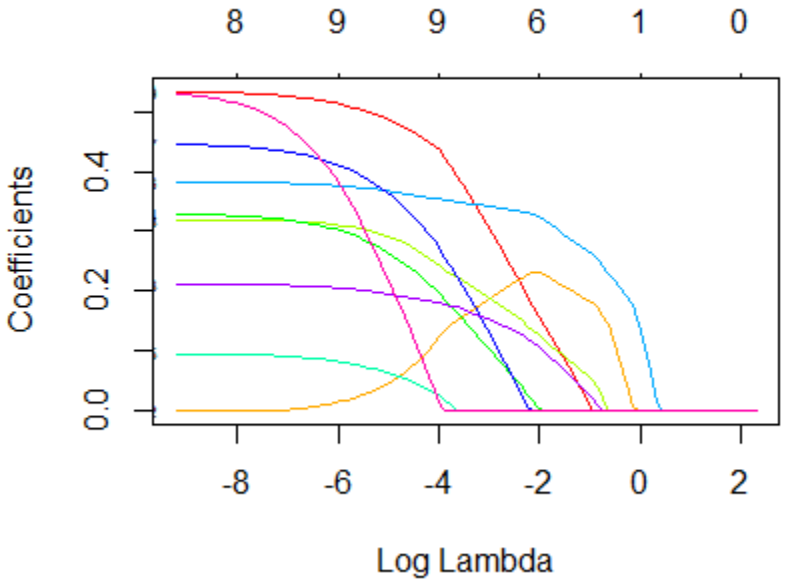

Fig 3

From Fig 3 we can understand thar all values reaches 0 at Log Lambda = 0.5.

Mitosis is a statistically insignificant variable as it reaches very close to zero

```
>min_lambda = lasso_cv_fit$lambda.min
>min_lambda
```

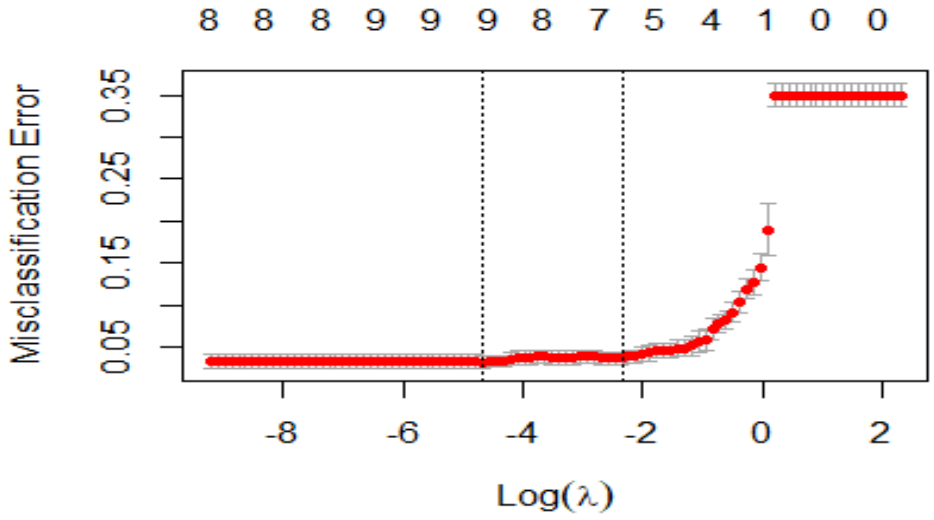

Fig 4

From Fig 4 we can clearly verify that the value of min lambda is **0.009326033**

Now we understand the coefficients of the lasso regression at the optimal lambda value using the below formula:

```
>coef(lasso_cv, s = lasso_cv$lambda.min)
```

From the coefficient matrix we can deduce that Cl.thickness , Bare.nuclei , Bl.cromatin are most statistically significant variables in predicting the Class - benign or malignant.

We should also focus on the fact that even though we applied Lasso penalty, no coefficient of a variable was equal to 0 , hence this classifier contains all variables.

LDA(Linear Discriminant Analysis):

It is basically a dimensionality reduction technique. Using the Linear combinations of predictors, LDA tries to predict the class of the given observations.

We use the **MASS** library to apply lda to our data.

```
>lda_fit = lda(Class ~ ., data = nbc)
>lda_fit
```

When we analyse the lda_fit we get a few inferences:

Cl.thickness and Bare.nuclei have larger value of coefficients. This means that these variables are statistically more significant in predicting Class.

Cross Validation using test error:

For all three classifiers, we perform the test error to figure out which classifier among them has the least error and highest accuracy and can be used a a genralized classifier for new data to predict the Class - benign or malignant.

Firstly we have divided the data into training and testing data using K-fold method.
We have used K=10.

We then pass each of the classifier into the training data and then into the testing data to calculate the error for each classifier. If we find a huge difference in the error rate between different classifiers then we select the classifier with the least error.

```
                                          Model Error Rate
1                  Logistic Regression with BIC 0.03680765
2 Logistic Regression with LASSO regularization 0.03689233
3                  Linear Discriminant Analysis 0.03230458
```

We can see from the above code that our error is almost same for all the classifiers. Hence we use the simplest classifier with low error rate which is the logistic regression with BIC. This classifier is a 5 variable model which is less complex and combined with a low error rate , there is less chances of overfitting.