Breast Cancer Prediction using Statistical Learning

Nishaal Ajmera

29/11/2020

Introduction

Breast Cancer is the most prevalent cancer in women across the globe. This cancer begins when cancerous cells develop in the breast tissue forming a tumour. This tumour can be either benign where it will not spread or malignant where the cells can spread to other tissues.
In this project, "BreastCancer" data set is obtained from Wisconsin which contains 9 cytological characteristics of the tissue sample from 699 women.
The goals of this project are:
* to build a classifier for the "Class" benign or malignant of a tissue sample based the cytological characteristics
* to assess which cytological characteristics are most significant to classify the tissue samples

Data Mining

This data contains 699 rows, 9 predictor variables and 1 response variable `Class`

Data Wrangling

Table 1: Top 6 rows of Breast Cancer data

Cl.thickness	Cell.size	Cell.shape	Marg.adhesion	Epith.c.size
0.1977598	-0.7016978	-0.7412304	-0.6388973	-0.5552016
0.1977598	0.2770488	0.2625905	0.7574766	1.6939247
-0.5112687	-0.7016978	-0.7412304	-0.6388973	-0.5552016
0.5522740	1.5820442	1.6010185	-0.6388973	-0.1053763
-0.1567545	-0.7016978	-0.7412304	0.0592897	-0.5552016
1.2613024	2.2345419	2.2702324	1.8047571	1.6939247

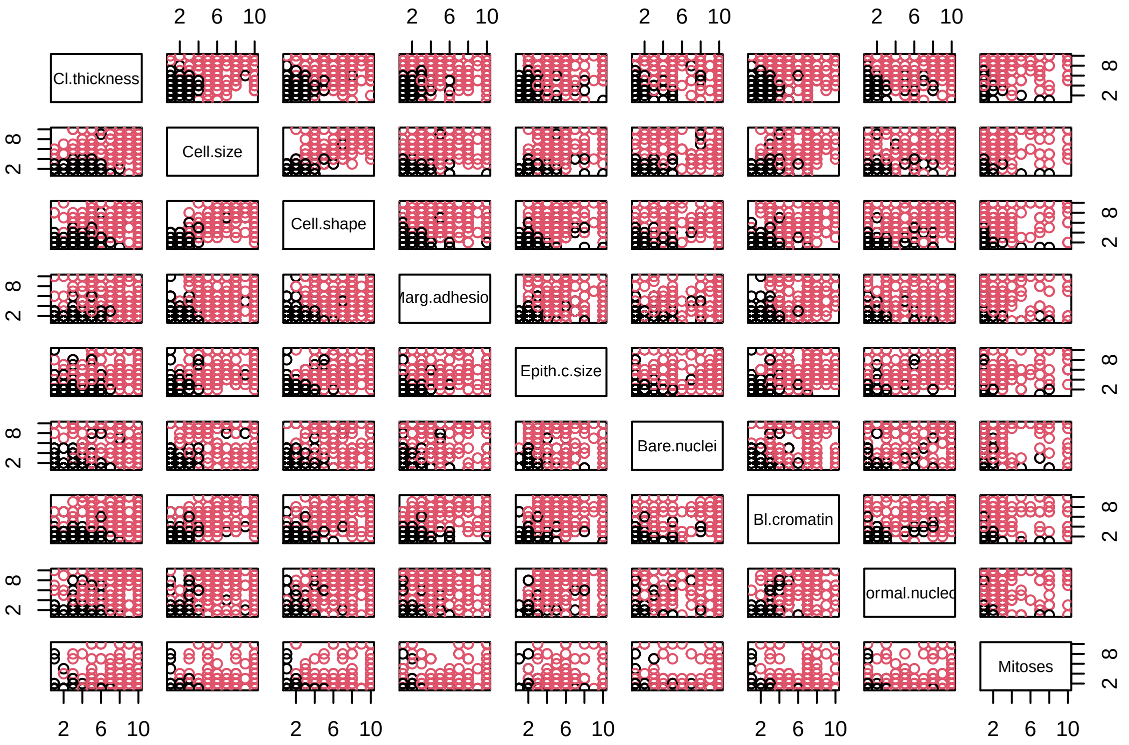
Table 2: Top 6 rows of Breast Cancer data

Bare.nuclei	Bl.cromatin	Normal.nucleoli	Mitoses
-0.6983413	-0.181694	-0.6124785	-0.3481446
1.7715689	-0.181694	-0.2848960	-0.3481446
-0.4239068	-0.181694	-0.6124785	-0.3481446
0.1249621	-0.181694	1.3530163	-0.3481446
-0.6983413	-0.181694	-0.6124785	-0.3481446

Bare.nuclei	Bl.cromatin	Normal.nucleoli	Mitoses
1.7715689	2.267589	1.3530163	-0.3481446

The data has been modified to include the only the 9 predictor and response variable. Rows that contained missing values were omitted. The data now contains 683 tissue samples

Graphical Summary



For almost all variables the ratings given for benign is lower suggesting that those samples are healthier. There is a linear correlation between `Cell.size` and `Cell.shape`,suggesting that the bigger the cell size the more irregular the cell shape.

Numerical Summary

Table 3: Benign sample means

	x
Cl.thickness	2.963964
Cell.size	1.306306
Cell.shape	1.414414
Marg.adhesion	1.346847
Epith.c.size	2.108108
Bare.nuclei	1.346847
Bl.cromatin	2.083333
Normal.nucleoli	1.261261

	x
Mitoses	1.065315
Class	0.000000

Malignant Sample means

Table 4: Malignant sample means

	x
Cl.thickness	7.188284
Cell.size	6.577406
Cell.shape	6.560670
Marg.adhesion	5.585774
Epith.c.size	5.326360
Bare.nuclei	7.627615
Bl.cromatin	5.974895
Normal.nucleoli	5.857741
Mitoses	2.602511
Class	1.000000

The means for all the predictor variables with malignant samples are higher than the means for benign samples. This suggests that malignant cells are very unhealthy looking and rogue

Measure of Scatter

Table 5: Variance by columns

	x
Cl.thickness	7.956694
Cell.size	9.395113
Cell.shape	8.931615
Marg.adhesion	8.205716
Epith.c.size	4.942109
Bare.nuclei	13.277695
Bl.cromatin	6.001013
Normal.nucleoli	9.318772
Mitoses	3.002160

Bare.nuclei is highly spread compared to the other variables. Mitoses has the smallest variance suggesting that it is more centered around the mean compared to the other variables

Correlation matrix

Table 6: Correlation matrix

	Cl.thickness	Cell.size	Cell.shape	Marg.adhesion
Cl.thickness	1.0000000	0.6424815	0.6534700	0.4878287
Cell.size	0.6424815	1.0000000	0.9072282	0.7069770
Cell.shape	0.6534700	0.9072282	1.0000000	0.6859481
Marg.adhesion	0.4878287	0.7069770	0.6859481	1.0000000
Epith.c.size	0.5235960	0.7535440	0.7224624	0.5945478
Bare.nuclei	0.5930914	0.6917088	0.7138775	0.6706483
Bl.cromatin	0.5537424	0.7555592	0.7353435	0.6685671
Normal.nucleoli	0.5340659	0.7193460	0.7179634	0.6031211
Mitoses	0.3509572	0.4607547	0.4412576	0.4188983

Table 7: Correlation matrix

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

This matrix quantifies the strength of the relationship between the covariates. `Cell.size` and `Cell.shape` are highly correlated (0.907). One of these variables could be eliminated to avoid collinearity.

Covariance matrix of standardized data

Table 8: Correlation matrix of standardized data

	Cl.thickness	Cell.size	Cell.shape	Marg.adhesion
Cl.thickness	1.0000000	0.6424815	0.6534700	0.4878287
Cell.size	0.6424815	1.0000000	0.9072282	0.7069770
Cell.shape	0.6534700	0.9072282	1.0000000	0.6859481
Marg.adhesion	0.4878287	0.7069770	0.6859481	1.0000000
Epith.c.size	0.5235960	0.7535440	0.7224624	0.5945478
Bare.nuclei	0.5930914	0.6917088	0.7138775	0.6706483
Bl.cromatin	0.5537424	0.7555592	0.7353435	0.6685671
Normal.nucleoli	0.5340659	0.7193460	0.7179634	0.6031211
Mitoses	0.3509572	0.4607547	0.4412576	0.4188983

Table 9: Correlation matrix of standardized data

	Epith.c.size	Bare.nuclei	Bl.cromatin	Normal.nucleoli	Mitoses
Cl.thickness	0.5235960	0.5930914	0.5537424	0.5340659	0.3509572

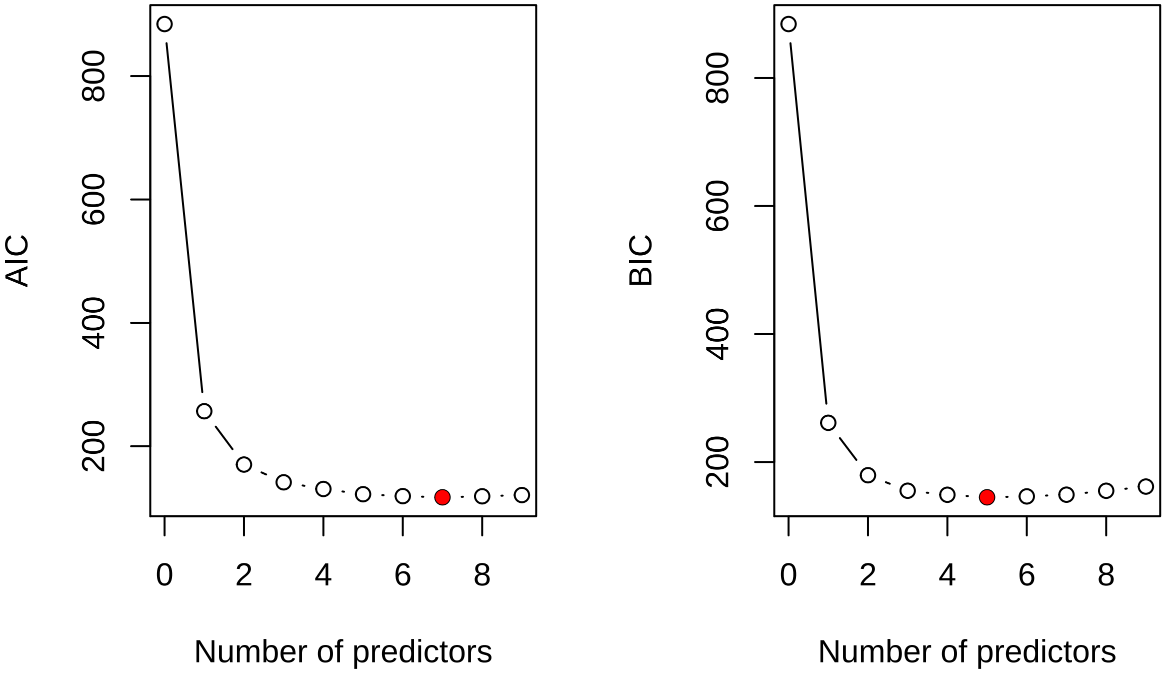
	Epith.c.size	Bare.nuclei	Bl.cromatin	Normal.nucleoli	Mitoses
Cell.size	0.7535440	0.6917088	0.7555592	0.7193460	0.4607547
Cell.shape	0.7224624	0.7138775	0.7353435	0.7179634	0.4412576
Marg.adhesion	0.5945478	0.6706483	0.6685671	0.6031211	0.4188983
Epith.c.size	1.0000000	0.5857161	0.6181279	0.6289264	0.4805833
Bare.nuclei	0.5857161	1.0000000	0.6806149	0.5842802	0.3392104
Bl.cromatin	0.6181279	0.6806149	1.0000000	0.6656015	0.3460109
Normal.nucleoli	0.6289264	0.5842802	0.6656015	1.0000000	0.4337573
Mitoses	0.4805833	0.3392104	0.3460109	0.4337573	1.0000000

Standardized data covariance matrix is the same as correlation matrix of original data

Classifiers

1. Best Subset Selection of Logistic Regression Classifier

Best subset selection is used with two model comparison criterions (AIC and BIC) to select the best subset model. The aim is to select the model with variables that gives a good compromise between the mdoel with smallest AIC and model with smallest BIC



Model with 6 predictor variables looks like a good compromise

Model with 6 predictor variables is selected

The predictors variables selected are `Cl.thickness`, `Cell.shape` ,`Marg.adhesion` , `Bare.nuclei` , `Bl.chromatin` , `Normal.nucleoli`

All the regression coefficients for the subset with 6 predictor variables show significance

Table 10: Regression coefficients

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.2592045	0.2903572	-4.336743	0.0000145
Cl.thickness	1.7560138	0.3867812	4.540070	0.0000056
Cell.shape	1.0445414	0.4932028	2.117874	0.0341858
Marg.adhesion	0.9668875	0.3311709	2.919603	0.0035048
Bare.nuclei	1.3793829	0.3418244	4.035355	0.0000545
Bl.cromatin	1.1546299	0.4069335	2.837392	0.0045484
Normal.nucleoli	0.7423195	0.3313810	2.240079	0.0250858

K-fold Cross validation to calculate out of sample misclassification error

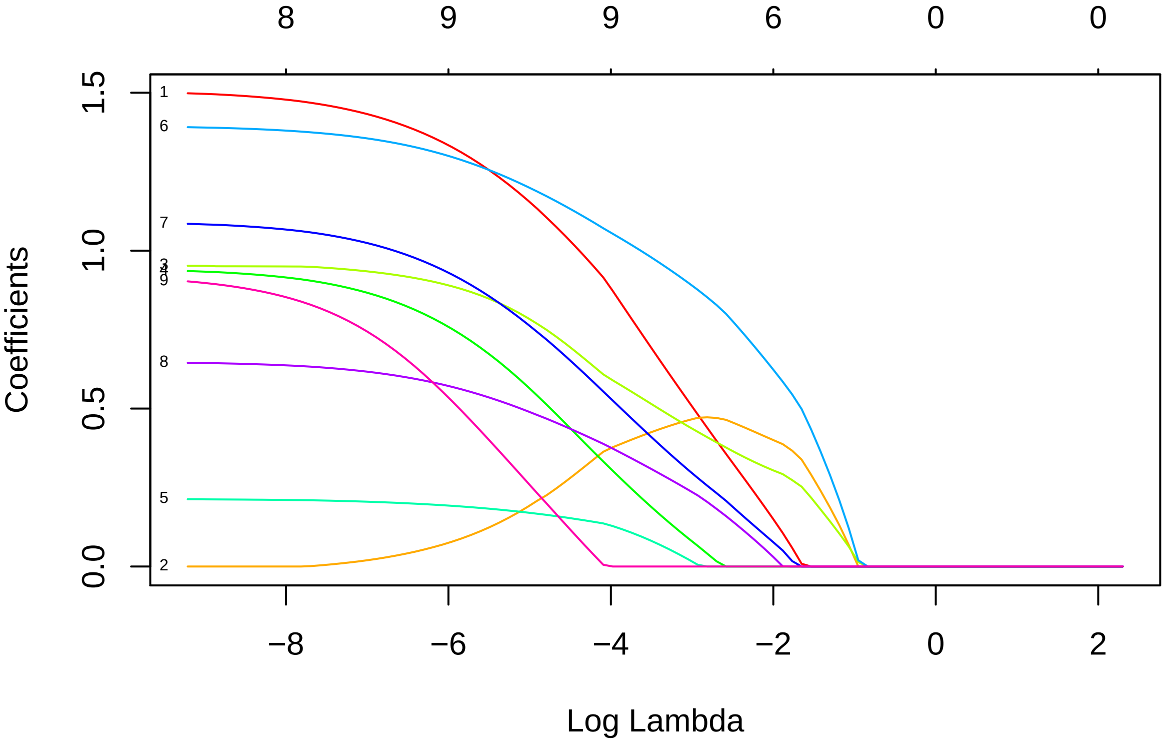
K=10 and the`fold_index` used is the same for all the K-fold cross validation going forward for fair comparison across classifiers.

Misclassifcation error for Best Subset with 6 predictor variables using Logistic Regression

```
## [1] 0.03074671
```

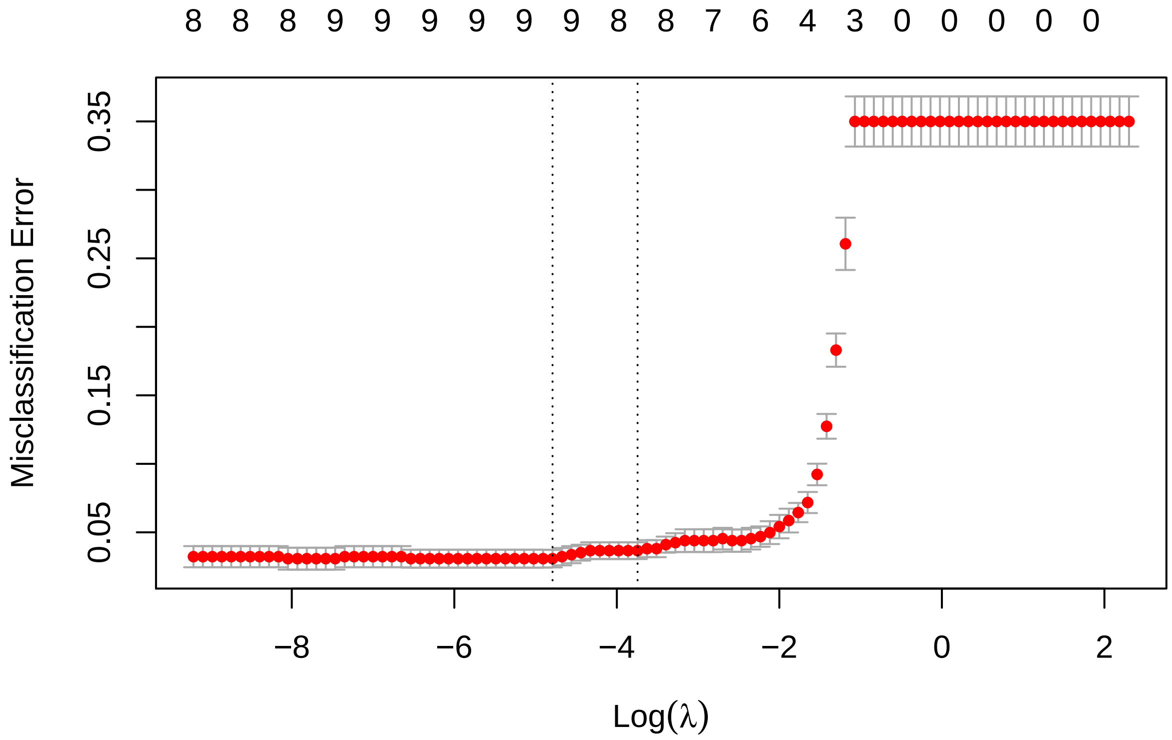
Regularized logistic regression with LASSO penalty

LASSO penalty is applied to the logistic regression model with all the variables to perform covariate selection.



From the plot it can be observed that as the lambda increases the coefficients shrink to 0. The optimum lambda will be selected using cross validation with the same `fold_index`

K-fold Cross Validation Misclassification Error plot



Optimum value of lambda

```
## [1] 0.008302176
```

The LASSO penalty coefficients for optimum value of lambda

Table 11: Coefficients with LASSO

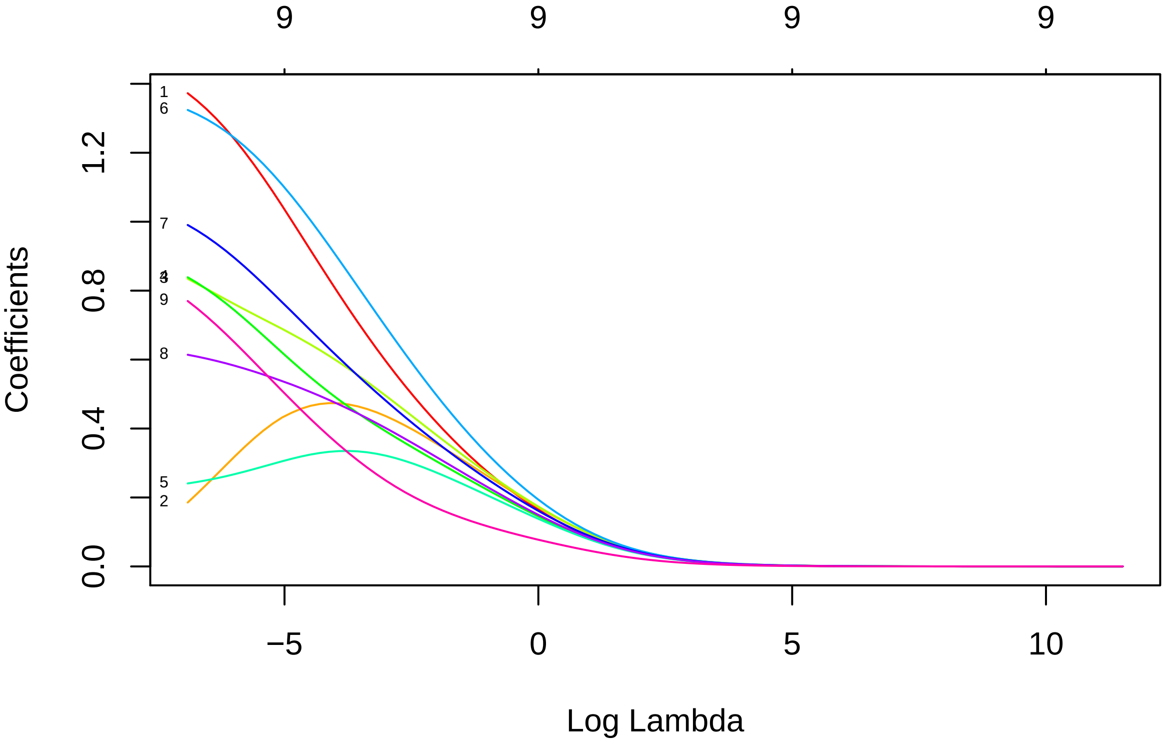
	1
(Intercept)	-1.0631822
Cl.thickness	1.1045803
Cell.size	0.2257127
Cell.shape	0.7491367
Marg.adhesion	0.5124781
Epith.c.size	0.1634143
Bare.nuclei	1.1722658
Bl.cromatin	0.7181726
Normal.nucleoli	0.4679572
Mitoses	0.1993772

None of the variables were eliminated as none shrunk to zero completely. Therefore we will perform ridge regression and test which model has lower test error.

Misclassifcation error Logistic Regression with LASSO penalty

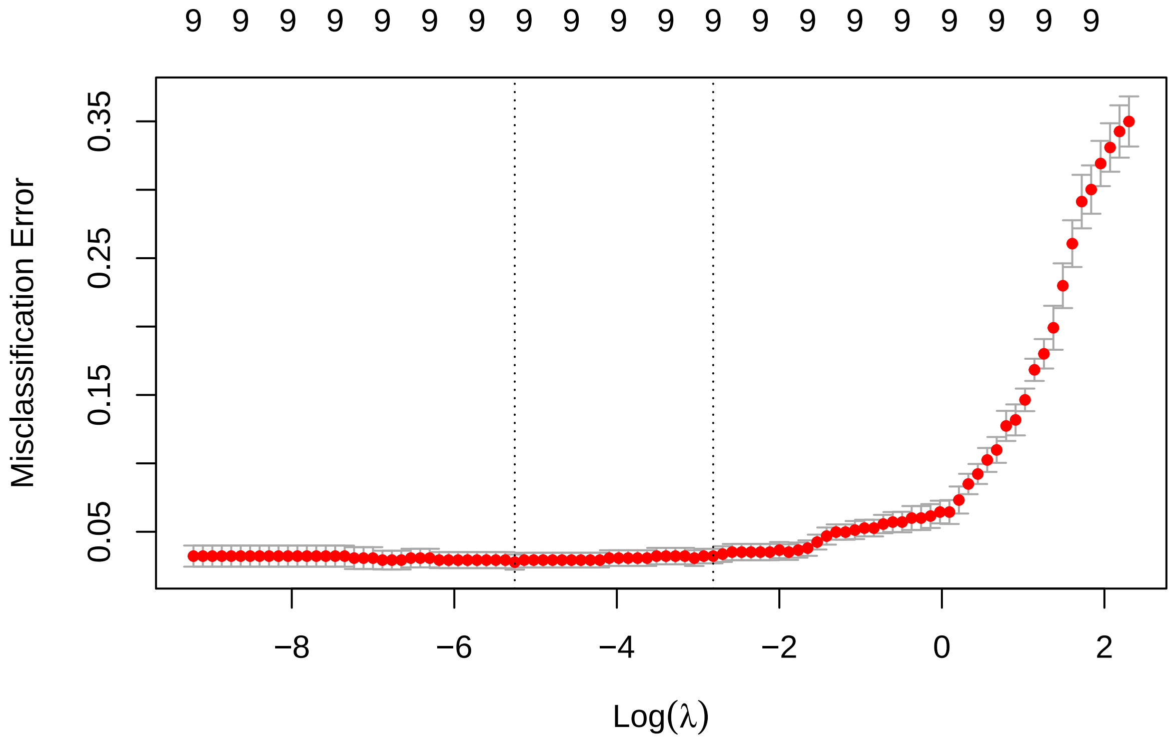
```
## [1] 0.03074671
```

Regularized logistic regression with Ridge penalty



From the plot it can be observed that as the lambda increases the coefficients shrink to 0. The optimum lambda will be selected using cross validation with the same `fold_index`

K-fold Cross Validation Misclassification Error plot for Ridge regression



Optimum value of lambda

```
## [1] 0.005214008
```

The Ridge penalty coefficients for the optimum value of lambda

Table 12: Coefficient of Ridge

	1
(Intercept)	-1.0293773
Cl.thickness	1.0931273
Cell.size	0.4114847
Cell.shape	0.7051444
Marg.adhesion	0.6484986
Epith.c.size	0.2964435
Bare.nuclei	1.1422749
Bl.cromatin	0.7967752
Normal.nucleoli	0.5486101
Mitoses	0.5410290

Misclassifcation error Logistic Regression with Ridge penalty

```
## [1] 0.02781845
```

Table 13: LDA group means

	Cl.thickness	Cell.shape	Marg.adhesion	Bare.nuclei	Bl.cromatin	Normal.nucleoli
0	-0.5240440	-0.6025644	-0.5178153	-0.6031546	-0.555890	-0.5268939
1	0.9735377	1.1194084	0.9619665	1.1205047	1.032699	0.9788322

Table 14: QDA group means

	Cl.thickness	Cell.shape	Marg.adhesion	Bare.nuclei	Bl.cromatin	Normal.nucleoli
0	-0.5240440	-0.6025644	-0.5178153	-0.6031546	-0.555890	-0.5268939
1	0.9735377	1.1194084	0.9619665	1.1205047	1.032699	0.9788322

Linear Discriminant Analysis

LDA is performed on the 6 significant variables selected through Best Subset selection.

The estimated group means for the benign tissue sample variables are much lower than the malignant tissue samples.

Misclassification Error for LDA

Misclassification error is calculated using K-fold cross validation with K=10 and the same `fold_index`

```
## [1] 0.04245974
```

Quadratic Discriminant Analysis

QDA is performed on the 6 significant variables selected through Best Subset selection.The test error will be compared to LDA.

The estimated group means for the benign tissue sample variables are much lower than the malignant tissue samples.

Misclassification Error for QDA

Misclassification error is calculated using K-fold cross validation with K=10 and the same `fold_index`

```
## [1] 0.04978038
```

Test errors using K-fold crossvalidation

The test errors have been computed using K=10 with the same `fold_index`

```
df=data.frame(Model=c("BSS","LASSO","Ridge","LDA","QDA"),Test_errors=c(test_error_BSS,test_error_LASSO,t
kable(df,caption="Test errors of models")
```

Table 15: Test errors of models

Model	Test_errors
BSS	0.0307467
LASSO	0.0307467
Ridge	0.0278184
LDA	0.0424597
QDA	0.0497804

Conclusion

The classifier that is selected is the logistic regression model with ridge penalty using all the variables as predictors because it has the lowest missclassification error.