MAS8404_Project_210431461

210431461 | 21/10/22

Introduction

At the University of Wisconsin Hospital, Dr. Wolberg (Wolberg et. al (1992)) collected breast tissue samples from women using fine needle aspiration cytology (FNAC)(Newcastle University (2022)). Histological examination of the tissue collected by this procedure allows for a physician to determine whether or not the tissue is benign or malignant. Our objective is to build a classifier that determines whether a tissue sample is likely to be benign or malignant based on its cytological characteristics.

Data Description

Dr. Wolberg reported his data chronologically to the `mlbench` R package (Wolberg et. al (1992)). The FNAC procedure allows for the identification of nine cytological characteristics (Newcastle University (2022)): clump thickness (Cl.thickness / CT), uniformity of cell size (Cell.size / CS), uniformity of cell shape (Cell.shape / CSh), marginal adhesion (Marg.adhesion / MA), single epithelial cell size (Epith.c.size / E), bare nuclei (Bare.nuclei / BN), bland chromatin (Bl.chromatin / BC), normal nucleoli (Normal.nucleoli / NN) and mitoses (Mitoses / M). These characteristics for each tissue sample are measured on a discrete scale of one to ten, where smaller numbers indicate that the sample is healthier (Newcastle University (2022)). To aid our analysis, the ordinal variables of this scale have been converted to quantitative variables.

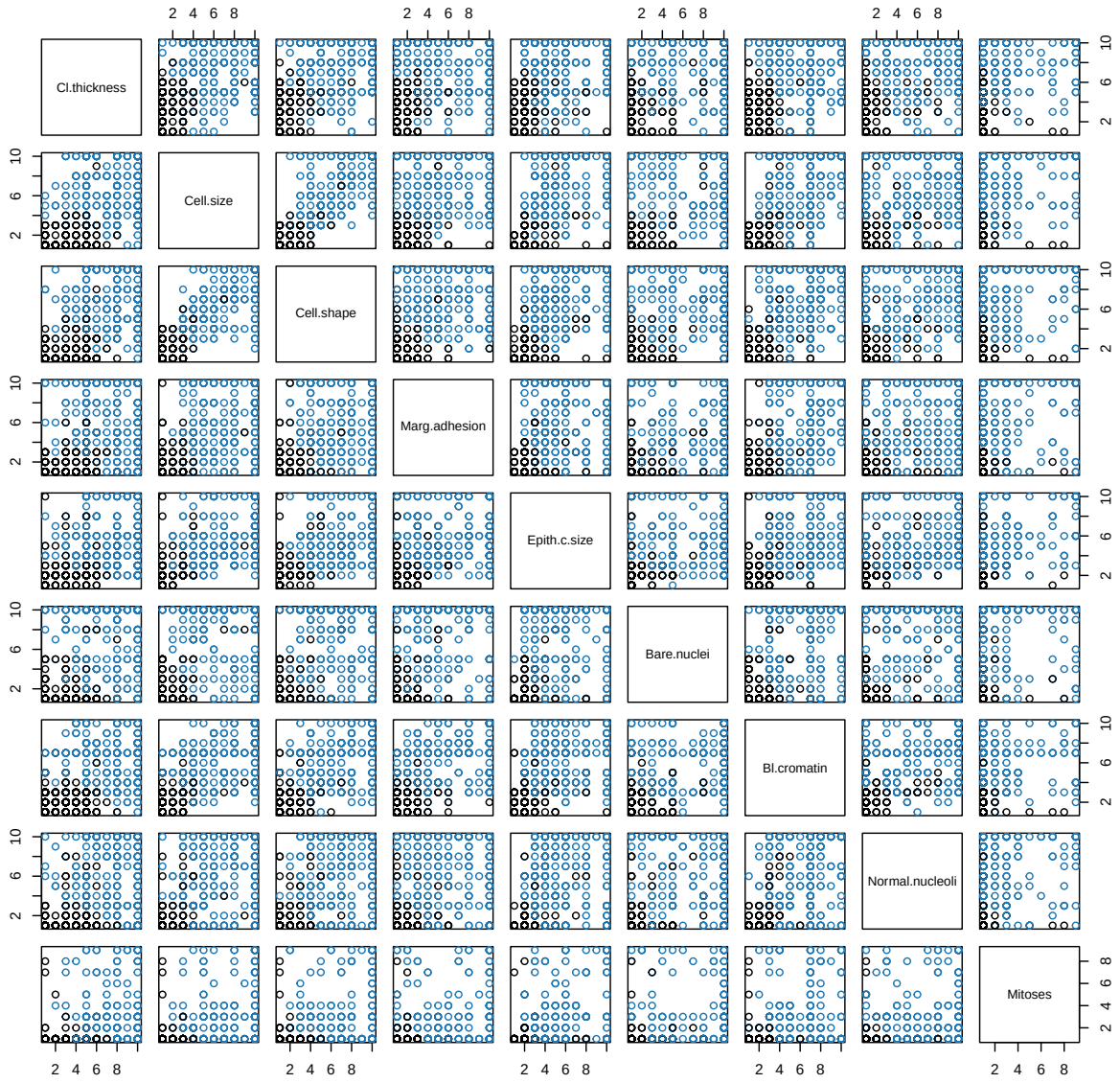
This report explores data from a sample of 699 women in the `BreastCancer` data set; please note that 16 of the 699 observations have been removed due to missing attribute values. It is assumed that this is a random sample of women experiencing symptoms of breast cancer (Newcastle University (2022)). Each woman is represented by a sample code number (Id) that reflects the chronological grouping of this data (Wolberg et. al (1992)). The data set also includes the result of further histological examination (Class), which confirms whether each woman's tissue sample was begin or malignant; this has been converted into a binary variable of 0 or 1, respectively.

Data Exploration

Of the tissues samples from 683 women, 239 are confirmed as malignant; in our sample, 53.83% of the women who are experiencing breast cancer symptoms have malignant tissue.

Whether the tissue is benign or malignant is considered the response variable, and the cytological characteristics of tissue are considered the predictor variables. The characteristics of malignant tissue tend to have higher numbers on the scale one to ten; this is visualized by the density distributions provided in `Appendix A`. All nine cytological characteristics of the tissue samples are presented below in a scatter plot matrix, where blue indicates the tissue is malignant.

Plot 1: Scatter Plot Matrix of All Cytological Characteristics

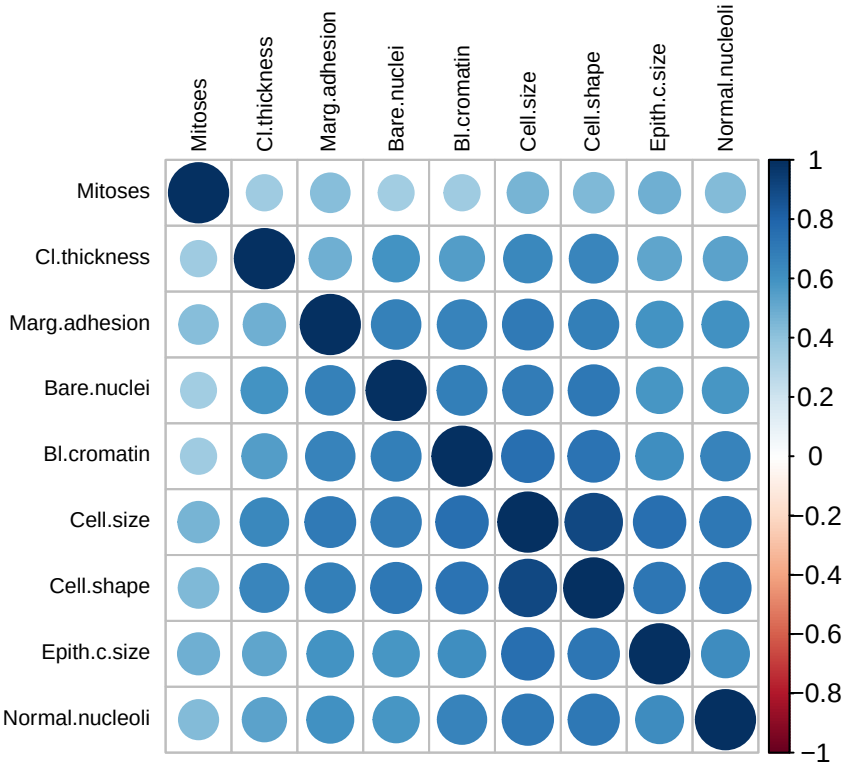


This plot visualises the relationship between the characteristics; it suggests that there is a strong linear relationship between CS and CSh, indicating that larger cells have a more significant shape and more are likely to be malignant. There also appears to be a weak

linear relationship between CS and CSh with E and BC, as well as a weak linear relationship between CT and CSh. Further investigation of these relationships would be beneficial.

The correlations between other characteristics are not initially obvious and require deeper analysis. A sample correlation matrix quantifies the strength of the linear relationship and, for all characteristics, is visualised by the correlation plot below.

Plot 2: Cytological Characteristics Correlation



In this plot, the correlation is more significant when the circle is both darker and larger. This plot confirms the strong linear relationship between CS and CSh, which suggests that any classifier is unlikely to need both characteristics. These characteristics also appear to have stronger relationships with CT, MA, E, BN, BC, and NN. M does not appear to have a significant correlation with any of the characteristics. These insights indicate that an accurate classifier may only need a selection of the predictor variables.

Boxplots of each cytological characteristic are provided in `Appendix B`; these highlight that there is a distinctive difference between the characteristics of benign and malignant tissue. The mean, median and standard deviation (SD) of our data is, therefore, likely to be skewed unless the data is filtered by its Class (benign or malignant). The filtered summary statistics are presented in Table 1 on the next page.

Table 1: Summary Statistics by Class

	Benign			Malignant		
Variable	Mean	SD	Median	Mean	SD	Median
CT	2.96	1.67	3.00	7.19	2.44	8.00
CS	1,31	0.86	1.00	6.58	2.72	6.00
CSh	1.14	0.96	1.00	6.56	2.57	6.00
MA	1.35	0.92	1.00	5.59	3.20	5.00
E	2.11	0.88	2.00	5.33	2.44	5.00
BN	1.35	1.18	1.00	7.63	3.12	10.00
BC	2.08	1.06	2.00	5.97	2.28	7.00
NN	1.26	0.95	1.00	5.86	3.35	6.00
M	1.07	0.51	1.00	2.54	2.40	1.00

This table clearly demonstrates that there is a significant difference between the characteristics of benign and malignant tissue. This is especially true for CT, BN and BC, which suggests that they may be the most important predictor variables. It is interesting that the SD of MA, BN and NN is noticeably larger than the other characteristics. If these characteristics are selected for the classifier, their variance may impact the accuracy of our classifier model.

Given this, it is also important to explore how varied the data is for the cytological characteristics. There are two single measures of multivariate scatter that help us generalise this; these are the generalised variance, which for our data is 70.70, and the total variation, which for our data is 47432.00. The large values tell us that there is a high degree of scatter about the sample means of each variable. To aid our analysis, therefore, it may be beneficial to standardise our data, despite it being a common, discrete scale for all the cytological characteristics.

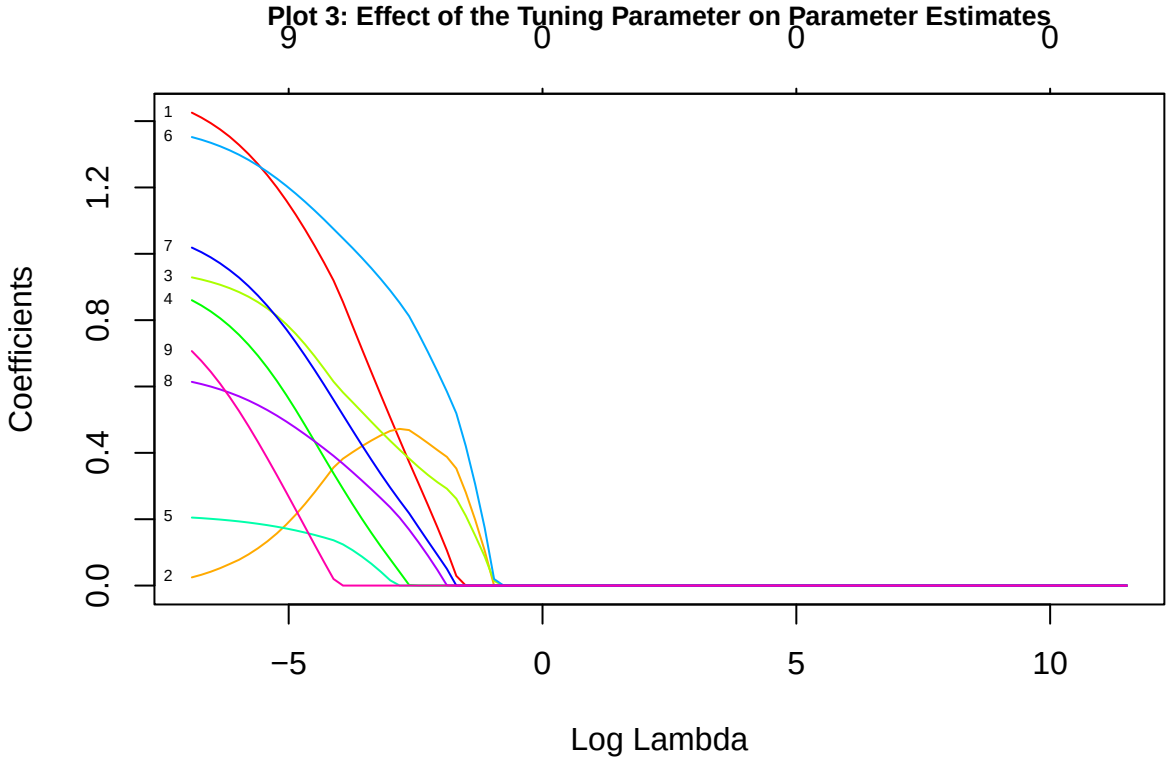
To understand which characteristics most influence the variance, principal component analysis (PCA) was conducted. Three principal components account for 80.12% of the variance, and six account for 92.85% of the variance. The first principal component can be interpreted as an average measure of CS and CSh, which accounts for 65.60% of the variation. The second corresponds to M; together the first and second principal components account for 74.13% of the variation. The third principal component corresponds to CT. Although PCA is a distinct analysis from best subset selection, it is interesting to note that CS, CSh, M and CT are the principal components of our data.

Subset Selection

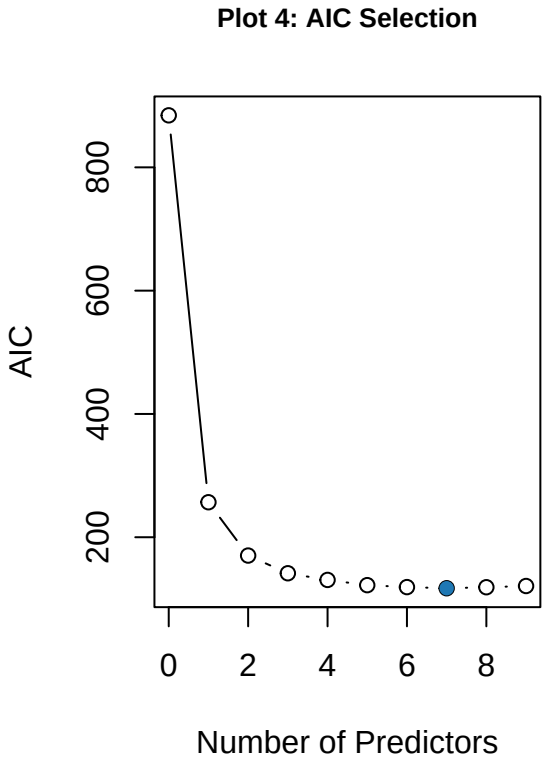
The exploratory data analysis suggests that some predictor variables are likely to be better at predicting our response variable than others. If this proves to be correct, this will allow us to learn the effects of fewer predictor variables more precisely.

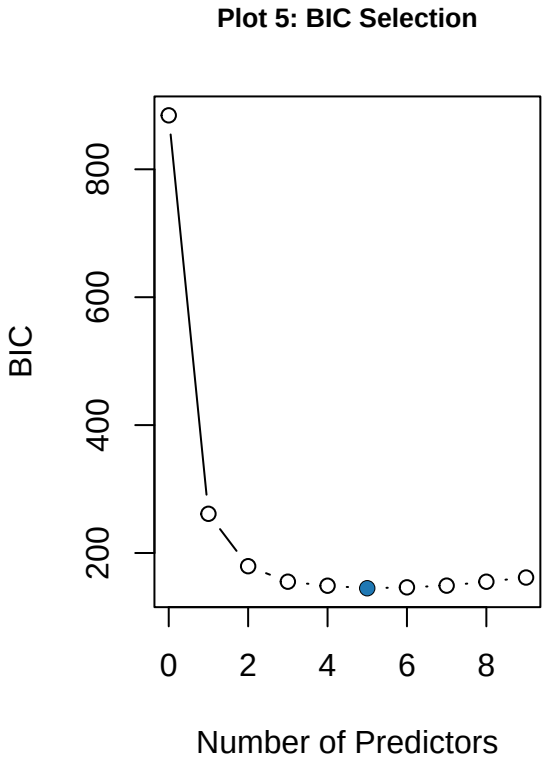
As our response variable is binary (benign or malignant), applying logistic regression is an appropriate approach to identify the best subset of predictor variables. After fitting the logistic regression model for `Class` to standardised data it is clear that CT, BN, MA and BC have a coefficient which is significantly different to zero when testing at the 5% level.

This result is broadly aligned to the variable selection performed by a LASSO regression. In the plot below, the last five variables to drop out are BN, CSh, CS, CT and NN; apart from CS, these predictor variables align with the results of the previous logistic regression. In the LASSO model, all the predictor values are retained as parameter estimates associated with the optimal value of the tuning parameter. These estimates highlight that CT, BN, CSh, BC, MA, and NN are the least shrunk towards zero and, therefore the most influential.



To confirm how many predictor variables would be the best subset selection, the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) are applied. The results are presented in the two plots on the next page.





In both plots, the blue dot is number of predictor variables identified by the best subset selection approach. It appears that a model with six variables is likely to be a good compromise between five and seven. As determined by the logistic regression subset selection, the cytological characteristics for these six variables are CT, CSh, MA, BN, BC, and NN. These align closely to the strongly correlated variables identified in our correlation matrix above.

Modelling

To increase the reliability of our approach, the data is divided in two to allow for out-of-sample validation. In this project, 80% of the data is randomly allocated to a training data set to construct our classifier, with the remaining 20% becoming the testing data set to compute our test errors.

Logistic Regression

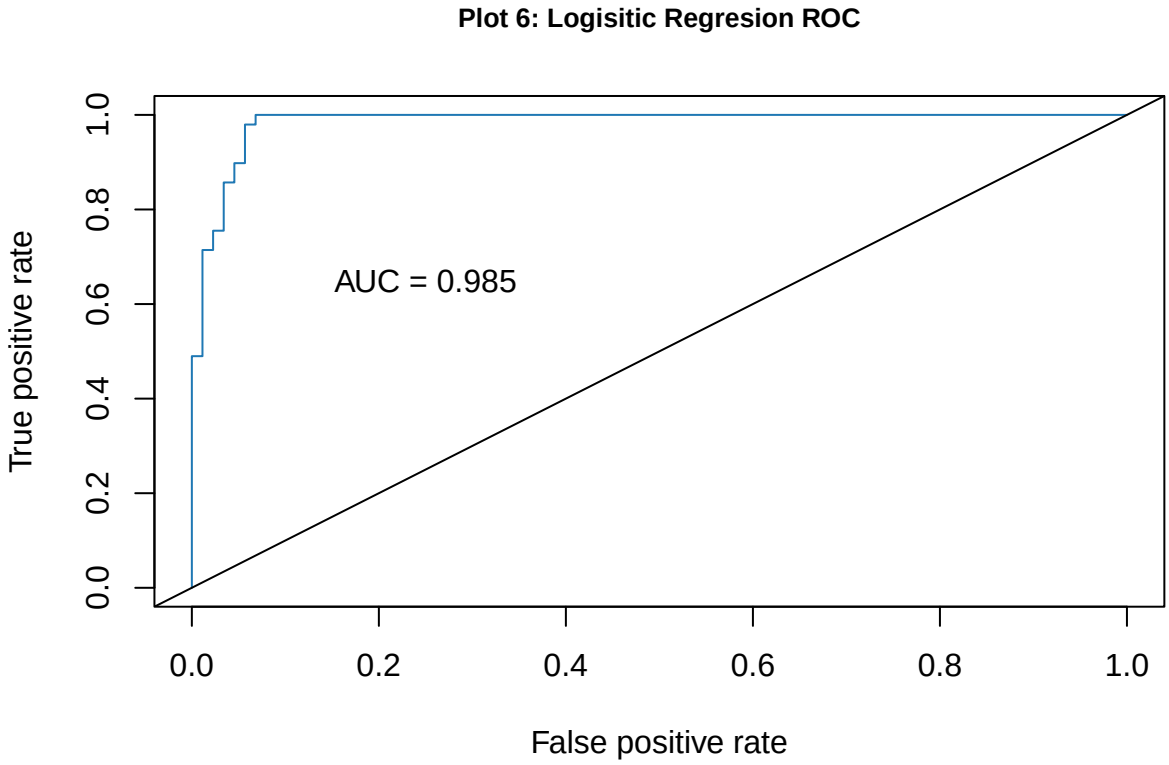
Logistic regression is used to assign observations to discrete response variables. When applied to our the data of our six predictor variables, the maximum likelihood estimates of the regression coefficients are presented on the next page.

Table 2: Estimates of Regression Coefficients for Logistic Regression

	Intercept	CT	CSh	MA	BN	BC	NN
Estimates	-1.28	2.08	1.14	1.39	1.69	1.52	0.87

With 'in-sample validation', the training error is 2.02%. This is not very interesting as only the test error measures how well the method performs on previously unseen data (Newcastle University (2022)). When using our train and test data sets for out-of-sample validation, the test error is 5.85%. This is slightly larger than the training error, which may indicate that unnecessary predictor variables have been included.

The accuracy of our model is visualised by the Receiver Operating Characteristic (ROC) curve below.

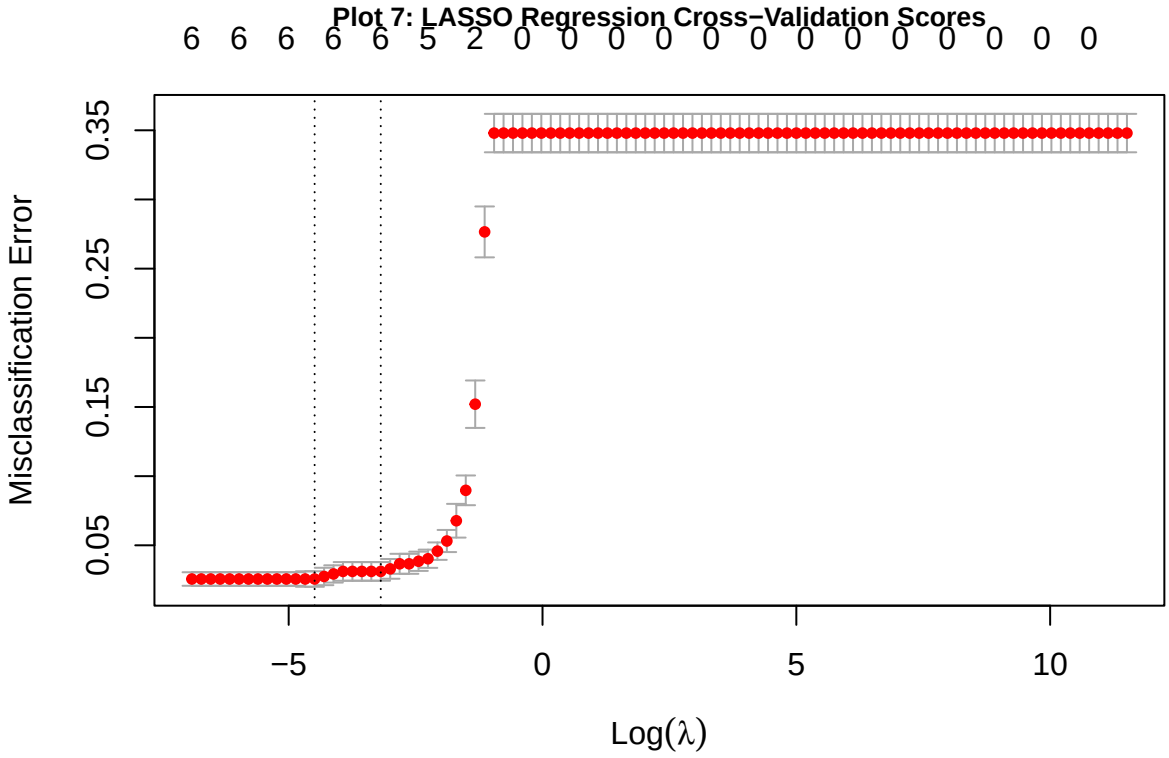


The measure of accuracy is the area under the ROC curve. This curve is far from the diagonal and is quite close to the perfect accuracy area of 1.00. Helpfully, the plot specifies the accuracy of our model as 0.985.

LASSO Regression

Regularisation methods, such as LASSO, are shrinkage methods; they work to minimise the loss function and shrink the maximum likelihood estimates of regression coefficients to zero. LASSO performs subset selection in addition to shrinkage; given that the test error

was higher than the training error for logistic regression, this method is selected over ridge regression. The cross-validation scores of the LASSO regression applied to the same data as above is visualised below.

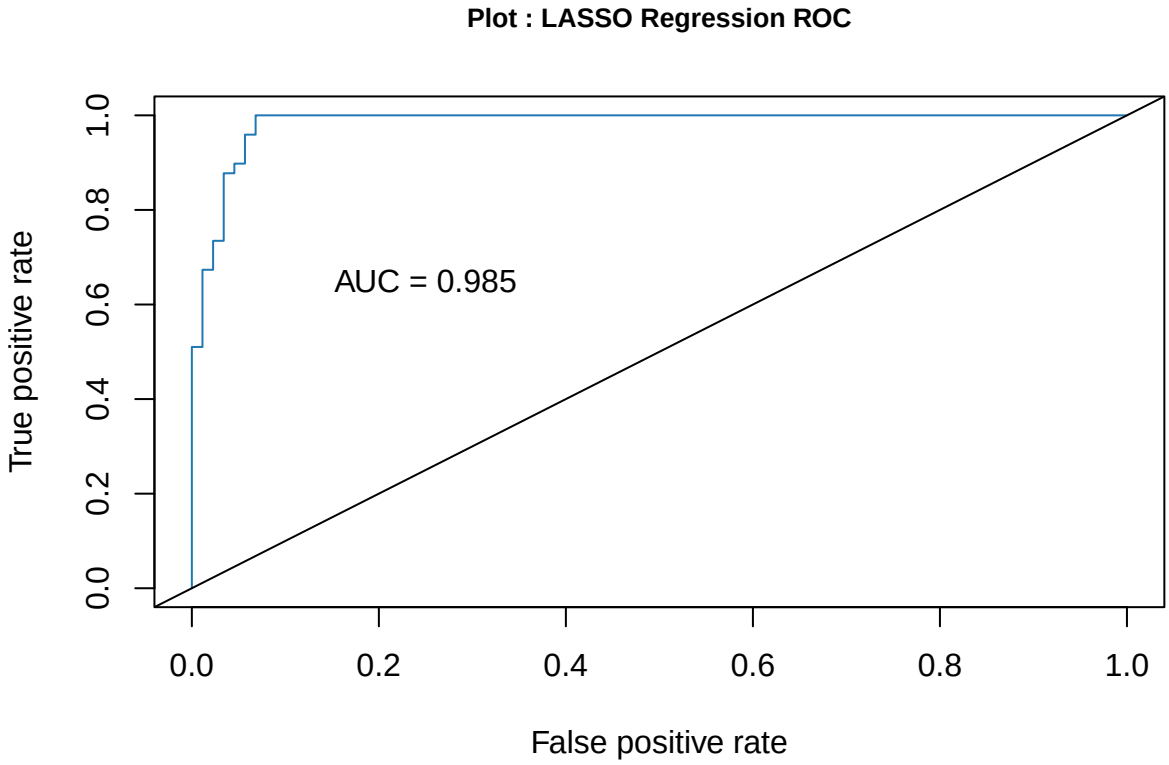


For each value of lambda, this plot shows us the mean MSE across the k folds with error bars which cover the mean plus or minus one standard error. [What does this show us?]

To choose an appropriate value for the turning parameter of LASSO, 10-fold cross-validation is used. This allows the optimal turning parameter and mean MSE to be calculated, which are 0.002104904 and 0.02380952, respectively. The regression coefficients obtained by performing the LASSO with the optimal turning parameter are shown in the table below.

When using our train and test data sets for out-of-sample validation, the test error is 5.85%. This is the same test error as the logistic regression. The full test error value for the logistic regression is 0.05847953 and, for the LASSO regression, it is 0.05839416. This shows us that the LASSO model is very slightly more accuracte.

The accuracy of our model is also visualised by the ROC curve below.



The measure of accuracy is the area under the ROC curve. This curve is far from the diagonal and is quite close to the perfect accuracy area of 1.00. Helpfully, the plot specifies the accuracy of our model as 0.985.

Discriminant Analysis

Discriminant analysis is a technique used to classify observations into groups that do not overlap. Linear Discriminant Analysis (LDA) assumes that the common covariance matrix of the predictor variables is the same. Quadratic Discriminant Analysis (QDA) does not assume this. However, in practice, whether the common covariance is generally unknown.

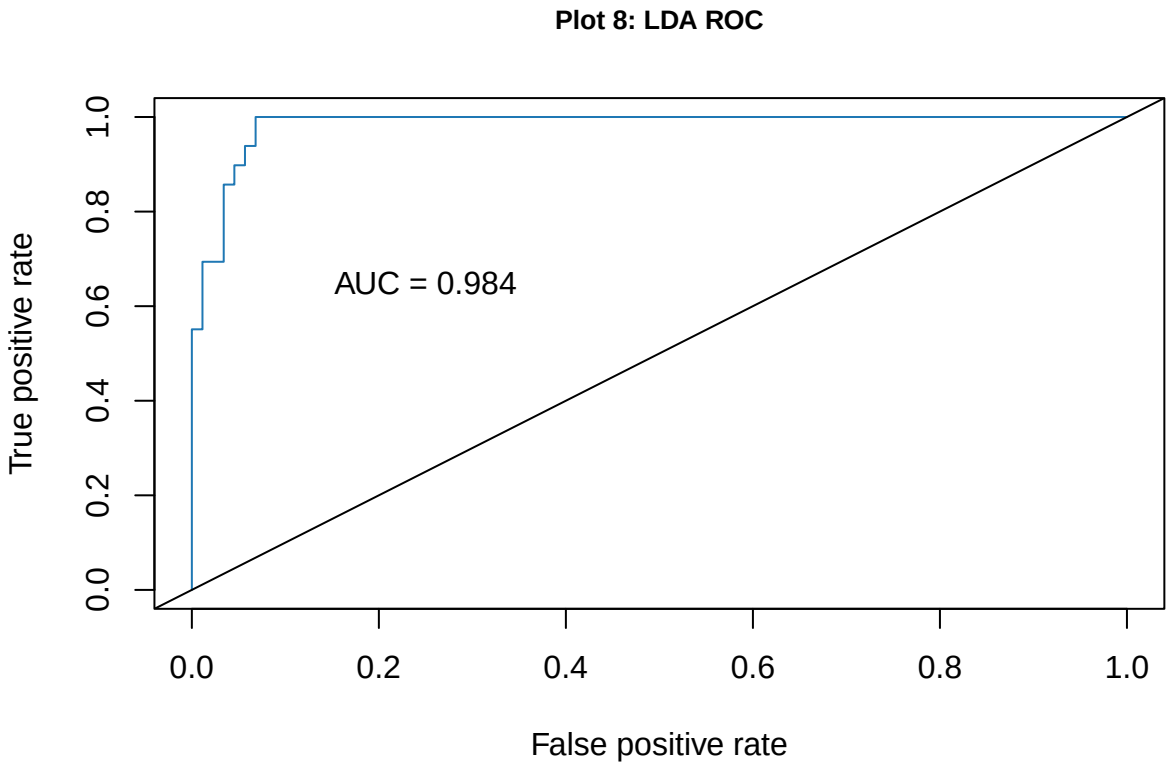
The density distributions in `Appendix A` and boxplots in `Appendix B` suggest that our data is not normally distributed and indicates non-equal variances (respectively). QDA is likely to be more appropriate than LDA. The group means generated by both LDA and QDA are the same, and so only one version of the group means is presented in the table on the next page.

Table 3: LDA and QDA Group Means

	CT	CSh	MA	BN	BC	NN
Benign	-0.56	-0.61	-0.53	-0.61	-0.57	-0.53
Malignant	0.95	1.13	0.96	1.15	1.03	0.99

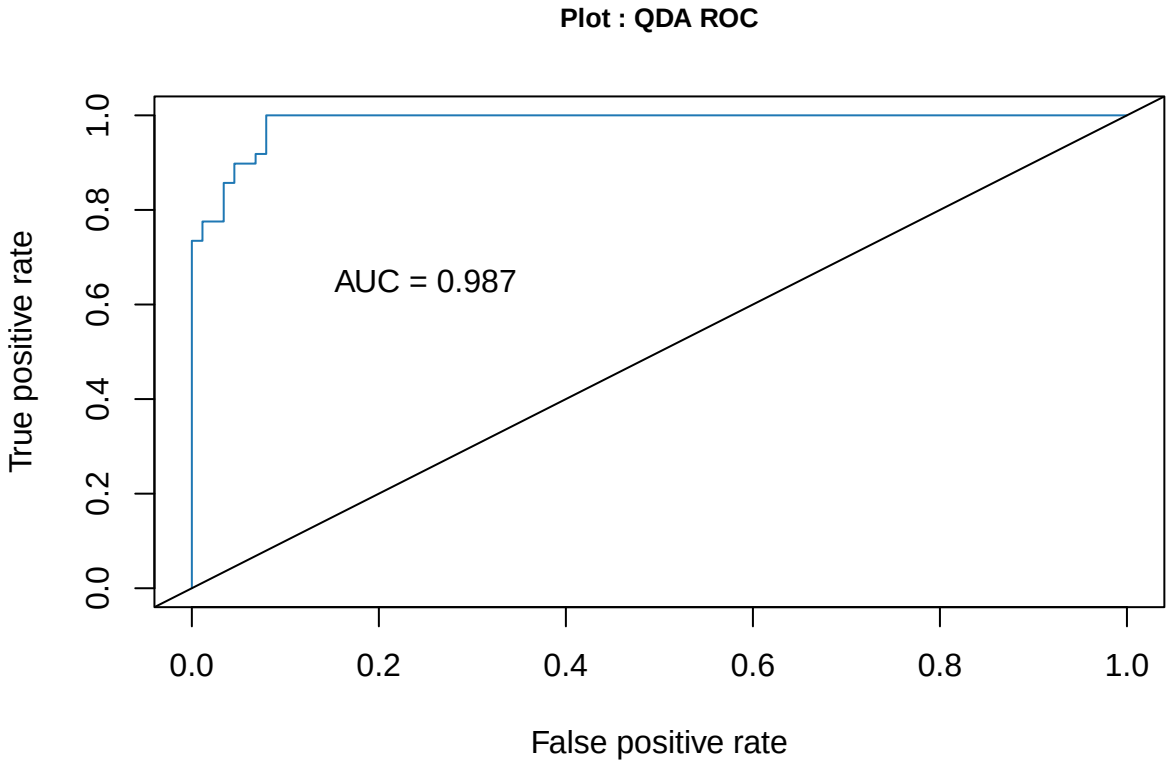
When using our train and test data sets for out-of-sample validation, the test error for LDA is 7.30% and QDA is 6.57%. As expected QDA has the lower test error.

The accuracy of our LDA model is visualised by the ROC curve below.



The measure of accuracy is the area under the ROC curve and the accuracy of our model is 0.984.

The accuracy of our QDA model is visualised by the ROC curve below.



The measure of accuracy is the area under the ROC curve and the accuracy of our model is 0.987. Interestingly, the ROC curve suggests the QDA model is the most accuracte despite its test error being larger than that of the logistic or LASSO regressions.

Evaluation {UPDATE}

Based on the predictive performance of the three models applied, the logistic regression has the lowest test error at 5.85%. It if followed by QDA (6.57%), LDA (7.30%) and, lastly, the LASSO regression with a test error of 25.75%. I had hoped to be able to compare all the models with ROC plots, and unfortunately there was an issue with the code for LASSO and the decriminant analyses. Smaller test errors indicate more accurate models, so our logistic regression would be the best classifier.

Our best classifier does not include all nine cytological characteristics. Using the methods of AIC and BIC for best subset selection, it was determined that six predictor variables was likely to provide the most accurate model. There were: CT, CSh, MA, BN, BC, and NN.

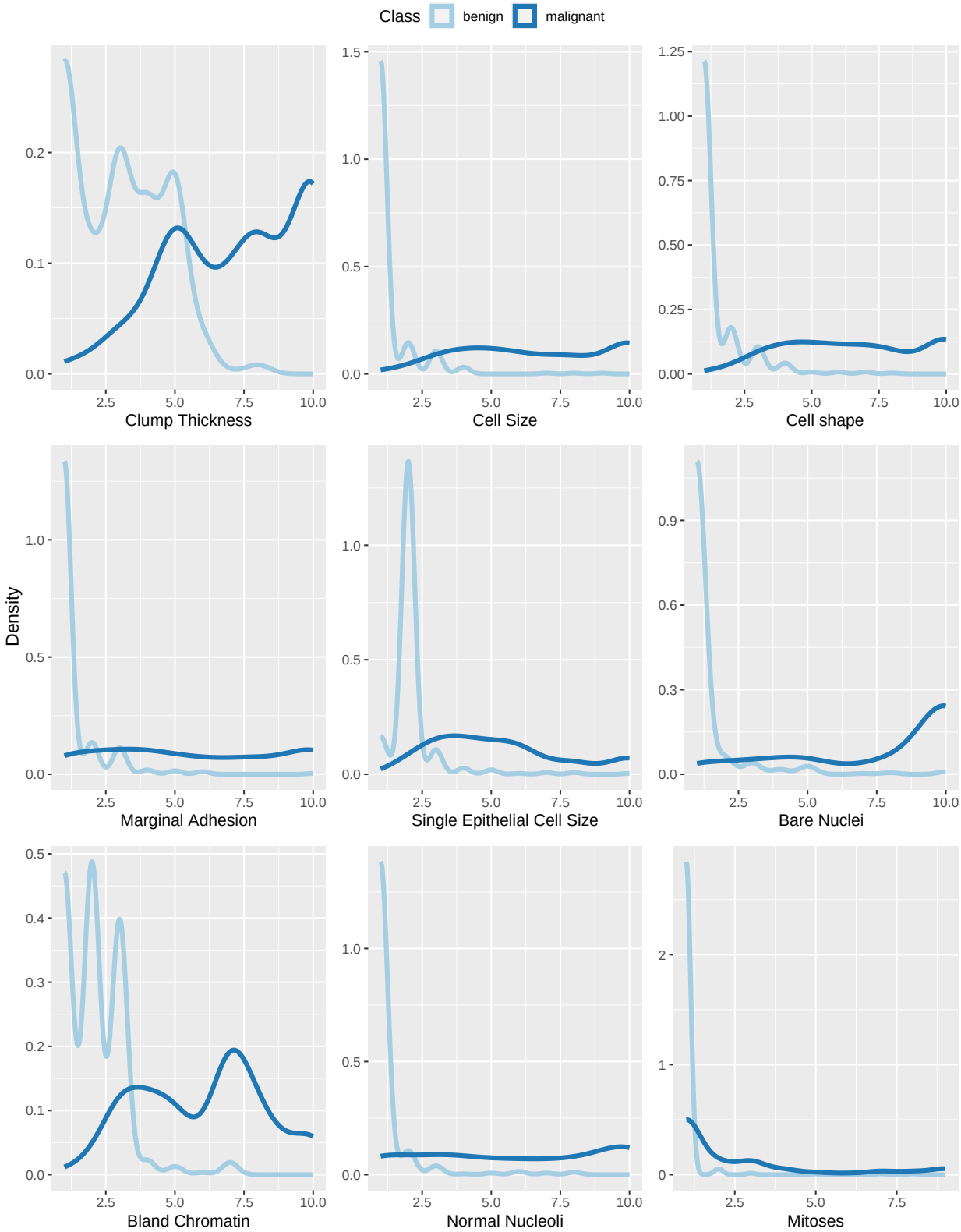
However, as mentioned in the logistic regression section, the test error was slightly larger than the training error. This indicates unnecessary predictor variables have been included. Cell Shape (CSh) had the largest p-value of the chosen six variables, so the classifier may benefit from re-running the logistic regression without this variable.

Similarly, it was interesting that the LASSO regression did conduct any further subset se-

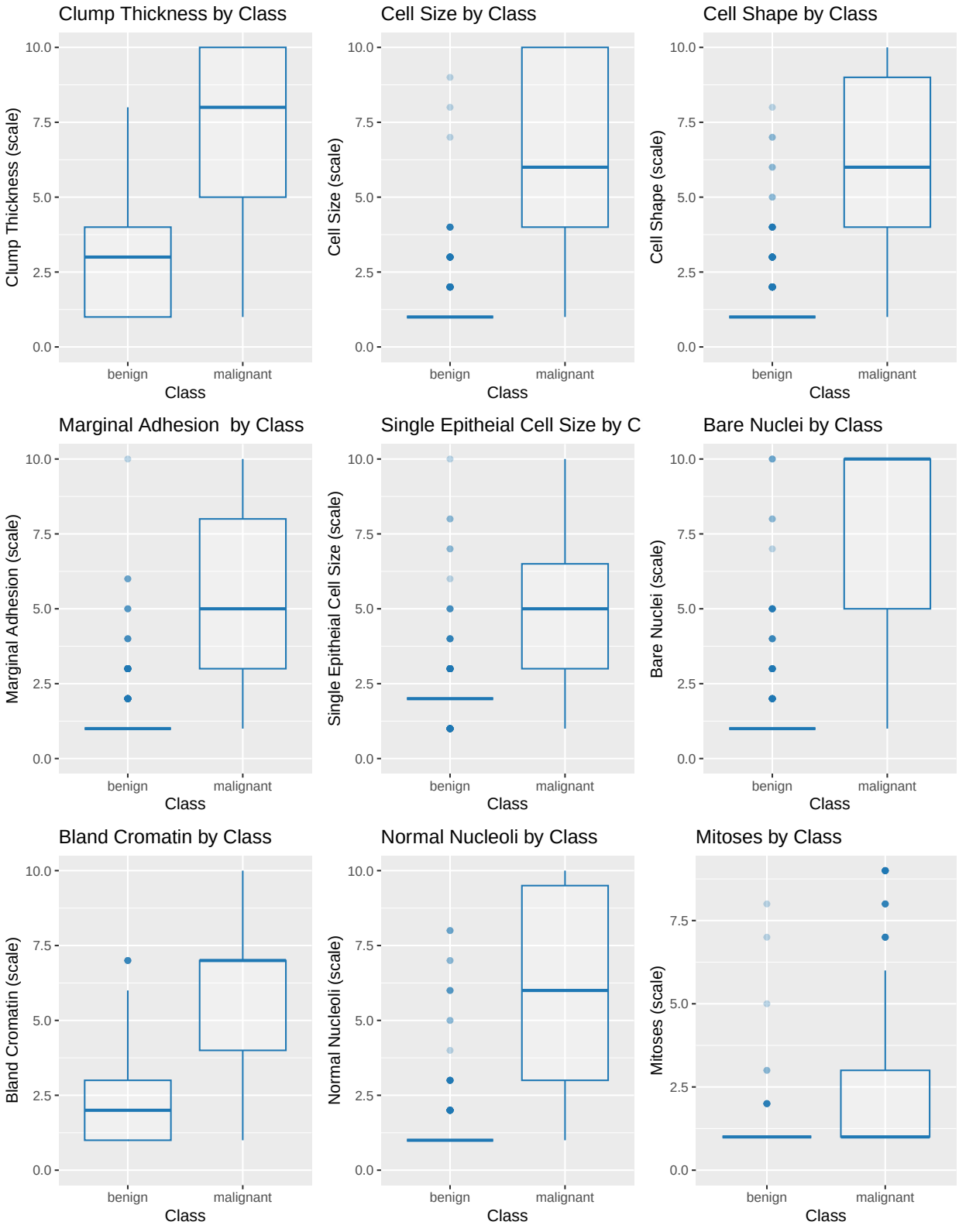
lection from the six variables. It would be interesting to re-run the LASSO regression on the original standardised data, before best subset selection with the logistic regression, to determine whether this would improve the accuracy of the model.

All of the models applied in this report have test error rates that would be impractical in a clinical setting. Further investigation would be beneficial to improve the accuracy of the models; this could include different methods of best subset selection, dividing the test and train data differently (perhaps by only using cross-validation k-folds), and, ideally, access to a much larger data set to avoid over-fitting the models.

Appendix A: Denisty Distributions of Cytological Characteristics by Class



Appendix B: Boxplots of Cytological Characteristics by Class



Bibliography

Newcastle University, Stefen Grunewalder on behalf of. 2022. "MAS8404: Project." https://ncl.instructure.com/courses/46421/files/6177247?wrap=1.

Wolberg et. al, Dr. 1992. "R Documentation: Wisconsin Breast Cancer Database." https://www.rdocumentation.org/packages/mlbench/versions/2.1-3/topics/BreastCancer.