Breast Cancer Classification Analysis

Pranav Sunil Raja

2024-11-19

1. Abstract

This study dives into the examination of data collected from 699 women in Wisconsin who underwent a biopsy known as FNAC (Fine Needle Aspiration Cytology) to assess the breast tissue. Nine characteristics were measured on a scale of 1 to 10 which indicates cell health. Assuming these women represent a random subset experiencing breast cancer symptoms, the project will extensively analyze this data-set. The main objective is to determine if these characteristics alone can accurately classify tissue samples as benign or malignant. It involves fitting a unsupervised & supervised learning model in our analysis which aims to evaluate the reliability of these characteristics in distinguishing between benign and malignant breast tissue. This successful outcome of this analysis is going significantly help breast cancer diagnosis, aiding in more informed treatment decisions.

2. Data Exploration

Data Exploration and preparation involves converting the factors to appropriate format. In order to address the missing variables/ attributes are removed. The next step is to convert class variables from categorical to numerical, with 'benign' denoted as 0 and 'malignant' as 1. Finally, the data-set got reduced and contains 444 benign observations and 239 malignant.

We transform the class variable into 0 (benign) and 1 (malignant) for our analysis

2.1 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
```

This provides insight about 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 'Mitoses' exhibits lowest mean and broader ranges, hinting at potentially significant variability within the dataset.

2.2 Scatter plot matrix

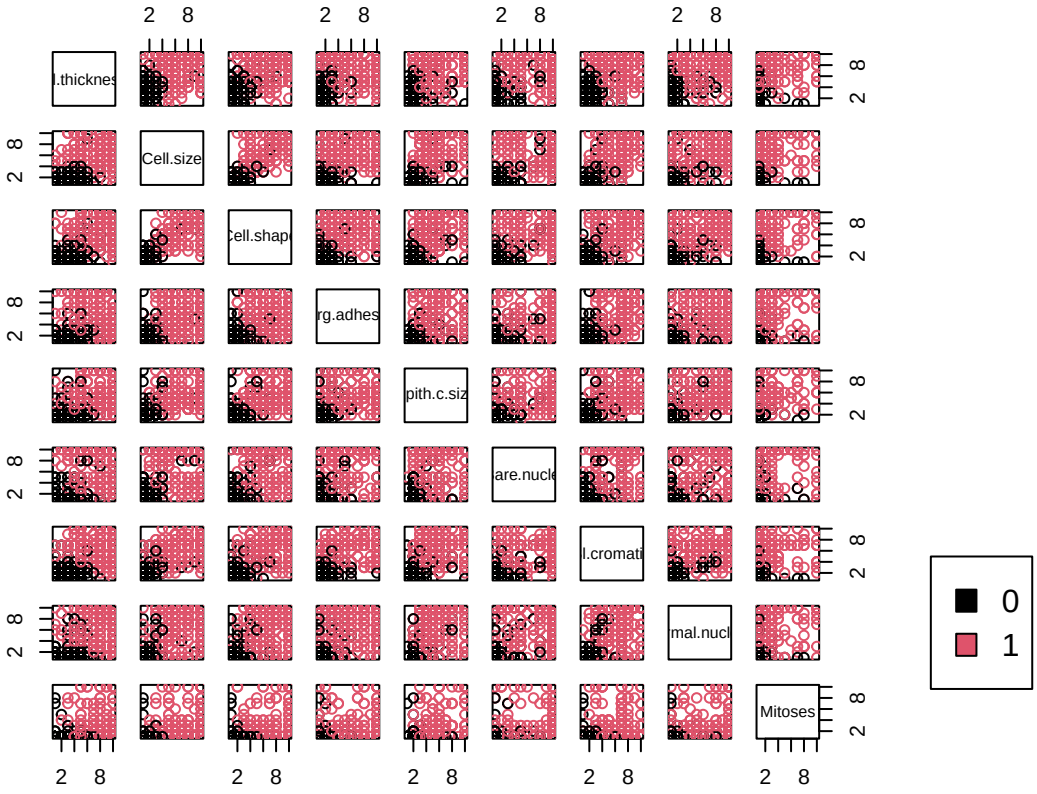

Figure 1: Scatter Plot Matrix

This Scatterplot matrix (Figure 1) reveals a distinct separation between the two classes across response variables, highlighting a clear difference. However, weaker separations are noticeable in normal. nucleoli,

bare.nuclei, marg.adhesion, and epith.c.size, suggests overlapping values between classes in these specific variables. On the other hand we see a positive correlation between cell.size, cell.shape & '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. We also find lower covariance values 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. These findings offer valuable insights.

2.3 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 shows strong positive correlations with predictor variables 'Cl.thickness', 'Cell.size', 'Cell.shape', 'Marg.adhesion', 'Epith.c.size', 'Bare.nuclei', and 'Bl.cromatin'. 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 relationship with class because it has a lower correlation.

Correlation Among Predictor Variables:

Among the predictor variables, we can notice a few strong between 'Cell.size', 'Cell.shape', 'Bare.nuclei' and 'Bl.cromatin'. These exhibit correlations which suggest 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 which implies a strong relationship between these two and the same is observed between 'Cell.size' or 'Cell.shape' and 'Bl.cromatin' aswell.

2.4 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 signify the spread of data points within each Predictor variable. Higher SD is observed in 'Normal.nucleoli', 'Bare.nuclei' and 'Cell.size'. This suggests greater variability in their values across the dataset, which indicates a wider spread from their respective means. Conversely, 'Mitoses' exhibits lower variability, with data points clustered closer to its mean.

3. Exploratory Data Analysis: Unsupervised Learning

In order to understand the dataset deeper, we apply unsupervised machine learning methods to identify patterns and relationships in data. The ultimate goal is to assess whether unusual tissue can be classified as malignant or benign based on its features.

3.1 K-means clusterring

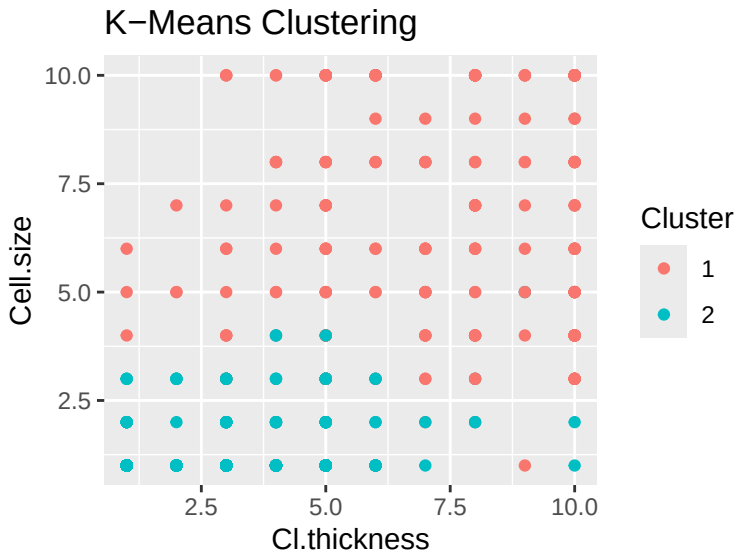

Figure 2: Representation of K-means

In the above plot (Figure 2), the two clusters correspond to benign and malignant tissues. This clustering algorithm has grouped data points based on similarities in their features. Features like "Cell.size" and "Cl.thickness" helped separate the two classes reasonably well, while some overlap exists between clusters too.

K-means clustering show how data was grouped based on feature similarities, like "withinss" indicating how compact the clusters are and "betweenss" showing how distinct they are. A high ratio of "betweenss"

to the “totss” suggests the clusters capture meaningful patterns. However, since K-Means doesn’t use the actual labels (benign or malignant), it can’t optimize for classification accuracy.

On the other hand, supervised methods directly use these labels to learn patterns and make precise predictions. For something as critical as diagnosing breast cancer, supervised learning is the better choice because it focuses on accuracy and reliability when distinguishing between malignant and benign tissues. Medical diagnostic techniques influences subtle tissue measurements like “Cl.thickness”, “Cell.size”, and “Bare.nuclei” to differentiate between malignant and benign tissues, but these individual features alone are insufficient for reliable diagnosis.

4. Supervised Learning

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.

4.1 Best subset selection with BIC

We can apply best subset selection using BIC using the bestglm package. BIC: Penalizes complexity more than AIC and often selects smaller models compared to AIC.

```
##
## 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 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.

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%.

4.2 Regularized Logistic regression with Lasso penalty

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

Plot function can be used to examine how the coefficients of each variable change as the tuning parameter is increased (Figure 3). 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.

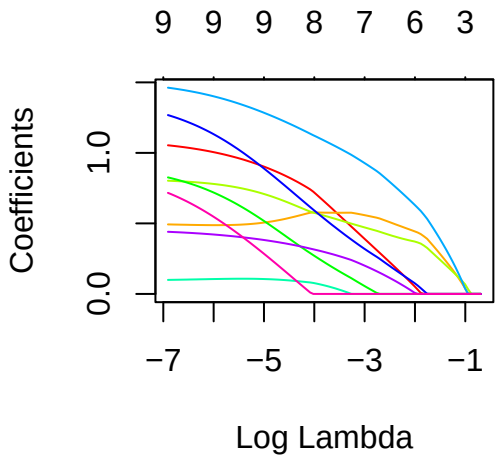

Figure 3: The effect of varying the tuning parameter in the logistic regression model with LASSO penalty for the Weekly data.

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 is 2.5%.

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.

4.3 Bayes classifier for Linear Disciminant Analysis

```
## 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
```

In the LDA model all the variables have been used. We can observe 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 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.

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.

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.

5. Cross validation & Conclusion

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 approach makes the comparison fair as same datasets are used for training and testing for all the models implemented. 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.

Among the three different models used for the Breast Cancer dataset to detect the nature of 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 model comprises of five predictor variables: Cl.thickness, Cell.size, Marg.adhesion, Bare.nuclei, and Bl.cromatin. These variables show a strong positive correlation with a 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 variablesor 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 extra variables, when included in the model, do not provide any massive 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.