Midterm Project

Ekta Chaudhary 26/03/2020

Introduction

Breast cancer is an uncontrolled growth of breast cells. It is the second most common cancer diagnosed in women in the United States. It can be benign (not dangerous to health) or malignant (has the potential to be dangerous). Benign tumors are not considered cancerous: their cells are close to normal in appearance, they grow slowly, and they do not invade nearby tissues or spread to other parts of the body. Malignant tumors are cancerous. Left unchecked, malignant cells eventually can spread beyond the original tumor to other parts of the body. It is important to identify factors that can predict the malignancy of the cancer.
We are using cellular factors like Clump Thickness, Uniformity of Cell Size, Uniformity of Cell Shape, Marginal Adhesion, Single Epithelial Cell Size, Bare Nuclei, Bland Chromatin Normal Nucleoli, Mitoses to predict the class of cancer i.e., Benign or Malignant.

Dataset

The dataset we have used is a data set provided by UCI Machine Learning Repository. This breast cancer databases was obtained from the University of Wisconsin Hospitals, Madison from Dr.William H. Wolberg. This dataset has 699 observations and 11 columns, including the class of cancer which is our outcome variable. The first column denotes the ID of the patient. Each column but the class in the raw data is a ordinal variable taking values from 1-10. Class is a binary variable with 2 values i.e., 2 for benign and 4 for malignant.

Data Cleaning

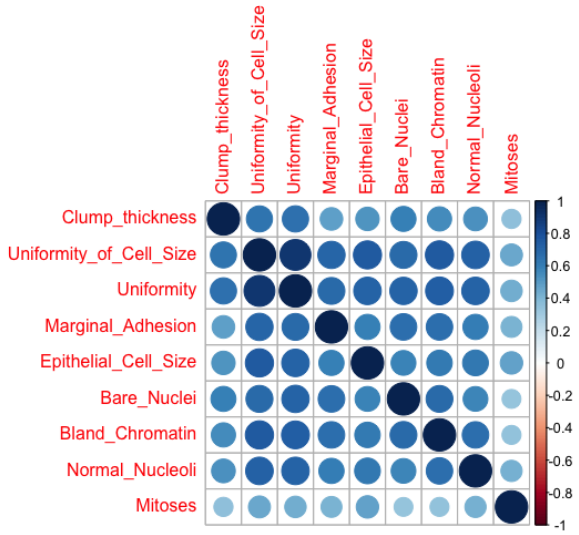
There were 16 observations in Groups 1 to 6 that contain a single missing (i.e., unavailable) attribute value, denoted by "?". The missing values were imputed using KNN Imputation. In this method of imputation, the missing values of an attribute are imputed using the given number of attributes that are most similar to the attribute whose values are missing. Using the str method that allows us to know the data type of each variable.

```
'data.frame':                   699 obs. of  11 variables:
 $ id_number            : num  1000025 1002945 1015425 1016277 1017023 ...
 $ Clump_thickness      : num  5 5 3 6 4 8 1 2 2 4 ...
 $ Uniformity_of_Cell_Size : num  1 4 1 8 1 10 1 1 1 2 ...
 $ Uniformity_of_Cell_Shape: num  1 4 1 8 1 10 1 2 1 1 ...
 $ Marginal_Adhesion     : num  1 5 1 1 3 8 1 1 1 1 ...
 $ Epithelial_Cell_Size  : num  2 7 2 3 2 7 2 2 2 2 ...
 $ Bare_Nuclei           : num  1 10 2 4 1 10 10 1 1 1 ...
 $ Bland_Chromatin       : num  3 3 3 3 3 9 3 3 1 2 ...
 $ Normal_Nucleoli       : num  1 2 1 7 1 7 1 1 1 1 ...
 $ Mitoses               : num  1 1 1 1 1 1 1 1 5 1 ...
 $ Class_cancer          : num  2 2 2 2 2 4 2 2 2 2 ...
```

The str method will allows us to know the data type of each variable. As we can see the data type of all the columns is numeric. The column Class_cancer takes two values: 2 for benign and 4 for malignant. Using ifelse on the column Class_cancer → If Class_cancer = 4 then it's malignant, else its benign. Next, converted the Class_cancer into factors.

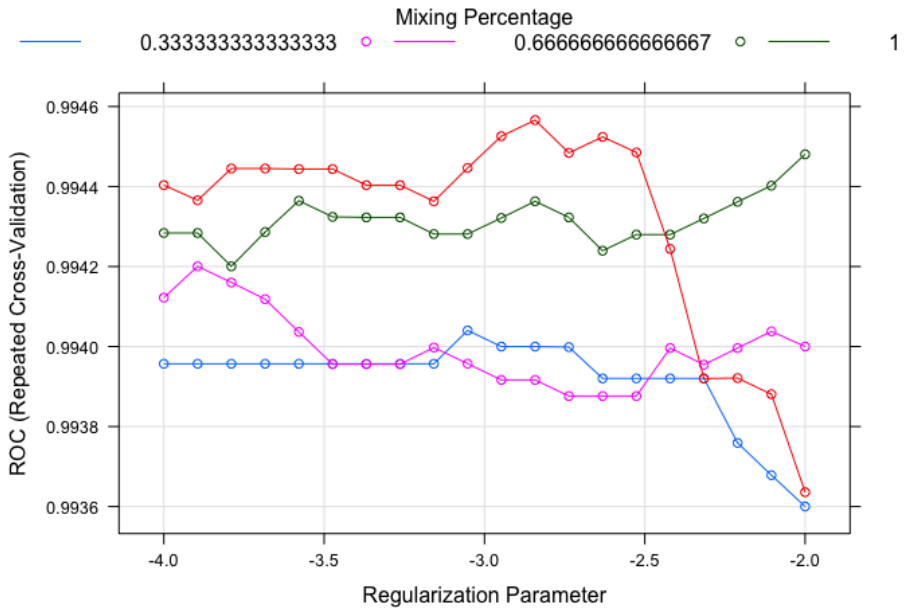
Exploratory analysis/visualization

A Correlation plot was created using all the variables. Uniformity of Cell Size is highly correlated to the Uniformity of Cell Shape.



Models

The predictor variables included in the models are Clump Thickness, Uniformity of Cell Size, Uniformity of Cell Shape, Marginal Adhesion, Single Epithelial Cell Size, Bare Nuclei, Bland Chromatin, Normal Nucleoli and Mitoses. We used regularized logistic regression, LDA and KNN to predict the class of Breast Cancer in the given population. We used the `train` function to select the optimal tuning parameters. The best Lambda is 0.0583027.



The variables that play an important roles in predicting the response, based on the p-value are: Bare nuclei, Clump thickness, Uniformity of Cell Shape, Mitoses in this order. All these

have a positive effect on the outcome i.e., as these increase, the Breast Cancer is more likely to be Malignant.

```
##                              Estimate Std. Error    z value     Pr(>|z|)
## (Intercept)                -9.0107637  1.2121934 -7.4334372 1.058115e-13
## Clump_thickness             0.4221055  0.1655169  2.5502258 1.076532e-02
## Uniformity_of_Cell_Size     0.1457212  0.2364887  0.6161865 5.377715e-01
## Uniformity_of_Cell_Shape    0.6370116  0.2746305  2.3195228 2.036671e-02
## Marginal_Adhesion           0.1479078  0.1437487  1.0289337 3.035108e-01
## Epithelial_Cell_Size       -0.2071035  0.2176655 -0.9514763 3.413627e-01
## Bare_Nuclei                 0.5172205  0.1221590  4.2339955 2.295755e-05
## Bland_Chromatin             0.2023960  0.2142662  0.9446009 3.448627e-01
## Normal_Nucleoli             0.1863717  0.1252358  1.4881661 1.367071e-01
## Mitoses                     0.6735681  0.3364145  2.0021970 4.526355e-02
```

Training performance:

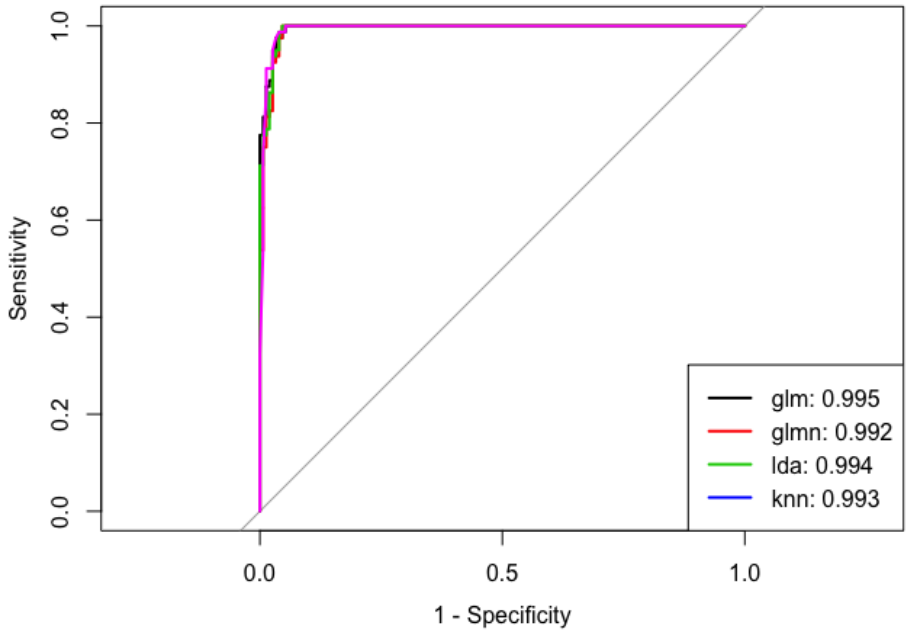
```
Models: GLM, GLMNET, LDA, KNN
Number of resamples: 50

ROC
            Min.   1st Qu.    Median      Mean 3rd Qu. Max. NA's
GLM    0.9556452 0.9883233 0.9979503 0.9935506       1    1    0
GLMNET 0.9576613 0.9917339 0.9979839 0.9945660       1    1    0
LDA    0.9556452 0.9905637 0.9979839 0.9941979       1    1    0
KNN    0.9506048 0.9881048 0.9964767 0.9910701       1    1    0

Sens
            Min.   1st Qu.   Median      Mean 3rd Qu. Max. NA's
GLM    0.9333333 0.9666667 0.983871 0.9777419       1    1    0
GLMNET 0.9354839 0.9677419 1.000000 0.9875699       1    1    0
LDA    0.9333333 0.9677419 1.000000 0.9855914       1    1    0
KNN    0.9333333 0.9666667 0.983871 0.9803656       1    1    0

Spec
         Min.   1st Qu.    Median      Mean   3rd Qu. Max. NA's
GLM    0.8125 0.8750000 0.9375000 0.9363235 1.0000000    1    0
GLMNET 0.6875 0.8235294 0.8750000 0.8942647 0.9375000    1    0
LDA    0.6875 0.8750000 0.8786765 0.9003676 0.9375000    1    0
KNN    0.7500 0.8750000 0.9375000 0.9188971 0.9852941    1    0
```

Test Performance:



Based on the training performance and the test performance, LDA is the best model to predict the class of Breast Cancer using the predictors.

Conclusions:

Based on the analysis, we can say that the cellular factors like Clump Thickness, Uniformity of Cell Size, Uniformity of Cell Shape, Marginal Adhesion, Single Epithelial Cell Size, Bare Nuclei, Bland Chromatin, Normal Nucleoli, Mitoses can be used to predict the class of cancer i.e., Benign or Malignant.

Next steps:

In the further analysis, we would want to remove the redundant variables that are highly correlated to each other.