Project1

Atharva Dharane

## Assignment

This assignment builds upon the R/RStudio class and expands the n-fold cross-validation example.

1. for the assignment use the second dataset called TCGA\_breast\_cancer\_ERpositive\_vs\_ERnegative\_PAM50.tsv that shows ER assignment for each sample (Positive vs. Negative).
2. compute 5-fold and 10-fold cross-validation estimates of prediction accuracies of ER using all genes by utilizing logistic regression and compare with NNC (2x2 table).
3. modify the the R markdown document template to report your computation and results in a table format.
4. comment on the quality of results
5. In the second part of the assignment use Project1fs.R to process a large data set by first removing all genes with sd < 1 and subsequently use Feature selection to pick top 50 genes vs top 100 genes for cross-validation based on the t-test statistic.
6. For extra credit – please replace centroid based classifier with one utilizing logistic or lasso regression similarly to the first part of the assignment and report on any difficulties.

For the assignment use Project1.Rmd file which has a number “FIXME:” labels indicating where your intervention is required.

There is a companion Project1.R where you can test and debug your code before adding it to Project1.Rmd. For extra points use lasso regression on the large dataset instead of logistic regression.x

The assignment is due on – February 25, 2021 midnight.

The submission should be zip compressed file named “project1-[your UC username].zip” (e.g. “project1-lastnfi.zip”) which includes project1.Rmd, project1.docx and any supporting R files. The zip file should be uploaded Canvas The assignment entry in Canvas will be created shortly.

## Reading data

Please add R code that reads data here - reading file: TCGA\_breast\_cancer\_ERpositive\_vs\_ERnegative\_PAM50.tsv

## Data successfully Read.

## user system elapsed   
## 0.11 0.00 0.11

## Computation

Please add R code that computes the results

##   
## Missclasification Rates for n-fold 5 for centroid are:   
## [1] 0.08212560 0.06730769 0.06763285 0.05314010 0.05797101  
## Mean of Misclassification rate : 0.06563545   
## S.D of Misclassification Rate: 0.01110843

## Missclasification Rates for n-fold 5 for GLM are:   
## [1] 0.10 0.09 0.19 0.13 0.19  
## Mean of Misclassification rate : 0.14   
## S.D of Misclassification Rate: 0.04795832

## Missclasification Rates for n-fold 5 for Lasso are:   
## [1] 0.13265306 0.13265306 0.12244898 0.07142857 0.16326531  
## Mean of Misclassification rate : 0.1244898   
## S.D of Misclassification Rate: 0.03337842

## Missclasification Rates for n-fold 10 for centroid are:   
## [1] 0.02912621 0.07619048 0.09615385 0.04807692 0.05769231 0.10576923  
## [7] 0.07766990 0.08737864 0.01941748 0.04854369  
## Mean of Misclassification rate : 0.06460187   
## S.D of Misclassification Rate: 0.02866078

## Missclasification Rates for n-fold 10 for GLM are:   
## [1] 0.07 0.08 0.10 0.07 0.07 0.04 0.09 0.05 0.08 0.06  
## Mean of Misclassification rate : 0.071   
## S.D of Misclassification Rate: 0.01791957

## Missclasification Rates for n-fold 10 for Lasso are:   
## [1] 0.05102041 0.09183673 0.09183673 0.05102041 0.04081633 0.13265306  
## [7] 0.05102041 0.05102041 0.04081633 0.04081633  
## Mean of Misclassification rate : 0.06428571   
## S.D of Misclassification Rate: 0.0308194

## Computation of KNNc and Logistic Regression using 5-fold and 10-fold cross-validation successfully completed.

## user system elapsed   
## 1.81 0.06 1.89

Note that the echo = FALSE parameter was added to the code chunk to prevent printing of the R code.

## Results

These are our results:

### 5-fold And 10-fold cross validation

## GLM kNNC  
## 5-fold mean= 0.14 sd= 0.04796 mean= 0.06564 sd= 0.01111  
## 10-fold mean= 0.071 sd= 0.01792 mean= 0.0646 sd= 0.02866

## Lasso  
## 5-fold mean= 0.12449 sd= 0.03338  
## 10-fold mean= 0.06429 sd= 0.03082

## Discussion

Based on the results, following can be inferred:

1. Logistic Regression Model -

* When we use glm model the mean of the misclassification rate decreases considerably (from 0.128 to 0.07) from 5-fold cross-validation to 10-fold cross-validation.
* Thus, we can say that the accuracy (1-misclassification\_rate) increases when we increase the number of folds.
* The reason behind this can be, as in 10-folds the amount of training data increases (9 parts) as compared to 5-folds (4 parts), thus being able to fit data more accurately.
* Also, we can see that , the standard deviation also decreases when we move from 5-fold to 10-fold, thus showing the increase variance of the prediction and the model is converging nicely.

1. Centroid Method -

* As visible from from the centroid method result, the mean and the standard deviation more or less remain the same during both the k-fold runs.
* This mey be due to the fact that, the centroid method is not a structured algorithm, thus the type or amount of data it receives during its training does not have much impact on its prediction accuracy.
* But, for this data the centroid algorithm shows much a promising accuracy (~93%) thus, indicating that the data is discretely divided in to groups and each group is considerably distant from another.

1. Lasso Method -

* Lasso regresssion works pretty much same as the glm model but only better.
* For 10-fold cross validation lasso regression model gives better accuracy than the glm model with less sd.
* One of the reasons for this may be, as this data contains lots of attributes, after penalising (adding absolute term) most of these become zero, thus only keeping the relevant ones.

1. General Comparison -

* This overall, it can be inferred as : Centroid method works best on the given data capturing all the patterns and predicting with an accuracy of almost 93%.
* The GLM model also shows promising results during 10-fold cross validation.
* Thus, based on the results, it can be said that KNNc can be used without any doubt to predict the ERStatus.

# Part 2

Change eval=TRUE when ready to include Project1fs.R

## Results of Part 2 are as follows:

## Top 50 genes which were selected are:

[1] "15544" "1108" "15607" "16034" "1461" "16167" "16427" "8000" "3"   
## [10] "548" "12642" "9037" "11143" "4777" "5994" "5462" "15802" "12829"  
## [19] "11953" "5095" "13346" "15525" "1501" "5979" "11" "17054" "8665"   
## [28] "9987" "7127" "8464" "9327" "16050" "15796" "8221" "1136" "4446"   
## [37] "1970" "4643" "15926" "14860" "13299" "17465" "17486" "5425" "7860"   
## [46] "713" "5560" "9799" "16816" "16064"

## Misclassification rates for each fold are:

[1] 0.11764706 0.05882353 0.04901961 0.08823529 0.07843137  
## [1] "mean=(0.0784) sd=(0.0268)"