Project1

Taylor Senters

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_Positive_vs_Negative_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.

The assignment is due on – February 12, 2026 midnight.

Part 1

Reading data

reading file: TCGA_breast_cancer_ERPositive_vs_ERNegative_PAM50.tsv

```
##    user  system elapsed
##   0.038   0.002   0.040
```

Computation

```
##    user  system elapsed
##   0.302   0.013   0.315
```

Results

5-fold and 10-fold Cross Validation Estimates for GLM and kNNC

	GLM_Mean	GLM_SD	kNNC_Mean	kNNC_SD
5-Fold	0.0676	0.0232	0.0656	0.0111
10-Fold	0.0694	0.0200	0.0637	0.0277

Part 1 Discussion

Both GLM and kNNC performed well with error rates around 6-7%. For the 5-fold Cross Validation, kNNC had a slightly lower error rate at 6.56% compared to the GLM error rate of 6.76%. For the 10-fold Cross Validation, kNNC again performed slightly better at 6.37% compared to GLM at 6.94%.The standard deviations for both kNNC and GLM for the 5-fold and 10-fold cross validations were all low, ranging from 0.0111 to 0.0277, indicating the predictions were consistent.

Part 2

```
## Loading required package: Matrix

## Loaded glmnet 4.1-10

## Warning: glm.fit: algorithm did not converge

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: algorithm did not converge

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: algorithm did not converge

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: algorithm did not converge

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: algorithm did not converge

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: algorithm did not converge

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: algorithm did not converge

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
## Warning: glm.fit: algorithm did not converge

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: algorithm did not converge

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: algorithm did not converge

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

Genes	Alg	Result
Top 50	Centroid	0.0433 sd=(0.0299)
Top 50	GLM	0.1577 sd=(0.046)
Top 50	Lasso	0.0497 sd=(0.0257)
Top 100	Centroid	0.0432 sd=(0.0295)
Top 100	GLM	0.1767 sd=(0.0785)
Top 100	Lasso	0.0527 sd=(0.0374)

Part 2 Discussion

For Part 2, all genes with a standard deviation less than 1 were removed, and then a t-test was used to determine the top 50 and top 100 genes. Error rates and standard deviations were then computed for kNNC, GLM, and Lasso regression for both the top 50 genes and top 100 genes, as the table above shows. Specifically for the GLM, many warnings were output stating the algorithm did not converge and fitted probabilities numerically 0 or 1 occurred. As seen in the table, GLM had the worst error rates ranging from 15-17% and higher standard deviations at 0.046 and 0.0785, indicating the model was overfitting and unstable.Implementing Lasso regression helped stabilize the model via regularization to prevent overfitting, and therefore Lasso regression had much lower error rates around 5%, much closer to the kNNC error rates around 4%.