Project1

David Lewis

Assignment

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.

Reading data

Please add R code that reads data here - reading file: TCGA_breast_cancer_ERpositive_vs_ERnegative_PAM50.tsv

```
##    user  system elapsed
##   0.141   0.009   0.149
```

Computation

Please add R code that computes the results

```
##    user  system elapsed
##   1.196   0.079   1.276
```

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 cross validation

	GLM	kNNC
5-fold	mean= 0.0676 sd= 0.0232	mean= 0.0656 sd= 0.0111
10-fold	mean= 0.0694 sd= 0.02	mean= 0.0637 sd= 0.0277

Discussion

This is what I found out

Part 2

Change eval=TRUE when ready to include Project1fs.R

```
## [1] "top 50 genes"
```

x
0.0618 sd=(0.038)

```
## [1] "top 100 genes"
```

x
0.0622 sd=(0.0496)