NAME: Akshit Singhvi

UIN: 324 006 782

Comp Proj-2

ECEN- 649

4/25/2016

Assumptions

For DLDA classification rule, $P(Y = 1) = \frac{n_1}{N} = \frac{23}{60}$ and $P(Y = 1) = \frac{n_1}{N} = \frac{23}{60}$

Code

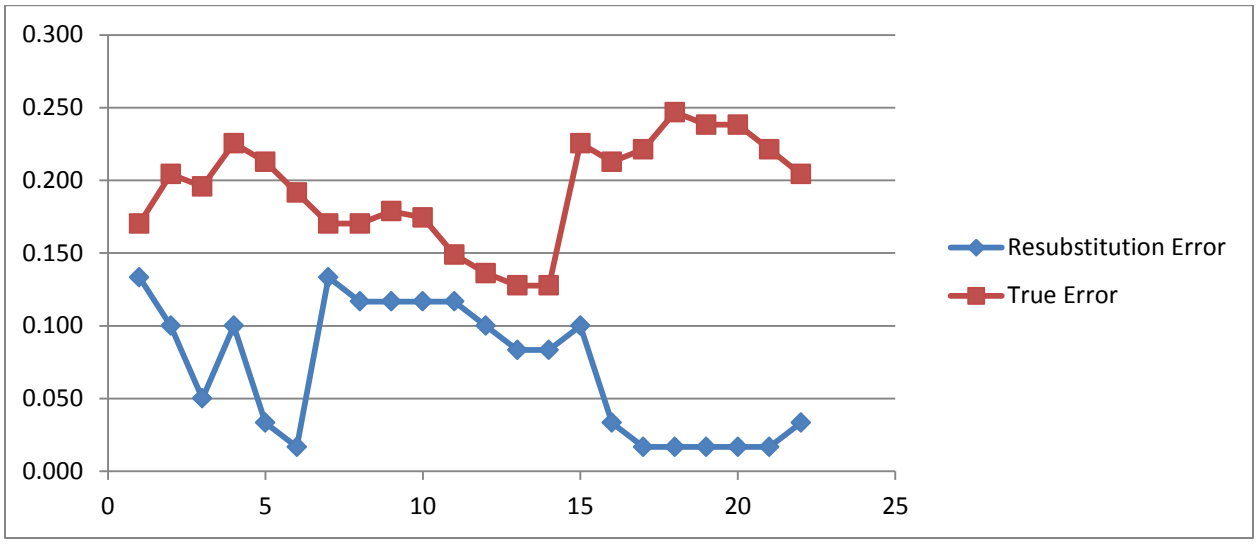
MATLAB code attached in file named *final_code.m*

Classification Rule	Feature Selection	Feature Set	Resub Error	True Error
DLDA	Exhaustive-1	'ORC6L'	0.133	0.170
DLDA	Exhaustive-2	'G3' 'G5'	0.100	0.204
DLDA	Exhaustive-3	'MMP9' 'IGFBP5.1' 'CENPA'	0.050	0.196
3NN	Exhaustive-1	'CENPA'	0.100	0.226
3NN	Exhaustive-2	'G5' 'CENPA'	0.033	0.213
3NN	Exhaustive-3	'G1' 'FGF18' 'ORC6L'	0.017	0.191
DLDA	SFS-1	'ORC6L'	0.133	0.170
DLDA	SFS-2	'ORC6L' 'G4'	0.117	0.170
DLDA	SFS-3	'ORC6L' 'G4' 'FLT1'	0.117	0.179
DLDA	SFS-4	'ORC6L' 'G4' 'FLT1' 'ALDH4'	0.117	0.174
DLDA	SFS-5	'ORC6L' 'G4' 'FLT1' 'ALDH4' 'KIAA1442'	0.117	0.149
DLDA	SFS-6	'ORC6L' 'G4' 'FLT1' 'ALDH4' 'KIAA1442' 'IGFBP5.1'	0.100	0.136
DLDA	SFS-7	'ORC6L' 'G4' 'FLT1' 'ALDH4' 'KIAA1442' 'IGFBP5.1' 'MMP9'	0.083	0.128
DLDA	SFS-8	'ORC6L' 'G4' 'FLT1' 'ALDH4' 'KIAA1442' 'IGFBP5.1' 'MMP9' 'G2'	0.083	0.128
3NN	SFS-1	'CENPA'	0.100	0.226
3NN	SFS-2	'CENPA' 'G5'	0.033	0.213
3NN	SFS-3	'CENPA' 'G5' 'PECI.1'	0.017	0.221
3NN	SFS-4	'CENPA' 'G5' 'PECI.1' 'G15'	0.017	0.247
3NN	SFS-5	'CENPA' 'G5' 'PECI.1' 'G15' 'COL4A2'	0.017	0.238
3NN	SFS-6	'CENPA' 'G5' 'PECI.1' 'G15' 'COL4A2' 'G8'	0.017	0.238
3NN	SFS-7	'CENPA' 'G5' 'PECI.1' 'G15' 'COL4A2' 'G8' 'DC13'	0.017	0.221
3NN	SFS-8	'CENPA' 'G5' 'PECI.1' 'G15' 'COL4A2' 'G8' 'DC13' 'G4'	0.033	0.204

Results and Discussion

Q) How do you compare the error estimators and feature selection methods used based on the gene sets found and the estimates of the true error?

A) The resubstitution error calculated doesn't act as a very good indicator of the True Error. For example, the following graph shows both the errors for all 22 feature sets. As it can be seen from the graph, Resubstitution error is typically very low for 3NN- SFS feature sets, but the true error is large. On Contrary, Resubstitution error is very high for DLDA- SFS, but the true error is less.



Q) How do you think the results might change if there were more training samples available or different classification rules?

A) The results might improve with more training samples available or with a different classification rule.

If more training samples are available, we can use better error estimates like cross validation or bolstered resubstituion error estimate which are estimate of the true error than the naïve resubstituion method, which suffers from high bias.

Even for same number of samples, using a classification rule like NMC (Nearest Mean Classifier) can provide better results because number of samples are very less here as compared to the feature vector size and hence any complex rule would result in over-fitting of data. So, using a simple rule like NMC might provide better results than DLDA.