Student ID: 1805115

Name: Md. Azizur Rahman Anik

Instructions on how to run

The prompt for running each dataset for logistic regression and adaboost with k= 5, 10, 15, 20 are made in the main function. Comment the functions calls except the desired one and the result can be found in result.txt file. The adult train and test file must be a csv file.

Telco Customer Churn Dataset:

Performance Measure	Training	Test
Accuracy	0.804224	0.804116
Sensitivity / Recall	0.569362	0.508620
Specificity	0.891077	0.901036
Precision	0.659056	0.627659
False Discovery Rate	0.340943	0.341853
F1 Score	0.610934	0.372340

Number of Boosting Rounds	Training	Test
5	0.780795	0.786373
10	0.777955	0.783534
15	0.781860	0.787792
20(17 Hypotheses)	0.775825	0.782824

Adult Dataset:

Performance Measure	Training	Test
Accuracy	0.850644	0.851851
Sensitivity / Recall	0.585767	0.580343
Specificity	0.934665	0.935826
Precision	0.739851	0.736633
False Discovery Rate	0.260148	0.263366
F1 Score	0.653854	0.649214

Number of Boosting Rounds	Training	Test
5	0.813882	0.815183
10	0.818304	0.818254
15	0.817014	0.816718
20	0.818058	0.818008

Credit Card Fraud Detection Dataset (Subset):

Performance Measure	Training	Test
Accuracy	0.995119	0.996096
Sensitivity / Recall	0.810880	0.849056
Specificity	0.999562	1
Precision	0.978125	1
False Discovery Rate	0.021875	0
F1 Score	0.886685	0.918367

Number of Boosting Rounds	Training	Test
5	0.994692	0.995364
10 (7 Hypotheses)	0.994692	0.995364
15(7 Hypotheses)	0.994692	0.995364
20(7 Hypotheses)	0.994692	0.995364

Observations:

1) For dataset 3, increasing the boosting rounds provided same result as same number of hypotheses was taken and others were discarded due to “if error > 0.5 then continue” condition. As total positive labeled data << total negative labeled data, this might be an issue for this kind of result.
2) As the number of weak learners is quite low, the actual effect of adaboost is not observed in the results achieved. The results tend to be similar.