CSE472 (Machine Learning Sessional)

Assignment 2: Logistic Regression and AdaBoost for Classification

Student_ID: 1805064

Constant SEED used: 40

How to Run:

- First, make sure that numpy, pandas and sklearn is installed in your device
- To install, you can write the command “pip install numpy”
- After that, to read the datasets, you need to download the 1st and 3rd dataset and absolute path of the dataset needs to be added to pd.read_csv in line 33 and 154.
- For dataset 2 no need to download, it will work just as given
- Now, to run dataset 1, we need to uncomment line 384-388 for **data_preprocessing** steps, if we want to use feature selection we need to uncomment line 392-393 and input k value in **top_k_information_gain()** and also in **logistic_regression_train()** function before epoch variable
- After that, for training and performance measure we need to comment out line 426-432
- For **Adaboosting,** we need to uncomment line 458-467
- For another 2 datasets similar process should be followed
- When data is calculated, dataset 1 and 3 is calculated without feature selection and dataset 2 is calculated with top 15 features selected.

Telco Customer Churn Dataset:

Performance measure	Training	Test
Accuracy	76.11 %	74.8 %
True positive rate (sensitivity, recall, hit rate)	0.7791	0.744
True negative rate (specificity)	0.7546	0.7495
Positive predictive value (precision)	0.5339	0.5186
False discovery rate	0.4661	0.4814
F1 score	0.6336	0.6112

Number of boosting rounds	Training	Test
5	78.56 %	76.65 %
10	78.26 %	77.08 %
15	78.1 %	78.14 %
20	78.19 %	77.22 %

Adult Dataset:

Performance measure	Training	Test
Accuracy	74.87 %	75.49 %
True positive rate (sensitivity, recall, hit rate)	0.8282	0.8427
True negative rate (specificity)	0.7235	0.7277
Positive predictive value (precision)	0.4872	0.4891
False discovery rate	0.5128	0.5109
F1 score	0.6135	0.6189

Number of boosting rounds	Training	Test
5	83.35 %	83.39 %
10	83.36 %	83.4 %
15	83.14 %	83.34 %
20	83.39 %	83.53 %

Credit Card Fraud Dataset:

Performance measure	Training	Test
Accuracy	99.48 %	99.56 %
True positive rate (sensitivity, recall, hit rate)	0.7892	0.8447
True negative rate (specificity)	0.9998	0.9995
Positive predictive value (precision)	0.9903	0.9775
False discovery rate	0.0097	0.0225
F1 score	0.8784	0.9062

Number of boosting rounds	Training	Test
5	99.48 %	99.56 %
10	99.48 %	99.56 %
15	99.48 %	99.56 %
20	99.46 %	99.56 %

Observation:

- For dataset1 and dataset2 we can see that the performance is increased after adaboosting
- But for dataset3 as the logistic regression model gives a performance much closer to 100% adaboosting is not necessarily needed for further improvement
- For dataset2, without feature selection much time is needed as dataset size is huge.