CSE472 (Machine Learning Sessional)
Assignment 1: Logistic Regression and AdaBoost for Classification

How to run the script:

Command in windows os for running the script : **python 1605006.py**
For training Telco_Customer and Adult Dataset, I have two parameters in the function named **boosting round,k** and **number of top features**. By default, training_on_telco_dataset() is run in **line no 579**. For training on Adult dataset,**line no 580** must be uncommented. However, for training on CreditCard Dataset, we have another parameter which is the **count of negative class samples**. For training on CreditCard Dataset, **line no 581** must be uncommented. I have used 20000 and 6000 negative samples accordingly. By running this script, various performance measures of Logistic Regression and Adaboost will be printed.

Hyperparameters of this experiment is given follow:
Learning rate : 0.1
Maximum iteration: 1000
Minimum error threshold: 0.5

Performance Evaluation:

Logistic Regression on Telco_Customer Dataset

Performance Measure	Training	Test
Accuracy	80.4224%	80.0568%
True positive rate (sensitivity, recall, hit rate)	54.2971%	51.9022%
True negative rate (specificity)	89.9105%	90.0096%
Positive predictive value (precision)	66.1526%	64.7458%
False discovery rate	33.8474%	35.2542%
F1 score	59.6414%	57.6169%

Accuracy of Adaboost on Telco_Customer Dataset(top 8 features)

Number of boosting rounds	Training	Test
5	73.3582%	73.8822%
10	78.328%	77.7857%
15	78.47%	77.7147%
20	78.47%	77.7147%

Logistic Regression on Adult Dataset

Performance Measure	Training	Test
Accuracy	83.4219%	83.7725%
True positive rate (sensitivity, recall, hit rate)	52.0087%	52.2101%
True negative rate (specificity)	93.3859%	93.5344%
Positive predictive value (precision)	71.3811%	71.4083%
False discovery rate	28.6189%	28.5917%
F1 score	60.1741%	60.3184%

Accuracy of Adaboost on Adult Dataset (top 8 features)

Number of boosting rounds	Training	Test
5	75.919%	76.3774%
10	75.919%	76.3774%
15	75.919%	76.3774%
20	75.919%	76.3774%

Logistic Regression on CreditCard Dataset(20000 negative class samples)

Performance Measure	Training	Test
Accuracy	97.9747%	97.9995%
True positive rate (sensitivity, recall, hit rate)	16.1616%	14.5833%
True negative rate (specificity)	100%	100%
Positive predictive value (precision)	100%	100%
False discovery rate	0%	0%
F1 score	27.8261%	25.4545%

Accuracy of Adaboost on CreditCard Dataset (30 features)

Number of boosting rounds	Training	Test
5	97.5843%	97.658%
10	97.5843%	97.658%
15	97.5843%	97.658%
20	97.5843%	97.658%

Logistic Regression on CreditCard Dataset(6000 negative class samples)

Performance Measure	Training	Test
Accuracy	96.8997%	96.9207%
True positive rate (sensitivity, recall, hit rate)	58.2902%	62.2642%
True negative rate (specificity)	100%	100%
Positive predictive value (precision)	100%	100%
False discovery rate	0%	0%
F1 score	73.6498%	76.7442%

Accuracy of Adaboost on CreditCard Dataset (30 features)

Number of boosting rounds	Training	Test
5	92.5669%	91.8399%
10	92.5669%	91.8399%
15	92.5669%	91.8399%
20	92.5669%	91.8399%

Observations:

1. In the CreditCard Dataset, if we use the full dataset, then the true positive rate(recall) tends to zero because the machine can not properly classify true predictions because there are only 492 positive class samples and 2,84,315 negative class samples. So, I have used 20,000 negative class samples for this experiment. But, the True positive rate is still low(around 16%). So, I have also used 6,000 negative samples and show that the True Positive Rate is now better(around 58%) than that of the previous experiment.

2. When we use only Logistic Regression without early stopping, there is more accuracy. When we use Adaboost, then Logistic Regression is used with early stopping criteria with 50% accuracy of base learner. By using Adaboost, this 50% accuracy is increased by around 75% for the 1st two dataset and 92% for 3rd dataset. But we see that overall accuracy of Logistic Regression performs better without early stopping. Adaboost performs better than weak/base learner(LR with early stopping) because it ensembles many weak learners. But it is not guaranteed that Adaboost outperforms normal Logistic Regression.