Machine Learning Assignment Report

Sk. Sabit Bin Mosaddek (1805106)

How to run:

To run, we can follow the following format (the arguments are mostly self explanatory),

Dataset1:

```
python3 ./1805106.py --dataset_no 1 \
                     --input_file 'dataset1/WA_Fn-UseC_-Telco-Customer-Churn.csv' \
                     --epochs 1000 \
                     --number_of_learners 5 \
                     --feature_count 15 \
                     --mini_batch_size 100 \
                     --missing_value 'mean' \
                     --learning_rate 100 \
                     --k_fold 1 \
                     --dataset_size 10000 \
                     --seed 1
```

Dataset2 (here the train and test files path are concatenated with '$'):

```
python3 ./1805106.py --dataset_no 2 \
                     --input_file 'dataset2/adult.data$dataset2/adult.test' \
                     --epochs 100 \
                     --number_of_learners 5 \
                     --feature_count 20 \
                     --mini_batch_size 100 \
                     --missing_value 'mean' \
                     --learning_rate 100 \
                     --k_fold 1 \
                     --dataset_size 40000
```

Dataset 3:

```
python3 ./1805106.py --dataset_no 3 \
                     --input_file 'dataset3/creditcard.csv' \
                     --epochs 1000 \
                     --number_of_learners 5 \
                     --feature_count 10 \
                     --mini_batch_size 100 \
                     --missing_value 'mean' \
                     --learning_rate 100 \
                     --k_fold 1 \
                     --dataset_size 20000 \
                     --all_positive True
```

Result Analysis

Dataset 1:

Performance measure of Logistic Regression implementation:

Performance measure	Training	Test
Accuracy	0.796592	0.818311
True positive rate (sensitivity, recall, hit rate)	0.516876	0.511173
True negative rate (specificity)	0.899103	0.922931
Positive predictive value (precision)	0.652464	0.693182
False discovery rate	0.347536	0.306818
F1 score	0.576809	0.588424

The **accuracy** of Adaboost implementation with Logistic Regression:

Number of boosting rounds	Training	Test
5	0.792865	0.814053
10	0.790557	0.810504
15	0.790735	0.809084
20	0.791800	0.809794

Observation: Logistic Regression may have fit the model well enough and when we introduced more learners using adaboost, it may happen that some bad data has diverted the result.

Dataset 2:

performance measure of Logistic Regression implementation:

Performance measure	Training	Test
Accuracy	0.768680	0.770837
True positive rate (sensitivity, recall, hit rate)	0.180334	0.183567
True negative rate (specificity)	0.955299	0.952473
Positive predictive value (precision)	0.561334	0.544333
False discovery rate	0.438666	0.455667
F1 score	0.272973	0.274548

The **accuracy** of Adaboost implementation with Logistic Regression:

Number of boosting rounds	Training	Test
5	0.797365	0.798845
10	0.802402	0.803390
15	0.813335	0.814262
20	0.815331	0.814815

Observation: Adaboost has provided a slight improvement

Dataset 3:

performance measure of Logistic Regression implementation:

Performance measure	Training	Test
Accuracy	0.983099	0.981404
True positive rate (sensitivity, recall, hit rate)	0.139949	0.060606
True negative rate (specificity)	1.000000	1.000000
Positive predictive value (precision)	1.000000	1.000000
False discovery rate	0.000000	0.000000
F1 score	0.245536	0.114286

The **accuracy** of Adaboost implementation with Logistic Regression:

Number of boosting rounds	Training	Test
5	0.989349	0.990202
10	0.989349	0.990202
15	0.989349	0.990202
20	0.989499	0.990402

Observation: At first, Adaboost has provided an improvement but then it has almost got saturated.