Report On Assignment 2

Logistic Regression and AdaBoost for Classification

Date: 9 December 2023

Name: Ishika Tarin

ID: 1805092

Section: B2

Subject: CSE472

Machine Learning Sessional

Instructions on how to run the script

- Following command is for how to run the .py file:

```
PS E:\BUET CSE  18\4-2\CSE 472\Offline 2> python -u 1805092.py
```

To run the .py file, *python –u 1805092.py* should be given as command in terminal.

- Following instructions are for how to run the preprocessing steps:

```
469     train_x, test_x, train_y, test_y = preprocessTelco()
470     train_x, test_x, train_y, test_y = preprocessAdult()
471     train_x, test_x, train_y, test_y = preprocessCreditCard()
```

- To preprocess dataset 1 , from the code snippet, line no 470 and 471 should be commented, leaving only *train_x, test_x, train_y, test_y = preprocessTelco()*

- To preprocess dataset 2 , from the code snippet, line no 469 and 471 should be commented, leaving only *train_x, test_x, train_y, test_y = preprocessAdult()*

- To preprocess dataset 1 , from the code snippet, line no 469 and 470 should be commented, leaving only *train_x, test_x, train_y, test_y = preprocessCreditcard()*

- To run only Weaklearner /logistic regression, codes from 476 to 530 should be uncommented and codes from 535 to 575 should be commented.

- To run only Adaboost with logistic regression with different k values [5,10,15,20] , codes from 476 to 530 should be commented and codes from 535 to 575 should be uncommented.
-

Performance Measures

In Weak learner: Learning rate = 0.01, Number of iterations = 8000

Logistic Regression on training set and test set for given dataset 1 (Telco dataset)

Performance Measure	Training	Test
Accuracy	**0.8019**	**0.7857**
True positive rate (sensitivity, recall, hit rate)	**0.5302**	**0.4863**
True negative rate (specificity)	**0.9009**	**0.8900**
Positive predictive value (precision)	**0.6611**	**0.6062**
False discovery rate	**0.0991**	**0.1100**
F1 score	**0.5885**	**0.5396**

Accuracy of Adaboost implementation with Logistic Regression on training set and test set for given dataset 1 (Telco dataset)

Number of boosting rounds	Training	Test
5	**0.7437**	**0.7140**
10	**0.6228**	**0.6267**
15	**0.5659**	**0.5422**
20	**0.6692**	**0.6508**

Logistic Regression on training set and test set for given dataset 2 (Adult dataset)

Performance Measure	Training	Test
Accuracy	**0.8403**	**0.8393**
True positive rate (sensitivity, recall, hit rate)	**0.5429**	**0.5364**
True negative rate (specificity)	**0.9346**	**0.9329**
Positive predictive value (precision)	**0.7248**	**0.7121**
False discovery rate	**0.0654**	**0.0671**
F1 score	**0.6208**	**0.6119**

Accuracy of Adaboost implementation with Logistic Regression on training set and test set for given dataset 2 (Adult dataset)

Number of boosting rounds	Training	Test
5	**0.6864**	**0.6847**
10	**0.6117**	**0.5999**
15	**0.7496**	**0.7592**
20	**0.6967**	**0.6922**

Logistic Regression on training set and test set for given dataset 3 (Credit card dataset)

Performance Measure	Training	Test
Accuracy	**0.9951**	**0.9944**
True positive rate (sensitivity, recall, hit rate)	**0.8030**	**0.7604**
True negative rate (specificity)	**0.9998**	**1.0000**
Positive predictive value (precision)	**0.9907**	**1.0000**
False discovery rate	**0.0002**	**0.0000**
F1 score	**0.8870**	**0.8639**

Accuracy of Adaboost implementation with Logistic Regression on training set and test set for given dataset 3 (Credit card dataset)

Number of boosting rounds	Training	Test
5	**0.6876**	**0.6924**
10	**0.5857**	**0.5858**
15	**0.6546**	**0.6506**
20	**0.7241**	**0.7265**

Observations for the logistic Regression implementation on both training and test sets

- All models show high accuracy, indicating that the majority of predictions are correct.
- Dataset 3 (Credit card) has the highest sensitivity, suggesting a good ability to detect positive cases. It also has extremely high specificity, indicating a strong ability to correctly identify negative cases.
- Dataset 3 (Credit card) achieves near-perfect precision, meaning the positive predictions are highly reliable.
- All models have low false discovery rates, indicating a low rate of false positive predictions.
- F1 scores, which consider both precision and recall, are reasonably high across all datasets.
- From the observation, it can be concluded that, dataset 3, dealing with credit card transactions, exhibits a significant class imbalance where fraudulent transactions (positive class) are much less frequent than non-fraudulent transactions (negative class). In such cases, the model has achieved high accuracy by simply predicting the majority class. However, other metrics like sensitivity and precision become crucial for understanding the model's performance, especially in detecting the rare positive cases.

Observations for the Adaboost implementation with Logistic Regression

- The relationship between the number of boosting rounds and accuracy varies across datasets, indicating that the optimal number of rounds might be dataset dependent.
- It is seen that, from telco dataset, the performance on the training set decreases as the number of boosting rounds increases, suggesting potential

overfitting. But in adult dataset, there is no consistent trend in accuracy with changes in the number of boosting rounds. Again, the model's performance on the credit card dataset seems less affected by the number of boosting rounds compared to the other datasets.

Overall Observations

- AdaBoost is an ensemble method that combines multiple weak learners (in this case, Logistic Regression) to create a strong learner. As the number of boosting rounds increases, there is a risk of overfitting the training data, which may lead to a drop in performance on the test set.
- Different datasets have different characteristics, and the optimal algorithm or ensemble method can vary. AdaBoost may not always outperform Logistic Regression, especially if the dataset does not meet the assumptions or requirements that favor boosting algorithms.
- AdaBoost can be sensitive to noisy data and outliers. If the datasets contain noise or outliers, AdaBoost may struggle to handle them effectively, impacting its overall performance.
- AdaBoost relies on combining weak learners, which may be less effective if the relationships between features are not well captured by the weak learners or if there are complex interactions that Logistic Regression is better at modeling.

That’s why accuracy may drop in adaboost in comparison with using only logistic regression.