Course No. : CSE 472

Course Title : Machine Learning Sessional

Assignment-2 : Logistic Regression and AdaBoost for Classification

Submitted By:
Name: Sheikh Hasanul Banna
Roll: 1805094
Section: B2

How to run :

- Install numpy : pip install numpy
- Install pandas : pip install pandas
- Install Sklearn : pip install -U scikit-learn
- Install scipy : pip install scipy
- Run the Python file. It will ask for prompt to choose which dataset you want to use.
- After selecting dataset, you will be asked for prompt to choose the number of features you want to use.
- If you had selected Creditcard dataset, you will be further asked to choose between small sample(around 20k) or large sample.
- After that the execution will begin. First logistic regression will be executed and you will see the accuracy, sensitivity, specificity, precision, false discovery rate and f1 score of training set first and then test set.
  After logistic regression is done, AdaBoost will begin to compute accuracy for K=5,10,15,20.

Data Table:

Feature used : 10

Logistic Regression

For Telco Churn dataset:

Performance Measure	Training	Test
Accuracy	0.7325168	0.7267565
True positive rate (sensitivity)	0.81	0.7859078
True negative rate (specificity)	0.7044025	0.705769
Positive predictive value (precision)	0.498563	0.4865771
False discovery rate	0.5014361	0.513422
F1 score	0.6172212	0.601036

For Adult Salary Dataset:

Performance Measure	Training	Test
Accuracy	0.740786	0.746330
True positive rate (sensitivity)	0.824129	0.836453
True negative rate (specificity)	0.714349	0.718455
Positive predictive value (precision)	0.477852	0.478862
False discovery rate	0.522147	0.521137
F1 score	0.604942	0.609049

For Creditcard Dataset(small,around 20000 samples):

Performance Measure	Training	Test
Accuracy	0.994753	0.996340
True positive rate (sensitivity)	0.798982	0.848484
True negative rate (specificity)	0.999562	1.0
Positive predictive value (precision)	0.978193	1.0
False discovery rate	0.021806	0.0
F1 score	0.879551	0.918032

For Creditcard Dataset(full):

Performance Measure	Training	Test
Accuracy	0.998841	0.998929
True positive rate (sensitivity)	0.407124	0.474747
True negative rate (specificity)	0.999863	0.999841
Positive predictive value (precision)	0.837696	0.839285
False discovery rate	0.162303	0.160714
F1 score	0.547945	0.606451

AdaBoost Implementation

For Telco Churn Dataset:

Number of boosting rounds	Accuracy(Training)	Accuracy(Test)
5	0.788959	0.781405
10	0.783990	0.780695
15	0.785764	0.781405
20	0.781405	0.781405

For Adult Salary Dataset:

Number of boosting rounds	Accuracy(Training)	Accuracy(Test)
5	0.831879	0.832565
10	0.833261	0.833855
15	0.833077	0.833548
20	0.835933	0.836005

For Creditcard Dataset(small, around 20000 samples):

Number of boosting rounds	Accuracy(Training)	Accuracy(Test)
5	0.994753	0.996340
10	0.994753	0.996340
15	0.994753	0.996340
20	0.994753	0.996340

For Creditcard Dataset(full):

Number of boosting rounds	Accuracy(Training)	Accuracy(Test)
5	0.998801	0.998788
10	0.998801	0.998788
15	0.998801	0.998788
20	0.998801	0.998788

Observations:

- Increasing value of boosting round does not always increase accuracy.
- For adult salary and credit card datasets, logistic regression yielded better result for test set than training set, but for telco dataset, training set yielded better result than test set.
- Changing the value of learning rate can have drastic improvement in some cases. Instead of constant learning rate, adaptive learning rate could be used to improve results.
- Changing the seed value so that training and test sets can be split differently sometimes shows major improvement.
- Because Creditcard dataset had a major difference in positive and negative results, the models could label more data as negative, and as there were very little positive results, negative values had overwhelming majority, resulting high accuracy and low false positives.