INDIANA UNIVERISTY BLOOMINGTON

PA 2: Naïve Bayes / Logistic Regression

I 526 Applied Machine Learning

SHORT REPORT

Submitted by

Nihar Khetan
Masters Candidate in Computer Science
School of Informatics and Computing
Indiana University
nkhetan@indiana.edu

Ghanshyam Malu
Masters Candidate in Computer Science
School of Informatics and Computing
Indiana University
gmalu@indiana.edu

Xiao Liang
Masters Candidate in Health Informatics.
School of Informatics and Computing
Indiana university
liang25@indiana.edu

Under the Guidance of:

Professor Sriraam Natarajan
Asst. Professor, School of Informatics and Computing
Indiana University

Contents

1. Logistic Regression

Index	Learning Rate	*Convergence Threshold	Accuracy
1	0.001	0.001	100%
2	0.005	0.001	82.9%
3	0.006	0.001	80%
4	0.07	0.001	82.9%
5	0.01	0.001	80%
6	0.1	0.001	88.57%

*Convergence Threshold: is the difference in norm of weight vectors when the model converges

Confusion Matrix

Index 1:

		Predicted Value	
		1	***0***
Expected Values	***1***	21	0
	0	0	14

Index 2:

		Predicted Value	
		1	***0***
Expected Values	***1***	15	6
	0	0	14

Index 3:

		Predicted Value	
		1	***0***
Expected Values	***1***	14	7
	0	0	14

Index 4:

		Predicted Value	
		1	***0***
Expected Values	***1***	15	6
	0	0	14

Index 5:

		Predicted Value	
		1	***0***
Expected Values	***1***	14	7
	0	0	14

Index 6:

		Predicted Value	
		1	***0***
Expected Values	***1***	18	3
	0	1	13

Effect of Learning Rate

When learning rate is low logistic regression has more tendency to settle to a local minima which is required. However when we increase learning rate then model has less chance to settle to a local minima hence accuracy decreases.

To summarize : at learning rate greater than 0.001 we get lesser accuracy.

Weka Results

Correctly Classified Instances 35 100 %

a b <-- classified as

21 0 | a = 0

0 14 | b = 1

How weka results are different form our results?

In Weka we did simple logistic regression where we did not choose eeta (0.001) that is the learning rate. For our model we had two parameters one was the learning rate and the other was convergence threshold. We kept convergence threshold fixed to 0.001 and varies eeta to get different results. We got the best result for eeta = 0.001 which was 100% accuracy and was same as weka.

In to summarize we got same results as weka when learning rate was 0.001.

2. Naïve Bayes

With Laplacian Correction

```
Incorrect Classification Count: 0   Correct Classification Count: 35

   ***********************************************************************
                       Accuracy is 100.000 %
   ***********************************************************************
```

Confusion Matrix

		Predicted Value	
		1	***0***
Expected Values	***1***	21	0
	0	0	14

Weka results

Correctly Classified Instances 35 100 %

=== Confusion Matrix ===

a b <-- classified as

21 0 | a = 0

0 14 | b = 1

How weka results are different form our results?

Weka results and our results with Laplacian Correction are the same. However, if the Laplacian Correction is not done then our model gives 2 misclassifications.

3. Why does one method does better than other, speculate

In this case, as the dataset is small both classifiers perform equally well. But if dataset was large Logistic regression would have performed better.

In Logistic Regression we might be overfitting as the dataset is very small. In logistic regression weights can be optimized better if the size of the dataset is large however with small dataset it has tendency to overfit.

In this dataset as we are unaware of the features, they might be dependent on each other thus violating the key assumptions of Naïve Bayes, thus Naïve Bayes classifier has a tendency to underfit but Laplacian Correction takes care of it.