Programming Assignment 2 – Applied Machine Learning

Results

After coding out the algorithms, we first run Logistic Regression on the simplified dataset. We obtained 100% accuracy on the test set. We then tried running the Naïve Bayes code we wrote on the simplified data set, we again obtained 100% accuracy. These results stumped us, as Prof. Natarajan told us that if you get 100% accuracy the result is wrong.

We started scanning the code for bugs, out code seemed alright, so we then we put the datasets in WEKA. Surprisingly, for Naïve Bayes, even WEKA gave 100% accuracy. However, for Logistic Regression, WEKA gave an accuracy of 94.28%.

Confusion Matrices

Confusion matrix for Naïve Bayes algorithm with Laplace Correction on the simplified dataset using our code.

	1	0	Predicted Result
1	14	0	
0	0	21	

Expected Result

Confusion matrix for Logistic Regression algorithm on the simplified dataset using our code.

	1	0	Predicted Result
1	14	0	
0	0	21	

Expected Result

Confusion matrix for Naïve Bayes algorithm with Laplace Correction in WEKA on the simplified dataset.

	1	0	Predicted Result
1	14	0	
0	0	21	

Expected Result

Confusion matrix for Logistic Regression algorithm in WEKA on the simplified dataset.

	1	0	Predicted Result
1	14	0	
0	2	19	

Expected Result

When we were getting these results, we ran the program with one vs rest approach for other class labels.

For class label = 2, Naïve Bayes gave an accuracy of 88.5% and Logistic Regression gave an accuracy of 80%.

Confusion Matrix for Logistic Regression.

	2	0	Predicted Result
2	0	7	
0	0	28	

Expected Result

Confusion Matrix for Naïve Bayes. Confusion Matrix for Logistic Regression.

	2	0	Predicted Result
2	0	2	
0	2	31	

Expected Result

Varying the Learning Rates

We now vary the learning rate for logistic regression and see what difference it makes in the results. For simplicity, we keep a fixed threshold = 0.01

With learning rate = 0.000001, the algorithm converges after 1 iteration and gives the confusion matrix as below.

Confusion Matrix for Logistic Regression.

	1	0	
			Predicted Result
1	14	0	
0	0	21	

Expected Result

With learning rate = 0.01, the algorithm converges after 76 iterations and gives the confusion matrix as below.

Confusion Matrix for Logistic Regression.

	1	0	
			Predicted Result
1	14	0	
0	0	21	

Expected Result

With learning rate = 1, the algorithm converges after 9 iterations and gives the confusion matrix as below.

Confusion Matrix for Logistic Regression.

	1	0	
			Predicted Result
1	14	0	
0	0	21	

Expected Result

For all the above learning rates, logistic regression converges but with different number of iterations.

Observations

In both the algorithms we implemented, we obtained an accuracy of 100%, which was unusual but when we run the algorithms in WEKA on the given training and test dataset we obtained similar results. In WEKA, Naïve Bayes gives an accuracy of 100% and Logistic Regression gives an accuracy of 94.29%. Hence Naïve Bayes works better than Logistic Regression in this case because Logistic Regression overfits in the case of small datasets. (Jordan) The data set given was too small and with a lot of features. The training set did not provide a good variety of tuples for training, we observed that whenever feature number 4 (counting from 1) was 1, the actual label was 1. Even an extremely simple

classifier like a decision stump can correctly classify the given test data because even the test data follows the pattern of the training set ie: the actual label is 1 whenever feature number 4 is 1. Apart from this observation, we also read papers (Jordan) (Schapire & Blei) which confirm our conclusion that for small data sets naïve Bayes works better than logistic regression but with large datasets, logistic regression works better.

References

Jordan, A. N. (n.d.). *On Discriminativ evs Generative Classifiers: A comparison of Logistic Regression and naive Bayes.* Retrieved from http://ai.stanford.edu/~ang/papers/nips01-discriminativegenerative.pdf

Natarajan, P. (n.d.). Class slides.

Schapire, R., & Blei, D. (n.d.). *COS 424: Interacting with Data.* Retrieved from http://www.cs.princeton.edu/courses/archive/spr07/cos424/scribe_notes/0410.pdf

By: Huzefa Dargahwala, Rahul Pasunuri