CS 6375.001
Machine Learning

By

Prof. Anjum Chida

ASSIGNMENT – 3

VEDANT PARESH SHAH VXS200021

Naive Bayes for Text Classification

naive.py : Implements the multinomial Naive Bayes algorithm for text classification

Total Test files (478) – Ham files (348) and Spam files (130)

Before Removing the stop words: Naïve Bayes

Total Word Count: 9186

Naive Bayes Accuracy for Spam and Ham Emails Classification: 92.19%

After Removing the stop words: Naïve Bayes

Total Word Count: 9068

Naive Bayes Accuracy for Spam and Ham Emails Classification: 92.31%

Observation:

Overall Total Accuracy on Ham and Spam files marginally increased after removing stop words. Marginal increase in accuracy of the output shows that stop words in the files are not too common and their conditional probabilities don't have much effect on classification probabilities.

Before Removing the stop words: Naïve Bayes

Accuracy 0.9219214600635702

After Removing the stop words: Naïve Bayes

Accuracy 0.9231868643222761

Logistic Regression for Text Classification

logistic.py : Implement the MCAP Logistic Regression algorithm with L2 regularization

Total Test files (478) – Ham files (348) and Spam files (130)

Learning Rate = λ, η and number of iterations = n are used to calculate the accuracy of the text classification using logistic regression.

η =0.01 throughout the program

Total Word Count: 9186 Total Word Count: 9068

Learning Rate and Number of Iterations	**Before Removing the stop words**	**After Removing the stop words**
λ =0.1, n=50	0.897489539748954	0.8472803347280334
λ =0.01, n=50	0.8744769874476988	0.8640167364016736
λ =0.001, n=50	0.805439330543933	0.8179916317991632

Observation:

As we increase the value of λ the accuracy of text classification through Logistic Regression increases. Also as we increase number of iterations the we get better accuracy of text classification. Overall Total Accuracy on Ham and Spam files increased after removing stop words. Smaller values of the regularization parameter have no effect on classification results. Only significant values for lambda brought significant changes in classification results.

As the value of lamda increases from λ =0.001 to λ = 0.1 for n = 50:-

The Accuracy Before Removing the stop words: - 80.54% to 89.74%

The Accuracy After Removing the stop words: - 81.79% to 84.72%

Total Word Count: 9186 Total Word Count: 9068

Learning Rate and Number of Iterations	Before Removing the stop words	After Removing the stop words
λ =0.1, n=10	0.8870292887029289	0.8451882845188284
λ =0.01, n=10	0.8221757322175732	0.8514644351464435
λ =0.001, n=10	0.6736401673640168	0.4267782426778247

Total Word Count: 9186 Total Word Count: 9068

Learning Rate and Number of Iterations	Before Removing the stop words	After Removing the stop words
λ =0.1, n=5	0.797071129707113	0.8347280334728033
λ =0.01, n=5	0.8368200836820083	0.8389121338912134
λ =0.001, n=5	0.5815899581589958	0.3326359832635983

Total Word Count: 9186 Total Word Count: 9068

Learning Rate and Number of Iterations	Before Removing the stop words	After Removing the stop words
λ =0.1, n=2	0.7301255230125523	0.8096234309623431
λ =0.01, n=2	0.7866108786610879	0.6673640167364017
λ =0.001, n=2	0.2740585774058577	0.2740585774058577

Total Word Count: 9186 Total Word Count: 9068

Learning Rate and Number of Iterations	Before Removing the stop words	After Removing the stop words
λ =0.1, n=1	0.7384937238493724	0.7845188284518828
λ =0.01, n=1	0.7719665271966527	0.502092050209205
λ =0.001, n=1	0.2719665271966527	0.2719665271966527