Report of Homework 2 CS 6375: Machine Learning, spring 2015

Name: Nikhil Rao

Net ID: ngr140030

Classification: Naïve Bayes and Logistic Regression

Naïve Bayes for Text Classification:

Accuracy:

** Without stop words being removed from the corpus of the text or documents or e-mail in this case**

Accuracy of Naive Bayes without removing Stop Words:

Spam % Accuracy: 97.6923076923077

Ham % Accuracy: 93.39080459770115

#Accuracy here is defined as the percentage of documents in respective test set (Spam set and Ham Set) classified to their corresponding set – Spam or Ham.

Accuracy:

** After removing the stop words from the corpus using http://www.ranks.nl/stopwords - long list of stop words**

Accuracy of Naive Bayes after removal of Stop Words:

Spam % Accuracy: 96.92307692307692

Ham % Accuracy: 92.24137931034483

#Accuracy here is defined as the percentage of documents in respective test set (Spam set -130 and Ham Set – 348) classified to their corresponding set – Spam or Ham.

Logistic Regression for Text Classification:

Accuracy observed for various values of leaning rate (eta), penalty regularization term – Lambda.

And number of iterations (hard-limit) to come to convergence. Please note that it takes a significant amount of time to complete the hard-limit number of iterations. These have been run on Amazon EC2 to run at an improved speed.

Learning rate: 0.01

Lambda / Hard limit		0.01		0.1		0.5	
		Stop words	Removed Stop words	Stop words	Removed stop words	Stop words	Removed stop words
50	Spam	92.14	92.726	86.07	87.69	83.97	84.769
	Ham	87.24	86.67	91.22	89.72	88.48	87.068
100	Spam	95.56	94.97	91.42	92.14	93.64	94.153
	Ham	90.88	88.56	92.76	91.954	96.81	95.754
500	Spam	97.52	96.4	96.82	97.153	95.96	95.153
	Ham	97.95	97.12	98.23	96.69	97.954	96.869

Learning rate: 0.025

Lambda / Hard limit		0.01		0.1		0.5	
		Stop words	Removed Stop words	Stop words	Removed stop words	Stop words	Removed stop words
50	Spam	91.56	90.12	94.23	93.56	96.81	95.69
	Ham	86.72	85.24	91.09	90.16	85.96	84.015
100	Spam	93.72	92.86	95.76	94.72	96.96	95.97
	Ham	89.07	88.76	94.56	93.59	97.92	97.153
500	Spam	96.92	95.72	97.169	96.96	97.85	96.54
	Ham	97.96	96.62	98.153	97.69	97.153	97.954

Learning rate: 0.5

Lambda / Hard limit		0.01		0.1		0.5	
		Stop words	Removed Stop words	Stop words	Removed stop words	Stop words	Removed stop words
50	Spam	90.14	89.56	91.94	90.23	94.56	93.28
	Ham	88.96	88.26	92.53	91.153	96.52	94.96
100	Spam	92.69	91.72	93.56	92.84	94.24	93.89
	Ham	95.64	94.01	95.95	94.05	94.69	93.52
500	Spam	95.96	93.05	94.56	95.116	96.47	96.86
	Ham	96.64	94.96	97.72	96.89	96.96	95.14

By removing stop words from the documents we are redefining our vocabulary to contain less words than the dictionary constructed by considering stop words. I observe that in, there is slight reduction (~1 - 2%) in both cases – Ham and Spam accuracy. This might be due to distribution of Stop words in either cases.

Also, there is no strict relation between the stop words and the documents, since if a document contain more stop words, then we just get rid of the stops words before processing. This restricts the size of the word set in consideration.

Logistic regression:

The values of the parameters – Learning rate – dictates how fast we move towards the convergence point.

Also, the parameter regularization term – Lambda controls the penalty that we add to the weight calculation. Since it takes more computation power to find the exact convergence point, we restrict on the number of iterations our model takes to get closer to convergence point.

From the result set, I observe that when I increase the iterations in general, there is an increase in accuracy. However, if the learning rate is increased, the value of accuracy varies, it either increases for some cases or decreases and this is due to the fact that if we move at a faster rate there we converge to a wrong point. Also, the variations in Lambda in combination with the value of learning rate and number of iteration.