Machine Learning

CS 6375

Naïve Bayes and Logistic Regression

Submitted by:

Parth Mehta

Net ID: pjm150030

As a setup, we have spam and ham file and I have calculated accuracy with the help of Naïve Bayes and Logistic Regression both.

I have implemented it with stop words and without stop words.

NAÏVE BAYES

Accuracy with Stop Words:

Spam Accuracy: 84.61%

Ham Accuracy: 97.98%

Accuracy after removal of Stop Words:

Spam Accuracy: 86.92%

Ham Accuracy: 96.55%

As we can see, naïve bayes doesn’t show a big change in accuracy because this algorithm is based on count of words occurred in HAM file, count of words occurred in SPAM file. There are not many stop words in the training data provided to us. So, there is not much impact of stop words in our number of features. But, we can see there is an increase in accuracy in Spam after removing stop words. It means that there are more stop words in spam train data and after removing stop words, we removed noise from data.

LOGISTIC REGRESSION

Accuracy Calculation with Stop Words:

Iterations	Learning Rate	Lambda	Spam Accuracy	Ham Accuracy
100	0.01	0	85.38%	94.54%
100	0.01	1	86.15%	94.82%
100	0.01	5	12.30%	99.71%
100	0.01	10	8.46%	99.71%
100	0.01	15	83.84%	87.93%
100	0.01	20	5.38%	99.71%
100	0.01	25	4.61%	99.71%

Accuracy Calculation without Stop Words:

Iterations	Learning Rate	Lambda	Spam Accuracy	Ham Accuracy
100	0.01	0	86.15%	94.35%
100	0.01	1	88.46%	96.83%
100	0.01	5	90.0%	97.41%
100	0.01	10	60.76%	99.42%
100	0.01	15	96.92%	18.67%
100	0.01	20	16.92%	99.71%
100	0.01	25	17.69%	99.71%

As we can see, there is an increase in HAM and SPAM accuracy after removing the stop words. So, removal of Stop words are helping in improving the accuracy. We can see that sometime there is sudden decrease in the spam Accuracy after increasing value of lambda. This is the case when after penalizing weights the decision boundary has included spam data on ham side. This can be the case of under fitting as the decision boundary is not proper.