NAÏVE BAYES ALGORITHM

The Naïve Bayes algorithm is tested for five different Laplace Smoothing constants and the results are shown as under.

For Laplace Smoothing factor = 1

Accuracy before removing stop words

The accuracy over 130.0 spam files is: 98.46153846153847 %

The accuracy over 348.0 ham files is: 96.26436781609196 %

Accuracy after removing the stop words

The accuracy over 130.0 spam files is: 99.23076923076923 %

The accuracy over 348.0 ham files is: 95.97701149425288 %

For Laplace Smoothing Factor = 2

Accuracy before removing stop words

The accuracy over 130.0 spam files is: 99.23076923076923 %

The accuracy over 348.0 ham files is: 95.6896551724138 %

Accuracy after removing the stop words

The accuracy over 130.0 spam files is: 99.23076923076923 %

The accuracy over 348.0 ham files is: 94.82758620689656 %

For Laplace Smoothing Factor = 3

Accuracy before removing stop words

The accuracy over 130.0 spam files is: 99.23076923076923 %

The accuracy over 348.0 ham files is: 94.82758620689656 %

Accuracy after removing the stop words

The accuracy over 130.0 spam files is: 99.23076923076923 %

The accuracy over 348.0 ham files is: 93.39080459770115 %

For Laplace Smoothing constant = 4

Accuracy before removing stop words

The accuracy over 130.0 spam files is: 99.23076923076923 %

The accuracy over 348.0 ham files is: 93.10344827586206 %

Accuracy after removing the stop words

The accuracy over 130.0 spam files is: 99.23076923076923 %

The accuracy over 348.0 ham files is: 91.66666666666666 %

For Laplace Smoothing constant = 5

Accuracy before removing stop words

The accuracy over 130.0 spam files is: 99.23076923076923 %

The accuracy over 348.0 ham files is: 92.24137931034483 %

Accuracy after removing the stop words

The accuracy over 130.0 spam files is: 99.23076923076923 %

The accuracy over 348.0 ham files is: 91.37931034482759 %

The highest average of accuracy was observed for Laplace smoothing factor = 2. With the greater factor the overall accuracy tends to fall down. Also it appears there is a higher percentage of stop words in the spam files resulting in an increase in the percentage accuracy. There is a slight fall in the percentage of accuracy of determining the Ham files.

LOGISTICAL REGRESSION

As an input to the program, we provide the number of iterations, regularization factor (λ) and the learning rate. In addition to this weight (w0) is a random number in a range between 0 and 2. Other initial weights have been randomized in a range between 0 and 40.

Enter the number of iterations, regularization factor and the learning rate

5 30 0.01

Accuracy before removing the stop words

The accuracy over spam files is: 94.61538

The accuracy over ham files is: 99.425285

Accuracy after removing the stop words

The accuracy over spam files is: 5.3846154

The accuracy over ham files is: 42.24138

Where 5 is the number of iterations, 30 is the regularization factor and 0.01 is the learning rate.

Enter the number of iterations, regularization factor and the learning rate

10 20 0.01

Accuracy before removing the stop words

The accuracy over spam files is: 94.61538

The accuracy over ham files is: 98.85057

Accuracy after removing the stop words

The accuracy over spam files is: 0.0

The accuracy over ham files is: 26.436783

Since the weights are still under gradient descent, the accuracy across the files increases. However, there is a steep decline in the accuracy when stop words are removed relying heavily on the accordance of such words.

Enter the number of iterations, regularization factor and the learning rate

20 30 0.02

Accuracy before removing the stop words

The accuracy over spam files is: 95.38461

The accuracy over ham files is: 99.13793

Accuracy after removing the stop words

The accuracy over spam files is: 79.23077

The accuracy over ham files is: 73.27586

As observed, the accuracy increased when the number of iterations was increased to 20. The regularization factor is increased to 30 and the learning rate is reduced to 0.02. With this combination also increases when the stop words were removed from both training and test data. It is also observed that for the same data set and same user input, the accuracy increases again.

Enter the number of iterations, regularization factor and the learning rate

20 30 0.02

Accuracy before removing the stop words

The accuracy over spam files is: 95.38461

The accuracy over ham files is: 99.425285

Accuracy after removing the stop words

The accuracy over spam files is: 76.92308

The accuracy over ham files is: 76.43678

Enter the number of iterations, regularization factor and the learning rate

25 20 0.04

Accuracy before removing the stop words

The accuracy over spam files is: 96.15385

The accuracy over ham files is: 99.425285

Accuracy after removing the stop words

The accuracy over spam files is: 87.69231

The accuracy over ham files is: 84.195404

Enter the number of iterations, regularization factor and the learning rate

30 30 0.01

Accuracy before removing the stop words

The accuracy over spam files is: 94.61538

The accuracy over ham files is: 98.85057

Accuracy after removing the stop words

The accuracy over spam files is: 25.384617

The accuracy over ham files is: 48.563217

Now with the increase of the number of iterations the accuracy slowly starts decreasing. Also with an increase in learning rate, the accuracy after the removal of stop words was observed to be reduced.

Enter the number of iterations, regularization factor and the learning rate

35 50 0.01

Accuracy before removing the stop words

The accuracy over spam files is: 94.61538

The accuracy over ham files is: 98.85057

Accuracy after removing the stop words

The accuracy over spam files is: 0.0

The accuracy over ham files is: 17.816092

Enter the number of iterations, regularization factor and the learning rate

35 40 0.02

Accuracy before removing the stop words

The accuracy over spam files is: 95.38461

The accuracy over ham files is: 98.85057

Accuracy after removing the stop words

The accuracy over spam files is: 72.30769

The accuracy over ham files is: 68.67816

Enter the number of iterations, regularization factor and the learning rate

100 20 0.02

Accuracy before removing the stop words

The accuracy over spam files is: 0.0

The accuracy over ham files is: 0.0

Accuracy after removing the stop words

The accuracy over spam files is: 53.846157

The accuracy over ham files is: 60.91954

The accuracy is 0 in case of stop words not being removed from the word list. It was observed that for larger iterations the probability sum tends to not applicable as the calculation increases and does not fit in the float variables. Similarly the accuracy after the removal of stop words also tends to fall with a higher number of iteration.

It was found that the global minima was found to exist somewhere in the 50^{th} iteration. After which both the scenarios tend to fall sharply.