

ML Assignment-1

Group Details

Tangeda Sai Sharan	2017A7PS0241H
Sanjiv Yelthimar Shenoy	2017A7PS0224H
Manish Kumar Bachhu	2017A7PS0036H

Naïve Bayes – Question-2

We are submitting 2 files for this question. Both work individually. One uses stemming for an external library and the other one doesn't.

Process:

1. We first did some basic pre-processing such as removing punctuations and making everything into lower case.
2. This step is different for the 2 files we submit for this question. We stemmed the words in one of the files using Porter's stemming algorithm. In the other one we didn't.
3. Then we did laplace scaling to even things out in the probabilities and reduce the effect false negatives in our results. False negatives were a menace as if even if 1 word in the test sentence belonged to a negative class, then automatically the entire sentence would be termed negative.(since we are multiplying probabilities in the numerator, even if one is 0, the entire fraction becomes 0).
4. After this we applied the naïve bayes algorithm to get the result.

Accuracy and F-Score Values without Stemming

Average Accuracy: 0.808 +- 0.0289137

Average F Score: 0.81267 +- 0.03214

Accuracy is individual folds of the 5-fold cross validation:

0.815, 0.855, 0.765, 0.800, 0.805

F Score is individual folds of the 5-fold cross validation:
0.828, 0.869, 0.781, 0.794, 0.791

Accuracy and F-Score Values with Stemming

Average Accuracy: 0.825 +- 0.0255

Average F Score: 0.826 +- 0.033

Accuracy is individual folds of the 5-fold cross validation:
0.845, 0.865, 0.80, 0.810, 0.805

F Score is individual folds of the 5-fold cross validation:
0.853, 0.876, 0.811, 0.802, 0.789

Conclusion:

Stemming increased the accuracy by 1.7% and F-Score by 1.4%. As they say every little bit counts.

We got an accuracy above 80%. A random method of selecting 0 or 1 would yield 50% probability. Our model performs significantly better.