UTD

THE UNIVERSITY
OF TEXAS AT DALLAS

An Report
On

Machine Learning :
Naive Bayes vs Logistic Regression
Assignment - 3

By

Soumyadeep Choudhury
Net ID :- sxc180056
UTD ID :- 2021439916

ASSIGNMENT RESULTS

A. Data Input (Training and Testing)

Number of Training Samples for Spam :	**123**
Number of Training Samples for Ham :	**340**
Number of Test Samples for Spam :	**130**
Number of Test Samples for Ham :	**348**
Vocabulary Size with Stop Words :	**6196**
Vocabulary Size without Stop Words :	**6309**

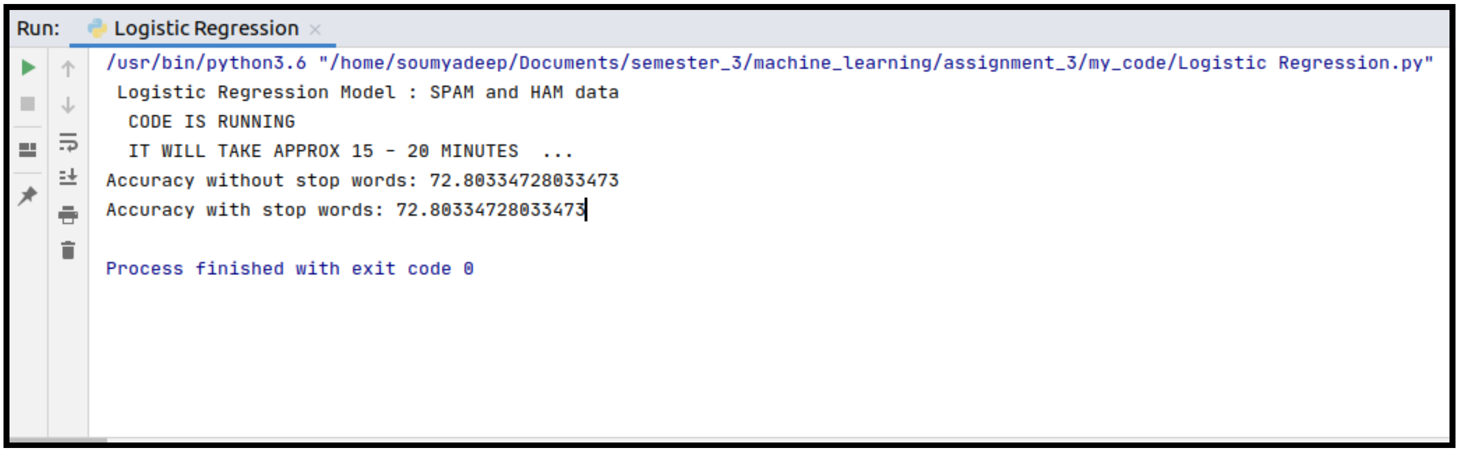
B. Logistic Regression Statistics

- Prediction Accuracy with Stop Words :

Learning Rate : **0.1** ; Regularization(L2) : **0.1** ->	**72.80334**
Learning Rate : **0.1** ; Regularization(L2) : **0.5** ->	**72.91327**
Learning Rate : **0.1** ; Regularization(L2) : **0.75** ->	**71.65622**
Learning Rate : **0.5** ; Regularization(L2) : **0.1** ->	**72.36367**
Learning Rate : **0.5** ; Regularization(L2) : **0.5** ->	**72.10385**
Learning Rate : **0.5** ; Regularization(L2) : **0.75** ->	**71.94555**
Learning Rate : **0.75** ; Regularization(L2) : **0.1** ->	**72.92547**
Learning Rate : **0.75** ; Regularization(L2) : **0.5** ->	**73.07882**
Learning Rate : **0.75** ; Regularization(L2) : **0.75** ->	**73.06533**

- Prediction Accuracy without Stop Words :

Learning Rate : **0.1** ; Regularization(L2) : **0.1** ->	**72.80334**
Learning Rate : **0.1** ; Regularization(L2) : **0.5** ->	**72.91327**
Learning Rate : **0.1** ; Regularization(L2) : **0.75** ->	**71.65622**
Learning Rate : **0.5** ; Regularization(L2) : **0.1** ->	**72.36367**
Learning Rate : **0.5** ; Regularization(L2) : **0.5** ->	**72.10385**
Learning Rate : **0.5** ; Regularization(L2) : **0.75** ->	**71.94555**
Learning Rate : **0.75** ; Regularization(L2) : **0.1** ->	**72.92547**
Learning Rate : **0.75** ; Regularization(L2) : **0.5** ->	**73.07882**
Learning Rate : **0.75** ; Regularization(L2) : **0.75** ->	**73.06533**

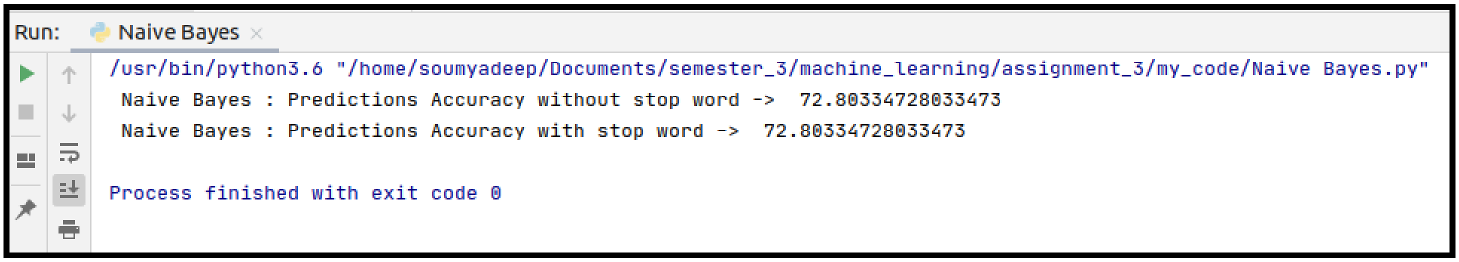


```
/usr/bin/python3.6 "/home/soumyadeep/Documents/semester_3/machine_learning/assignment_3/my_code/Logistic Regression.py"
 Logistic Regression Model : SPAM and HAM data
  CODE IS RUNNING
  IT WILL TAKE APPROX 15 - 20 MINUTES  ...
Accuracy without stop words: 72.80334728033473
Accuracy with stop words: 72.80334728033473

Process finished with exit code 0
```


C. *__Naive Bayes Statistics__*

- Prediction Accuracy with Stop Words : **72.80334**
- Prediction Accuracy without Stop Words : **72.80334**



```
/usr/bin/python3.6 "/home/soumyadeep/Documents/semester_3/machine_learning/assignment_3/my_code/Naive Bayes.py"
 Naive Bayes : Predictions Accuracy without stop word ->  72.80334728033473
 Naive Bayes : Predictions Accuracy with stop word ->  72.80334728033473

Process finished with exit code 0
```

D. <u>Conclusion</u>

-> From the experiment, we can observe that there is very less or very minor change in the accuracy with the variation in L2 regularization and learning rate as tuning factor. It is because the occurrence or frequency of these words are less variational or same in both ham/spam words. As a result the probability of having predicted ham/spam with respect to these tuning s decreases significantly. However, the slight variation is due to the computational approximation during matmul function. Also, with higher learning rate and regularization, the data tends to lose its normal distribution of the data in the probabilistic approach.

-> Secondly, in this experiment we observe that there is no change in accuracy of the model for Naive Bayes and Logistic Regression models. It's due to the less variation in spam/ham words distribution in the learning memory space. However, the Naive Bayes trains and executes very much faster than Logistic Regression algorithm. Naive Bayes takes seconds whereas Logistic Regression takes approx 15-20 mins to train and execute completely.

------------------------------------ END --------------------------------------