MACHINE LEARNING ASSIGNMENT – 3

1. For question 1 and 3, Naïve Bayes Algorithm with and without using stopwords.

Compilation: go the folder containing the files in cmd and execute

python NB.py <Path of spam training files> <Path of ham training files> <Path of test spam files> <Path of ham test files> <stopWords file path> <yes/no to remove stop-words>

**My Results for Naïve Bayes:**

* Q1 - python NB.py train\spam train\ham test\spam test\ham stopWords.txt yes

The Accuracy for Naive Bayes = 94.56066945606695

* Q3 - python NB.py train\spam train\ham test\spam test\ham stopWords.txt no

The Accuracy for Naive Bayes = 94.76987447698745

So, the accuracy decreases after removing the stop words.

1. For question 2 and 3, Logistic Regression Algorithm with and without using stopwords

Compilation: go the folder containing the files in cmd and execute

python LR.py <Path of spam training files> <Path of ham training files> <Path of test spam files> <Path of ham test files> <stopWords file path> <yes/no to remove stop-words>

**My Results for Logistic Regression:**

* η-learning rate=0.1
* λ- regularization factor=0.1
* number of iterations = 100

η, λ and number of iterations are hard coded in the LR.py, so we can change these values in code.

* Q2 - python LR.py train\spam train\ham test\spam test\ham stopWords.txt yes

The Accuracy of Logistic Regression is: 95.60669456066945

* Q3 - python LR.py train\spam train\ham test\spam test\ham stopWords.txt no

The Accuracy of Logistic Regression is: 94.35146443514645

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| λ | η | Number of iterations | Accuracy without stop words | Accuracy with stop words |
| 0.1 | 0.1 | 100 | 95.60669456066946 | 94.35146443514645 |
| 0.01 | 0.1 | 100 | 95.60669456066946 | 94.76987447698745 |
| 0.001 | 0.1 | 100 | 95.39748953974896 | 94.97907949790795 |
| 0.005 | 0.1 | 100 | 95.60669456066946 | 94.76987447698745 |
| 0.005 | 0.01 | 100 | 95.18828451882845 | 96.02510460251046 |
| 0.1 | 0.1 | 300 | 94.56066945606695 | 94.56066945606695 |
| 0.01 | 0.1 | 300 | 95.60669456066946 | 94.76987447698745 |
| 0.001 | 0.1 | 300 | 94.56066945606695 | 94.56066945606695 |
| 0.005 | 0.1 | 300 | 95.81589958158996 | 94.56066945606695 |
| 0.005 | 0.01 | 300 | 95.81589958158996 | 94.76987447698745 |

* From the above table if **η** is very small then changes in the value of **λ** results in negligible/no change in accuracy
* After removing the stop words, the accuracy of LR decreases in most of the cases but increases for high values of λ because λ is the penalty on higher values to avoid overfitting.
* Before removing the stop words, the accuracy of LR increases, but decreases for high values of λ because some stop words can’t be used for classification of a mail as spam or ham but they still interfere with the calculations.