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>

**Results for Naïve Bayes:**

* Q1 - python NaiveBayes.py train\spam train\ham test\spam test\ham stopWords.txt yes
  + - The Accuracy for Naive Bayes = 94.56066945606695
* Q3 - python NaiveBayes.py train\spam train\ham test\spam test\ham stopWords.txt no
  + - The Accuracy for Naive Bayes = 94.76987447698745

Removing stop words reduces accuracy.

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>

**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 LogisticRegression.py train\spam train\ham test\spam test\ham stopWords.txt yes
  + - The Accuracy of Logistic Regression is: 95.60669456066945
* Q3 - python LogisticRegression.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 |

* If the value of **η** is very small then changes in the value of **λ** results in negligible or no change in accuracy
* If we remove the stop words,
  + For most cases, the accuracy of LogisticRegression decreases
  + But for high values of λ, accuracy increases because λ is the penalty on higher values to avoid overfitting.
* The accuracy of LR increases, before removing stop words.

It decreases for high values of λ because some stop words can’t be used for classification of mail as spam or ham but they hamper the calculations.