**Machine Learning 6375.002**

**Assignment 2**

**Problem:**

Implementation of Email classifier by

* Naïve Bayes Algorithm and
* Logistic Regression

**Naïve Bayes Implementation:**

Naïve Bayes algorithm based on the conditionally independent property is implemented. The implementation is done by removing the stop words and by having them. The removal of stop words has a good impact on the implementation and there-by increasing the accuracy of the system.

The Naïve Bayes equation:

P(Y=ham/X1= x1, X2=x2,…. Xn=xn) = P(X=x1/Y= ham) \* P(X=x2/Y= ham) \*……. P(X=xn/Y= ham) \*P(Y= ham)

Similarly probability of spam is measured. The class that has maximum value is assigned as the class for the test sample.

Various stages of development:

|  |  |  |
| --- | --- | --- |
| Features considered | Accuracy with stop words | Accuracy without stop words |
| All the words are considered as features without any preprocessing | 15.28% | 73.76% |
| Features are compared with case sensitivity but unnecessary punctuations are removed | 53.41% | 83.54% |
| Features are made case insensitive | 95.60% | 95.60% |

**Logistic Regression Implementation:**

Logistic regression is as follows

h(x) = 1/(1+exp(-Summation(wixi))), i ranges from 1 to n

P(Y=Ham/X) = 1/1+exp-(w0+ Summation(wixi)), i ranges from 1 to n

P(Y=Spam/X) = 1 – P(Y=Ham/X)

wi - weights assigned to the feature xi

xi – feature (word in this problem specification)

If h(x)>=0.5, then the sample belongs to class ham

Else

The sample belongs to class spam

The weights (wi) of the features are calculated as

w0 = w0 – (a/m) Summation(h(x)-y)x­0

wi = wi – (a/m) (Summation((h(x)-y)x0) + l \*wi)

a - Learning rate

l – regularization factor

Experimental Results:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Learning Rate | Regularization Factor | Iterations | Accuracy with stopwords | Accuracy without stopwords |
| 0.001 | 5 | 5 | 72.80% | 73.43% |
| 0.001 | 1 | 5 | 72.80% | 73.43% |
| 0.0025 | 0.1 | 20 | 73.01% | 73.96% |

Conclusion:

The assignment helped to understand the important classification algorithms and posted a challenge to implement it efficiently.