Report

Machine Learning CS 6375.001

Assignment 2

Name: Ameya S. Gamre

Net Id: asg160330

Naïve Bayes for Spam/Ham classification

1. **Implementation**
   - Implemented the multinomial Naïve Bayes algorithm for text classification into spam and ham. First the training file was input from the command line, all words were read from every file and a vocabulary was made containing one instance of each word. The actual labels of the training set were recorded and the spam count and ham count was calculated for each word occurring in the spam and ham documents combined.
   - Prior for ham and spam is calculated by finding the ratio of training documents of a category to the total number of documents. (Nc/N)
   - For every word in the vocabulary, Laplace smoothing is done to make sure that each word is pretended to be seen atleast once.
   - To predict the accuracy of the classifier, we run the classifier against a bunch of test files in the test directory.
   - For each file, if the word is present in the vocabulary, we calculate the sum of log likelihood of each word in the document for spam and ham using the probabilities of the vocabulary word and the priors of ham or spam. The output label depends on the maximum log likelihood.
   - If predicted label is same as actual label, the accuracy is incremented.
   - The above steps were repeated once again after stop words were removed to calculate the new accuracy.

   Input: Training set, Test Set

   Output: Accuracy of Naïve Bayes Classifier for the test set before and after excluding the stop words.

2. **Output**

   Without stop words being removed from the corpus of documents containing emails

   Accuracy without removal of stop words: 79.0794979079

   After removing stop words from the corpus of the documents containing emails

   Accuracy after removal of stop words: 85.9832635983

   Accuracy increases as we remove the stop words as these are words that do not contribute to classifying the document as spam or ham. This way, we can focus more on the words which are significant in the classification.

Logistic Regression Classifier

1) Implementation

- Implemented the logistic regression algorithm for text classification into spam and ham. First the training file was input from the command line, all words were read from every file and a vocabulary was made containing one instance of each word. The actual labels of the training set were recorded and the spam count and ham count was calculated for each word occurring in the spam and ham documents combined.
- To train the logistic regression classifier, the vocabulary words are used to form the feature vector for each instance. Length of the feature vector is same as number of words in the vocabulary.
- Initialized weight vector containing 1 more than the number of vocabulary words (extra for the bias) to zero.
- Running 150 iterations, the logistic regression classifier was trained using batch gradient ascent rule.
- For each iteration, the error in each instance was calculated by the formula, Error = y – y' where y is the actual label and y' is the predicted label.
- Then for that iteration, the weight of each word in the vocabulary is updated by multiplying the xi for each instance and updating the w for each word in the vocabulary.
- To predict the accuracy of the classifier, a set of test files are used. A feature vector representing the frequency of each word in the vocabulary found in the document is made using each test file and then P(Y/X) is calculated using the weight vector formed from training.

Input: Training file, test file

Output: Accuracy of logistic regression classifier before and after excluding the stop words.

2) **Output**

When
η = 0.001
λ = 2
Accuracy without removal of stop words: 92.8870292887
Accuracy after removal of stop words: 93.5146443515

When
η = 0.005
λ = 1
Accuracy without removal of stop words: 94.5606694561
Accuracy after removal of stop words: 95.1882845188
When

η = 0.005
λ = 2
Accuracy without removal of stop words: 94.769874477
Accuracy after removal of stop words: 94.5606694561

When
η = 0.005
λ = 0.5
Accuracy without removal of stop words: 93.9330543933
Accuracy after removal of stop words: 95.1882845188

The accuracy doesn't seem to have changed a lot when stop words were removed. Though the stop words are low information features, there should have been atleast slight changes. However, if the weights associated with these stop words didn't contribute too much to finding the actual label, the accuracy wouldn't change much.

Conclusion

The Naïve Bayes and Logistic regression classifier was successfully implemented to classify the test data into ham and spam classes.