Project 2: Bayesian learning for classifying netnews text articles

Rathna Sindura Chikkam

Procedure:

For this project to perform Naïve Bayes classification on text articles, I have followed Bag of Words approach to compartmentalize the data. Below are the steps:

1. First, I have read the file names of all the files present in 20_newsgroup folder and have added them into one big master list of filenames and their categories.
2. From the master list of files and their respective categories, I am using train_test_split cross validation package from Scikit-learn to split the entire data into train and test document sets along with their train and test categories. I have used the split size to be 0.5 as mentioned in the project description. I have also written few helper function likes flatten, doc_tokenize, hasNumbers that helps in list of words formation, tokenzing each document and performing regex validation to find numbers in words.
3. In Step 3, I am performing preprocessing on the bag of words formed from train set of documents that we had split earlier. I have used nltk – natural language toolkit to remove stop words that have no significance in category classification, have removed words of length 1 and 2, have performed alphanumeric words removal and have made sure that no words has any digit in it.
    a. I am also performing a nltk frequency count of words and their number of appearances in order to remove least common words and pull the first 10,000 words as features – I have performed a feature selection of choosing first 10,000 most frequently occurred words from the entire dataset.
4. From the 4th step, I have pulled a dictionary set from every document present in train and test split of data.
5. In the 5th step, I have written helper functions to fit the train data set, find class probability while performing classification on test data set, and have used accuracy_score and classification_report packages from Scikit-learn to know the accuracy of our Naïve bayes claasifier.

Accuracy obtained is: 77% approximately as shown below:

```
In [1]: runfile('E:/Academics/Sem3/ML/Projects/Project2/Project2_rxc3518/NaiveBayes_NewsGroup.py', wdir='E:/Academics/Sem3/
ML/Projects/Project2/Project2_rxc3518')
[nltk_data] Downloading package stopwords to
[nltk_data]     C:\Users\sindu\AppData\Roaming\nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
0.7708770877087708
                        precision    recall  f1-score   support

                    -1       0.00      0.00      0.00         0
           alt.atheism       0.81      0.74      0.77       484
         comp.graphics       0.65      0.86      0.74       502
 comp.os.ms-windows.misc       0.94      0.49      0.65       504
comp.sys.ibm.pc.hardware       0.77      0.76      0.76       473
   comp.sys.mac.hardware       0.95      0.69      0.80       479
          comp.windows.x       0.79      0.87      0.83       509
            misc.forsale       0.94      0.62      0.75       501
               rec.autos       0.97      0.64      0.77       528
         rec.motorcycles       0.99      0.74      0.85       501
      rec.sport.baseball       0.99      0.79      0.88       501
        rec.sport.hockey       0.94      0.96      0.95       501
               sci.crypt       0.71      0.97      0.82       488
         sci.electronics       0.94      0.54      0.69       519
                 sci.med       0.98      0.84      0.90       504
               sci.space       0.84      0.89      0.87       482
  soc.religion.christian       0.99      0.97      0.98       516
      talk.politics.guns       0.88      0.63      0.73       504
   talk.politics.mideast       0.91      0.95      0.93       520
      talk.politics.misc       0.28      0.92      0.43       481
      talk.religion.misc       0.71      0.56      0.63       502

            avg / total       0.85      0.77      0.79      9999

C:\ProgramData\Anaconda3\lib\site-packages\sklearn\metrics\classification.py:1137: UndefinedMetricWarning: Recall and F-score
```

References:

https://machinelearningmastery.com/naive-bayes-classifier-scratch-python/

https://towardsdatascience.com/unfolding-na%C3%AFve-bayes-from-scratch-2e86dcae4b01

https://github.com/gokriznastic/20-newsgroups-text_classification/blob/master/Multinomial%20Naive%20Bayes-%20BOW%20with%20TF.ipynb