Program 1

PR1 Medical Text Classification

By **Chaithra Lakshmi Sathyanarayana**

F1-score: 0.7507

Approach

Extract data -> Clean data (Preprocessing) -> Train the model -> Predict for new data based on the training data

Medical abstracts were given with the abstract and classes. Based on the abstract the diseases were classified into 5 classes. Two files were given: train.dat and test.dat

The file train.dat has both class and abstract. The file test.dat has only abstract. Main steps are given below.

1. Extract data from training and test datasets
2. Create a bag of words matrix using training and test data
3. Clean the data to remove punctuations, numbers and other data that is deemed unnecessary
4. Normalize the matrix based on the inverse frequency of the words
5. Form a matrix which has the cosine distance between each test row and training row
6. From this matrix, for each test row, get the closest k training records, and based on the which label is repeated highest, assign that label to the test row.

Methodology

1. The first step is extraction of training and test data. The training data was given in train.dat and test data was given in test.dat.

2. I separated the class and abstract in training data. Then, I created a new dataset, which had all abstracts from training and test files.

3. After combining the training and test files, the data was cleaned. All words with length less than 3, punctuations, brackets and numbers were removed. The data was stemmed. I tried with and without removing the stop words using nltk library.

4. After this the a new matrix was formed based on “bag of words” model. Here all unique words in the dataset is collected and in each document the frequency of the particular word is filled.

5. Since, some words might be repeated across the dataset and might not be unique, to ensure these words don’t get more weightage while calculating distance between two documents, normalization is performed. This will be a csr matrix. First form a csr matrix for inverse document frequency and on this matrix L2-normalization is performed.

6. For each document from test data, the cosine distance for all documents in training data is calculated and is stored as a matrix.

7. For prediction knn algorithm and cosine distance was used. In this matrix, for each test document, find the top k similar training documents. In this, based on the number of times the classes are repeated, choose the class that is repeated most, and assign this class for the test row.

8. I tried the entire process for different data preprocessing techniques and k-values.I used 10-fold validation to determine optimal k value. With this got best result for k=10.

9. To test this, I removed a few random samples from train.dat used these values as testing data. When the entire process was repeated, for different preprocessing techniques and k-values, I found that k=10 worked better for me. Whichever gave me the best prediction, I submitted the code for that.

Conclusion:

1. As the values of k increases, the accuracy of prediction improves. At some value of k, based on the methodology used, prediction accuracy reaches a maxima and after that as k increases accuracy decreases.
2. The quality of data cleansing has a major impact on accuracy of prediction. Care should be taken to remove unnecessary text in the file, and retain the necessary ones.