MACHINE LEARNING

Assignment 5: Naive Bayes based Newsgroup Classification

Due date: Monday, 31 October 2016, 11:55 PM
Requested files: train.py, test1.py, test2.py (Download)
Type of work: Individual work
Problem Statement: You will be given a train file to train you Naive Bayes classifier. The file consists on n articles one per line.

Each line starts with the class of label followed by the list of words appearing in the article. Your task is to implement a Naive Bayes classification algorithm to classify each article into one of the classes.

There are 2 parts to the assignment:

1) In this part, you are required to implement a standard naive bayes algorithm and you will be tested upon if you have implemented the algorithm correctly on public and private testset(different from the given training set). We have implemented a Naive bayes Algorithm. You must implement the standard naive bayes as we are checking you scripts against ours and allowing only 2% accuracy difference. In other words, if our implementation gives an accuracy of k%, you algorithm must give accuracy between (k-2)% and (k+2)%.

2) In this part, you have to optimise the standard naive bayes algorithm with your own features or using different functions to get a better classifcation. You will be given marks acording to the relative score you are geting.

NOTE: Use Laplace smoothing in both cases, as there may be a chance that a word is not present in the training dataset but is present in the test dataset.

INPUT and OUTPUT Format:

Training:

The training file contains m articles each described in 1 line. The line starts with the class of the article followed by a list of words appearing in the article seperate by space(' ').

You will be required to output your model file.

train will take two input arguments trainfile and modelfile. trainfile is the input training file and the modelfile is the output model file which you will produce.

Testing

The testfile contains k articles each described in one line. Each line will contain a list of words appearing in the article and you will be required to output the class of the article in the output label file.

test will take 3 arguments testfile, modelfile and labelfile. testfile is the test file on which the testing will take place and the modelfile is the one that is generated by you in the above part and you will be outputting the labels in the label file.

Public test data: https://drive.google.com/open?id=0B2cnziR69ttRRGNzRkRyanlUZkk

General Instructions:

You have to code this assignment in python. **Please don't use any library functions**.

The marking is 50% for both the parts.

You are strictly advised to follow the input and output instructions. Not following them will result in high penalty.

Contact Details:

Please post on piazza for any queries. If you get no reply from any of the TAs within a day, then mail your doubts with subject "ML Assignment-5" to the following email-id

1) Tapas Jain: tapasjain01@gmail.com, cs5120487@iitd.ac.in [Regarding any theoretical or understanding of assignment]

2) Rohit Jindal: mcs162668@iitd.ac.in [Regarding any moodle glitches]

Requested files

train.py

```
1 #File for training
```

test1.py

```
1 #Code for Part1
2 #   Standard Naive bayes code
```

test2.py

```
1 #Code for Part2
2 #   Improved Naive bayes code
```

ADMINISTRATION