Classification of Spam Emails

Course project

DS 503 Machine Learning for Data Science

Deadline: 10 May 2019

Description of Dataset:

Email dataset, ("emaildataset.zip") is uploaded which contains numerous emails that are divided into 10 subdirectories (part1, ..., part10). These correspond to the 10 partitions of the corpus that were used in the 10-fold experiments. In each repetition, one part was reserved for testing and the other 9 were used for training.

Each one of the 10 subdirectories contains both spam and legitimate messages, one message in each file. Files whose names have the form spmsg*.txt are spam messages. All other files are legitimate messages.

Implementation:

In the class, we had discussed in detail about the practical aspects of developing machine learning system in general and spam detector in particular. This was also demonstrated by running python code as an example. You need to follow the same steps. Code and the two lectures regarding this are uploaded (Lecture 10, Lecture 11, ClassCode.zip). Lectures are taken from Coursera website. Please note that you have to test the performance with and without stemming. Also, with and without stop words. For stemming a tutorial can be found at the following link here. For stop words check out this tutorial.

You need to implement the technique mentioned in the attached Lecture 11 (slide 6 onwards). You are required to follow each step and report your findings. These findings need to be submitted report along with your code. Error analysis after each change is must and must be mentioned in the report. Effect on error with and without stemming and stop words must be part of your analysis. You are also required to report Precision, Recall and F1-Score on the provided dataset. I shall accept only python 3 compatible code. You are required to only use the tools studied in the class. Tools include Linear/Non-linear/Logistic regression, Backpropagation NN (no deep nets), Decision Trees, SVM.

Submission:

You need to submit a compressed archive of your working folder. Folder structure should be as follows. Please make sure case and naming convention is followed.

```
PROJECT-18I-0000
----Readme.txt
----code.py
----report.pdf
----support file 1
```

----support file 2
----…

Readme.txt file must have special instructions (if any) along with the name of packages required to run your code. During testing, I shall navigate to the folder (e.g. PROJECT-18I-0000) and execute your code using following command.

```
> python3 code.py email.txt
0
> python3 code.py email.txt
1
```

Where "email.txt" is in the same format as the provided dataset. Your code has to classify it as spam or not. **0 will be outputted for non-spam and 1 for spam**. NOTHING ELSE MUST BE PRINTED ON THE SCREEN. [0], ["0"], "[1]", spam, and so on are ALL INVALID outputs.
After completing your work, compress the working folder using the name convention **PROJECT-18I-000.zip** and upload it. Please make sure that all your files are in one folder and then compress the folder.

Evaluation and Testing of Code:

There will be two evaluations, one of report and the second of your code. Report will be evaluated by me after looking at the contents and the details of your reporting.

Your code will be tested using the methodology adopted during the assignment no 2. I shall test your code on the unseen dataset. Marks will be given on the basis of the accuracy of your code. For this automatic python script will be used. Please make sure, no warning, error or anything else is printed on the screen.

Report Template

Your report must be written in the following format. Each section must report your attempted model details, error analysis, improvements made and accuracy. Hopefully the accuracy of the attempts will improve with time and last one will be the best among the rest.

1. Attempt No 1
   - I. *Model details:*
   - II. *Error Analysis:*
   - III. *Improvement suggested:*
   - IV. *Precision, Recall, F1-Score*
2. Attempt No 2
   - I. *Model details:*
   - II. *Error Analysis:*
   - III. *Improvement suggested:*
   - IV. *Precision, Recall, F1-Score*