CS 403/725
Assignment 2
Spring 2016
IIT Bombay

Spam Detection

Problem Statement

Just open gmail and go to spam section, you will get thousand of messages there. Did you ever wonder how does google separate the normal mails (called ham) from spams? If yes, then don't you think it would be good to develop your spam detector. Learning a new language might not be fun unless we pivot it to some application.

The aim of this assignment is to implement a machine learning algorithm(Neural Network with logistic activation function) .

In this assignment, we are going to detect **if an email is a spam email or a ham**. We have discussed this problem in lecture 1 itself, PFA slides at the end of this document.

Dataset

Various features have been extracted from emails to create a dataset.

Link to dataset:
Training Data -
https://drive.google.com/file/d/0B0fWfrET8KqAT1hyYy1xcS1Yb1U/view?usp=sharing

Testing Data (Without labels)
https://drive.google.com/file/d/0B0fWfrET8KqAY2dia2l0TThHcjg/view?usp=sharing

For qualitative description of the dataset, please refer to:
https://drive.google.com/file/d/0B8xqoZByW_iAYzBMZ1VOcUgzdFk/view?usp=sharing

In the dataset the email is mapped to a certain number of features. You will have to do feature engineering over the data and then train the data on it. The constraint is to implement **Neural**

network with Logistic Function only. You will have to code this on your own and no libraries will be allowed. Please find screenshots of lecture 1 at the end which will give you a way to approach the problem.

The data set has been divided into two sets namely train and test set.

Evaluation:

You will be evaluated based on the accuracy of the labels you get on **testing** data set. We will open a kaggle competition soon, so that you check the performance of your system. The submissions on kaggle is not compulsory but strongly advised. The final submission will be on moodle.

Deliverables:

- You have to submit a python code which takes training input from "Train.csv" and testing data (without labels) from "TestX.csv" file and output should be in "TestY.csv" in the same folder.

Programing language:

Python will be allowed to maintain uniformity(**python2** or **python3**). You are **not allowed** to use the already available libraries for the implementation of neural network (standard input output and math library are allowed).

Important Note:

- Your code **should use** the dataset provided by us for the training. If that is not done, you will be awarded **zero marks** for the assignment
- Copying of assignments will be dealt with very seriously
- Post your queries on tutorspace

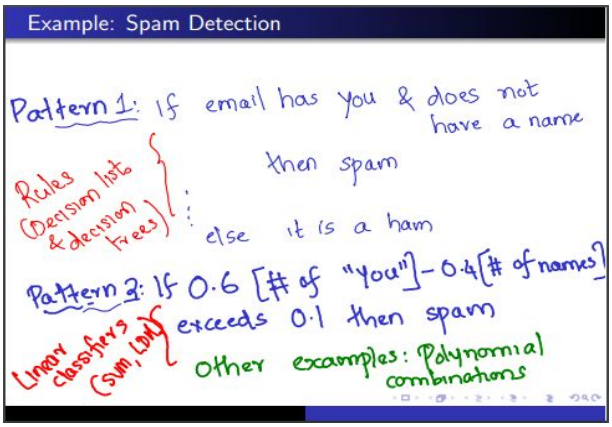
Submission format:

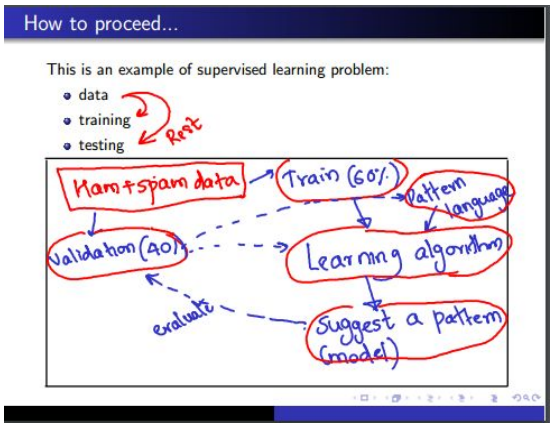
- For final submission, submit your python code in a file **rollno_2.py/rollno_3.py** (for **python2 (2.6 onwards) / python3** respectively) and small readme file README giving brief description of your work. Compress these in tar.gz format with name rollno.tar.gz
- Final evaluation will be done in an environment only with standard python libraries

Timeline:

Final Submission: 10th April

Appendix





References:

(1) https://archive.ics.uci.edu/ml/datasets/Spambase

(2) http://23.253.82.180/course/307/857/2212 (Lecture)