COMP 527 - 2017 - CA Assignment 1
Data Classification
Implementing Perceptron algorithm

Assessment Information

Assignment Number	1 (of 2)
Weighting	12%
Assignment Circulated	1st February 2016
Deadline	10th March 2017, 15:00 UK Time (UTC)
Submission Mode	Electronic via Departmental submission system
Learning outcome assessed	(1) A critical awareness of current problems and research issues in data mining. (3) The ability to consistently apply knowledge concerning current data mining research issues in an original manner and produce work which is at the forefront of current developments in the sub-discipline of data mining.
Purpose of assessment Marking criteria	This assignment assesses the understanding of the Perceptron algorithm by implementing a binary Perceptron for text clustering. Marks for each question are indicated under the corresponding question.
Submission necessary in order to satisfy Module requirements?	No
Late Submission Penalty	Standard UoL Policy.

1 Objectives

This assignment requires you to implement the Perceptron algorithm using the Python programming language.

> **Note that *no credit will be given for implementing any other types of classification algorithms or using an existing library for classification instead of implementing it by yourself. However, you are allowed to use numpy and scipy libraries for accessing data structures such as numpy.array or scipy.sparse. But it is not a requirement of the assignment to use numpy or scipy. You must provide a README file describing how to run your code to re-produce your results.***

2 Text Classification using Binary Perceptron Algorithm

Download the *CA1data.zip* file from the COMP 527 web site and decompress it. Inside, you will find four files: *train.positive*, *train.negative*, *test.positive*, and *test.negative*. These files correspond to the positive and negative train/test reviews we will be using in this assignment. Each line in each file represents a review using a set of features. We will be using both unigram and bigram (concatenated using two underscores) features to represent a review. A review is represented using a bag-of-features. Moreover, each feature is counted only once, giving a boolean valued feature representation (i.e. a set of features for each review).

Questions/Tasks

(1) Explain the Perceptron algorithm for the binary classification case, providing its pseudo code. **(20 marks)**

(2) Write a program to load the train/test instances (positive/negative) from the train/test files. **(20 marks)**

(3) Implement a binary Perceptron classifier and measure the classification accuracy on the test instances. Classification accuracy is defined as the percentage of the total number of correctly classified instances to the total number of test instances. **(40 marks)**

(4) Plot the train error rate and test error rate against the number of iterations. According to your plot, what would be the ideal number of iterations to terminate the training? **(20 marks)**

3 Deadline and Submission Instructions

- Deadline for submitting this assignment is **10th March 2017, 15:00 UK time (UTC)**.
- Submit
  (a) the source code for all your programs,
  (b) a README file (plain text) describing how to compile/run your code to produce the various results required by the assignment, and
  (c) a PDF file providing the answers to part (1) and (4).

Compress all of the above files into a single tar ball (tgz) file and specify the filename as *studentid.tgz* (replace "studentid" by your departmental student id). It is extremely important that you provide all the files described above and not just the source code! (If you are unable to create a tgz file then create a zip file)

- Submission is via the departmental submission system accessible (from within the department) from
  `http://intranet.csc.liv.ac.uk/cgi-bin/submit.pl?module=COMP527`