Big Data (MHI222956) Coursework: Practical Data Analysis

1. Overview

As a part of the assignment for the Big Data module, this coursework comes a weight of 40% of the final mark. The objective of this coursework is to help you fully understand some of the algorithms covered in the module (e.g., Support Vector Machine, Neural Networks) by doing some practical data analysis work. A report (1000 – 1500 words) should be submitted for this coursework. The report is a summary of your practical data analysis work, which should address the following and anything else you feel relevant.

- Description of the problem
- Construction and tuning the classifier
- Testing results
- Discussion

2. Submission

The report should be submitted as a PDF/Word file, which should be clearly labelled "Big Data Coursework 2017-18" and contain the submitting student's name. e.g., Harry Potter's report should be named as: "*Big Data Coursework 2017-18 H Potter*"

The cover page of the report must contain the submitting student's name, student ID, programme name, and the following declaration of ownership:

"I declare that all work submitted for this coursework is the work of <insert name of the author> alone unless stated otherwise."

As a part of the plagiarism check, the marker may randomly ask students to demonstrate the classification system developed by him/her.

The report should be submitted through the Big Data Coursework Submission link under Assignment. And the deadline for this submission is

Friday, 15 Dec 2017

3. Datasets

The data that you will need to complete this coursework are available in the Coursework Dataset folder under Assignments. There are two zip files: one for the Bank dataset, the other for the Mushroom dataset. Each zip file contains a .txt file explaining the attributes in the dataset, and a data file which is in .csv format. The evaluation data (to be used in demonstration) kept by the marker are in the same format as the data analysed by you.

4. Tasks and Distribution of marks

- Description of the problem (20 marks)
  Students need to analyse one of the two provided datasets. A short description of the problem should be given. The description should include description of the attributes and class in the dataset, number of instances, distribution of attributes, distribution of class, missing value in attributes, etc.

- Construction and tuning the classifier (40 marks)
  Students are required to use either SVM (Support Vector Machine) or NN (Neural Networks) to create a classifier for the analysis of the selected dataset. Details about the data pre-processing, model structure, how the model is tuned should be explained. The definition of the training data and testing data should be given. The model needs to be implemented using Python. Screenshots and core scripts of the developed system should be included in the report.

- Testing results (25 marks)
  Using the trained model/classifier, testing data specified by the student should be classified. Results of the classification (including accuracy) need to be illustrated with execution screenshots and explained in the report.

- Discussion (15 marks)
  The students are required to give discussions on significant of data pre-processing, effectiveness of model tuning, and general performance of the selected algorithm for the specific dataset, etc.