Data Mining
COC 131

Coursework assignment

Credit value: 100% of the module

This coursework is divided into two parts. The first part should be submitted as a consultancy report and the second part as a report on the tasks that you are to carry out giving results and their interpretation on a data set that you have been provided to analyse using the Data Mining Software WEKA.

Part A. Consultancy Report (40 marks)
You have recently been appointed as a Consultant for a very large company working in one of the following sectors (please select one sector of your interest):

- Music
- Health care

The company is new to the use of Data Mining and wishes to apply such technology to improve their operation. However, they need to make an informed decision by finding out more about what it is and what has been done first.

Your task as a consultant is to provide them with a report (no more than 6 pages) about data mining to help them make informed decisions. Your report should cover:

1. An introduction of what data mining is, covering topics such as purpose, process and techniques.

   [10 marks]

2. A review of data mining applications that have been developed for the sector. The report should provide a broad coverage of the different types of applications. Where information is available give indications of whether these applications are research prototypes or are in practical use.

   [18 marks]

3. Discuss how data mining could be introduced into the company by considering the data mining process and the risks associated with it.

   [12 marks]

Marking Criteria for Consultancy Report

The assignment will be assessed on the following basis:

1. Presentation
    - Is the coursework well structured and presented?
    - Is the standard of English satisfactory?

2. Technical content
    - Is the topic well researched with appropriate references?
    - Is the topic well covered, addressing the key issues?
    - Is the technical content presented at the right level?
    - Is the discussion focused and clear?
    - Are the claims and arguments based on evidence?

Part B. WEKA Exercises (60 marks)

WEKA is a collection of machine learning algorithms for data mining tasks available from (http://www.cs.waikato.ac.nz/ml/WEKA). WEKA contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well-suited for developing new machine learning schemes and you can program your own work. In this assignment you are not expected to write any software code but use existing tools provided by WEKA. As a first step, download and install WEKA on your computer. After that perform the following tasks on the data set provided and produce results and their interpretation in the form of a report.

(i) Use linear regression and two other methods to rank each feature (attribute) on its ability to predict the class variable. Tabulate the results and describe how you obtained them. Based on the results, discuss the relative importance of the features and how you would rank them.

[15 marks]

(ii) Apply two chosen classifiers to the original data set using 10-fold cross-validation and also apply the same two classifiers to a modified data set containing features of your choice based on the results of (i). Perform *t*-tests to determine whether the differences in performance of the classifiers and data sets are statistically significant from one another. Discuss your findings.

[15 marks]

(iii) Use the equal-width binning strategy to produce a modified data set. Apply the better performing classifier used in (ii) to the modified data set. Discuss the differences in performance between the results in (ii) and (iii) for the chosen classifier and the merits of using the binning strategy.

[20 marks]

(iv) Apply k-means clustering to the original data set. Set the number of desired clusters to the number of classes already known. Assign each instance a class in the clustering output. Discuss the difference between such classification and the available ground-truth.

[10 marks]

Guidelines for WEKA Exercises

Your report on WEKA Exercises should be no more than 8 pages.
The report should contain a section for each of the task. Each section should describe:
(a) Methodology and Data/Tools used;
(b) Experimental Results;
(c) Interpretation of these results.

Marking Criteria for WEKA Exercise

The assignment will be assessed on the following basis:

1. Presentation
   - Is the coursework well structured?
   - Are the results presented clearly?

2. Technical Content
   - Have you used an appropriate process for generating results?
   - Are the purposes of the tasks and the results understood and well discussed?

Dr Parisa Derakhshan