Project

There are two projects contributing 50% to the total assessments of this unit. The projects are to be submitted to cssubmit during the semester.

- **Project 1 submission** is an individual effort on data preprocessing, warehouse design and implementation. The deadline of submission is on Friday 23:59 pm 2nd April (cssubmit). It is worth 25% of the total assessments.
- **Project 2 submission** can be an individual or a paired effort (i.e. to complete the final submission in a group of 1 or 2 people, ideally both from CITS3401), due on Friday 11:59 pm 21st May (cssubmit). It is worth 25% of the total assessments.

Marking scheme and more details of the projects have been available in the following. The overall objectives of the projects are to build a data warehouse from real-world datasets, and to carry out basic data mining activities including association rule mining, classification and clustering.

Project 2 Pattern Discovery and Building Predictive Models

Project 2 aims to produce clean, reduced or transformed data for pattern discovery and predicative analysis. In this project, we will assess the data cleaning and predictive model building skills. You can use either **Weka** or other data analytic toolsets (e.g., R or Python) familiar to you.

Datasets and Problem Domain

For this project, we would like to use the mobile price classification dataset as the source of data. The target of this project is to predict whether the price of a mobile phone is high or not. A copy of the necessary files is here

Your tasks

1. Data cleaning and analysis
   - a. Read through the table and the table column descriptions. Understand the meaning of each column in the table.
   - b. Distinguish the type of each attribute (e.g., nominal/categorical, numerical). You may need to discretise some attributes, when completing Task 2, 3 or 4.
   - c. Determine whether an attribute is relevant to your target variable. You may remove some attributes if they are not helpful for Task 2, 3, or 4. You might create separate data files for Task 2, 3 and 4.
   - d. Identify inconsistent data and take actions using the knowledge you have learnt in this unit.
   - **Note**: You may use different data processing procedures, when working on different tasks to get better results.
2. Association rule mining
   - Select a subset of the attributes (or all the attributes) to mine interesting patterns. To rank the degree of interesting of the rules extracted, use support, confidence and lift.
   - Explain the top k rules (according to lift or confidence) that have the "price_category" on the right-hand-side, where k >= 1.
   - Explain the meaning of the k rules in plain English.
   - Given the rules, what recommendation will you give to a company willing to design a high price mobile phone (e.g., should the mobile phone equipped with bluetooth)?
3. Classification
   - Use the "price_category" as the target variable and train two classifiers based on different machine learning algorithms (e.g. classifier 1 based on a decision tree; classifier 2 based on SVMs).
   - Evaluate the classifiers based on some evaluation metrics (e.g., accuracy). You may use 10-fold cross-validation for the evaluation.
4. Clustering
   - Run a clustering algorithm of your choice and explain how the results can be interpreted with respect to the target variable.
5. Data reduction
   - Perform numerosity reduction and perform attribute reduction.
   - Train the two classifiers in Task 3 on the reduced data.
   - Answer the question: "Does data reduction improve the quality of the classifiers"?
6. Attribute selection
   - Select the top-10 most important attributes manually based on your understanding of the problem; select the top-10 most important attributes based on Information Gain.
   - Which attribute selection method is better and why?

File to submit

A zip file that contains:

1. A report in PDF containing the six tasks listed above. If you work in team, only one submission is needed and the contribution of each team member should be clearly mentioned. Clearly indicate the name and student number of yourself and the team member.
2. All the codes (e.g., python, SQL script) and/or screenshots (for Excel, or other data processing software) of data cleaning and process procedures.
3. Intermediate and final result files for all data processing procedures.

The file needs to be submitted to cssubmit.
Plagiarism is strictly prohibited. Don't submit codes downloaded from the Internet.

Marking scheme (Pattern Discovery and Predictive Analytics)

[30 marks]
[5 marks] Explain the data processing operations (e.g., remove some attributes and action on inconsistent data) that you have done.
[5 marks] Explain and interpret the top k association rules mined; based on the association rules, provide a recommendation for a company willing to design a high price mobile phone.
[5 marks] Explain how you train the classifiers and your evaluation results.
[5 marks] Clustering and interpretation of the clustering result (with respect to the target variable)
[5 marks] Explain the data reduction you have performed; compare the classifiers trained on reduced data with the classifiers trained on the original data.
[5 marks] Your answer to Task 6.

Project 1 Building a Data Warehouse