CSCM35, CSLM35 Big Data and Data Mining

by Dr. Jingjing Deng

Released on 4th Mar 2020

Submission

In this coursework, you will be given a practical data mining task to complete which consists of software programming and technical report writing. Source code must be organised and formatted neatly. Sufficient and clear comments are very welcome and necessary for markers to assess your work. The technical report must be submitted in **PDF** format. Submissions and feedback will be done via Blackboard-Tunitin system. Plagiarism will not be tolerated. Zip all your files with the following naming convention for submission:

- [Student Number]-[Last Name][First Initial]-[Coursework][Number].zip
- For example: *123456-DengJ-Coursework1.zip*

Policy

- To be completed by students working individually.
- Feedback: individual feedback is given on Blackboard within two weeks of deadline.
- Learning outcome: The tasks in this coursework are based on both your practical work in the lab sessions and your understanding of the theories and methods of data mining. Thus, through this coursework, you are expected to demonstrate both practical skills and theoretical knowledge that you have learned in this module. You will also formally present your understandings through technical writing. It is an opportunity to apply analytical and critical thinking, as well as practical implementation.
- Unfair practice: This work is to be attempted individually. You may get help from your lecturer, academic tutor and lab tutor, but you may not collaborate with your peers. Copy and paste from the Internet is not allowed. Using external code without proper referencing is also considered as breaching academic integrity.
- Submission deadline: The report and your Python 3 implementation need to be submitted electronically to Blackboard by the deadline.

CSCM35/CSLM35 Coursework 1 Complete by 1st/04/2020

Task: Data Mining Practice – Associated Rule Mining

Association rule learning is a rule-based machine learning method for discovering interesting relations between variables in large databases. It is intended to identify strong rules discovered in databases using some measures of interestingness.[1] Your tasks are to:

1. Understand the problem of mining common pattern and data by performing basic visualisation and statistical analysis;
2. Apply associated rule mining algorithm to the data and analyse the outcomes;
3. Write a report to describe your method and discuss the key findings that you discover from the experimental results.

Data: Online Retail Data Set

This is a transnational data set which contains all the transactions occurring between 01/12/2010 and 09/12/2011 for a UK-based and registered non-store online retail. The company mainly sells unique all-occasion gifts. Many customers of the company are wholesalers. The following Python packages and materials might be helpful to complete your task:

- Dataset File IO: Pandas, `https://pandas.pydata.org/`.
- Data Visualisation: Matplotlib, `https://matplotlib.org/`;
- Top 50 matplotlib Visualizations – The Master Plots, `https://bit.ly/2BBlWLE`;
- Machine Learning: Scikit-Learn, `https://scikit-learn.org/stable/`, `http://rasbt.github.io/mlxtend/`;
- Numerical Computing: `http://www.numpy.org/`, `https://www.scipy.org/`.

Assessment: Demo Code and Technical Report [30 marks in total]

You are required to write a 5-page report and supporting Python 3 program to summarise the proposed solutions, the initial results and findings. The report is expected to demonstrate your understanding of data mining in depth and what application can be derived from those methods given the dataset. Writing a convincing proposal also requires a good demo which supports your arguments and conclusions. Your report should contain the following sections:

1. Introduction [**4 marks**]: Provide an overview of the problem and your proposed solutions.
2. Proposed Solutions [**15 marks**]: Present and discuss your proposed solutions in detail. This should cover your understanding of the problem and data, visual and statistical analysis of the sample dataset, how to apply data mining methods to the problem, and what outcomes you expect to achieve. This section may contain multiple sub-sections.

[1] `https://en.wikipedia.org/wiki/Association_rule_learning`

3. Discussion and Conclusion [**4 marks**]: Provide a summary for your proposals, the initial results or prototype if you have implemented one and your critical analysis.
4. References [**2 marks**]: Include references where appropriate. **The reference section is not included in the page limit.**
5. Demo Code [**5 marks**]: Attach demo code in your submission including data visualisation, statistical analysis and prototype of your solution. A *"README.txt"* file describing how to run your codes is required.

Page Limit: The report should be **no more than 4 pages excluding reference**. Font size should be **no smaller than size 10**, and the text area is approximately 9.5x6 inches. You may use images but do so with care; do not use images to fill up the pages. You may use an additional cover sheet, which has your name and student number. **Reports that exceed the specified page limit will result in penalties: 3 marks deduction for every over-length page.**

Source Code: Submit your *Python3 Jupyter Notebook* code to Blackboard, together with your report, in a **Single Zip** file.