CSCM35 Big Data and Data Mining

by Dr. Jingjing Deng

Released on 14th Feb 2022

CSCM35 Coursework 2 Complete by 4th/Apr/2022

Assessment: Associated Rule Mining

Association rule mining 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. Implement associated rule mining algorithms (You are **NOT** allowed to use any external packages for associated rule mining);
2. Apply your algorithms to the proof-of-concept dataset and the large dataset that are provided together with this instruction;
3. Writ a technical report to present and discuss the implementation and experimental results of your algorithms.

You will be assessed based on the following criteria:

1. Python Implementation **[25 marks in total]**:
   (a) Basic I/O [2 marks]: Implement basic I/O function that can read the data from the dataset and write the results to a file.
   (b) Frequent Itemset [2 marks]: Find all possible 2, 3, 4 and 5-itemsets given the parameter of minimum-support.
   (c) Apriori Algorithm [4 marks]: Use Apriori algorithm for finding frequent itemsets.
   (d) Associated Rule [4 marks]: Find interesting association rules from the frequent itemsets given the parameter of minimum-confidence.
   (e) FP-Growth Algorithm [4 marks]: Use FP-Growth algorithm for finding frequent itemsets.
   (f) Reusability and Coding Style [9 marks]: The code should be well structured that can be easily maintained and ported to new applications. There should be sufficient comments for the essential parts to make the implementation easy to read and understand.
2. Technical Report **[25 marks in total]**:

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

(a) Introduction [5 marks]: Provide a concise but general overview of the algorithms and experiments that you have implemented. This should also cover a brief summary of what have found via evaluating your method on the given datasets.

(b) Experiment & Discussion [10 marks]: Apply your associated rule mining algorithms to the dataset and show some interesting rules. You should consider and evaluate the run-time performance of your implementation. The efficiency and scalability should also be considered and reflected in the implementation and technical report.

(c) Writing Style & Citation [10 marks]: Language usage and report format should be in a professional standard and meet the academic writing criteria. References should be included and cited where appropriate. A guide of citation style can be found at library guide[2].

Submission

In this coursework, you will be given a programming task to implement data mining algorithms. The work must be uploaded to Canvas before the deadline stated on the instruction. Source codes must be written in Python Jupyter Notebook and formatted neatly with sufficient and clear comments. Submissions will be done via Canvas system. Plagiarism will not be tolerated. Zip all your files (include source code and technical report in PDF format, but do NOT include the dataset) with the following naming convention for submission:

- [Student Number]-[Last Name][First Initial]-[Coursework][Number].zip
- For example: *123456-DengJ-Coursework2.zip*

Policy

- To be completed by students working individually.
- Word Limit: The report should be **no more than 1000 words excluding references**. You may use images, figures and tables but do so with care; do not use them to fill up the pages. You may use an additional cover sheet, which has your name and student number. **Reports that exceed the word limit will result in penalties (5 marks deduction for every over-length page).**
- Feedback: A general comprehensive feedback will be given to all students.
- 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 should not ask for help from your peers, lecturer, academic tutor and lab tutor regarding the implementation of the algorithm. Copy and paste from the Internet is not allowed. Using external code

[2] `https://libguides.reading.ac.uk/citing-references/referencingstyles`

without proper referencing is also considered as breaching academic integrity. [3]

- Submission deadline: Your implementation in *Python Jupyter* and technical report need to be submitted electronically to Canvas by the deadline. If you are not able to meet the deadline due to extenuation circumstance, you should apply for the extension formally by following the department, faculty and University regulations. [4]

[3] https://myuni.swansea.ac.uk/academic-life/academic-regulations/assessment-and-progress/academic-misconduct-procedure/

[4] https://myuni.swansea.ac.uk/academic-life/extenuating-circumstances/