COMPSCI361: Machine Learning

Assignment 4: Association Rule Mining

Due: June 4, 2021, 23:59 NZT.

This is worth 10% of your final grade.

The main purpose of this project is to get an in depth understanding of how the Apriori algorithm works.

Task 1: Given the dataset below whereby each line is representing a transaction, what is the frequent itemsets and rules produced, if *minsup* is set to 0.15 and *minconf* is set to 0.80. Show your working.

Trans ID	Transactions
1	A, B, C, D, F
2	A, B, C, D
3	A, B, C, D
4	A, B
5	B, C, E

Task 2: Your task is to write a Apriori program (in python using Jupyter notebook), that takes as parameters:

- *minsup* - minimum support,
- *minconf* - minimum confidence,
- *minlift* - minimum lift, and
- the name of file of transactions (whose format is comma separated value as that of the supermarket.csv downloaded from Canvas). Each line within the data file represents a transaction, where items are separated by commas. Imagine you are going to a grocery store, your transaction at the check-out counter would be one of the lines.

It will produce all association rules which can be mined from the transaction file which satisfy the minimum support, lift, and confidence requirements. The rules should be output sorted first by the number of items that they contain (in decreasing order), then by the lift value, then by the confidence, and finally by their support (also in decreasing order).

You can use libraries e.g. Pandas, NumPy but you may NOT use any prebuilt Apriori packages.

Task 3: Now run your implementation using the data from the Task 1. Show that you can produce the same output as Task 1. This can be the output from your Jupyter notebook.

Task 4: Your task is to investigate a dataset and perform an association rule mining task.

- Run your Apriori code on the data downloaded from Canvas (supermarket.csv). Try different parameters *minsup* e.g. 0.10, 0.15, 0.20. TIP: Please note that this may take awhile if your code is inefficient.
- Generate rules (you can try different measures (*minsup*, *minlift*, *minconf* to see which gives you more useful and interesting results)). Explain why the rules are interesting to you. TIP: Don't over think this. Just describe what the measures for the rule and why you think it is interesting.

COMPSCI361: Machine Learning

Task 5: Modify your program to take in an additional constraint and parameter called *minrelativesup*. In addition to the *minsup* constraint, add the following pruning constraint at each candidate generation of *k* level, where $k > 2$. Given *k*-itemset denoted as S_k. A frequent itemset S_k is a *k*-itemset whose *support/maxsubset* ≥ minRelativeSup. Here *maxsubset* is the maximum support for the itemsets in S_{k-1}' where $S_{k-1}' = \{s | s \subset S_k, |s| = k-1\}$. Here *s* is an itemset of size *k-1* where $k > 2$.

- Discuss how would this change the itemset and rules generated.
- Would this change the anti-monotonic property in Apriori? Discuss your answer.

You will need to start another copy of your Apriori algorithm, such that it does not interfere with Task 2. Please describe how you have done this in your report and link it to your code in the Jupyter Notebook.

What to submit?

A copy of your Jupyter notebook and a final report. They must be deposited to Canvas. Please name your report file "Your_UPI.pdf". Your report should be no more than two pages long. As a rough guide of page length, you may use font Times New Roman with size 12pt and single spacing.

This includes any images or references you may choose to show or use. Reproducible machine learning is one of the criteria of this assignment. So, you need to report processing and parameters for recreating your results. You should include a description of the different runs (if you had carried out multiple run), and why you needed to make changes from your initial choices, in the report. One simple question that will indicate whether you have fulfilled minimum requirement of reproducibility is "Can someone reproduce your results based on your explanation?".

Grading rubric

1 mark for the correct output in Task 1.

1 mark for correct implementation of *Apriori (Candidate Generation and Rule Generation)* – Task 2.

1 mark for correct output formatting as defined in the assignment - Task 2.

1 mark for the correct output in Task 3.

1 mark for justification of interesting rules in the dataset provided (0.5 discussion of each interestingness measure and rule) Task 4.

1 mark for correct implementation of *minrelativesup* constraint - Task 5.

1 mark for discussion of how the *minrelativesup* would impact/change the output of *Apriori* - Task 5.

1 mark for discussion of any changes to the anti-monotonic properties in *Apriori* - Task 5.

1 mark for reproducibility, are all the parameters defined and results repeatable.

1 mark for clarity of the report.