COMP40370 – Practical 4

ASSOCIATION RULES

Prof. Tahar KECHADI

Academic year 2021-2022

After studying the first two main phases of the knowledge discovery process (data mining process), which are data collection/exploration and data pre-processing, Association rules mining is the first approach in the data analysis phase. The main aim of this practical is to learn how to apply some popular association rules algorithms on some datasets. Python mlxtend library along with previously introduced libraries are required to complete this practical. The datasets needed to complete the practical are described below.

Assignment Files

Data files

- ./practical04.pdf — assignment questions (this file)
- ./specs/ gpa_question1.csv — data file for Question 1
- ./specs/bank_data_question2.csv — data file for Question 2
- ./test_practical4.py — test file to validate your python program.

Question 1: Association rules with Apriori

The file "gpa_question1.csv" contains a data sample of University students. The main objective of this exercise is to extract interesting association rules, if there is any, from this file. As in any data mining process, we start by pre-processing the dataset before its analysis.

1. Preprocess the data (if need) so that it is free from missing values, noise, outliers.
2. Use the Apriori algorithm to generate frequent itemsets from the input data, with a minimum support equals to 0.15. In your answer, comment on the number of frequent itemsets and their sizes.
3. Does the attribute "count" have an impact on the Apriori algorithm's results? Justifier your answer.
4. Sort the itemsets according to the support in descending order. Save the generated itemsets into ./output/question1_out_apriori.csv. Include the support column in your output file.
5. Using these frequent itemsets, find all association rules with a minimum confidence equals to 0.6.
6. Sort the itemsets according to the confidence in descending order. Save the generated rules into ./output/question1_out_rules06.csv. Include the support and confidence columns in your output file.

7. Using the same frequent itemsets as in 5), find all association rules that satisfy a minimum confidence of 0.9. Include a short description for major 5 rules in your report.
8. Sort the itemsets according to the confidence in descending order. Save the generated rules into ./output/question1_out_rules09.csv in the same format as in the previous questions.

Question 2: Association rules with FP-Growth

The file ./specs/bank_data_question2.csv contains customer records from the marketing department of a financial firm. The data contains the following fields.

id:	a unique identification number of a customer
age:	age of customer in years (numeric)
sex:	customer's gender (MALE or FEMALE)
region:	inner city / rural / suburban / town
income:	income of customer (numeric)
married:	YES / NO
children:	number of children (numeric)
car:	owns a car → YES / NO
save acct:	if the customer has a saving account - YES / NO
current acct:	if the customer has a current account - YESY / NO
mortgage:	if the customer has a mortgage - YES / NO
pep:	if the customer has signed for a Personal Equity Plan after the last mailing - YES / NO

1. Data Pre-processing:
    a. Which attributes should be selected for data mining task? Justify your answer. (Hint: explain why you exclude, if any, some attributes).
    b. Discretize the numeric attributes into 3 bins of equal width.
2. Assume that the minimum support is equal to 20%. Use the FP-Growth algorithm to generate frequent all frequent itemsets. Comment in your report on the frequent itemsets, number, size, and usefulness.
3. Sort the itemsets according to the support in descending order. Save the generated itemsets into ./output/question2_out_fpgrowth.csv
4. Generate all the rules associated to these frequent itemsets.
5. Which confidence values that can return a set of rules of size at least equal to 10 (i.e., the number of rules). Explain in your report how did you identify these confidence values.
6. Sort the itemsets according to the confidence in descending order. Save the generated rules into ./output/question2_out_rules.csv
7. Identify 4 most interesting rules, explaining, for each rule:
    - why you believe it is interesting, based on the company's business objectives,
    - the recommendations that might help the company to better understand behaviour of its customers or its marketing campaign.

    Note that the most interesting rules should provide some non-trivial and actionable knowledge based on the underlying business objectives.

Expected output and submission data

Your submission should be a single archive file (zip, tar, tgz, . . .) containing one folder called output and the following files:

- ./run.py — main Python script (your python program)
- ./report.pdf — PDF report (max 3 pages)
- ./output/question1_out_apriori.csv — frequent itemsets for Q 1
- ./output/question1_out_rules06.csv — association rules for Q 1 (C = 0.6)
- ./output/question1_out_rules09.csv — association rules for Q 1 (C = 0.9)
- ./output/question2_out_fpgrowth.csv — frequent itemsets for Q 2
- ./output/question2_out_rules.csv — association rules for Q 2
- ./test_practical4.py — test file to validate your python program

Programming requirements and tools

You may need to use the following packages for this assignment:

- pandas 1.3+
- mlxtend 0.19+

The documentation of mlxtend can be found here: http://rasbt.github.io/mlxtend. In particular, in the following you will find user guides URLs of the required algorithms for this practical:

- Apriori: http://rasbt.github.io/mlxtend/user_guide/frequent_patterns/apriori
- FP-Growth: http://rasbt.github.io/mlxtend/user_guide/frequent_patterns/fpgrowth
- Association rules: http://rasbt.github.io/mlxtend/user_guide/frequent_patterns/association_rules

Keep in mind that you can save the data frames generated by the Apriori, FPgrowth, and other algorithms directly into csv files.

Note that the data attributes required by the mlxtend functions should be binomial. You can use the pandas get dummies function, in case you do not want to use the mlxtend's encoders.

o0o

The final deadline for the submission is Thursday, 14th of October at 23:00. All submissions must be done in Brightspace.