CS 422: Data Mining

Department of Computer Science
Illinois Institute of Technology
Vijay K. Gurbani, Ph.D.

Fall 2022: Homework 5 (10 points)

Due date: Sunday, October 30 11:59:59 PM Chicago Time

Please read all of the parts of the homework carefully before attempting any question. If you detect any ambiguities in the instructions, please let me know right away instead of waiting until after the homework has been graded.

1. Exercises (4 points divided evenly among the questions) Please submit a PDF file containing answers to these questions. Any other file format that can be read will lead to a loss of 0.5 point. Non-PDF files that cannot be opened and read for grading will lead to a loss of all points allocated to this exercise.

1.1 Tan, Ch. 5 (Association Analysis)

Questions 15

1.2 Zaki, Chapter 8 (Frequent Pattern Mining)

Questions 1(a), 4

2. Practicum problems (6 points) Please label your answers clearly, see Homework 0 R notebook for an example (Homework 0 R notebook is available in "Blackboard → Assignment and Projects → Homework 0".) Each answer must be preceded by the R markdown as shown in the Homework 0 R notebook (### Part 2.1-A-ii, for example). Failure to clearly label the answers in the submitted R notebook will lead to a loss of 2 points per problem below.

2.1 Association Analysis

In this assignment you will be using the *Extended Bakery* dataset, which describes transactions from a chain of bakery shops that sell a variety of drinks and baked goods.

The dataset is presented as a series of transactions, 1,000, 5,000, 20,000 and 75,000 transactions, stored in files named tr-1k.csv, tr-5k.csv, tr-20k.csv and tr-75k.csv, respectively. Each file contains the data in a sparse vector format, i.e., each line of the file has the following format:

1, 7, 15, 44, 49
2, 1, 19
...

The first column is the transaction ID and the subsequent columns contain a list of purchased goods from the bakery represented by their product ID code. In the example above, the first line implies that transaction ID one contained four items: 7, 15, 44, and 49. The mapping of the product ID to product name is provided in the products.csv file.

(a) **[1 points]** For each series of transaction files (i.e., tr-5k.csv, tr-20k.csv, …) create a canonical representation of the transaction file. A canonical representation for each dataset will be a file that contains a list of product names (not IDs) on a line, each product separated by a comma and a newline ends the line. So, as an example, the first two lines shown above (7, 15, 44, 49; and 1, 19) would correspond to the following canonical representation, respectively:

> Coffee Eclair, Blackberry Tart, Bottled Water, Single Espresso
> Lemon Cake, Lemon Tart
> …

Save the canonical representation in files with the canonical suffix, i.e., tr-5k-canonical.csv, and so on. Use these files for the rest of the work. **Include these files in the archive that you upload to Blackboard.**

You can use any programming language of your choice to do part (a).

(b) **[3 point]** Given the database of transactions, where each transaction is a list of items, find rules that associate the presence of one set of items with that of another set of items. Ideally, we only want to find rules that are substantiated by the data; we want to avoid spurious associations.

Find association rules that exceed specific values of *minimum support* and *minimum confidence*. You are free to experiment with different values until you find something that produces meaningful results. However, be aware that if you specify minimum support or confidence very low, your R process may appear to "hang" as the many itemsets are mined. In such a case, restart R. Use a structured approach by first reading the documentation to understand what the default value of *minsup* and *minconf* is, and then experiment from there.

Recall that finding rules requires two steps: finding the *frequent itemsets* and discovering strong *association rules* within them. You will use the R **arules** package as shown in class.

Your output should contain the following:

- For each frequent itemset:
    1. All items in it, described by the product names.
    2. The support of the itemset.
- For each rule:
    1. The antecedent.
    2. The consequent.
    3. The support of the rule.
    4. The confidence of the rule.

(c) **[1 point]** Given the above output, respond to the following question: Compare the rules you obtained for each different subset (1,000 – 75,000 transactions). How does the number of transactions affect the results you observed? (Write the answer in your R markup file, easily identified.)

(d) **[1 point]** Answer the following questions for the 75,000 transactions dataset using the same support level as determined in (b):

> (i) What is the most frequently purchased item or itemset?

(ii) What is the least frequently purchased item or itemset?