CS699 A1 – Spring 2025

Homework 8

Due: 4/7

**Problem 1 (20 points)**. Consider the following transactional database.

|  |  |
| --- | --- |
| TID | Items |
| 100 | 2, 4, 5, 6 |
| 200 | 1, 4, 5, 7 |
| 300 | 2, 4, 5 |
| 400 | 1, 2, 4, 5, 6, 7 |
| 500 | 1, 2, 6 |

1. Mine all frequent itemsets using the Apriori algorithm that we discussed in the class. Show all candidate itemsets and frequent itemsets. You should follow the step by step process that we discussed in the class (i.e., C1 → L1 → C2 → L2 → …). You don't need to show the pruning steps. Minimum support = 30% (or 2 or more transactions). To save your time, L1 is given below:

L1:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Itemset | 1 | 2 | 4 | 5 | 6 | 7 |
| Count | 3 | 4 | 4 | 4 | 3 | 2 |

1. Sort all frequent 4-itemsets by their item number. Then, select the first frequent 4-itemset from the sorted list of frequent 4-itemsets and mine all strong rules from this itemset that have the format {W, X} => {Y, Z}, where W, X, Y, and Z are individual items. Assume that minimum confidence = 80%.

**Problem 2 (10 points).** This problem is about the collaborative filtering that we discussed in the class. Consider the following matrix, which shows ratings of four items by four users:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | I1 | I2 | I3 | | I4 |
| U1 | 2 | 2 | 4 | | 3 |
| U2 | 1 | 3 | ***?*** |  | 5 |
| U3 | 2 | 4 | 1 | | 1 |
| U4 | 5 | 2 | 1 | | 3 |

1. Predict the rating of Item 3 (I3) by User 2 (U2) using UBCF.
2. Predict the rating of Item 3 (I3) by User 2 (U2) using IBCF.

You must use the method we discussed in the class and you must do all calculations yourself without using any data mining or data analysis software.

**Problem 3 (10 points).** Consider the following contingency table.

|  |  |  |
| --- | --- | --- |
|  | *C* (buys coffee = Yes) | *C* (buys coffee = No) |
| *T* (buys tea = Yes) | 42 | 64 |
| *T* (buys tea = No) | 168 | 121 |

Compute the *lift*, *all-confidence*, *cosine*, and *Kulczynski* measure, and determine whether buying coffee and buying tea are positively correlated, negatively correlated, or not correlated.

**Problem 4 (10 points).** Suppose you have the following transactional database:

|  |  |  |
| --- | --- | --- |
| CID | Time | Items |
| 1 | 1 | 10,30 |
| 1 | 4 | 80 |
| 1 | 6 | 60, 70 |
| 2 | 3 | 10 |
| 2 | 7 | 30,40 |
| 2 | 20 | 60,70,80 |
| 3 | 2 | 30,50 |
| 3 | 8 | 70,80 |
| 4 | 2 | 70 |
| 4 | 10 | 80 |
| 5 | 1 | 30, 50 |
| 5 | 5 | 70 |

Show all frequent 2-sequences along with their supports. Assume that the minimum support is 40% (or two data-sequences). You must not use any data mining software including R. You must find all frequent 2-sequences yourself.

**Problem 5 (10 points)**. This problem is about mining frequent sequences and induced temporal rules from a transactional database using R. Use *hw8\_p5.csv* data for this problem. Write a R program for this problem and name it *hw8.R*.

1. Mine all frequent sequences with minimum support = 40% and include the screenshot of the summary in your submission.

(1)-1. How many frequent 1-sequences are there? (1)-2. How many frequent 2-sequences are there? (1)-3. How many frequent 3-sequences are there? (1)-4. How many frequent 4-sequences are there?

1. From the above frequent sequences, mine all induced temporal rules. How many rules are there?
2. Show top 5 rules based on *confidence*.
3. Show top 5 rules based on *lift*.

**Submission:**

You must submit a single archive file that includes the following two files:

* + *Hw8.pdf* or *hw8.docx*, which includes answers to all problems.
  + *hw8*.*R* which includes the R code for Problem 5.

Name the archive file *LastName\_FirstName*\_*HW8.EXT*. Here, “*EXT*” is an appropriate archive file extension (e.g., *zip* or *rar*) and submit it to Blackboard.