**CS 536/CS 432 – Data Mining**

**Assignment 2**

**Due: March 12 (Tuesday) at 12 midnight**

**Instructions:** Submit a soft-copy report to the submission folder on LMS. Include report and code needed to reproduce your results. **Note:** You should discuss results/outcomes of each part in detail in your report and provide code for Q3 part A and pictures for Q1 (if done on notebook) in your submission. Zip the folder and name it as Rollnumber\_Name. In case of plagiarism (in any of the part), whole assignment will be graded zero.

1. **Apriori and FP-Growth Algorithms (40 points)**

Consider the following transactional database:

|  |  |
| --- | --- |
| TID | Items |
| 1 | BD |
| 2 | ABD |
| 3 | AC |
| 4 | EF |
| 5 | CDEF |
| 6 | BE |
| 7 | AE |
| 8 | AEF |
| 9 | ADE |
| 10 | AE |
| 11 | BDF |
| 12 | DE |
| 13 | DFF |
| 14 | CDE |

1. Find all frequent itemsets using the Apriori algorithm. Assume minimum support count is 2.
2. Find all frequent itemsets using the FP-growth algorithm. Assume minimum support count is 2.
3. Identify all closed and max itemsets.
4. Generate all strong association rules from the longest closed pattern(s) found in the database. Assume minimum confidence is 70%.
5. **Frequent Itemset Mining Using Rapid Miner (40 points)**

Experiment with RapidMiner’s implementation of Apriori and FP-growth algorithms. Apply these algorithms to the Adult income data set (available from LMS)

Preprocessing needed like missing values and discretize numerical values into 4 bins for all three numerical columns)

**For instance,** discretizing educational-num with four bins, defined boundaries set min = 0 and max = 16 because the statistics showed educational-num is distributed between 0 to 16. Resultantly it will have be nominal having bins with ranges 0-4,4-8,8-12,8-12

1. For Apriori generate rules and itemsets for (i) default parameter values, (ii) confidence = 0.7; rules = 50, (iii) minimum support is 0.05. Summarize the results and discuss/interpret them w.r.t income of individuals and their information.

(**Search for Weka in Operators. It will prompt to Install weka extension. Download and install it. Then search for W-Apriori operator to use in this part)**

1. For FP-growth, generate itemsets for (i) default parameter values, (ii) minimum support = 0.1, (iii) find min number of itemsets is unchecked, minimum support = 0.1. Summarize and interpret the interesting results.
2. From results in (a), separate out all strong classification rules, i.e., rules that contain the class attribute (income) on the right-hand-side. (If any)
3. Mention any other interesting result by any approach or settings. Like I found “If a person earns more than $50000, he is very likely to be a married man with large number of years of education”
4. **Download the spambase dataset from LMS. (20 points)**

**(Python or R)**

* 1. Divide the dataset into 4 equal bins and find the correlated attributes from each bin. Compare the results from each bin for few highly +ve & -ve correlated attributes. Just write as attr1, attr2… for indicating attribute/columns.(Provide code file for this part)
  2. Mention how correlation among features can help us in dimensionality reduction. You can briefly mention any other various techniques of your choice for dimensionality reduction. Please report your technique in document.