Data Mining and Visualization module [DMV]

Final exam – November 2022

Duration: 2 hours

All paper documents authorized. Please read the subject carefully. Write legibly. You can answer to the questions in English or in French. A tentative allocation of points to exercises is given, however it may change at the time of grading.

Exercise 1 [7 points]

Consider the sequence database below, defined over the items {a,b, c, d, e, f, g}:

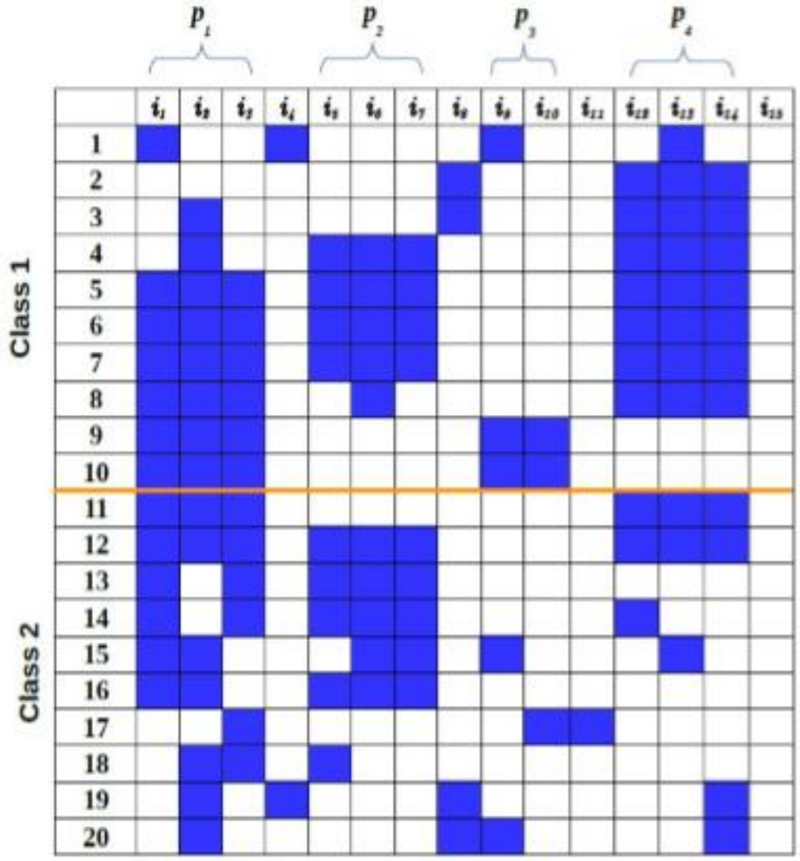
Sequence ID	Sequence
S1	<c(ef)(ab)fg>
S2	<a(ab)(ef)f>
S3	<(ab)c(df)a >
S4	<a(bd)b(fg)d>
S5	<d(ef)bcd(abc)>
S6	<(ab)(efg)>

Question 1 [5 points]. Apply the PrefixSpan algorithm on this data and report the frequent sequential patterns found along with their frequency. The minimum support threshold is 60%. All the steps of the algorithm must be detailed.

Question 2 [1 point]. In the found frequent sequential patterns, which one are closed? Justify your answer.

Question 3 [1 point]. Same question for the frequent sequential patterns that are maximal. Justify your answer.

Exercise 2 [4 points]



		P_1				P_2				P_3			P_4			
		i_1	i_2	i_3	i_4	i_5	i_6	i_7	i_8	i_9	i_{10}	i_{11}	i_{12}	i_{13}	i_{14}	i_{15}
Class 1	1	■			■					■				■		
	2								■				■	■	■	
	3		■						■				■	■	■	
	4		■			■	■	■					■	■	■	
	5	■	■	■		■	■	■					■	■	■	
	6	■	■	■		■	■	■					■	■	■	
	7	■	■	■		■	■	■					■	■	■	
	8	■	■	■			■						■	■	■	
	9	■	■	■						■	■					
	10	■	■	■						■	■					
Class 2	11	■	■	■									■	■	■	
	12	■	■	■		■	■	■					■	■	■	
	13	■		■		■	■	■								
	14	■		■		■	■	■					■			
	15	■	■				■	■		■				■		
	16	■	■			■	■	■								
	17			■							■	■				
	18		■	■		■										
	19		■		■				■						■	
	20		■						■	■					■	

The figure above shows a transaction dataset with 15 items (columns) and 20 transactions (rows). Filled squares indicate the presence of an item in a transaction. The dataset is divided into two classes: transaction 1 to 10 are in class 1, and transactions 11 to 20 are in class 2.

Question 1 [1 point]. 4 patterns *P_1, P_2, P_3, P_4* are shown in the figure. For each pattern, give its support in the total dataset, and in class 1 and class 2.

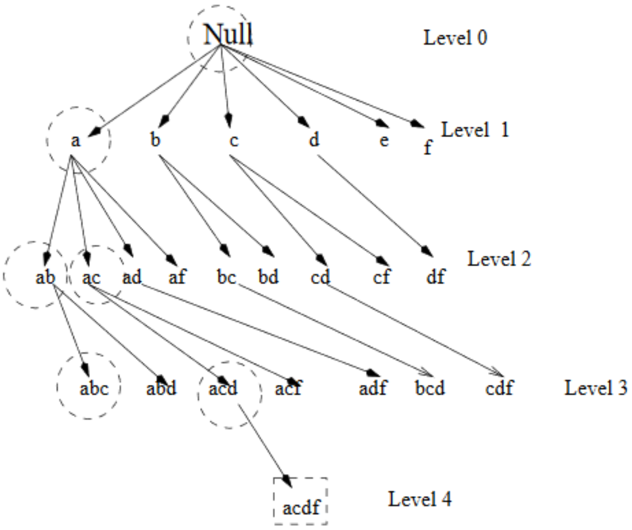
Question 2 [3 points]. Are any of these patterns discriminative patterns for class 1? Justify your answer by explaining your choice of discriminance measure, and by detailing your computation of this measure for each pattern.

If you had to return only one discriminative pattern to the user, which one would it be? Justify your answer.

Exercise 3 [5 points]

During the course, we have studied the FP-Growth algorithm, which is based on the FP-Tree data structure. In their journal paper of 2002, the authors of FP-Growth compare their work with another algorithm that use a tree-based data structure, called *TreeProjection* (Aggarwal et al, 2001).

TreeProjection relies on a lexicographic order of items. It constructs iteratively a lexicographic tree of the frequent itemsets, as in the figure below (*source: Aggarwal et al. Circles on nodes are not important for this exercise.*):



Question 1 [1 point]. At each node of the tree, and similarly to FP-Growth and other algorithms, TreeProjection constructs a projected database (also called *conditional database*) in order to perform support counting. Note that the difference with FP-Growth is that the conditional database is not expressed with an FP-Tree: it is just a multiset of transaction, as the original database.

Recall why a projected database is it advantageous for pattern mining algorithms (max ~10 lines). As an example, what will be the common characteristics of all transactions of the projected database for node *ab*? (nb: you don't need the database to answer that question).

Question 2 [4 points]. Each node of the tree represents a pattern *P*. TreeProjection then computes *E(P)*, the set of possible extensions of *P* (i.e. all the items that could be added to *P* to make a bigger frequent itemset). In order to determine the child nodes of *P* in the lexicographic tree, a support counting phase takes place. It is performed with a matrix of size *E(P)*E(P)* where each row and column represents an item of *E(P)*. The cell *[i, j]* of the matrix receives the support count of pattern $P \cup \{i, j\}$, which is equals to the support count of *{i,j}* in the projected database of *P*. As the matrix is symmetric, only the lower triangular part of the matrix is computed.

1. [1 point] Let $T(P) = [t_1,\ldots,t_n]$ be the projected database of *P*. By construction, each t_i only contains items of *E(P)*. What is the time complexity of the support counting step (construction of the above matrix)?
2. [1 point] In their paper, the authors of TreeProjection make a special section on memory requirements, and state that: "*The memory requirement of TreeProjection is equal to the sum of the memory requirement for triangular matrices* [rq: matrices described above] *at each level (k-1)-node of the tree. At first sight, this might seem as a rather large requirement. However, this is proportional to the number of candidate (k+1)-itemsets at that level. […] Thus, the memory requirements […] are quite comparable to other schemes in the literature*".
   Why is the memory requirement indeed large? What is the potential problem with the TreeProjection algorithm?
3. [2 points] Compare the lexicographic tree of TreeProjection with the FP-tree of FP-Growth. Which one should use the less memory? Justify your answer.

Exercise 4 [3 points]

Propose a visualization application to display the user's shopping behavior of a given supermarket. The main objective from the management perspective is to improve the positioning of the different top selling items in order to maximize the traversal of customers in the supermarket. Propose at least two visualizations that summarize the shopping behavior of customers to help the decision process. For each visualization discuss the what? why? and how?

You can suppose that you have access to all sales data of the supermarket.

Remember that the visualizations should not only include static visualizations but interactive ones, and they could be linked together.

For example, you could consider the relationship between the top selling items, the selling statistics for each item or the overall customer coverage over the supermarket.