Institutionen för datavetenskap
Linköpings universitet

EXAM
732A02 Data Mining – Clustering and Association Analysis
April 29, 1-5pm

Teachers: Patrick Lambrix, José M Pena

Grades: Requirement for grade C: 15 points. Requirement for grade A: 23 points

Instructions:
- Start each question at a new page.
- Write at one side of a page.
- Add name and personal number on each page.
- Write clearly.
- If you make assumptions about a question, that are not explicitly stated, you need to write these down. (These assumptions cannot change the exercise or question.)

Help: dictionary

GOOD LUCK!

1. Apriori algoritm (3p+2p=5p)

a. Run the Aprori algorithm on the following transaction database with minimum support equal to 2 transactions. Explain step by step the execution.

Transaction id	Items
1	A, B, C
2	A, B, C
3	C, D
4	C, D

b. Sketch a proof of the correctness of the Apriori algorithm. It suffices to sketch a proof showing that all the frequent itemsets of size k are among the candidates of size k.

2. FP algorithm (2p+1p=3p)

a. Run the FP algorithm on the transaction database in exercise 1 with minimum support equal to 2 transactions. Explain step by step the execution.

b. Discuss the advantages and disadvantages of the FP algorithm with respect to the Apriori algorithm.

3. Constraints (2p+2p+2p=6p)

a. Run the Apriori algorithm on the transaction database in exercise 1 with minimum support equal to 2 transactions and the constraint that the sum of the prices of the items in an itemset must be greater than 1 (do not simply run the algorithm and afterwards consider the constraint but incorporate the constraint into the algorithm). Explain step by step the execution.

Item	Price
A	1
B	2
C	1
D	1

b. Run the Apriori algorithm on the transaction database in exercise 1 with minimum support equal to 2 transactions and the constraint that the sum of the prices of the items in an itemset must be equal or smaller than 2 (do not simply run the algorithm and afterwards consider the constraint but incorporate the constraint into the algorithm). Explain step by step the execution.

c. Give an example of a convertible monotone constraint that is not monotone. Give an example of a convertible antimonotone constraint that is not antimonotone. How would you incorporate convertible monotone and antimonotone constraints into the Apriori algorithm ?

4. Clustering by Partitioning (4p+1p=5p)

a. Describe the principles and ideas regarding PAM. Explain the different steps of the algorithm.

b. Given the graph representation of the clustering problem where a node represents k mediods (and thus a potential solution for the clustering), and nodes are neighbors if the sets of objects represented by the nodes differ by one object. Finding a solution for the clustering problem can then be seen as a search in the graph.
What is the difference between PAM and CLARANS regarding searching this graph?

5. Hierarchical clustering (4p)

Describe the principles and ideas regarding Agglomorative Hierarchical Clustering.
Show the different steps of the algorithm using the dissimilarity matrix below
and complete link clustering. Give partial results after each step.

	1	2	3	4	5
1	0				
2	2	0			
3	4	3	0		
4	10	7	9	0	
5	8	5	6	1	0

6. Density and grid-based clustering (3p)

Describe the principles and ideas regarding the CLIQUE algorithm.
What is the main purpose of the algorithm? For what kind of purpose would you use this algorithm? Explain the major steps. What are the strenghts and weaknesses of the algorithm?

7. Potpourri (2p+1p+1p=4p)

a. Given the following two objects.
- What is the distance between the objects if all variables are symmetric?
- What is the distance between the objects if all variables are asymmetric?

	Attribute1	Attribute2	Attribute3	Attribute4
Object1	1	1	0	0
Object2	1	0	1	0

b. Give an example of an ordinal variable. Give an example of a nominal variable.

c. When is a point p directly density-reachable from a point q w.r.t Eps and MinPts?