Information page for written examinations at Linköping University

Examination date	2016-08-23
Room (1)	G32
Time	8-12
Course code	732A31 + TDDD41
Exam code	TEN1
Course name **Exam name**	Data Mining - Clustering and Association Analysis (Data Mining - Clustering and Association Analysis) Written Examination (Skriftlig tentamen)
Department	IDA
Number of questions in the examination	7
Teacher responsible/contact person during the exam time	Patrick Lambrix, Jose Pena
Contact number during the exam time	2605, 1651
Visit to the examination room approximately	9:30, 11:00
Name and contact details to the course administrator (name + phone nr + mail)	Elin Brödje, 4767, elin.brodje@liu.se
Equipment permitted	dictionary
Other important information	
Number of exams in the bag	

Institutionen för datavetenskap
Linköpings universitet

EXAM

732A31 and TDDD41
Data Mining –
Clustering and Association Analysis
August 23, 2016, kl 8-12

Teachers: Patrick Lambrix, José M Pena

Instructions:

- Start each question at a new page.
- Write at one side of a 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. Clustering by partitioning (4+1=5p)

a. - Describe the algorithm for CLARANS.
- Assume the data set {0, 3, 4, 10}. Assume we use Euclidean distance and k = 2. Draw the graph representation of the clustering problem. Then start at an arbitrary node and show one iteration (i.e. for one value of numlocal) of the CLARANS algorithm with maxneighbor = 2. Make sure to show all computations the algorihtm makes.

b. Which, if any, of the algorithms PAM/CLARA/CLARANS guarantees a global optimum for the cost function?

2. Hierarchical clustering (3+3=6p)

a. 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.

	A	B	C	D	E
A	0				
B	8	0			
C	5	10	0		
D	1	7	3	0	
E	6	2	4	9	0

a. Describe the principles and ideas regarding the ROCK algorithm. For what kind of data is ROCK particularly suited? Explain the major steps of the algorithm. Make sure to define neighbor, common neighbor, link and goodness measure.

3. Density-based clustering (2p)

a. DBSCAN: Consider the following statement: if p is density-connected to q wrt Eps and Minpts then p is density-reachable from q wrt Eps and Minpts. Is this statement true? If yes, then prove. If no, then give a counterexample.

b. What is the main idea behind OPTICS?

4. Different types of data and their distance measures (2+1=3p)

a. What is the distance between Item K and Item L?

	A	B	C	D	E	F	G
Item K	(5,2)	(7,5,1)	Y	Y	N	N	8
Item L	(5,-1)	(4,4,1)	N	Y	Y	N	no-value-available

Attribute A is interval-based and Euclidean distance is used.
Attribute B is interval-based and Manhattan distance is used.
Attributes C and D are binary symmetric variables.
Attributes E and F are binary asymmetric variables.
Attribute G is interval-based.

b. Show how an interval-based measure can be defined for ordinal variables.

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

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

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

b. Repeat the exercise 5a with the following additional constraint: Find the frequent itemsets that contain the item A. Explain step by step the execution. Make clear when and how the constraint is used. Incorporate the constraint into the algorithm, i.e. do not simply run the algorithm and afterwards consider the constraint.

c. Repeat the exercise 5a with the following additional constraint: Find the frequent itemsets that do not contain the item A. Explain step by step the execution. Make clear when and how the constraint is used. Incorporate the constraint into the algorithm, i.e. do not simply run the algorithm and afterwards consider the constraint.

d. Sketch a proof of the correctness of the Apriori algorithm.

6. FP growth algorithm (2p+1p+1p+1p=5p)

a. Describe the FP growth algorithm. Do not use examples.

b. Explain how you incorporate a monotone constraint into the FP growth algorithm.

c. Explan how you incorporate an antimonotone constraint into the FP growth algorithm.

d. What is the main advantage of the FP growth algorithm over the Apriori algorithm ?

7. Constraints and (1p+1p+1p+1p=4p)

a. Let D1 denote the following definition of convertible monotone constraint. A constraint is convertible monotone if there exists an ordering R1 of the items such that when an itemset violates the constraint so does any **suffix** of the itemset. Let D2 denote the following definition of convertible monotone constraint. A constraint is convertible monotone if there exists an ordering R2 of the items such that when an itemset violates the constraint so does any **prefix** of the itemset. Are D1 and D2 equivalent, i.e. does any constraint either satisfy both of them or none ? Explain your answer.

b. Assume that a constraint satisfies both D1 and D2. Are R1 and R2 the same ?

c. Give an example of a constraint that is both convertible monotone and convertible antimonotone (i.e. strongly convertible). Explain your answer.

d. Give an example of an association rule with lift greater than one and another example of a rule with lift smaller than one.