Information page for written examinations at Linköping University

Examination date	2013-03-13
Room (1) If the exam is given in different rooms you have to attach an information paper for each room and mark intended place	TER3
Time	8-12
Course code	TDDD41
Exam code	TEN1
Course name **Exam name**	Data Mining - Clustering and Association Analysis 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 approx.	9.30, 11.15
Name and contact details to the course administrator (name + phone nr + mail)	Carita Lilja, 1463, carita.lilja@liu.se
Equipment permitted	dictionary
Other important information	
Which type of paper should be used, cross-ruled or lined	
Number of exams in the bag	

Institutionen för datavetenskap
Linköpings universitet

EXAM
732A31and TDDD41
Data Mining –
Clustering and Association Analysis
March 13, 2013, 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. Apriori algoritm (2p+1p+1p+2p=6p)

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

Transaction id	Items
1	A, B
2	A, B
3	A, B
4	A, C
5	A, C
6	A, C
7	A, D
8	A, D
9	A, D
10	A, B, E
11	A, C, E
12	A, B, E
13	A, C, E
14	A, D, E

b. Repeat the exercise 1a with the following additional constraint: Find the frequent itemsets that contain the items A and/or B. 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 1a with the following additional constraint: Find the frequent itemsets that do not contain the items A and B (both). 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.

2. FP grow algorithm (2p+1p+1p+1p=5p)

a. Repeat the exercise 1a with the FP grow algorithm. Explain step by step the execution.

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

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

d. What is the main advantage that the FP grow algorithm has over the Apriori algorithm ?

3. Constraints and lift (1p+1p+1p=3p)

a. Give an example of a constraint that is both monotone and antimonotone. If you think it is not possible, explain why.
b. Consider the constraint avg(S)+range(S) where S is an itemset, avg(S) returns the average price of the itemset S, and range(S) returns the difference in price between the most expensive and the cheapest items in S. Is this constraint monotone, antimonotone, convertible monotone, convertible antimonotone, or none ?
c. Give an example of a rule with lift greater than one and another example of a rule with lift smaller than one.

4. Clustering by Partitioning (3p)

Given the graph representation of the clustering problem where n is the number of data points and k is the number of clusters.

(i) What does a node represent?
(ii) How can this graph be used for finding a solution for the clustering problem?
(iii) When are two nodes neighbors and how many neighbors does a node have?
(iv) Considering PAM, CLARA and CLARANS, the graph for which algorithm/algorithms contains/contain the most nodes? Explain.
(v) For a given node, which of CLARANS and PAM visits the most neighbors? Explain.
(vi) Which of PAM, CLARA and CLARANS guarantees to find a globally optimal solution? (Obs: not necessarily 1 answer.)

5. 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 single link clustering. Give partial results after each step.

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

b. Describe the principles and ideas regarding the CHAMELEON algorithm..In your description, make sure to define the important notions and define how clusters are formed.

6. Density-based clustering (2p+1p+2p=5p)

a. Describe the principles and ideas regarding the DBSCAN algorithm. In your description, make sure to give a sketch of the algorithm and to define core point, direct density-reachable, density-reachable, and density-connected.
b. What is the main idea behind OPTICS?
c. Describe the principles and ideas regarding the DENCLUE algorithm. In your description, make sure to define the important notions and define how clusters are formed. Also discuss whether arbitrary-shape clusters can be formed.

7. Distance measure (2p)

What is the distance between Item K and Item L?

	A	B	C	D	E	F	G
Item K	(5,6)	(1,1)	Y	N	Y	N	(1,1,1)
Item L	(5,5)	(2,3)	Y	N	N	N	no-value-available

Attributes A and G are 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.