Information page for written examinations at Linköping University

Examination date	2014-06-10
Room (1) If the exam is given in different rooms you have to attach an information paper for each room and mark intended place	TER1
Time	8-12
Course code	732A31
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 (9:30), Jose Pena (10.45)
Contact number during the exam time	2605 /1651
Visit to the examination room approx.	9:30, 10:45
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
June 10, 2014, 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 (2p+2p=4p)

a. Describe the principles and ideas regarding PAM.
- Give a sketch of the algorithm.
- Define swapping cost.

b. 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.

2. Hierarchical clustering (3p+3p=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	3	0			
3	5	10	0		
4	1	7	4	0	
5	6	2	8	9	0

b. Describe the principles and ideas regarding the ROCK algorithm. For what kind of data is this algorithm particularly suited? Explain the major steps. Further, give an example with 4 objects that shows what a *neighbor* and a *common neighbor* are in ROCK and how it is used to define *Link*.

3. Density-based clustering (1p+1p+2p=4p)

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?
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.

4. Distance measure (2p)

What is the distance between Item K and Item L?

	A	B	C	D	E	F	G
Item K	(3,10)	(4,4)	Y	N	Y	N	8
Item L	(3,40)	(5,7)	N	N	N	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.

5. Apriori algoritm (2p+1p+1p+2p=6p)

a. Describe the Apriori algorithm.

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

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

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

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

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

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

b. Repeat the exercise 6a with the following additional constraint: Find the frequent itemsets that do not contain the item C. 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. Let the items A, B, C and D have a price of respectively 1, 2, 3 and 4 units. Repeat the exercise 6a with the following additional constraint: Find the frequent itemsets whose total cost is smaller than 11. 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. What is the main advantage that the FP grow algorithm has over the Apriori algorithm ?

7. 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. Apply the Simple algorithm to the frequent itemset ABC on the database in exercise 6 in order to find association rules with confidence greater or equal than 50 %.

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