Information page for written examinations at Linköping University

Examination date	2016-06-08
Room (1)	TER3
Time	8-12
Course code	TDDD41 + 732A31
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 Jose Pena visits the exam room.
Contact number during the exam time	2605 / 1651
Visit to the examination room approximately	10:00
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	

Information page for written examinations at Linköping University

Examination date	2016-06-08
Room (1)	TER3
Time	8-12
Course code	732A31
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 Jose Pena visits the exam room.
Contact number during the exam time	2605 / 1651
Visit to the examination room approximately	10:00
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

732A31and TDDD41
Data Mining –
Clustering and Association Analysis
June 8, 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 (5p)

- Describe the principles and ideas regarding PAM.
  - Describe the algorithm.
  - Define swapping cost.
  - Draw an example of a data set in two dimensions where the swapping cost TCih is 0 and one where the swapping cost TCih is strictly negative.

- Why is PAM more robust than K-means in the presence of outliers?
- Why is CLARANS more efficient than PAM?

2. Hierarchical clustering (3+2=5p)

a. Describe the principles and ideas regarding Agglomerative 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	8	0			
3	3	10	0		
4	1	7	4	0	
5	2	6	5	9	0

b. Describe the principles and ideas regarding the CHAMELEON algorithm. Explain the major steps.

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

a. Describe the principles and ideas regarding the DBSCAN algorithm. In your description, make sure to describe the algorithm and to define core point, direct density-reachable, density-reachable, and density-connected.
b. What is the relationship between DBSCAN and OPTICS?

4. Different types of data and their distance measures (3p)

a. Give and explain the distance measure for objects with asymmetric binary variables using contingency tables.
b. Give and explain the distance measure for objects with variables of mixed types.

c. Can the formula in question b also be used for objects with only asymmetric variables? If no, explain why. If yes, state whether you would get the same results as with the method in question a and explain why or why not.

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

a. Describe the Apriori algorithm. Pay special attention to candidate generation and pruning due to subset checking. Do not use examples.

b. Describe how to incorporate a monotone constraint in the Apriori algorithm. Pay special attention to any efficiency gain obtained. Do not use examples.

c. Describe how to incorporate an antimonotone constraint in the Apriori algorithm. Pay special attention to any efficiency gain obtained. Do not use examples.

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

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

a. Describe the FP grow algorithm. Pay special attention to the mining of the tree. Do not use examples.

b. Describe how to incorporate a monotone constraint in the FP grow algorithm. Pay special attention to any efficiency gain obtained. Do not use examples.

c. Describe how to incorporate an antimonotone constraint in the FP grow algorithm. Pay special attention to any efficiency gain obtained. Do not use examples.

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

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

a. Give an example of a constraint that is convertible monotone but not monotone. Include an explanation.

b. Give an example of a constraint that is convertible antimonotone but not antimonotone. Include an explanation.

c. Give an example of a constraint that is neither convertible monotone nor convertible antimonotone. Include an explanation.

d. Define the lift of an association rule.