Information page for written examinations at Linköping University

Examination date	2013-08-26
Room (1) If the exam is given in different rooms you have to attach an information paper for each room and mark intended place	TER2
Time	8-12
Course code	TDDD41 + 732A31
Exam code	TEN1
Course name **Exam name**	Data Mining - Clustering and Association Analysis Skriftlig tentamen
Department	IDA
Number of questions in the examination	8
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	For pass you need half of the points.
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
August 26, 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. Describe the Apriori algorithm.

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

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

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

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

a. Describe the FP grow algorithm.

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

c. Describe 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 convertible monotone and antimonotone. If you think it is not possible, explain why.

b. Consider the constraint 3<sum(S)<10 where S is an itemset and sum(S) returns the total price of the itemset S. Assume that all the prices are positive. Is this constraint monotone, antimonotone, or neither ?

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 (1p + 3p = 4p)

Given the data set {0, 2, 3, 8}. Assume we use Euclidean distance. Assume we are using PAM with k=2 and currently have the cluster centers 2 and 3.

a. Give for each of the current clusters which data objects belong to the cluster.
b. Show the next iteration in the PAM algorithm. Give all steps in the calculation. Will there be a change in cluster centers? If so, give for each of the new clusters which data objects belong to the cluster.

5. Hierarchical clustering (4p)

Describe the principles and ideas regarding BIRCH.
- Give a sketch of the algorithm.
- Explain Clustering Feature Vector. Given a cluster with the data points (1,1), (1,2), (2,1) and (2,2), what is its clustering feature vector?
- Explain what a CF-tree is and how it is used in BIRCH.
- What parameters are used as input?

6. Clustering categorical data (4p)

Describe the principles and ideas regarding the ROCK algorithm.
Within your description, make sure to give a sketch of the algorithm and to define neighbor, common neighbor, link for objects, link for clusters, and G (goodness measure).

7. Data mining system (2p)

Give an architecture for a typical data mining system. Give the functionality for each of the components.

8. Distance measure (2p)

What is the distance between Item K and Item L?

	A	B	C	D	E	F	G
Item K	(50,1)	(2,3)	Y	N	N	N	7
Item L	(50,2)	(1,4)	Y	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.