Försättsblad till skriftlig tentamen vid Linköpings universitet

(fylls i av ansvarig)

Datum för tentamen	2013-06-04, kl 8-12
Sal	
Tid	T1 TER2
Kurskod	TDDD41 / 732A31
Provkod	TEN
Kursnamn/benämning	Data Mining - Clustering and Association Analysis
Institution	*IDA*
Antal uppgifter som ingår i tentamen	8
Antal sidor på tentamen (inkl. försättsbladet)	4 + försättsblad
Jour/Kursansvarig	Patrick Lambrix / Jose M Pena
Telefon under skrivtid	2605 / 1651
Besöker salen ca kl.	9:30; 11:00
Kursadministratör (namn + tfnnr + mailadress)	Carita Lilja, 1463, carita.lilja@liu.se
Tillåtna hjälpmedel	dictionary
Övrigt (exempel när resultat kan ses på webben, betygsgränser, visning, övriga salar tentan går i m.m.)	
Vilken typ av papper ska användas, rutigt eller linjerat	
Antal exemplar i påsen	18 + 6

Institutionen för datavetenskap
Linköpings universitet

EXAM
732A31and TDDD41
Data Mining –
Clustering and Association Analysis
June 4, 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 one transaction. Explain step by step the execution.

Transaction id	Items
1	A, B, C
2	A, B, D

b. Repeat the exercise 1a with the following additional constraint: Find the frequent itemsets that contain the itemset AB. 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 itemset ABC. 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. 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. Explain how you incorporate a monotone constraint into the FP grow algorithm.

c. Explain 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 sum(S)+min(S)>3 where S is an itemset, sum(S) returns the total price of the itemset S, and min(S) returns the price of the cheapest item in S. Note that some prices may be negative. 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 (3p)

Describe for each of the different kinds of clustering methods (i) partitioning approach, (ii) hierarchical approach, and (iii) density-based approach:

- Main ideas
- Possible input parameters
- Possible output (e.g. properties of the clusters)

5. Clustering by Partitioning (2p+1p=3p)

a. Describe the principles and ideas regarding CLARANS.
   (i) Give a sketch of the algorithm.
   (ii) Define swapping cost.

b. Compare CLARANS with PAM.

6. Hierarchical clustering (3p)

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

7. 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).

8. Distance measure (1p+1p+1p = 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.