Försättsblad till skriftlig tentamen vid Linköpings Universitet

(fylls i av ansvarig)

Datum för tentamen	*2009-06-10*
Sal	*TER2*
Tid	*14-18*
Kurskod	*TDDD41*
Provkod	*TEN1*
Kursnamn/benämning	*Data Mining - Clustering and Association Analysis*
Institution	*IDA*
Antal uppgifter som ingår i tentamen	*7*
Antal sidor på tentamen (inkl. försättsbladet)	*4*
Jour/Kursansvarig	*Jose M. Peña*
Telefon under skrivtid	*0708229596*
Besöker salen ca kl.	*15 och 17*
Kursadministratör **(namn + tfnnr + mailadress)**	*Elisabeth Qvarnström* *013-281706, eliqv@ida.liu.se*
Tillåtna hjälpmedel	*Ordbok*
Övrigt **(exempel när resultat kan ses på webben, betygsgränser, visning, övriga salar tentan går i m.m.)**	*Requirement for grade C: 15 points. Requirement for grade A: 23 points.*
Vilken typ av papper ska användas, rutigt eller linjerat	*Fri val*
Antal exemplar i påsen	4+2

Institutionen för datavetenskap
Linköpings universitet

EXAM
732A02 and TDDD41
Data Mining –
Clustering and Association Analysis
June 10, 2009, 2-6pm

Teachers: Patrick Lambrix, José M Pena

Grades: Requirement for grade C: 15 points. Requirement for grade A: 23 points.

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 (3p+2p+2p+1p=8p)

a. Run the Aprori algorithm on the following transaction 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
3	X, Y, Z
4	X, Y, W
5	Z, W

b. Run the Apriori algorithm on the transaction database above to find all the itemsets with support equal to one transaction (i.e. support must be exactly one, not greater or equal to one). Tips: Rephrase this requirement as a (anti)monotonic constraint. Explain step by step the execution. Make clear when the constraint is checked and what the consequences of the checking are, if any. Do not simply run the algorithm and afterwards consider the constraint but incorporate the constraint into the algorithm.

c. Run the Apriori algorithm on the transaction database above to find all the itemsets whose items have all equal support (i.e. the itemset RST is in the output if and only if R, S and T have all the same support). Tips: Rephrase this requirement as a (anti)monotonic constraint. Explain step by step the execution. Make clear when the constraint is checked and what the consequences of the checking are, if any. Do not simply run the algorithm and afterwards consider the constraint but incorporate the constraint into the algorithm.

d. Sketch a proof of the correctness of the Apriori algorithm. It suffices to sketch a proof showing that all the frequent itemsets of size k are among the candidates of size k.

2. FP algorithm (3p+1p=4p)

a. Run the FP algorithm on the transaction database in exercise 1 with minimum support equal to one transaction and the constraint that the sum of the prices of the items in an itemset must be strictly smaller than 10. Make clear when the constraint is checked and what the consequences of the checking are, if any. Do not simply run the algorithm and afterwards consider the constraint but incorporate the constraint into the algorithm.

Item	Price
A	10
B	10
C	10
D	10
X	1
Y	1
Z	1
W	1

b. Sketch a proof of the correctness of the FP grow algorithm.

3. Constraints (1p+1p=2p)

a. Give an example of a convertible monotone constraint that is not monotone.
b. Give an example of a convertible antimonotone constraint that is not antimonotone.

4. Clustering by Partitioning (4p+1p=5p)

a. Describe the principles and ideas regarding CLARANS. Explain the different steps of the algorithm. Explain in particular the notion of 'cost' and give an example.

b. Compare CLARANS with PAM.

5. Hierarchical clustering (4p)

Describe the principles and ideas regarding Agglomorative Hierarchical Clustering.
Show the different steps of the algorithm using the dissimilarity matrix below
and complete link clustering. Give partial results after each step.

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

6. Density-based clustering (3p)

Describe the principles and ideas regarding the DBSCAN algorithm. Explain the major steps. Define the notions of core point, border point, density-reachable and density-connected.

7. Potpourri (2p+2p=4p)

a. Explain the difference between symmetric and asymmetric binary variables. How is this difference used in the definition of distance measures.

b. Explain the difference between a centroid and a medioid. What is the advantage of using medioids instead of centroids in the partioning approaches for clustering?