Försättsblad till skriftlig tentamen vid Linköpings universitet

(fylls i av ansvarig)

Datum för tentamen	08/06/2011
Sal	U3, U1
Tid	14-18
Kurskod	732A31, TDDD41 (732A02)
Provkod	TEN
Kursnamn/benämning	Data Mining - Clustering and Association Analysis
Institution	*IDA*
Antal uppgifter som ingår i tentamen	7
Antal sidor på tentamen (inkl. försättsbladet)	4 + cover page
Jour/Kursansvarig	Patrick Lambrix, Jose M Pena
Telefon under skrivtid	2605, 1651
Besöker salen ca kl.	15.15, 16.45
Kursadministratör (namn + tfnnr + mailadress)	
Tillåtna hjälpmedel	dictionary
Övrigt (exempel när resultat kan ses på webben, betygsgränser, visning, övriga salar tentan går i m.m.)	For pass you need 15 points.
Vilken typ av papper ska användas, rutigt eller linjerat	
Antal exemplar i påsen	

Institutionen för datavetenskap
Linköpings universitet

EXAM
732A31, 732A02 and TDDD41
Data Mining –
Clustering and Association Analysis
June 8, 2011, kl 14-18

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+1p+1p=6p)

a. Run the Aprori algorithm on the following transactional database with minimum support equal to two transactions. Explain step by step the execution.

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

b. Repeat the exercise 1a with the following additional constraint: Find the frequent itemsets that contain at least two items. 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 item E. 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.

e. Run the Simple Algorithm on the transactional database below to produce association rules for the itemset XYZ. Use a minimum confidence threshold equal to 100 %. Explain step by step the execution.

Transaction id	Items
1	X, Y, Z
2	X, Y
3	Y, Z

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

a. Run the FP algorithm on the transactional database in exercise 1a with minimum support equal to two transactions. Explain step by step the execution.

b. How do you incorporate a monotone constraint in the FP grow algorithm ?

c. How do you incorporate an antimonotone constraint in the FP grow algorithm ?

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

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

a. Let C1 and C2 be two monotone constraints. Let us define a new constraint C3 so that C3 holds for an itemset X if and only if both C1 and C2 hold for X. Let us define a new constraint C4 so that C4 holds for an itemset X if and only if C1 or C2 hold for X. Are C3 and C4 monotone, antimonotone, both, none, or we simply cannot know ? Explain your answer.

b. Let C1 be a convertible monotone constraint and C2 a convertible antimonotone constraint. Let us define a new constraint C3 so that C3 holds for an itemset X if and only if both C1 and C2 hold for X. Let us define a new constraint C4 so that C4 holds for an itemset X if and only if C1 or C2 hold for X. Are C3 and C4 convertible monotone, convertible antimonotone, both, none, or we simply cannot know ? Explain your answer.

c. Give an example of a constraint that is convertible monotone but not monotone. Explain your answer.

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

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

b. Why is PAM more robust than K-means in the presence of outliers?

c. 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.

5. Hierarchical clustering (4p+2p=6p)

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

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

6. Density-based clustering (3p)

Describe the principles and ideas regarding the DBSCAN algorithm.
- Give a sketch of the algorithm.
- Define core point, direct density-reachable, density-reachable, density-connected.
- What parameters are used as input?

7. Data mining concepts (1p+1p=2p)

a. The purpose of data mining is to extract interesting patterns from a huge amount of data. When is a pattern 'interesting' mean in this case?
b. Data in the real world can be dirty. Give 3 reasons and an example for each.