Försättsblad till skriftlig tentamen vid Linköpings Universitet

Datum för tentamen	2012-03-07
Sal (1) Om tentan går i flera salar ska du bifoga ett försättsblad till varje sal och ringa in vilken sal som avses	TER2
Tid	8-12
Kurskod	TDDD41
Provkod	TEN1
Kursnamn/benämning **Provnamn/benämning**	Data Mining - Clustering and Association Analysis Skriftlig tentamen
Institution	IDA
Antal uppgifter som ingår i tentamen	8
Jour/Kursansvarig Ange vem som besöker salen	Patrick Lambrix
Telefon under skrivtiden	2605
Besöker salen ca kl.	9^{30}, 11^{15}
Kursadministratör/kontaktperson (namn + tfnr + mailaddress)	carita.lilja@liu.se
Tillåtna hjälpmedel	
Övrigt	PASS GRADE: 15 points
Vilken typ av papper ska användas, rutigt eller linjerat	
Antal exemplar i påsen	

Institutionen för datavetenskap
Linköpings universitet

EXAM
732A31and TDDD41
Data Mining –
Clustering and Association Analysis
March 7, 2012, 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
3	A, B, E
4	A, C, D
5	A, C, E

b. Repeat the exercise 1a with the following additional constraint: Find the frequent itemsets that contain the item A. 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 A. 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. Repeat the exercise 1a with the FP grow algorithm. Explain step by step the execution.

b. Repeat the exercise 1b with the FP grow algorithm. Explain step by step the execution.

c. Repeat the exercise 1c with the FP grow algorithm. Explain step by step the execution.

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

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

a. Let D1 denote the following definition of convertible monotone constraint. A constraint is convertible monotone if there exists an ordering R1 of the items such that when an itemset violates the constraint so does any **suffix** of the itemset. Let D2 denote the following definition of convertible monotone constraint. A constraint is convertible monotone if there exists an ordering R2 of the items such that when an itemset violates the constraint so does any **prefix** of the itemset. Are D1 and D2 equivalent, i.e. does any constraint either satisfy both of them or none ? Explain your answer.

b. Assume that a constraint satisfies both D1 and D2. Are R1 and R2 the same ?

c. Give an example of a constraint that is both convertible monotone and convertible antimonotone (i.e. strongly convertible). Explain your answer.

4. Clustering by Partitioning (3p)

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.

(v) For a given node, which of CLARANS and PAM visits the most neighbors? Explain

(vi) Which of PAM, CLARA and CLARANS guarantees to find an optimal solution? (Obs: not necessarily 1 answer.)

5. Hierarchical clustering (4p)

Describe the principles and ideas regarding BIRCH.
- Give a sketch of the algorithm.
- Explain Cluster Feature Vector. Given a cluster with the data points (1,2), (1,3) and (2,2), what is its clusture feature vector?
- Explain what a CF-tree is and how it is used in BIRCH.
- What parameters are used as input?

6. Density-based clustering (5p)

Describe the principles and ideas regarding the OPTICS algorithm.
- What is the main purpose of the algorithm?
- What is the relation to DBSCAN?
- Give a sketch of the algorithm.
- Define core distance and reachability distance.
- What does the visualization of the output show?

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	(0,0)	(1,1)	Y	N	Y	N	(1,1,1)
Item L	(5,0)	(3,2)	Y	N	N	N	no-value-available

Attributes A and G are 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.