Försättsblad till skriftlig tentamen vid Linköpings Universitet

Datum för tentamen	2011-03-16
Sal (2) Om tentan går i flera salar ska du bifoga ett försättsblad till varje sal och ringa in vilken sal som avses	TER2 (TER3)
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	7
Jour/Kursansvarig Ange vem som besöker salen	Patrick Lambrix, José Pena
Telefon under skrivtiden	2605, 1651
Besöker salen ca kl.	09.15, 10.45
Kursadministratör/kontaktperson (namn + tfnr + mailaddress)	carita.lilja@liu.se 1463
Tillåtna hjälpmedel	Dictionary
Övrigt	
Vilken typ av papper ska användas, rutigt eller linjerat	Rutigt
Antal exemplar i påsen	30

Institutionen för datavetenskap
Linköpings universitet

EXAM
732A31, 732A02 and TDDD41
Data Mining –
Clustering and Association Analysis
March 16, 2011, 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+1p+1p=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 items A and B (both!). 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 or the item B. 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 1 transaction. 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. Given a transactional database, we want to find all the itemsets whose support is exactly 10. Specify the minimun support and constraints (if any) that are needed to run the Apriori or FP grow algorithm and find the desired itemsets.

b. Let C1 be a monotone constraint. Let us define another constraint C2 so that C2 holds for an itemset X if and only if C1 does not hold for X. Is C2 monotone, antimonotone, both, none, or we simply cannot know ? Explain your answer.

c. Give an example of a constraint that is both convertible monotone and convertible antimonotone. Explain your answer.

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

a. Describe the principles and ideas regarding PAM.
- Give a sketch of the algorithm.
- Define swapping cost.
- Draw an example where the swapping cost is 0 and one where the swapping cost is strictly negative.

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

c. Why is CLARANS more efficient than PAM?

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

a. Given the following data objects K, L, M and N with binary asymmetric variables A, B, C, D, E and F.
- Using the Jaccard coefficient similarity, give the similarities between each pair of objects.
- Given threshold 0.5, what are the common neighbors of K and M?

	A	B	C	D	E	F
Item K	Y	Y	Y	Y	N	N
Item L	Y	Y	Y	N	Y	N
Item M	Y	N	Y	Y	Y	N
Item N	Y	Y	Y	N	N	Y

b. Regarding ROCK
- What is Link defined between two clusters?
- Give and explain the goodness measure in ROCK. Also explain how it is used.

6. Density-based clustering (5p+2p=7p)

a. 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?

b. Regarding DENCLUE:
- Explain influence function and significant density attractor.
- How can you obtain arbitrary-shape clusters?