Försättsblad till skriftlig tentamen vid Linköpings universitet

(fylls i av ansvarig)

Datum för tentamen	2010-03-12
Sal	TER2
Tid	14-18
Kurskod	TDDD41 / 732A02
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)	5
Jour/Kursansvarig	Patrick Lambrix / Jose M Pena
Telefon under skrivtid	2605
Besöker salen ca kl.	15, 16.30
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 grade, half of the total points is required.
Vilken typ av papper ska användas, rutigt eller linjerat	
Antal exemplar i påsen	

Institutionen för datavetenskap
Linköpings universitet

EXAM
732A02 and TDDD41
Data Mining –
Clustering and Association Analysis
March 12, 2010, 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 transaction database with minimum support equal to 1 transaction. Explain step by step the execution.

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

b. In the exercise a, how many self-join operations did you perform ? And, how many times did you count the support of an itemset ? Explain your answers.

c. Repeat the exercises a and b with the following constraint: The number of items in an itemset must be equal or smaller than 2. Explain step by step the execution. Make clear when the constraint is used. Incorporate the constraint into the algorithm, i.e. do not simply run the algorithm and afterwards consider the constraint.

d. Repeat the exercise c with the following additional constraint: The itemset must contain either the item A or the item B or the item C. Note that the constraint from exercise c is also to be enforced, i.e. you have now two constraints. Explain step by step the execution. Make clear when the constraint is used. Incorporate the constraint into the algorithm, i.e. do not simply run the algorithm and afterwards consider the constraint.

e. Sketch a proof of the correctness of the Apriori algorithm.

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

a. Run the FP algorithm on the following transaction database with minimum support equal to 1 transaction. Explain step by step the execution.

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

b. Repeat the exercise a with the following constraint: The number of items in an itemset must be equal or smaller than 3. Explain step by step the execution. Make clear when the constraint is used. Incorporate the constraint into the algorithm, i.e. do not simply run the algorithm and afterwards consider the constraint.

c. Sketch a proof of the correctness of the FP 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. Consider the following contraint: The most expensive item in an itemset cannot be more expensive than the two cheapest items in the itemset together. Is the constraint convertible monotone, convertible antimonotone, both or none ? 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+2p=5p)

a. Describe the principles and ideas regarding K-Means. Explain the different steps of the algorithm. What are the weaknesses of K-means?

b. 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)

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 (4p+1p=5p)

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

b. What is the main idea behind OPTICS?

7. Distance measure (2p)

What is the distance between Item K and Item L?

	A	B	C	D	E	F	G
Item K	20	(0,0)	Y	N	Y	N	5
Item L	30	(1,1)	Y	N	Y	N	no-value-available

Attribute A is 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.
Attribute G is interval-based.