Försättsblad till skriftlig tentamen vid Linköpings universitet

	(fylls i av ansvarig)
Datum för tentamen	23-08-2010
Sal	TER3
Tid	8-12
Kurskod	TDDD41 och 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)	5
Jour/Kursansvarig	Patrick Lambrix / Jose Pena
Telefon under skrivtid	2605
Besöker salen ca kl.	9.15, 11.00
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 half of the points.
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
August 23, 2010, 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. Assume five items A, B, C, D and E. Assume two transactions. Assume that the items A, B and C are the only items in the first transaction, whereas the items A, B and D are the only items in the second transaction. Find all the itemsets with positive support by running the Apriori algorithm. Explain step by step the execution.

b. In the exercise a, how many candidate itemsets did you produce ? And, for how many of these candidate itemsets did you count support ? Explain your answers.

c. Assume the same items and transactions as in exercise a. Assume that the price of the items A, B, C, D and E is 1, 2, 3, 4 and 5, respectively. Run the Apriori algorithm to find all the itemsets with positive support and whose total price is larger than 5. Explain step by step the execution. Identify what type of constraint we are dealing with. 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 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	B
2	B
3	A, B
4	B, C
5	A, C

b. Assume that the price of each item is 1. Repeat the exercise a with the following constraint: The total price of an itemset must be smaller than 10. 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. Checking that an itemset satisfies the minimum support threshold is a monotone constraint. True or false ? Explain your answer.

b. Assume that we want to keep the running time of the Apriori algorithm as low as possible. Assume that we are dealing with a monotone constraint. Then, for any candidate itemset, what is best to check first the constraint or the minimum support threshold ? And what if the contraint is antimonotone ? Explain your answers.

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

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

a. Describe the principles and ideas regarding CLARANS.
   (i) Give a sketch of the algorithm..
   (ii) Define swapping cost.

b. Compare CLARANS with PAM.

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)

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

$$
\begin{array}{c|ccccc}
 & 1 & 2 & 3 & 4 & 5 \\
\hline
1 & 0 & & & & \\
2 & 2 & 0 & & & \\
3 & 5 & 3 & 0 & & \\
4 & 10 & 8 & 9 & 0 & \\
5 & 7 & 4 & 1 & 6 & 0
\end{array}
$$

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	gold	(0,0)	Y	N	Y	N	silver
Item L	bronze	(2,2)	N	N	N	N	no-value-available

Attributes A and G are ordinal variables with values gold/silver/bronze in that order.
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.