Försättsblad till skriftlig tentamen vid Linköpings universitet

(fylls i av ansvarig)

Datum för tentamen	August 20, 2012
Sal	TER 1
Tid	8-12
Kurskod	TDDD41
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)	3 + cover page
Jour/Kursansvarig	Patrick Lambrix, Jose Pena
Telefon under skrivtid	2605, 1651
Besöker salen ca kl.	9.45, 10.45
Kursadministratör (namn + tfnnr + mailadress)	Carita Lilja, 1463, Carita.Lilja@liu.se
Tillåtna hjälpmedel	dictionary
Övrigt (exempel när resultat kan ses på webben, betygsgränser, visning, övriga salar tentan går i m.m.)	For a pass grade, you need half of the max points.
Vilken typ av papper ska användas, rutigt eller linjerat	
Antal exemplar i påsen	10

Institutionen för datavetenskap
Linköpings universitet

EXAM
732A31 and TDDD41
Data Mining –
Clustering and Association Analysis
August 20, 2012, 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. Explain the Apriori algorithm. You may want to give the algorithm's pseudocode.

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

c. Explain how and where we incorporate a monotonic constraint into the Apriori algorithm.

d. Explain how and where we incorporate an antimonotonic constraint into the Apriori algorithm.

e. What role does the Apriori algorithm play in the search for association rules ?

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

a. Explain the FP grow algorithm. You may want to give the algorithm's pseudocode.

b. Explain how and where we incorporate a monotonic constraint into the FP grow algorithm.

c. Explain how and where we incorporate an antimonotonic constraint into 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. Give three examples of constraints that are monotone. Explain your answer.

b. Give three examples of constraints that are antimonotone. Explain your answer.

c. Give three examples of constraints that are neither monotone nor antimonotone but that
are convertible monotone and convertible antimonotone (i.e. strongly convertible).
Explain your answer.

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

a. Run the k-means algorithm on the data set {0, 2, 3, 8, 9, 10} with k=2 and use as initial cluster centers 2 and 3. Show step-by-step results. Give the clusters and cluster centers in each step.
b. What are the strengths and weaknesses of k-means?
c. Which weakness of k-means is addressed by using medoids and how?

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 cluster 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 (2p+1p+2p=5p)

a. Describe the principles and ideas regarding the DBSCAN algorithm. In your description, make sure to give a sketch of the algorithm and to define core point, direct density-reachable, density-reachable, and density-connected.
b. What is the main idea behind OPTICS?
c. Describe the principles and ideas regarding the DENCLUE algorithm. In your description, make sure to define the important notions and define how clusters are formed. Also discuss whether arbitrary-shape clusters can be formed.

7. Distance measure (2p)

What is the distance between Item K and Item L?

	A	B	C	D	E	F	G
Item K	(0,10)	(4,4)	Y	N	Y	N	5
Item L	(0,30)	(5,7)	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.