Försättsblad till skriftlig tentamen vid Linköpings universitet

Datum för tentamen	2015-06-09
Sal (1)	TER1
Tid	8-12
Kurskod	732A31
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
Telefon under skrivtiden	2605
Besöker salen ca klockan	9:30 / 11:00
Kursadministratör/kontaktperson (namn + tfnr + mailaddress)	Carita Lilja, 1463, carita.lilja@liu.se
Tillåtna hjälpmedel	dictionary
Övrigt	For pass, half of the Points are needed
Antal exemplar i påsen	

strengths/weaknesses
algorithm
~~strengths~~

Institutionen för datavetenskap
Linköpings universitet

EXAM
732A31and TDDD41
Data Mining –
Clustering and Association Analysis
June 9, 2015, 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. Clustering by partitioning (2p+2p=4p)

a. Describe the principles and ideas regarding PAM.
- Give a sketch of the algorithm.
- Define swapping cost. = TCih = $\sum$ summan av alla costs

b. Describe the principles and ideas regarding CLARA. TCih = $\sum_j C_{jih}$
- Give a sketch of the algorithm.
- What are the strengths and weaknesses of CLARA?

2. Hierarchical clustering (4p)

Describe the principles and ideas regarding BIRCH.
- Give a sketch of the algorithm.
- Explain Clustering Feature Vector. Given a cluster with the data points (0,0), (1,1) and (2,2), what is its clustering feature vector?
- Explain what a CF-tree is and how it is used in BIRCH. How is a CF-tree traversed?
- What parameters are used as input?

(P) **3. Clustering categorical data (4p)**

Describe the principles and ideas regarding the ROCK algorithm. Within your description, make sure to give a sketch of the algorithm and to define **and** give examples for neighbor, common neighbor, link for objects, link for clusters, and G (goodness measure).

4. Different types of data and their distance measures (2p+1p+1p=4p)

a. What is the distance between Item K and Item L?

	A	B	C	D	E	F	G
Item K	(20000,1)	(1,2)	Y	N	Y	N	55
Item L	(20000,100)	(3,5)	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.

(P) b. In the vector model for information retrieval documents are represented by vectors with positive real numbers. How is the similarity between two vectors defined? visa formeln i slide "Vector objects" 23

(P) c. Show how an interval-based measure can be defined for ordinal variables. Slide 20 normalisering

5. FP grow algoritm (2p+2p+2p=6p)

a. Run the FP grow 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	X, Y, Z
3	A, Y, C
4	X, B, Z

b. Repeat the exercise 5a with the following additional constraint: Find the frequent itemsets that do not contain the item Z. 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. Let the items A, B, C, X, Y and Z have a price of respectively -3, -2, -1, 1, 2 and 3 units. Repeat the exercise 5a with the following additional constraint: Find the frequent itemsets whose most expensive item has positive price. 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.

6. Apriori algorithm (2p+1p+1p+1p=5p)

a. Describe the Apriori algorithm, i.e. describe how it works in general not in a particular example.

b. Describe how to incorporate a monotone constraint into the Apriori algorithm.

c. Describe how to incorporate an antimonotone constraint into the Apriori algorithm.

d. Give an example where the Apriori algorithm fails, i.e. there is a frequent itemset in your example that is not discovered by the Apriori algorithm.

7. Constraints and lift (1p+1p+1p=3p)

Are the following statements true or false ?

a. If the itemsets AB and BC are frequent, then the itemset AC is frequent too.
b. Every constraint is monotone, antimonotone, convertible monotone or convertible antimonotone.
c. Adding items to the antecedent of an association rule increases the lift of the rule.