Institutionen för datavetenskap
Linköpings universitet

EXAM

732A61and TDDD41
Data Mining –
Clustering and Association Analysis

732A75 Advanced Data Mining

June 5, 2018, kl 8-12

Teachers: Patrick Lambrix, José M Pena

Instructions:

1. Start each question at a new page.
2. Write at one side of a page.
3. Write clearly.
4. 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 (3p+1p+1p=5p)

a. Describe the principles and ideas regarding PAM.
   i. Describe the algorithm.
   ii. Define swapping cost.
   iii. Draw an example of a data set in two dimensions where the swapping cost TCih is 0 and one where the swapping cost TCih is strictly negative.

b. Which, if any, of the algorithms PAM/CLARA/CLARANS guarantees a **local** optimum for the cost function? (For your answer, only write 'none' or the name(s) of the algorithms.)

c. Which, if any, of the algorithms PAM/CLARA/CLARANS guarantees a **global** optimum for the cost function? (For your answer, only write 'none' or the name(s) of the algorithms.)

2. Hierarchical clustering (3p)

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

	1	2	3	4	5
1	0				
2	8	0			
3	3	4	0		
4	1	7	9	0	
5	10	2	6	5	0

3. Clustering categorical data (4p)

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

4. Density-based clustering (1p+1p=2p)

a. DBSCAN: Consider the following statement: if p is density-connected to q wrt Eps and Minpts then p is density-reachable from q wrt Eps and Minpts. Is this statement true? If yes, then prove. If no, then give a counterexample.
b. What is the main idea behind OPTICS?

5. 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	(1,1)	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.

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

a. What is the apriori property ?

b. How do you produce candidates in the apriori algorithm ?

c. How do you incorporate an antimonotone constraint in the apriori algorithm ?

d. Prove formally the correctness of the apriori algorithm.

7. FP grow algorithm (5p)

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	C, B, A
2	D, C, A
3	A, B
4	A, B
5	A, D
6	A, D

8. Constraints, Rules and Causality (2p+2p+1p=5p)

a. Give an example of a constraint that is monotone, another that is antimonotone, another that is convertible monotone but not monotone, and another that is convertible antimonotone but not antimonotone.

b. Apply the association rule generation algorithm to the frequent itemset ABC on the database in the exercise above in order to produce association rules with confidence greater or equal than 50 %.

c. Give an example of when an association rule is a causal rule. You may want to specify the (in)dependencies among the random variables involved, as well as any assumption you make.