Information page for written examinations at Linköping University

Examination date	2019-03-23
Room (3)	G32(35) G33(1) G34(19)
Time	8-12
Edu. code	TDDD41
Module	TEN1
Edu. code name **Module name**	Data Mining - Clustering and Association Analysis (Data Mining - Clustering and Association Analysis) Written examination (Skriftlig tentamen)
Department	IDA
Number of questions in the examination	8
Teacher responsible/contact person during the exam time	Patrick Lambrix
Contact number during the exam time	2605 (question 1-5) / 1651 (question 6-8)
Visit to the examination room approximately	10:00
Name and contact details to the course administrator (name + phone nr + mail)	Veronica Kindeland Gunnarsson, 5634 veronica.kindeland.gunnarsson@liu.se
Equipment permitted	dictionary
Other important information	
Number of exams in the bag	

Institutionen för datavetenskap
Linköpings universitet

EXAM

732A61 and TDDD41 Data Mining - Clustering and Association Analysis

732A75 Advanced Data Mining

March 23, 2019, 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. Data mining process (2p)

Describe the knowledge discovery process. Give the different steps and explain what they do.

2. Clustering by partitioning (2+3=5p)

a. Given the graph representation of the clustering problem where n is the number of data objects and k is the number of clusters.

i. What does a node represent?

ii. When are two nodes neighbors and how many neighbors does a node have?

iii. Which of PAM, CLARA and CLARANS guarantees to find a local optimum?

iv. Which of PAM, CLARA and CLARANS guarantees to find a global optimum?

b. Given the data set {0, 3, 4, 10}. Assume we use Euclidean distance and k = 2. Draw the graph representation of the clustering problem. Then start at an arbitrary node and show one iteration of the PAM algorithm on the graph. Give all steps in the computation and show at what node that iteration ends

3. Hierarchical clustering (1+3=4p)

a. For the ROCK algorithm:

Given the similarity matrix below. What is link(A,B) if the threshold is 0.6?

	A	B	C	D	E
A	1				
B	0.9	1			
C	0.8	0.7	1		
D	0.1	0.2	0.5	1	
E	0.2	0	0.3	0.4	1

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

4. Density-based clustering (3p)

Describe the principles and ideas regarding the CHAMELEON algorithm. In your description, make sure to describe the algorithm and to define the important notions in the algorithm.

5. Different types of data and their distance measures (2p)

What is the distance between Item K and Item L? (no normalization needed)

	A	B	C	D	E	F	G
Item K	(10,50)	(3,1,1)	N	N	Y	N	9
Item L	(10,55)	(2,3,1)	N	Y	N	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.

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

a. Run the Apriori algorithm on the following transactional database with minimum support equal to two transactions. Explain step by step the execution.

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

b. Repeat the previous exercise with the following additional constraint: Find the frequent itemsets that contain the item D. Explain step by step the execution. 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. Run the Apriori algorithm on the following transactional database with minimum support equal to two transactions. Explain step by step the execution.

Transaction id	Items
1	A, B, E
2	A, B, F
3	A, C, E
4	A, C, F
5	B, C, E
6	B, C, F
7	B, D, E
8	B, D, F
9	A, B, E

d. Apply the rule generation algorithm to the frequent itemset ABE on the previous database to produce association rules with confidence greater or equal than 60 %.

7. FP grow algorithm (2p+2p=4p)

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, D
2	B, C, D, E

b. Give an example that illustrates what you think is the main advantage of the FP grow algorithm over the Apriori algorithm.

8. Miscellaneous (1p+1p+1p+1p+1p=5p)

a. Association rules represent causal relationships. True or false ?
b. The Apriori algorithm produces candidate frequent itemsets, whereas the FP grow algorithms does not do it. True or false ?
c. The FP grow algorithm does not produce candidate frequent itemsets, which may make it miss some frequent itemsets. This however pays off because it runs faster than the Apriori algorithm. True or false ?
d. The only constraints that can be both monotone and antimonotone are the constraint that is true for all itemsets, and the constraint that is false for all itemsets. True or false ?
e. If a constraint is convertible antimonotone for some ordering of the items, then it is convertible antimonotone for every ordering.