Information page for written examinations at Linköping University

Examination date	2019-08-27
Room (1)	TER4(7)
Time	8-12
Edu. code	732A75
Module	TEN1
Edu. code name **Module name**	Advanced Data Mining (Avancerad data mining) Written examination (Skriftlig tentamen)
Department	IDA
Number of questions in the examination	8
Teacher responsible/contact person during the exam time	Patrick Lambrix (2605) / Jose M Pena (1651)
Contact number during the exam time	2605 (Q1-4): 1651 (Q5-8)
Visit to the examination room approximately	10:15
Name and contact details to the course administrator (name + phone nr + mail)	Annelie Almquist, annelie.almquist@liu.se, tel 2934
Equipment permitted	dictionary
Other important information	
Number of exams in the bag	

Institutionen för datavetenskap
Linköpings universitet

EXAM

732A61and TDDD41
Data Mining –
Clustering and Association Analysis

732A75 Advanced Data Mining

August 27, 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. Clustering by partitioning (2+1+2=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. Considering PAM, CLARA and CLARANS, the graph for which algorithm/algorithms contains/contain the most nodes? Explain.

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

b. Given the data set {0, 7, 8, 10}. Assume we use Euclidean distance and k = 2. Draw the graph representation of the clustering problem.

c. 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.

2. Hierarchical clustering (3p)

Describe the principles and ideas regarding Agglomerative Hierarchical Clustering. Show the different steps of the algorithm using the dissimilarity matrix below and *single* 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

3. Density-based clustering (2+1=3p)

a. Describe the principles and ideas regarding the DBSCAN algorithm. In your description, make sure to describe the algorithm and to define core point, direct density-reachable, density-reachable, and density-connected.

b. What is the relationship between DBSCAN and OPTICS?

4. Different types of data and their distance measures (1+2+2=5p)

a. Give and explain the distance measure for objects with variables of mixed types.

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

	A	B	C	D	E	F	G
Item K	(50,500)	(2,1,0)	Y	N	Y	N	77
Item L	(50,505)	(1,3,0)	Y	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.

c. Asymmetric binary variables.

i. Give and explain the distance measure for objects with asymmetric binary variables using contingency tables.

ii. Can the formula in question a also be used for objects with only asymmetric variables? If no, explain why. If yes, state whether you would get the same results as with the method in question c.i and explain why or why not.

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

a. Run the Apriori 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	A, C, D
3	A, B
4	A, B
5	A, D
6	A, D

b. Repeat the exercise above with the following additional constraint: Find the frequent itemsets that do not contain the itemset CD. 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. Apply the rule generation algorithm to the frequent itemset ABC on the database above in order to produce association rules with confidence greater or equal than 50 %.

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

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

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

b. Explain how you incorporate monotone and antimonotone constraints in the FP grow algorithm.

c. What is the main advantage of the FP grow algorithm over the Apriori algorithm ? Give an example that illustrates your answer.

7. Constraints and Causality (2p+1p=3p)

a. Give an example of convertible monotone and convertible antimonotone constraints that are not monotone and antimonotone, respectively. Show that your constraints are really

convertible monotone and antimonotone. Do not give examples to show it, i.e. provide an abstract and formal argument.

b. Give an example of an association rule that is not a causal rule. Include an explanation of your reasoning.

8. Miscellaneous (4*0.5p=2p)

a. The Apriori algorithm produces candidate frequent itemsets, whereas the FP grow algorithms does not do it. True or false ?
b. 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 ?
c. 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 ?
d. If a constraint is convertible antimonotone for some ordering of the items, then it is convertible antimonotone for every ordering. True or false ?