Information page for written examinations at Linköping University

Examination date	2016-03-18
Room (1)	TER1
Time	8-12
Course code	732A31
Exam code	TEN1
Course name **Exam 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	7
Teacher responsible/contact person during the exam time	Patrick Lambrix / Jose Pena
Contact number during the exam time	2605 / 1651
Visit to the examination room approximately	9:30 / 10:45
Name and contact details to the course administrator (name + phone nr + mail)	Åsa Kärrman, 5760, asa.karrman@liu.se
Equipment permitted	dictionary
Other important information	
Number of exams in the bag	

Institutionen för datavetenskap
Linköpings universitet

EXAM

732A31and TDDD41
Data Mining –
Clustering and Association Analysis
March 18, 2016, 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 (3p+1p=4p)

a. 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.

b. Why is PAM more robust than K-means in the presence of outliers?

2. Hierarchical clustering (3+2=5p)

a. Describe the principles and ideas regarding BIRCH by answering the following:
- 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?

b. For the ROCK algorithm:
- What is Link defined between two clusters?
- Give and explain the goodness measure in ROCK. Also explain how it is used.

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

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 relationship between DBSCAN and OPTICS?

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

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

	A	B	C	D	E	F	G
Item K	(40,50)	(2,1)	Y	N	Y	N	8
Item L	(40,55)	(1,5)	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.

b. Assume we have categorical data. One method to define a distance between two data objects is (p-m)/p where p is the total number of categorical variables and m is the number of categorical variables for which there is a match between the objects. A second method is to introduce a new asymmetric binary variable for each of the possible values for each of the categorical variables. Give a formula for the distance between two objects in the second method in terms of p and m (where p and m have the same meaning as above; i.e. p is the number of categorical variables - *not* the number of introduced binary variables, and m is the number of matches in the categorical variables).

5. Apriori algoritm (2p+1p+1p+1p=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	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 contain the item A. 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 range is smaller than 3 (recall that the range is the price of the most expensive item minus the price of the cheapest item). 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.

d. Sketch a proof of the correctness of the Apriori algorithm. -2 +1+ 3

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

a. Describe the FP growth algorithm. Do not use examples.

b. Explain how you incorporate a monotone constraint into the FP growth algorithm.

c. Explan how you incorporate an antimonotone constraint into the FP growth algorithm.

d. What is the main advantage of the FP growth algorithm over the Apriori algorithm ?

7. Constraints and lift (2p+1p+1p=4p)

a. Give the definitions of monotone, antimonotone, and convertible monotone and antimonotone constraints.

b. Apply the Simple algorithm to the frequent itemset XBZ on the database in exercise 5 in order to find association rules with confidence greater than 50 %.

c. Give an example of an association rule with lift greater than one and another example of a rule with lift smaller than one.