Institutionen för datavetenskap
Linköpings universitet

EXAM

732A61and TDDD41
Data Mining –
Clustering and Association Analysis

732A75 Advanced Data Mining

March 17, 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 (2p+3p=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, 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.

2. Hierarchical clustering (3+3=6p)

a. Describe the principles and ideas regarding BIRCH by answering the following:
   i. Give a sketch of the algorithm.
   ii. Explain Cluster Feature Vector. Given a cluster with the data points (1,2), (1,3) and (2,2), what is its cluster feature vector?
   iii. Explain what a CF-tree is and how it is used in BIRCH.
   iv. What parameters are used as input?

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

3. Density-based clustering (2p)

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.

4. Different types of data and their distance measures (2+1=3p)

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

	A	B	C	D	E	F	G
Item K	(10,500)	(2,1,1)	Y	N	Y	N	8
Item L	(10,505)	(1,3,1)	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.

b. Show how an interval-based measure can be defined for ordinal variables.

5. Apriori algoritm (2p+2p+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	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 %.

6. FP grow algorithm (2p+2p+1p=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 ?

7. Constraints and Causality (3p+2p=5p)

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 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.