Institutionen för datavetenskap
Linköpings universitet

EXAM

732A61and TDDD41
Data Mining –
Clustering and Association Analysis

732A75 Advanced Data Mining

August 28, 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 (1+2+2=5p)

Given the data set {0, 3, 4, 10}. Assume we use Euclidean distance and k = 2.

a. Draw the graph representation of the clustering problem.

b. 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.

c. Assume numlocal = 1 and maxneighbor = 2. Start at the same node as in question *b* and show one iteration of the CLARANS algorithm on the graph. Give all steps in the computation and show at what node that iteration ends.

2. 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

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 (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	(100,500)	(2,1,1)	Y	N	Y	N	8
Item L	(100,505)	(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.

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 (3p+2p+2p=7p)

a. Run the Apriori algorithm on a transactional database of your own choice. Choose the database so that you can show the main steps of the algorithm, namely the candidate generation step, the pruning due to minimum support, and the pruning due to subset checking.

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

Transaction id	Items
1	A, B, C
2	A, C, D
3	A, B
4	A, B
5	A, D
6	A, D

c. Sketch a proof of correctness for the Apriori algorithm.

6. FP grow algorithm (3p+1p=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	C, B, A
2	D, C, A
3	A, B
4	A, B
5	A, D
6	A, D

b. What is the main advantage of the FP grow algorithm over the Apriori algorithm ?

7. Constraints and Causality (3p+1p=4p)

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.