Institutionen för datavetenskap
Linköpings universitet

EXAM

TDDD41 Data Mining – Clustering and Association Analysis

732A75 Advanced Data Mining

March 18, 2023, kl 8-12

Teachers: Patrick Lambrix (questions 1-4, tel: 013-28 26 05),
Josef Wilzén (questions 5-7, tel: 013-28 16 71)

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 (book, no notes, not electronic)

GOOD LUCK!

1. Clustering by partitioning (1+2+2=5p)

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.

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 the full running of the CLARANS algorithm on the graph. Give all steps in the computation and show at what node the algorithm 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 Agglomerative 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 (2+1=3p)

a. For the following statements say whether they are true or false.
If a statement is true, then prove it. (Observe that an example is not a proof.)
If a statement is false, then give a counterexample.

- If p and q are density connected wrt eps and MinPts, then q is density reachable from p wrt eps and MinPts.
- p is directly density reachable from p wrt eps and MinPts
- If p is density reachable from q wrt eps and MinPts, then p and q are density connected wrt eps and MinPts
- If p is directly density reachable from q wrt eps and MinPts, then p and q are density connected wrt eps and MinPts.

b. What does OPTICS do? What is the use of OPTICS?

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

a. 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.

b. Asymmetric binary variables.

Given a data set with only asymmetric binary variables where we assume that there are no data items that only have 0-values. Then we could use contingency tables to measure the distance between two data items and we could use the formula for different types (as in question a) but only use asymmetric binary variables. Would we get the same result? If yes, prove this (example is not enough). If no, give a counterexample.

5. Apriori algoritm (2+2=4p)

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	C, D, F
2	A, B, C, G
3	B, C, D
4	G
5	A, D, E, G
6	A, D, G
7	G
8	C, D

b. The prices for the items are presented in the table below. Repeat the exercise above with the following additional constraint: max(S) ≤ 6, where S is an itemset with prices. 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.

Item	A	B	C	D	E	F	G
Price	2	5	4	5	2	10	8

6. FP grow algorithm (4+3=7p)

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

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

b. Let the items have price according to the table. Run the FP grow algorithm with the constraint sum(S) ≥ 9, where S is an itemset with prices. Explain step by step the execution. Make it 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.

Item	Price
A	3
B	1
C	7
D	1
E	4
F	5

7. Rule generation (4p)

Apply the rule generation algorithm to the frequent itemset {A, B, C, D} on the database below in order to produce association rules with confidence equal or greater than 50 %. Explain step by step the execution.

Transaction id	Items
1	A, B, C, D
2	A, B, C
3	B, C, D
4	A, E, F