Institutionen för datavetenskap
Linköpings universitet

EXAM

TDDD41 Data Mining – Clustering and Association Analysis

732A75 Advanced Data Mining

August 22, 2023, kl 8-12

Teachers: Patrick Lambrix (question 1-5, tel: 013-28 26 05),
Johan Alenlöv (questions 6-8, tel: 0730-30 44 68)

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 (2+3=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? When are two nodes neighbors and how many neighbors does a node have?

   ii. Considering PAM, CLARA and CLARANS, the graph for which algorithm/algorithms contains/contain the most nodes? Explain your answer.

   iii. Which of PAM, CLARA and CLARANS guarantees to find a global optimum? Explain your answer.

   iv. Which of PAM, CLARA and CLARANS guarantees to find a local optimum? Explain your answer.

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 (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. ROCK (1+1=2p)

For the ROCK algorithm:

(i) Given the similarity matrix below. What is link(A,B) if the threshold is 0.5?

	A	B	C	D	E
A	1				
B	0.8	1			
C	0.7	0.3	1		
D	0.8	0.9	0.3	1	
E	0.1	0.2	0.3	0.4	1

(ii) Give the definition of Link between clusters.

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

a. Given data points 1, 2, 3, 7, 8, 15, 17, 19, 21, 25, 26. What clusters would DBSCAN generate if Minpts = 3 and Epsilon = 3?

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

5. 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	53	(3,1,3)	N	Y	N	Y	no-value-available
Item L	50	(1,4,3)	Y	Y	N	N	21

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.

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

a. Run the Apriori algorithm on the following transactional database with minimum support equal to three transaction. Explain clearly step by step the execution of the algorithm.

Transaction id	Items
1	C, D, F
2	A, B, C, D, G
3	B, C, D
4	G
5	A, B, D, E, G
6	A, D, G
7	E, D, G
8	B, 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. What kind of constraint is it? Explain clearly 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

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

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

Transaction id	Items
1	B, C, E, F
2	C, E, F
3	A, B, C, F
4	C, E, F
5	A, C, D, E
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. What kind of constraint is it? Explain clearly 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

8. 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 clearly 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
5	A, C