Institutionen för datavetenskap
Linköpings universitet

EXAM

TDDD41 Data Mining – Clustering and Association Analysis

732A75 Advanced Data Mining

August 24, 2021, kl 8-12

Teachers: Patrick Lambrix, Johan Alenlöv

This is an individual exam. No help from others is allowed. No helping others is allowed. No uploading (except to hand in your exam in LISAM) or downloading solutions is allowed. No communication regarding the exam is allowed (except with the teachers mentioned above).

You are allowed to use the course literature, the course slides and your own notes.

Answers to the exam questions may be sent to Urkund.

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.)
5. If you have questions mail Patrick.lambrix@liu.se for clustering and johan.alenlov@liu.se for association analysis.
6. Hand in via LISAM before 12:00. If you have problems handing in via LISAM, send your answers via e-mail to Patrick.lambrix@liu.se latest 12:05 and keep trying to upload afterwards. Handing in after this time will be considered as not handed in. (If you have permission for extended time, you can add the extension to these times.)

GOOD LUCK!

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

a. Given the data set {0, 6, 7, 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. Which of PAM/CLARA/CLARANS guarantees a local optimum for the original clustering problem? Explain your answer.

d. Which of PAM/CLARA/CLARANS guarantees a global optimum for the original clustering problem? Explain your answer.

2. Hierarchical clustering (2 + 1 = 3 p)

(i) Show the different steps of the Agglomerative Hierarchical Clustering 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

(ii) Would it be possible to optimize the computation above if you know that the threshold is 5? What is the result if the threshold is 5?

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

a. Given the data set { 1, 3, 5, 6, 7, 21, 22, 23, 30, 40 }. Assume you use DBSCAN with MinPts = 3 and Eps =3. (i) What are the resulting clusters? (ii) What are the border points? (iii) What are the outliers?

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 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 is directly density reachable from q 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.

5. 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	(15,2)	(1,4,1)	N	N	Y	N	204
Item L	(20,2)	(1,3,3)	N	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. Given a data set with only asymmetric binary variables. 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.

6. Apriori algoritm (3 + 2 + 2 = 7p)

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	G,H,I
2	G,I
3	G,H,J
4	H,I
5	H,J
6	G,I
7	G,H
8	G,I,J
9	H,I
10	G,J

b. Run the Apriori algorithm on the previous transactional database with minimum support equal to two transactions, and the following additional constraint: Find the frequent itemsets that contain two or more items. 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. Run the Apriori algorithm on the previous transactional database with minimum support equal to two transactions, and the following additional constraint: Find the frequent itemsets that DO NOT contain the itemset {I,J}. 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.

7. FP grow algorithm (4p)

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	B, A
4	B, A
5	D, A
6	D, A

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 greater or equal than 75 %. Explain step by step the execution.

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