Institutionen för datavetenskap
Linköpings universitet

EXAM

TDDD41 Data Mining – Clustering and Association Analysis

732A75 Advanced Data Mining

June 8, 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. 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 (3 + 2 = 5p)

a. Given the data set { (0,0), (0,2), (4,3), (5,3) }. Assume we use *Manhattan* distance and k = 2.

(i) Draw the graph representation of the clustering problem.

(ii) Start at an arbitrary node and show one (numlocal) iteration of the CLARANS algorithm with maxneighbor = 2 on the graph. Give all steps in the computation and show at what node that iteration ends.

b. For each of the questions below, answer yes/no and explain why.
- Does PAM guarantee to find a local optimum of the clustering problem?
- Does CLARANS guarantee to find a local optimum of the clustering problem?
- Does CLARA guarantee to find a local optimum of the original clustering problem?
- Which of PAM, CLARA, CLARANS use the largest graph?

2. Hierarchical clustering (2 + 2 + 1 = 5p)

a. Show the different steps of the Agglomerative Hierarchical Clustering algorithm using the dissimilarity matrix below and *single* link clustering. Give partial results after each step.

What are the clusters is if (i) the threshold is 5 and (ii) if the threshold is 9?

	1	2	3	4	5
1	0				
2	6	0			
3	10	9	0		
4	3	2	5	0	
5	1	7	4	8	0

b. For the ROCK algorithm:

Assume the similarity matrix below. Assume that cluster C1 contains the data objects A, B and E. Assume that cluster C2 contains the data objects C and D. What is Link(C1,C2)? Show your computations.

	A	B	C	D	E
A	1				
B	0.9	1			
C	0.8	0.7	1		
D	0.1	0.2	0.7	1	
E	0.2	0	0.3	0.4	1

c. For the BIRCH algorithm:

Explain Clustering Feature Vector. Given a cluster with the data points (0,0), (1,1) and (2,2), what is its clustering feature vector? How do you compute its centroid using the elements in the clustering feature vector?

4. Density-based clustering (2p)

Show how DBSCAN works on the data set { 1, 3, 5, 6, 7, 20, 22, 24, 30, 40 } with MinPts = 3 and Eps =3. Show all steps as well as the resulting clusters.

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	(15,20,4)	(1,4,1)	N	N	Y	N	2
Item L	(20,20,4)	(1,2,3)	N	Y	Y	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. Assume we have categorical data. One method to define a distance between two data objects is (p-m)/p where p is the total number of categorical variables and m is the number of categorical variables for which there is a match between the objects. A second method is to introduce a new asymmetric binary variable for each of the possible values for each of the categorical variables. Give a formula for the distance between two objects in the second method in terms of p and m (where p and m have the same meaning as above; i.e. p is the number of categorical variables - not the number of introduced binary variables, and m is the number of matches in the categorical variables). Show how you obtained the formula.

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

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	A, B, C
2	B, C
3	A,C
4	B, A
5	C, A

b. Run the Apriori algorithm on the following transactional database with minimum support equal to two transactions, and the following additional constraint: Find the frequent itemsets that contain the item D. 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.

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

c. Run the Apriori algorithm on the last transactional database (same as b) with minimum support equal to two transactions, and the following additional constraint: Find the frequent itemsets that (1) contain the item D and (2) do not contain the item B. 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 (4 + 3 = 7p)

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

Transaction id	Items
1	U, X, Y
2	U, Y, Z
3	Y
4	V, X, Y, Z
5	U, V, X
6	U, V, X
7	Y
8	Y
9	U, X
10	V, X, Y, Z

Items	Price
U	100
V	110
X	50
Y	70
Z	150

b. Run the FP grow algorithm on the previous transactional database with minimum support equal to two transactions and the following additional constraint: Find the frequent itemsets where the sum of the prices is less than or equal to 210. 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.

8. Rule Generation (2p)

Apply the rule generation algorithm on the frequent itemset ACEG on the database below in order to produce association rules with confidence greater or equal than 60%. Explain step by step the execution.

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