Institutionen för datavetenskap
Linköpings universitet

EXAM

TDDD41 Data Mining – Clustering and Association Analysis

732A75 Advanced Data Mining

June 8, 2022, kl 8-12

Teachers: Patrick Lambrix (013-282605, questions 1-5),
Johan Alenlöv (questions 6-8)

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 on time. Handing in late is considered as not handing in. So, start queuing early enough!**

Help: dictionary

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?
   ii. When are two nodes neighbors and how many neighbors does a node have?
   iii. Considering PAM, CLARA and CLARANS, the graph for which algorithm/algorithms contains/contain the most nodes? Explain.
   iv. Which of PAM, CLARA and CLARANS guarantees to find a global optimum?

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 (3+3=6p)

a. Describe the principles and ideas regarding BIRCH by answering the following:

i. Explain Cluster Feature Vector. Given a cluster with the data points (1,2), (1,3) and (2,2), what is its cluster feature vector?

ii. Explain what a CF-tree is and how it is used in BIRCH.

iii. Describe the algorithm for building the CF tree.

iv. What parameters are used as input?

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 (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	(10,10)	(3,1,1)	Y	N	Y	N	38
Item L	(10,14)	(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. Show how an interval-based measure can be defined for ordinal variables

5. 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, D
3	A, C
4	A, B, C
5	C, D
6	A,B,C,D

Item	Value
A	2
B	2
C	1
D	1

b. Assign each item in the above transactional database according to the table in a. Run the Apriori algorithm, on the previous transactional database, with the constraint sum(S) < 5, where S is an itemset. 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.

c. Using the values and transactional database and values given previously in a, run the Apriori algorithm with the constraint sum(S) > 2, where S is an itemset. 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.

6. **FP grow algorithm (4p)**

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	E,G
2	F,G
3	H,I
4	F,H
5	E,G
6	H,I

7. Constraints (3p)

Present a constraint that is neither monotone nor anti-monotone but is convertible anti-monotone. Prove that the presented constraint has all of the properties, that it is neither monotone not anti-monotone but is convertible anti-monotone.

8. Rule Generation (2p)

Apply the rule generation algorithm on the frequent itemset JKL 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	JKL
2	JK
3	JKL
4	KL
5	J
6	K