Institutionen för datavetenskap
Linköpings universitet

EXAM

TDDD41 Data Mining – Clustering and Association Analysis

732A75 Advanced Data Mining

August 23, 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 - no notes in the dictionary allowed.

GOOD LUCK!

1. Clustering by partitioning (3 + 2 = 5p)

a. Given the data set { (0,0), (0,1), (4,4), (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 iteration of the PAM algorithm 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 global optimum of the clustering problem?
- Does PAM guarantee to find a local optimum of the clustering problem?
- Does CLARA guarantee to find a local optimum of the original clustering problem?
- Does CLARANS guarantee to find a local optimum of the clustering problem?

2. Hierarchical clustering (2p)

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 4 and (ii) if the threshold is 8?

$$
\begin{array}{c|ccccc}
 & 1 & 2 & 3 & 4 & 5 \\
\hline
1 & 0 & & & & \\
2 & 9 & 0 & & & \\
3 & 10 & 6 & 0 & & \\
4 & 3 & 8 & 5 & 0 & \\
5 & 1 & 7 & 4 & 2 & 0
\end{array}
$$

3. ROCK (1 + 1 + 1 = 3p)

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	0.9	1			
C	0.8	0.7	1		
D	0.1	0.2	0.5	1	
E	0.2	0	0.3	0.4	1

(ii) If the number of elements in a cluster is n and the number of neighbors for each element in the cluster is m, what is the contribution of an object in the cluster to the expected link for the cluster? Explain why.

(iii) When defining the goodness measure for clusters C1 and C2 for ROCK, why do we not just use link(C1,C2), but divide this by the expected link between C1 and C2?

4. Density-based clustering (2p)

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 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 + 1 = 4p)

a. What is the distance between Item K and Item L? (no normalization needed)

	A	B	C	D	E	F	G
Item K	(10,500)	(2,1,1)	Y	N	Y	N	8
Item L	(10,505)	(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. Show how an interval-based measure can be defined for ordinal variables.

c. Give and explain the distance measure for objects with asymmetric binary variables using contingency tables.

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

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 the itemset {A }. 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 {C }. 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 + 2 = 6p)

a. 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	A
2	A, B, X, Z
3	A, X
4	A
5	B, C, X
6	C
7	A, B, C, Y
8	A, C
9	C, X, Y

b. Discuss in a few sentences the difference between the FP-Grow algorithm and the Apriori algorithm. Write down pros and cons for both algorithm. You may use the above transactional database as an example highlighting your pros and cons.

8. Rule Generation (3p)

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 60 %. Explain step by step the execution.

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