Institutionen för datavetenskap
Linköpings universitet

EXAM

732A61 and TDDD41
Data Mining –
Clustering and Association Analysis

732A75 Advanced Data Mining

March 26, 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), (0,1), (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 *complete* 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
```

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 0.6?

```
    | 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 directly density reachable from q wrt eps and MinPts, then q is directly density reachable from p 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 | (35,20,4)  (2,5,1)  Y  N  Y  N  77
  Item L | (40,20,4)  (1,2,1)  Y  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. 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 (2 + 2 + 2 + 2 = 8p)

    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, E
2	A, B, C, D
3	B
4	B, C, D
5	C, D
6	A, B, C
7	B, C
8	A, C, D
9	B, C, D
10	A, B, D
11	A, D

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, 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.

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, 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.

d. Apply the rule generation algorithm on the itemset {B, C, D} on the previous database in order to produce association rules with confidence greater or equal than 60 %. Explain step by step the execution.

7. FP grow algorithm (4p)

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	S, T
2	U, V, S
3	W, S, T
4	T, V
5	S
6	V
7	S, V, W

8. Constraints (1 + 1 + 1 = 3p)

a. Prove that the constraint "the average price of a set of items is smaller than 200" is neither monotone nor antimonotone.

b. Prove that the previous constraint is convertible monotone.

c. Prove that the previous constraint is convertible antimonotone as well