Institutionen för datavetenskap
Linköpings universitet

EXAM
732A61 and TDDD41
Data Mining –
Clustering and Association Analysis
August 22, 2017, kl 8-12

Teachers: Patrick Lambrix, José M Pena

Instructions:

- Start each question at a new page.
- Write at one side of a page.
- Write clearly.
- 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.)

Help: dictionary

GOOD LUCK!

1. Clustering (4.5p)

Describe for each of the different kinds of clustering methods (i) partitioning approach, (ii) hierarchical approach, and (iii) density-based approach:

- Main ideas
- Possible input parameters
- Possible output (e.g. properties of the clusters)

2. Clustering by Partitioning (2p+0.5p = 2.5p)

a. Describe the principles and ideas regarding CLARANS.
   (i) Give the algorithm.
   (ii) Define swapping cost.

b. Does PAM guarantee a global optimum regarding the cost fucnction?

3. Hierarchical clustering (3p)

Describe the principles and ideas regarding Agglomorative Hierarchical Clustering. Show the different steps of the algorithm using the dissimilarity matrix below and *single link* clustering. Give partial results after each step.

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

4. Clustering categorical data (4p)

Describe the principles and ideas regarding the ROCK algorithm.
Within your description, make sure to describe the algorithm and to define **and** give examples of neighbor, common neighbor, link for objects, link for clusters, and G (goodness measure).

5. Distance measure (2p)

What is the distance between Item K and Item L?

	A	B	C	D	E	F	G
Item K	gold	(0,0)	Y	N	Y	N	silver
Item L	bronze	(1,1)	N	N	N	N	no-value-available

Attributes A and G are ordinal variables with values gold/silver/bronze in that order.
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.

6. Apriori algoritm (2p+1p+1p+2p=6p)

a. Write down the pseudocode for the Apriori algorithm. All the steps in the pseudocode should contain enough details for a non-expert in the algorithm to be able to implement it.

b. Write down the pseudocode for the Apriori algorithm with a monotone constraint.

c. Write down the pseudocode for the Apriori algorithm with an antimonotone constraint.

d. Sketch a proof of the correctness of the Apriori algorithm.

7. FP grow algorithm (2p+1p+2p=5p)

a. Write down the pseudocode for the FP grow algorithm. All the steps in the pseudocode should contain enough details for a non-expert in the algorithm to be able to implement it.

b. Write down the pseudocode for the FP grow algorithm with a monotone constraint.

c. What is the main advantage that the FP grow algorithm has over the Apriori algorithm ? Illustrate the advantage with an example.

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

a. Give an example of a constraint that is both monotone and antimonotone. If you think it is not possible, explain why.

b. Consider the constraint sum(S)+min(S)>3 where S is an itemset, sum(S) returns the total price of the itemset S, and min(S) returns the price of the cheapest item in S. Note that some prices may be negative. Is this constraint monotone, antimonotone, convertible monotone, convertible antimonotone, or none ?

c. Give an example of a rule with lift greater than one and another example of a rule with lift smaller than one.