Institutionen för datavetenskap
Linköpings universitet

EXAM
732A31and TDDD41
Data Mining –
Clustering and Association Analysis
March 19, 2014, 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 by Partitioning (4p)

Given the data set {0, 2, 3, 8}. Assume we use Euclidean distance. Assume we are using PAM with k=2 and currently have the cluster centers 0 and 2.

(i) Give for each of the current clusters which data objects belong to the cluster.

(ii) Show the next iteration in the PAM algorithm. Give all steps in the calculation. Will there be a change in cluster centers? If so, give for each of the new clusters which data objects belong to the cluster.

2. Hierarchical clustering (3p+3p=6p)

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

b. Describe the principles and ideas regarding the CHAMELEON algorithm..In your description, make sure to define the important notions and define how clusters are formed.

3. Density-based clustering (2p+1p=3p)

a. Describe the principles and ideas regarding the DBSCAN algorithm. In your description, make sure to give a sketch of the algorithm and to define core point, direct density-reachable, density-reachable, and density-connected.

b. What is the main idea behind OPTICS?

4. Distance measure (1p+1p+1p=3p)

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

b. Give and explain the distance measure for objects with variables of mixed types.

c. Can the formula in question b also be used for objects with only asymmetric variables? If no, explain why. If yes, state whether you would get the same results as with the method in question a and explain why.

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

a. Run the Aprori algorithm on the following transactional database with minimum support equal to one transaction. Explain step by step the execution.

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

b. Repeat the exercise 1a with the following additional constraint: Find the frequent itemsets that do not contain the itemset CD. 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. Let the items A, B, C and D have a price of respectively 1, 2, 3 and 4 units. Repeat the exercise 1a with the following additional constraint: Find the frequent itemsets whose total cost is greater than 2. 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. Sketch a proof of the correctness of the Apriori algorithm.

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

a. Describe the FP grow algorithm.

b. Explain how you incorporate a monotone constraint into the FP grow algorithm.

c. Explan how you incorporate an antimonotone constraint into the FP grow algorithm.

d. What is the main advantage that the FP grow algorithm has over the Apriori algorithm ?

7. 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. Apply the Simple algorithm to the frequent itemset ABC on the database in exercise 1 in order to find association rules with confidence greater or equal than 50 %.
c. Give an example of an association rule with lift greater than one and another example of a rule with lift smaller than one.