Institutionen för datavetenskap
Linköpings universitet

EXAM
732A31and TDDD41
Data Mining – Clustering and Association Analysis
March 18, 2015, 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 (2p+3p=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, 2, 3, 8}. 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 (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	8	0			
3	2	10	0		
4	1	5	7	0	
5	9	3	6	4	0

3. Density-based clustering (2p+1p+1p=4p)

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. DBSCAN: Consider the following statement: if p is density-connected to q wrt Eps and Minpts then p is density-reachable from q wrt Eps and Minpts. Is this statement true? If yes, then prove. If no, then give a counterexample.
c. What is the main idea behind OPTICS?

4. Different types of data and their distance measures (2p+2p=4p)

a. What is the distance between Item K and Item L?

	A	B	C	D	E	F	G
Item K	(20,1)	(2,2)	Y	N	Y	N	8
Item L	(20,100)	(3,6)	N	N	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. Assume we have categorical data. One method to define a distance between two data objects is (p-m)/p where p is the total number of categorical variables and m is the number of categorical variables for which there is a match between the objects. A second method is to introduce a new asymmetric binary variable for each of the possible values for each of the categorical variables. Give a formula for the distance between two objects in the second method in terms of p and m (where p and m have the same meaning as above; i.e. p is the number of categorical variables - *not* the number of introduced binary variables, and m is the number of matches in the categorical variables).

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

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

Transaction id	Items
1	A, B, C
2	X, Y, Z
3	A, Y, C
4	X, B, Z

b. Repeat the exercise 5a with the following additional constraint: Find the frequent itemsets that do not contain the itemset AB. 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, X, Y and Z have a price of respectively -3, -2, -1, 1, 2 and 3 units. Repeat the exercise 5a with the following additional constraint: Find the frequent itemsets whose most expensive item has positive price. 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. Explain how you incorporate an antimonotone constraint into the FP grow algorithm.

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

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

a. Prove that a constraint C cannot be both monotone and antimonotone, unless C(A)=C(B) for all itemsets A and B. Note that you have to prove the statement and thus it does not suffice with giving an example.

b. Apply the Simple algorithm to the frequent itemset XBZ on the database in exercise 5 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.