Institutionen för datavetenskap
Linköpings universitet

EXAM
732A61, 732A31 and TDDD41
Data Mining –
Clustering and Association Analysis
March 18, 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 by partitioning (4p+1p=5p)

a. - Describe the algorithm for CLARANS.
- Given the data set {0, 3, 4, 10}. 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 (i.e., for one value of numlocal) of the CLARANS algorithm with maxneighbor = 2. Give all steps in the computation and show at what node that iteration ends.

b. Why is CLARA in general faster than PAM?

2. Hierarchical clustering (3+1=4p)

a. Describe the principles and ideas regarding Agglomerative 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	3	4	0		
4	1	7	9	0	
5	10	2	6	5	0

b. For the ROCK algorithm:

Given the *similarity matrix* below. What is link(A,B) if the threshold is 0.7?

	A	B	C	D	E
A	1				
B	0.9	1			
C	0.8	0.7	1		
D	0.1	0.3	0.6	1	
E	0	0.2	0.4	0.5	1

3. Density-based clustering (2+1=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 relationship between DBSCAN and 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	(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. 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 contain the item A. 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 range is smaller than 3 (recall that the range is the price of the most expensive item minus the price of the cheapest item). 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+2p+2p+1p=7p)

a. Repeat the exercise 5a but this time with the FP grow algorithm.

b. Repeat the exercise 5b but this time with the FP grow algorithm.

c. Repeat the exercise 5c but this time with the FP grow algorithm.

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

7. Rule generation (2p)

Apply the Simple algorithm to the frequent itemset XBZ on the database in exercise 5 in order to find association rules with confidence greater than 50 %.