Institutionen för datavetenskap
Linköpings universitet

EXAM

732A61, 732A31 and TDDD41
Data Mining –
Clustering and Association Analysis
June 7, 2017, kl 8-12

Teachers: Patrick Lambrix, José M Pena

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.)

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. Describe the principles and ideas regarding PAM.
   i. Describe the algorithm.
   ii. Define swapping cost.
   iii. Draw an example of a data set in two dimensions where the swapping cost TCih is 0 and one where the swapping cost TCih is strictly negative.

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

a. Describe the principles and ideas regarding BIRCH by answering the following:
   i. Give a sketch of the algorithm.
   ii. Explain Cluster Feature Vector. Given a cluster with the data points (1,2), (1,3) and (2,2), what is its cluster feature vector?
   iii. Explain what a CF-tree is and how it is used in BIRCH.
   iv. What parameters are used as input?

b. For the ROCK algorithm:

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	0.3	0.6	1	
E	0	0.2	0.4	0.5	1

3. Density-based clustering (2p)

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

4. Different types of data and their distance measures (3p+1p+1p=5p)

a. Asymmetric binary variables.
   i. Give and explain the distance measure for objects with asymmetric binary variables using contingency tables.
   ii. Give and explain the distance measure for objects with variables of mixed types.
   iii. 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 or why not.

b. In the vector model for information retrieval documents are represented by vectors with positive real numbers. How is the similarity between two vectors defined?

c. Show how an interval-based measure can be defined for ordinal 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+1p=3p)

a. Explain how you incorporate monotone and antimonotone constraints in the FP grow algorithm.

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

7. Constraints (2p+2p=4p)

a. Give an example of convertible monotone and convertible antimonotone constraints that are not monotone and antimonotone, respectively.

b. Show that your constraints are really convertible monotone and antimonotone. Do not give examples to show it, i.e. provide an abstract and formal argument.

8. 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 %.