Institutionen för datavetenskap
Linköpings universitet

EXAM

732A61and TDDD41
Data Mining –
Clustering and Association Analysis

732A75 Advanced Data Mining

June 11, 2019, 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 (1+2+2=5p)

a. Given the data set {0, 7, 8, 10}. Assume we use Euclidean distance and k = 2. Draw the graph representation of the clustering problem.

b. 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.

c. Assume numlocal = 1 and maxneighbor = 2. Start at the same node as in question *b* and show the running of the CLARANS algorithm on the graph. Give all steps in the computation and show at what node the program ends.

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

a. 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	0.2	0.5	1	
E	0.2	0	0.3	0.4	1

b. 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?

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

a. 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.

b. What is the relationship between DBSCAN and OPTICS?

4. Different types of data and their distance measures (2+1+1=4p)

a. What is the distance between Item K and Item L? (no normalization needed)

	A	B	C	D	E	F	G
Item K	(50,500)	(2,1,0)	Y	N	Y	N	77
Item L	(50,505)	(1,3,0)	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. Show how an interval-based distance measure can be defined for ordinal variables.

c. Show how a distance measure can be defined for categorical (or nominal) variables.

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

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	A, C, D
3	A, B
4	A, B
5	A, D
6	A, D

b. Prove formally the correctness of the Apriori algorithm.

c. Prove formally the correctness of the rule generation.

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

a. Run the FP grow 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 above with the following additional constraint: Find the frequent itemsets that contain the item B. 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. What is the main advantage of the FP grow algorithm over the Apriori algorithm ?

7. Constraints and Causality (3p+2p=5p)

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

b. Give an example of an association rule that is NOT a causal rule. Explain your answer.