Institutionen för datavetenskap
Linköpings universitet

EXAM

732A61and TDDD41
Data Mining –
Clustering and Association Analysis

732A75 Advanced Data Mining

May 10, 2020, kl 8-12

Teachers: Patrick Lambrix, José M Pena

This is an individual exam. No help from others is allowed. No uploading or downloading solutions is allowed. No communication regarding the exam is allowed.

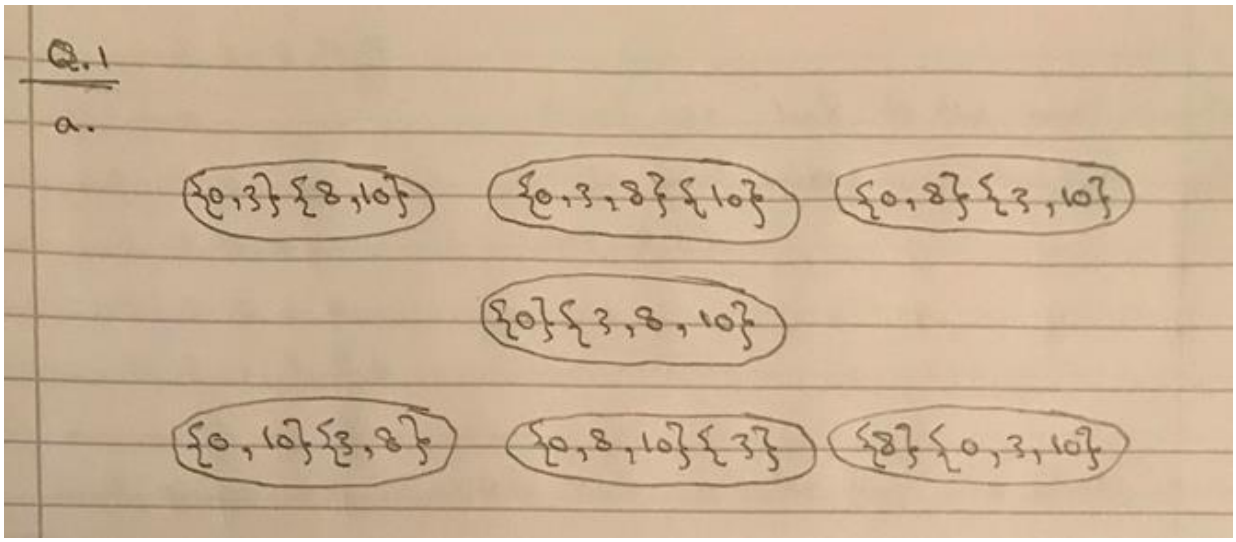
You are allowed to use the course literature, the course slides and your own notes.
Answers to the exam questions may be sent to Urkund.

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.)
5. Hand in a pdf file. (You could use Word for text; take pictures of drawings/calculations and insert in a Word file; and then export to pdf.)
6. Hand in via LISAM before 12:00. If you have problems handing in via LISAM, send your answers via e-mail to Patrick.lambrix@liu.se latest 12:05 and keep trying to upload afterwards. Handing in after this time will be considered as not handed in.

GOOD LUCK!

1. Clustering by partitioning (1+2+2=5p)

a. Given the data set {0, 3, 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.

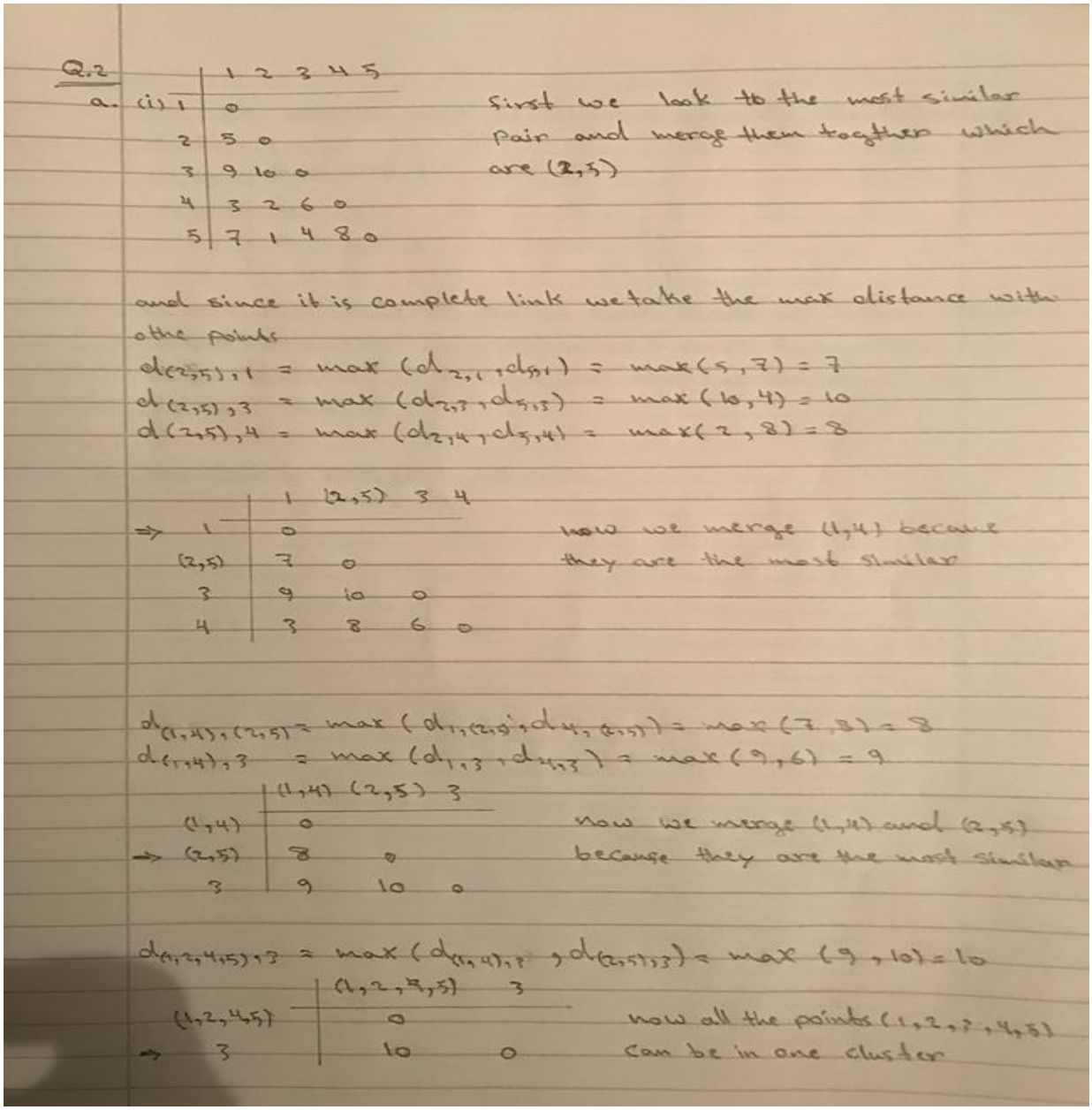
Starting at {0,3}{8,10} the PAM calculates the total distance to the medoids which is 5 and jump to the next node and compare the total distance if it is more it will skip it to the next one.. the best node is {0,3}{8,10} because it have the least total distance to the medoids (total cost)

c. For each of the questions below, answer yes/no and explain why.
- Does PAM guarantee to find a global optimum of the clustering problem?
Yes, because it takes all the data and it keep iterating until it reaches the global optimum.
- Does PAM guarantee to find a local optimum of the clustering problem?
No, unless the local optimum is the global optimum, it will keep iterating until it converges to the global optimum.
- Does CLARA guarantee to find a local optimum of the original clustering problem?
Yes, because it only takes some samples from the whole data so it will converges to a local optimum.
- Does CLARANS guarantee to find a local optimum of the clustering problem?
Yes, CLARANS can find the best local optimum because it keep randomly searching the whole graph as well as part of the graph until it converges to the best local optimum given the number of local minima & maximum number of neighbors to compare that it has provided with in the beginning.

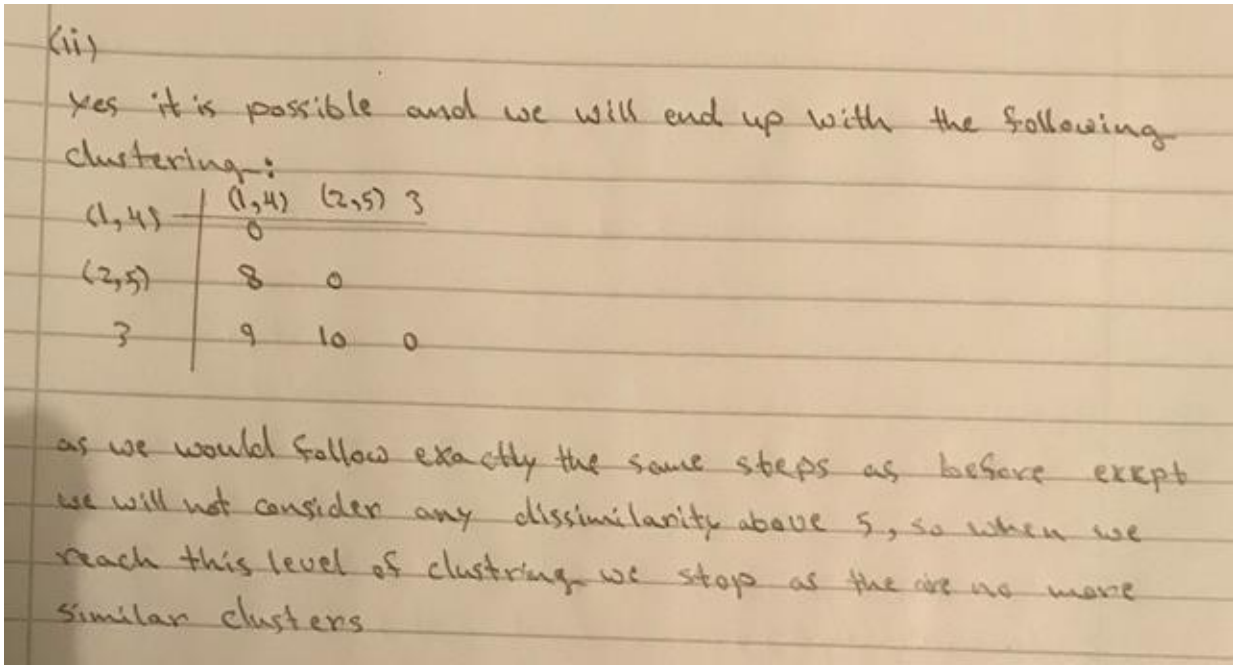
2. Hierarchical clustering (3+2=5p)

a. (i) Show the different steps of the Agglomerative Hierarchical Clustering algorithm using the dissimilarity matrix below and *complete* link clustering. Give partial results after each step.

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



(ii) Would it be possible to optimize the computation above if you know that the threshold is 5? What is the result if the threshold is 5?



b. For the ROCK algorithm:

(i) 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

The link is (AB), (AC) and (BC)

(ii) If the number of elements in a cluster is n and the number of neighbors for each element in the cluster is m, what is the contribution of an object in the cluster to the expected link for the cluster? Explain why.

3. Density-based clustering (2p)

For the following statements say whether they are true or false.
If a statement is true, then prove it. (Observe that an example is not a proof.)
If a statement is false, then give a counterexample.

- If p and q are density connected wrt eps and MinPts, then q and p are density connected wrt eps and MinPts.
  TRUE
- If p is density reachable from q wrt eps and MinPts, then q is density reachable from p wrt eps and MinPts.

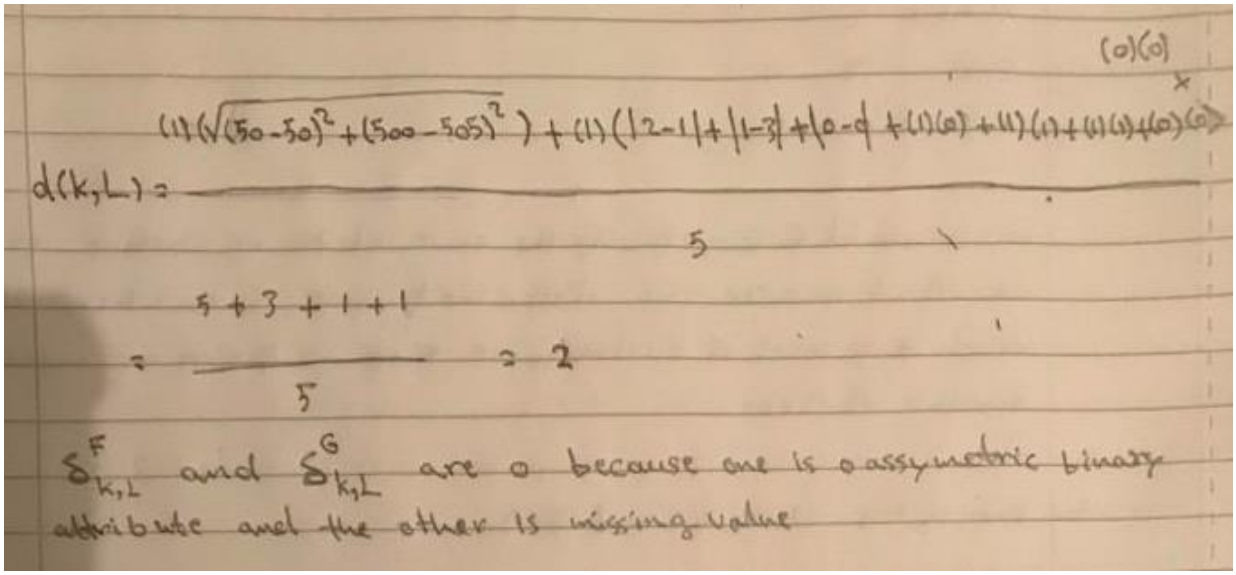
- If p is directly density reachable from q wrt eps and MinPts, then p is density reachable from q wrt eps and MinPts
- If p is directly density reachable from q wrt eps and MinPts, then p and q are density connected wrt eps and MinPts.

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

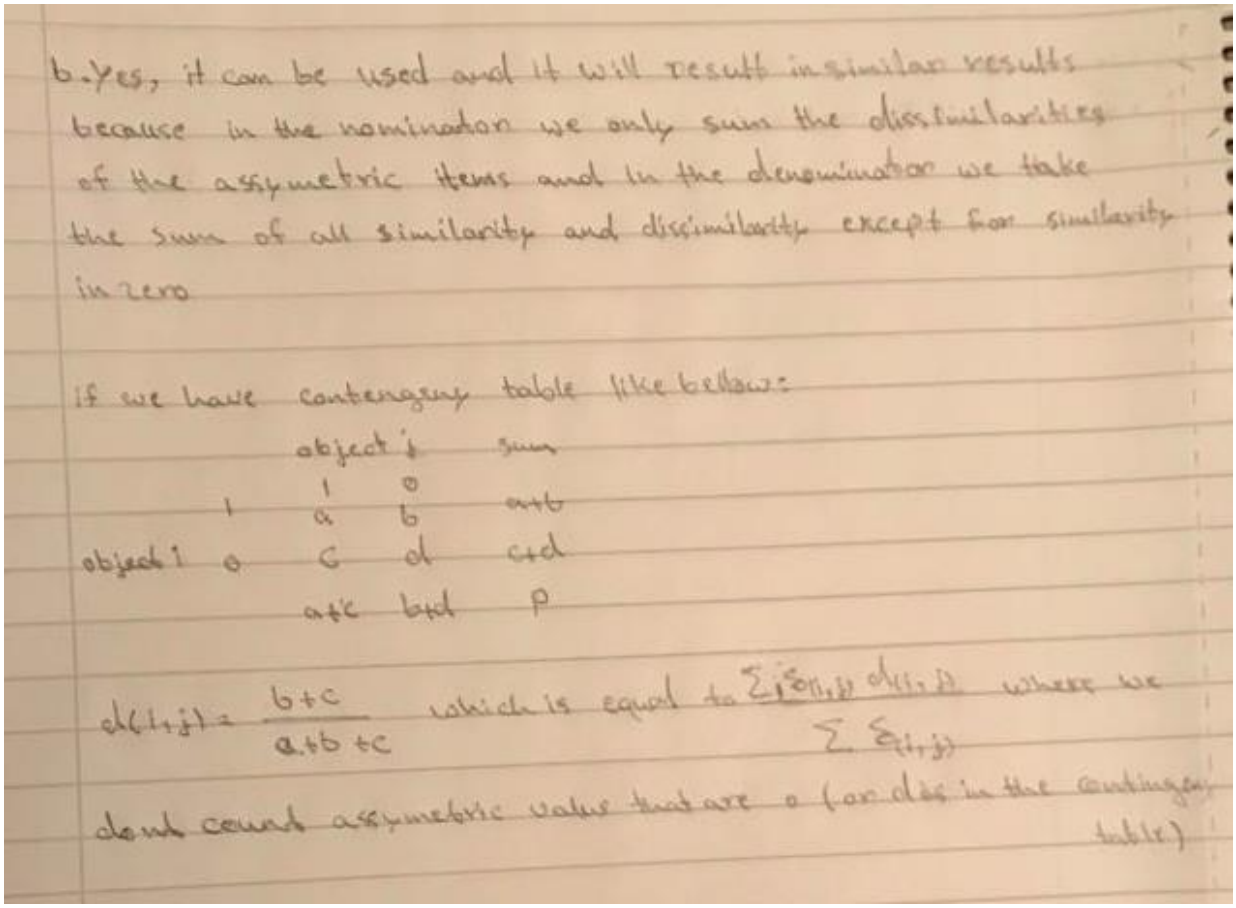
a. What is the distance between Item K and Item L? (no normalization needed)

	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. Can the formula for distance for objects with variables of mixed types (that you used in a) 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 using contingency tables and explain why or why not.



b. Yes, it can be used and it will result in similar results because in the nomination we only sum the dissimilarities of the asymetric items and in the denominator we take the sum of all similarity and dissimilarity except for similarity in zero

if we have contengency table like bellow:

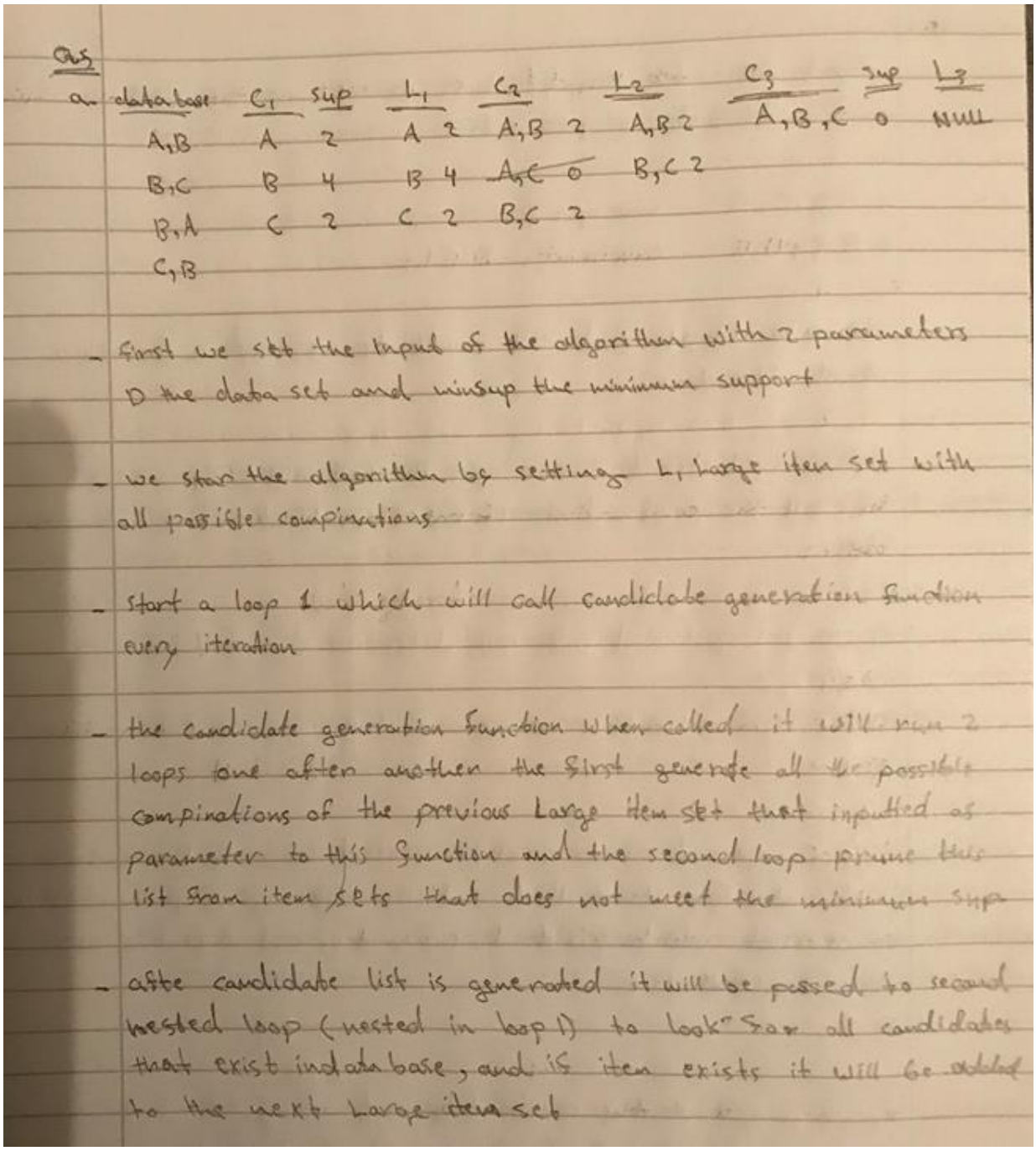
		object j		sum
		1	0	
object i	1	a	b	a+b
	0	c	d	c+d
		a+c	b+d	p

$d(i,j) = \frac{b+c}{a+b+c}$ which is equal to $\frac{\sum_{i,j} \delta_{(i,j)} d_{(i,j)}}{\sum \delta_{(i,j)}}$ where we don't count asymetric value that are 0 (or d in the contingency table)

5. Apriori algoritm (2p+2p+2p+2p+1p=9p)

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

Transaction id	Items
1	A,B
2	B,C
3	B,A
4	C,B



b. Run the Apriori algorithm on the following transactional database with minimum support equal to two transactions, and the following additional constraint: Find the frequent itemsets that contain the item D. 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.

Transaction id	Items
1	A, B, C, D
2	B, C, D, E

b.

- first we inter the data base as well as the minimum sup. and constraints as parameters

1 A, B, C, D minsup = 2
2 B, C, D, E Constraint = D ∈ L1

- we run loop1 and call candidate generation function each iteration.

- in the first iteration we get the following C_1 which has all the combinations of one item that exist in the data base :

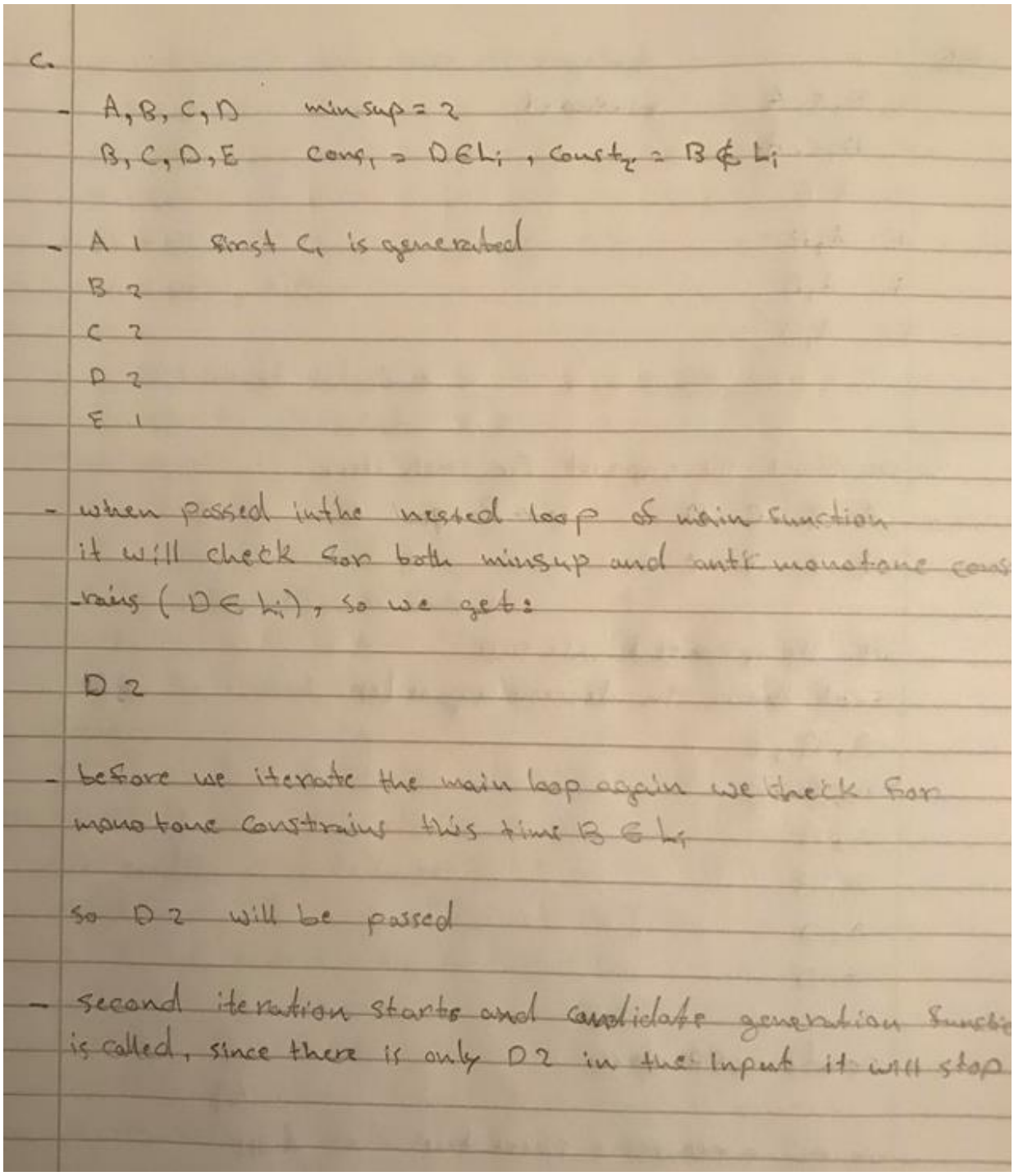
A 1
B 2
C 2
D 2
E 1

- this C_1 goes then to another nested loop in the main func. that look for items that satisfy both minsup and constraints which yields L_1 Large item set since it is antimonotone constraint.

D 2

- this then goes to the candidate generation function but it will stop since there are no more items to join

c. Run the Apriori algorithm on the last transactional database with minimum support equal to two transactions, and the following two additional constraints: Find the frequent itemsets that (1) contain the item D and (2) do not 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.



d. Apply the rule generation algorithm to the frequent itemset ABC on the database below in order to produce association rules with confidence greater or equal than 50 %. Explain step by step the execution.

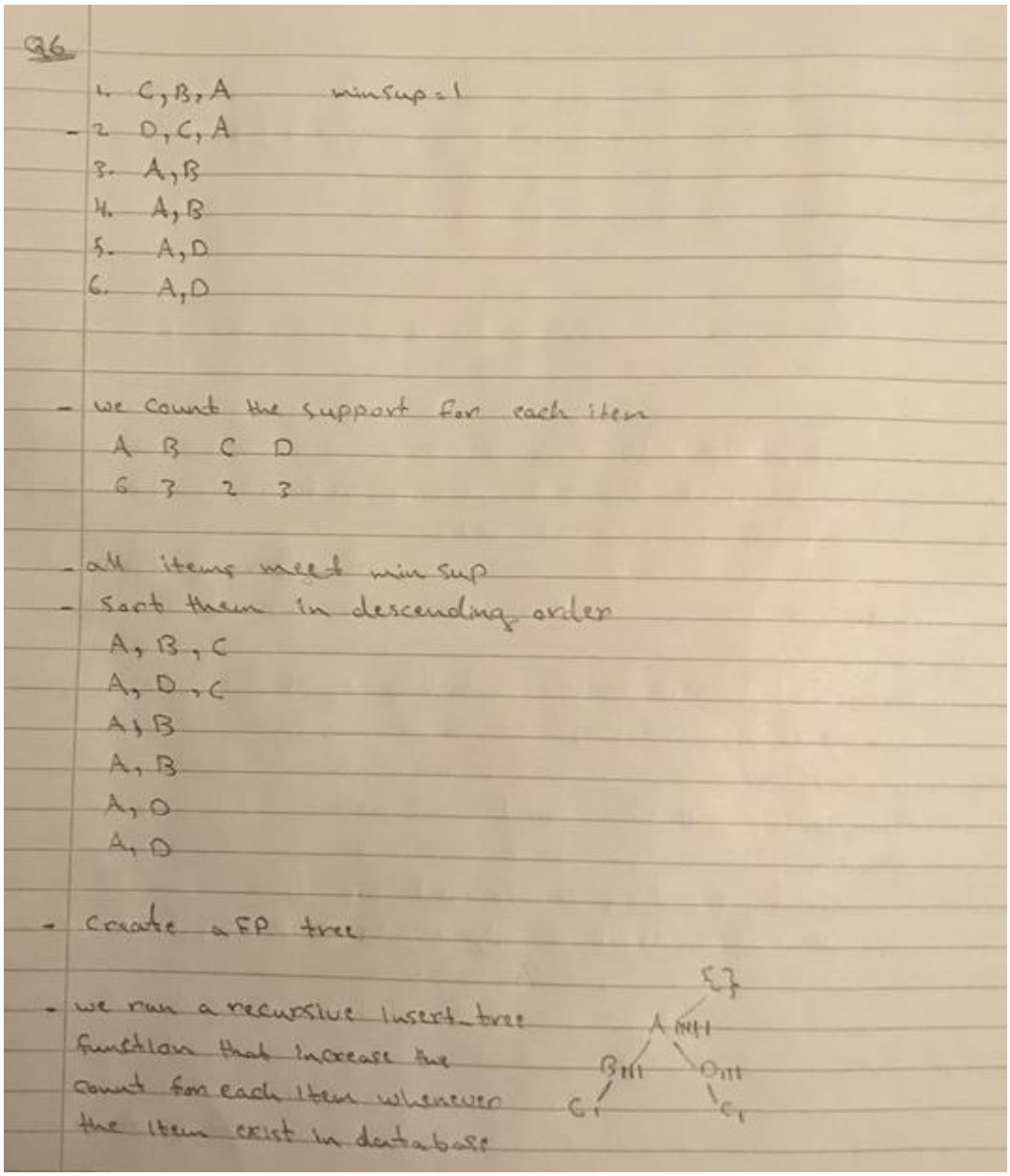
Transaction id	Items
1	C, B, A
2	D, C, A
3	A, B
4	A, B
5	A, D
6	A, D

e. In the light of the previous result, propose an improvement to the rule generation algorithm.

6. FP grow algorithm (3p)

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

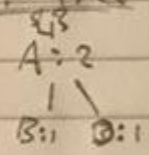


- a conditional data base is created

A: -
B: A:3
D: A:3
C: A,B:1 , A,D:1

- C- conditional data base
1. A,B → sorted A, B
2. A,D A, D

C FP tree



- B- conditional data base
1. A
2. A → sorted A →
3. A A

{}
|
A:3

- D conditional data base
1. A A
2. A → sorted A
3 A A

{}
|
A:3

- A- conditional data base
1. -

{}

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

a. Prove that the constraint "the average price of a set of items is greater than 10" is neither monotone nor antimonotone.
b. Prove that the previous constraint is convertible monotone. Check the slides for the definition of convertible monotone constraint.
c. Prove that the previous constraint is convertible antimonotone as well. Check the slides for the definition of convertible antimonotone constraint.