Q-1 (a) PAM

(i) PAM is better then K means as it uses medoids - It is good for small data sets -

Algo

1. Select k arbitrary nodes as cluster medoids -
2. Select ~~or~~ n arbitrary points i - For each selected object i and non selected object h calculate the total swapping cost TC_{ih}, do it for all items -

Repeat

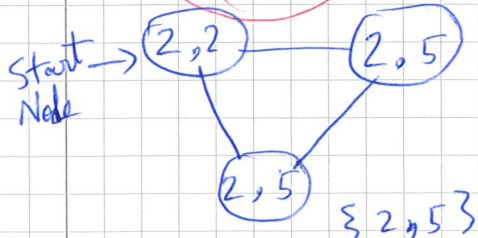
3. Select the ~~object node~~ objects with the minimum swapping cost TC_{ih} -
4. If $TC_{ih} < 0$ replace i by h and repeat the process for all the objects in the dataset -
5. ~~Set~~ set the current cost to the minimum cost and current node the best node so far -

Repeat for all items -

(ii) Swapping Cost - Is the cost got by replacing the selected item i and the non selected item h in the data set - TC_{ih} -

Is the cost we get if we replace two items we replace items with each other if the cost is minimum or $TC_{ih} < 0$ -

(iii) {2, 2, 5}

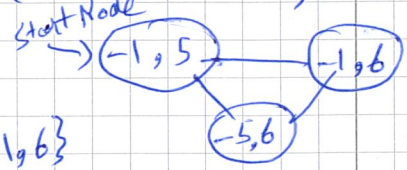


{2,5} ← Taking this node to check

	m_1	m_2
2	0	3
2	0	3
5	3	0

→ $TC_{ih} = 0$

{-1, -5, 6}



{-1, 6}

	m_1	m_2
-1	0	7
-5	-4	11
6	7	0

$TC_{ih} = -4$

Strictly Negative

LINKÖPINGS UNIVERSITET

AID-nummer: *AID-number:* 2203	Datum: *Date:* 18-06-05	Blad nummer: *Sheet number:* 2
Kurskod: *Course code:* 732A75	Provkod: *Exam code:* TENT	

Q-1-b

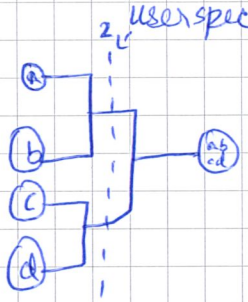
PAM, ~~CLARA~~ 0,5

Q-1-c

None 1

Q-2 Agglomerative Hierachical Clustering-

Agglomerative Hirachical clustering, AGNES is a bottom up hirachical clustering algorithm which uses dissimilarity matrixes to make clusters of the dataset. It combines the items based on the similarity between them to make a cluster until the user specified threshold.



It uses different types of links to combine clusters complete link, average link, medoids, centroid etc.

	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

① First we check the minimum item in the matrix which is 1 in (1, 4) so we will combine 1 and 4.

	(1,4)	2	3	5
(1,4)	0			
2	8	0		
3	9	4	0	
5	10	2	6	0

$dist((1,4),2) = Max(dist(1,4), dist(2))$
$= Max(dist(1,2), dist(4,2))$
$= Max(8, 7)$
$= 8$

$dist((1,4),3) = Max(dist(1,4), dist(3))$
$= Max(dist(1,3), dist(4,3))$
$= Max(3, 9) = 9$

: We are using Max because it is complete link clustering.

$dist((1,4),5) = Max(dist(1,4), dist\,5)$

$= Max(dist(1,5), dist(4,5))$

$= Max(10, 5) = 10$

② Now again we check for minimum value which is 2 in 2 and 5, so we combine (2,5)

	(1,4)	(2,5)	3
(1,4)	0		
(2,5)	10	0	
3	9	6	0

$dist((1,4),(2,5)) = Max(dist(1,4), dist(2,5))$

~~$= Max(dist(1,2), dist(4,5))$~~

$= Max(dist((1,4),2), dist((1,4),5))$

$= Max(8, 10)$

$= 10$

$dist(3,(2,5)) = Max(dist(3,2), dist(3,5))$

$= Max(4, 6)$

$= 6$

③ Now again checking for minimum value which is 6 so combining (3,(2,5)).

$Max(dist(1,4), dist(3,(2,5)) = Max(dist((1,4),3), dist((1,4),(2,5))$

$Max(dist(1,4), dist(3,2,5)) = Max(9, 10)$

$= 9$

	(1,4)	(3,2,5)
(1,4)	0	
(3,2,5)	10	0

Q.3 ROCK

Rock algorithm is a clustering algorithm used for categorical data because traditional hierarical algorithms do not do well with categorical data because categorical data sample space is discreet and we cannot apply distance measure on it-

ROCK algorithm uses cluster neighbor, common neighbors to combine the clusters -

In ROCK if two points ~~have~~ are similar to each other they are neighbors - If points have other elements ~~common~~ similar to them they are common neighbors -

Let $\langle a, b, c, d, e \rangle = \{\{a,b,c\}, \{a,b,d\}, \{a,b,e\}, \{a,c,d\}, \{a,c,e\}, \{a,d,e\}, \{b,c,d\}, \{b,c,e\}, \{b,d,e\}, \{c,d,e\}\}$

~~$\langle a, b, c, d \rangle = \{\{f,g,h\}, \{f,g,i\}, \{f,h,i\}$~~

$\langle a, b, e, f \rangle = \{\{a,b,e\}, \{a,b,f\}, \{a,e,f\}, \{b,e,f\}\}$

Let $T_1 = \langle a, b, c \rangle$ and $T_2 = \langle a, b, e \rangle$

Here $T_1 \cup T_2 = a, b, c, e$

$T_1 \cap T_2 = a, b.$

$sim(T_1, T_2) = \frac{a,b}{a,b,c,e} = \frac{2}{4} = \frac{1}{2}$

In the example $\{a,b,c\}, \{a,b,d\}, \{a,b,e\}$
$\{a,b,e\}, \{a,b,f\}$ are
neighbors as they have common a, b in them.
The points/objects are neighbors if

$$Sim(T_1, T_2) > t$$

$$Sim(T_1, T_2) = \frac{T_1 \cap T_2}{T_1 \cup T_2}$$

Links between the objects is the number common neighbors.

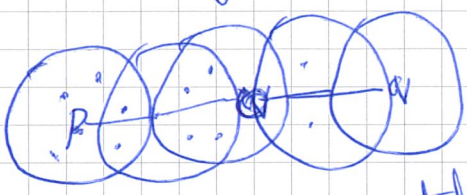
$$\text{Goodness Measure } (G_1) = \frac{\text{Number of links b/w clusters}}{\text{Number of expected links b/w the clusters}}$$

ROCK uses goodness Measure to combine clusters the ~~dust~~ items with more link are in the same cluster
If the two elements are neighbors there common neighbors are the elements ~~whit~~ which are similar to the nodes.

3,5

Q.4

If p is density Conected to q w.r.t Eps & Minpts then



If p is density ~~reachable~~ connected from q w.r.t. Eps & Minpts then it is that is p is directly density connected to o and q is directly density connected to point o - Both p and q are directly density connected to point o so they are density connected to each other - As p and q are density connected to each other so they are not density-reachable but they are directly density connected -?

For density reachable the points p and q should be density connected with the point o - There should be chain of points connecting p & q w.r.t. Eps and Minpts -

Q-4-b OPTICS

OpTics is used for finding the value of the epsilon - It is used to check how the epsilon will effect the cluster - It uses ordering of the objects - Optics ~~is~~ selected epsilon

value is used in the DBSCAN algorithm-

Q-5

	A	B	C	D	E	F	G
Item-K	gold	(0,0)	Y	N	Y	N	silver
Item-L	bronze	(1,1)	N	N	N	N	—
distance	1	1+1=2	1	0	1	0	~~0~~
delta	1	1	1	1	~~1~~	0	0

$M_{if} = \{1 \cdots M_f\}$

A and G are ordinal variables

{gold, silver, bronze} $m = 3$

gold → 1, silver → 2, bronze → 3

Gold $\Rightarrow \frac{r_f - 1}{M_f - 1} = \frac{1-1}{3-1} = \frac{0}{2} = 0$

Silver $\Rightarrow$ // $= \frac{2-1}{3-1} = \frac{1}{2}$

bronze $\Rightarrow$ // $= \frac{3-1}{3-1} = \frac{2}{2} = 1$

Taking Y = 1 and N = 0

E and F asymmetric

delta = 0 if x = 0 and f = asymetric or x = Missing

So $d(i,j) = \frac{\sum_{i=1}^{n} \delta \, d_{ij}}{\sum_{i=1}^{n} \delta}$

$= \frac{1+1+1+1}{5} = \frac{5}{5} = 1$

Q-6

(a) Apriori Property =

~~If a subset is infrequent all of its supersets will also be infrequent.~~

It states If a superset is frequent all of its not empty subsets will be frequent. 1

(b) We produce candidates in the apriori algorithm by combining the subset with each other after applying desending ordering to the item sets. How? 0

We generate candidates at each step or iteration of the apriori algorithm.

T.id	Items
1	A B C
2	A B E

→

A	2
B	2
C	1
E	1

→

AB	1
AC	1
AE	1
BC	1
BE	1
CE	0

→

ABC
ABE
ACE
BCE

→

(c) At the first step we check if the items satisfy the constraint we prune the items and do not take the item further. If the item do not satisfy the antimonotone constraint we take the item further. Because if an item do not satisfy the constraint none of its superset will satisfy the constraint in antimonotone constraint so we prune the item. 1

and check for the constraint again at the next step.

(d) Correctness of Apriori Algorithm =

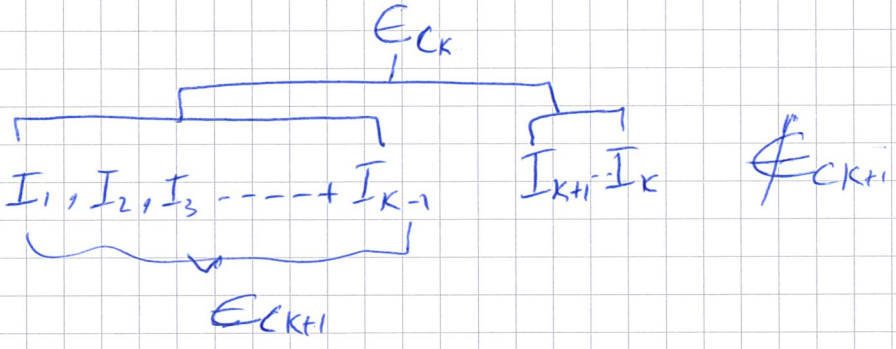
(1) Let $L_k \subseteq C_k$ ($k=0$) trivial case-

(2) Induction Hypothesis $L_{k+1} \subseteq C_{k+1}$ such that $1 \dots k$

(3) To prove $L_{k+1} \subseteq C_{k+1}$ -

Assume:

$$L_{k+1} \nsubseteq C_{k+1}$$



2

As self joining is producing $I_1 \dots I_2$ we are not able to prune the itemset so our assumtion $L_{k++} \nsubseteq C_{k+1}$ is denied -

Hence

$$L_{k+1} \subseteq C_{k+1} \text{ is True}$$

Q-7 FP grow Algorithm -

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

checking the frequency support of the items

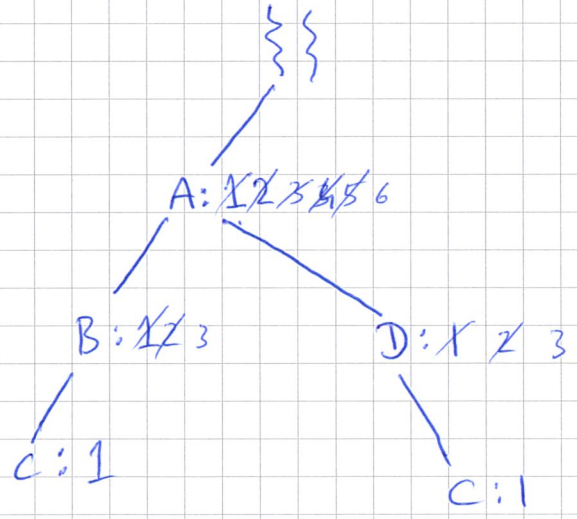
Items	Support
A	6
B	3
C	2
D	3

All of the items fullfill the minsupport requirment ≥ 1.

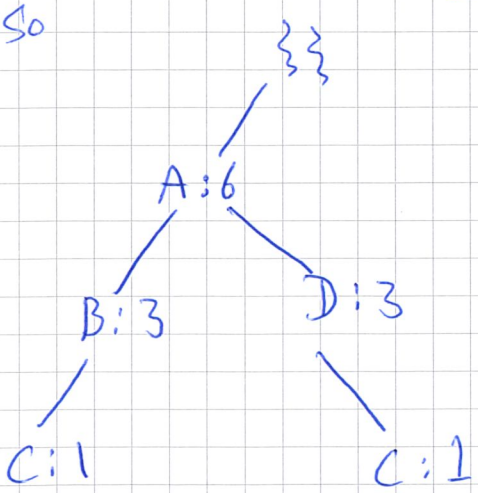
Arranging the items in the transaction database according to frequency -

$A \rightarrow B \rightarrow D \rightarrow C$

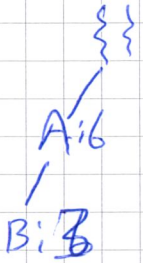
T.id	Items
1	A, B, C
2	A, D, C
3	A, B
4	A, B
5	A, D
6	A, D



So



=> Now we will check the α-conditional databases -



B-conditional database

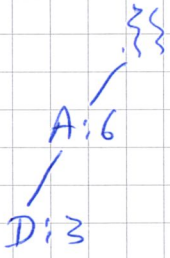
Items : A
Support : 1
Cond DB : AB : 3

AB - conditional database

Items : Null
ϕ

D- conditional database

Items : A
Support : 1
Cond DB : AD : 3



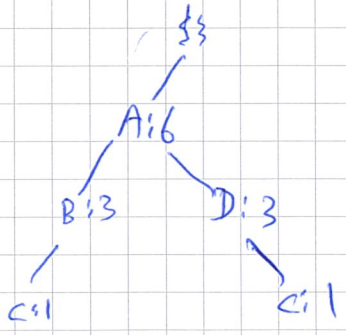
Output?
Under 4

AD - Conditional database

ϕ

C - conditional database

Items : A, B, D
Support : 2, 1, 1
Cond DB : AB : 1 , AD : 1
Cet : ABC , ADC



ABC - conditional database

Items : AB
Support : 1 1
Cond DB : ϕ

ADC - conditional database

Items : AD
Support : 1 1
Cond DB : ϕ

A-conditional database =

Items : {{

support : ∅

∅

{{

/

A:6

Q-8 (a)

Monotone

$sum(S) \geq V$ non-negative price

$range(S) \geq V$

$min(S) \leq V$

Antimonotone

$sum(S) \leq V$

$range(S) \leq V$

$min(S) \geq V$

Convertible Monotone But Not Monotone

$Avg(S) \geq V$ when items are in ~~decreasing~~ increasing order -

Convertible Antimonotone But Not Antimonotone

$Avg(S) \geq V$ when items are in ~~decreasing~~ increasing order -

(b) Frequent Itemset ABC

ABC

~~ABC~~ → AB → C $\quad \frac{sup(ABC)}{sup(AB)} = \frac{1}{3} = 0.33 = 33\%$

AC → B $\quad \frac{sup(ABC)}{sup(AB)} = \frac{1}{2} = 50\%$

BC → A $\quad \frac{sup(ABC)}{sup(BC)} = \frac{1}{1} = 100\%$

As AC → B and BC → A support are 50 % and 100 % respectively so we will not use them to check for others -

C → AB $\quad \frac{sup(ABC)}{sup(C)} = \frac{1}{2} = 0.5 = 50\%$

So the frequent Item ABC generate 3 association rules that satisfy condition of confidence greater or equal than 50 % 2

AC → B

BC → A

C → AB

(c) Association rule is a casual rule when the rule using items are independent on each other 0

Diper → Milk $\quad$ Milk → Diper