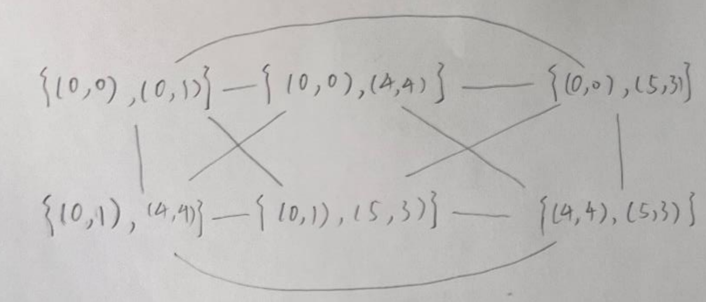
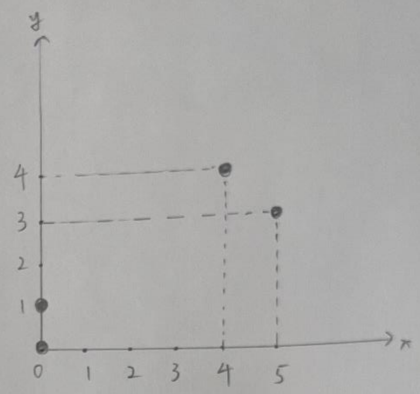
2022 3.

(i)





(ii) assume we start from $\{(0,0),(0,1)\}$

cost of $\{(0,0),(0,1)\} = d_{(0,1)-(4,4)} + d_{(0,1)-(5,3)}$
$= 7+7 = 14$

cost of $\{(0,0),(4,4)\} = d_{(0,0)-(0,1)} + d_{(4,4)-(5,3)}$
$= 1+2 = 3$

cost of $\{(0,0),(5,3)\} = d_{(0,0)-(0,1)} + d_{(5,3)-(4,4)}$
$= 1+2 = 3$

cost of $\{(0,1),(4,4)\} = d_{(0,1)-(0,0)} + d_{(4,4)-(5,3)}$
$= 1+2 = 3$

cost of $\{(0,1),(5,3)\} = d_{(0,1)-(0,0)} + d_{(5,3)-(4,4)}$
$= 1+2 = 3$

In the PAM, we will calculate all the neighbors and find one with minimum cost, here we have 4 neighboors equally cost 3, all of them represent a minimum cost, so we can randomly pick one from $\{(0,0),(4,4)\}$, $\{(0,0),(5,3)\}$, $\{(0,1),(4,4)\}$, $\{(0,1),(5,3)\}$

b.1) No. PAM always follow the local minimum, so it can only find local minimum.

2) Yes. As said above, PAM will follow optimal path in each iteration, so it can find local minimum.

3) No. CLARA is done on subdataset, so it can only find local optimum of subdata.

4) No. CLAPANS doesn't always follow the local minimum in each iteration, so it can not guarantee to find local optimum.

2. Hierarchical clustering

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

iteration 1

=>

	(1,5)	2	3	4
(1,5)	0			
2	7	0		
3	10	9	0	
4	8	2	5	0

iteration 2

=>

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

iteration 3

In iteration 1, we find $d_{(1,5)}=1$ is minimum.
So we combine (1,5) together

$d_{(1,5)-2} = \max[d_{(2,1)}, d_{(2,5)}] = 7$
$d_{(1,5)-3} = \max[d_{(3,1)}, d_{(3,5)}] = 10$
$d_{(1,5)-4} = \max[d_{(1,4)}, d_{(5,4)}] = 8$
} distance in iter 2

=>

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

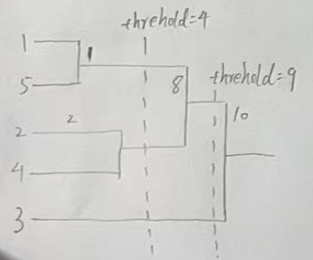
iteration 4

In iteration 2, we find $d_{(2,4)}=2$ is minimum
so, we combine (2,4) together.

$d_{(1,5)-(2,4)} = \max\{d_{(1,5)-2}, d_{(1,5)-4}\} = 8$
$d_{(2,4)-3} = \max\{d_{(2,3)}, d_{(4,3)}\} = 9$
} distance in iter3

In iteration 3, we find $d_{(1,5)-(2,4)}=8$ is minimum,
so, we combine [(1,5)(2,4)] together.

$d_{[(1,5),(2,4)]-3} = \max\{d_{(1,5)-3}, d_{(2,4)-3}\} = 10$ -> distance in iter4



(ii) From the graph in the left side, we can see

(i) threhold = 5 :
cluster 1 : (1,5)
cluster 2 : (2,4)
cluster 3 : (3)

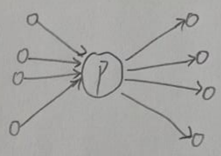
(ii) threhold = 9 :
cluster 1 : (1,2,4,5)
cluster 2 : (3)

3. Rock

(i) if the threshold = 0.6, then the neighbors of A are (A, B, C). The neighbors of are (A, B, C), so A, B have 3 common neighbors => link(A, B) = 3

(ii) The expected link is

$$n^{1+2m}$$



We consider the graph in the right side, for every point P in cluster C, it has p_i-th in-degree and p_j-th out-degree. Every point form in-degree and every point from out degree can be a pair whose common neighbor must contain P. So in this graph, P can contribute $p_i * p_j$ -th common neighbors.

Back to our problems, each neighbor can be in-degree or out-degree so for each point the expected link is $n^m \cdot n^m = n^{2m}$. And we have n-th points here, so the expected links are

$$n^{1+2m}$$

(iii) The expected link here works as a kind of normolization. If there is no such normolization the greatest goodness will be reached when we put all points into a single cluster. So, it is necessary to use this normolization.

4. Density-based clustering

1) False. If p and q are density connected, then it means there is a core "A" that p and q are density reachable from "A". So, "q is density from p" is not guaranteed.

2) False. "Directly density reachable" requires core point. There is no information showing p is a core point.

3) True. From the statement, we can know q is a core point. Meanwhile, there is a path from q to p: $p_1, p_2, \cdots p_n$, $p_1 = q$, $p_n = p$ and p_{i+1} is directly density reachable from p_i. So for p_{n-1}, both q and p are density reachable from p_{n-1}, so p and q are density connected.

4) True. If p and q are density connected, then there is an point "O" where p and q are density reachable from. So, we can see density connected is a symmetric condition.

5. Distance Measure

a.

$d_{KL}(A) = \sqrt{[3-(-2)]^2} = 5$ $\quad \delta_{KL}(A)=1$

$d_{KL}(B) = |2-(-1)|+|3-2| = 4$ $\quad \delta_{KL}(B)=1$

$d_{KL}(C) = 0$ $\quad \delta_{KL}(C)=1$

$d_{KL}(D) = 1$ $\quad \delta_{KL}(D)=1$

$d_{KL}(E) = 0$ $\quad \delta_{KL}(E)=1$

$d_{KL}(F) = 0$ $\quad \delta_{KL}(F)=0$

$d_{KL}(G) = 0$

$$d_{KL} = \frac{\sum \delta_{ij}(f)\cdot d_{ij}(f)}{\sum \delta_{ij}(f)} = \frac{5+4+1}{5} = 2$$

b. We will have the same answer.

If we us asymmetric variable only, we may have a table looks like below

	Y	N
Y	a	b
N	c	d

then the distance is $\frac{b+c}{a+b+c}$

If using the formula in question (a), the (Y,N) and (N,Y) will be counted into distance. For denominator, (Y,Y) (Y,N) (N,Y) will contribute. From the table, we have a-th pairs of (Y,Y), b-th pairs of (Y,N) and c-th pairs of (N,Y). So, the result would be

$\frac{b+c}{a+b+c}$.

Now, we can see they will have the same answer.

6. Aprior

a.

C_1

items	sup
A	4
B	4
C	4
D	4
E	2

$\longrightarrow L_1$

item	sup
A	4
B	4
C	4
D	4
E	2

$\longrightarrow C_2$

item	sup
AB	3
AC	3
AD	3
AE	2
BC	3
BD	3
BE	2
CD	3
~~CE~~	~~1~~
~~DE~~	~~1~~

$\longrightarrow L_2$

item	sup
AB	3
AC	3
AD	3
AE	2
BC	3
BD	3
BE	2
CD	3

$\longrightarrow C_3$

item	sup
ABC	2
~~ABD~~	~~1~~
ABE	2
ACD	2
~~ACE~~	~~1~~
~~ADE~~	~~1~~
BCD	2
~~BCE~~	~~1~~
~~BDE~~	~~1~~

$\longrightarrow L_3$

item	sup
ABC	2
ABE	2
ACD	2
BCD	2

$\longrightarrow C_4$

item	sup
~~ABCE~~	~~1~~
~~ABCD~~	~~1~~

$\longrightarrow L_4$

item	sup
null	null

output

L_1: {A}, {B}, {C}, {D}, {E}

L_2: {A,B}, {A,C}, {A,D}, {A,E}, {B,C}, {B,D}, {B,E}, {C,D}

L_3: {A,B,C}, {A,B,E}, {A,C,D}, {B,C,D}

6. b

This is antimonotone constraint.

C_1

item	sup	value
A	4	0
B	4	0
C	4	0
D	4	0
E	2	0

$\rightarrow L_1$

item	sup	range
A	4	0
B	4	0
C	4	0
D	4	0
E	2	0

$\rightarrow C_2$

item	sup	range
AB	3	1
AC	3	2
~~AD~~	~~3~~	~~3~~
~~AE~~	~~2~~	~~4~~
BC	3	1
BD	3	2
~~BE~~	~~2~~	~~3~~
CD	3	1
~~CE~~	~~1~~	~~2~~
~~DE~~	~~1~~	~~1~~

$\rightarrow L_2$

item	sup	range
AB	3	1
AC	3	2
BC	3	1
BD	3	2
CD	3	1

By using constraint, we move AD, AE, BE out of L2

$C_3 \rightarrow$

item	sup	range
ABC	2	2
~~ABD~~	~~1~~	~~3~~
~~ACD~~	~~2~~	~~3~~
BCD	2	2

$\rightarrow L_3$

item	sup	range
ABC	2	2
BCD	2	2

By applying constraint, we move ACD out.

$\rightarrow C_4$

item	sup	range
ABCD	1	3

$\rightarrow L_4$ empty

So: {A} {B} {C} {D} {E}
{AB}, {AC}, {BC}, {BD}, {CD}
{ABC}, {BCD}

C. This is monotone constraint

C_1

item	sup	min(S)
A	4	5
B	4	4
C	4	3
D	4	2
E	2	1

$\rightarrow L_1$

item	sup	min(S)
A	4	5
B	4	4
C	4	3
D	4	2
E	2	1

In L_1, by applying constraint, we can only output {D}, {E}

$\rightarrow C_2$

item	sup	min(S)
AB	3	4
AC	3	3
AD	3	2
AE	2	1
BC	3	3
BD	3	2
BE	2	1
CD	3	2
~~CE~~	~~1~~	~~1~~
~~DE~~	~~1~~	~~1~~

$\rightarrow L_2$

item	sup	min(S)
AB	3	4
AC	3	3
AD	3	2
AE	2	1
BC	3	3
BD	3	2
BE	2	1
CD	3	2

In L_2, by applying the constraint, we can only output {A,D}, {A,E}, {B,D}, {B,E}
{C,D}

$\rightarrow C_3$

item	sup	min(S)
ABC	2	3
~~ABD~~	~~1~~	~~2~~
ABE	2	1
ACD	2	2
~~ACE~~	~~1~~	~~1~~
~~ADE~~	~~1~~	~~1~~
BCD	2	2
~~BCE~~	~~1~~	~~1~~
~~BDE~~	~~1~~	~~1~~

$\rightarrow L_3$

item	sup	min(S)
ABC	2	3
ABE	2	1
ACD	2	2
BCD	2	2

In L_3, by applying constraint, we output {A,B,E}, {A,C,D}, {B,C,D}

$\rightarrow C_4$

item	sup	min(S)
~~ABCE~~	~~1~~	~~1~~
~~ABCD~~	~~1~~	~~2~~

$\rightarrow L_4$ empty

So, the output is {D}, {E}
{A,D}, {A,E}, {B,D}, {B,E}, {C,D}
{A,B,E}, {A,C,D}, {B,C,D}

7. FP grow algorithm

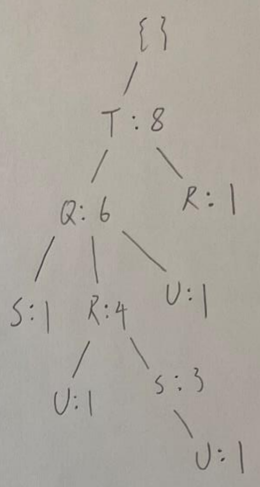
a.

item	sup
T	8
Q	6
S	4
U	3
R	5
~~P~~	~~1~~

(1)

Tid	items
1	T
2	T Q S
3	T Q U
4	T R
5	T Q R U
6	T Q R S
7	T Q R S U
8	T Q R S

We can iterate each Tid and build the tree.



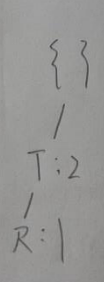
b.

Using the table (1) from question a. This is an antimonotone constraint

Tid	items	Range
1	T	0
~~2~~	~~T Q S~~	~~3~~
~~3~~	~~T Q U~~	~~4~~
4	T R	2
~~5~~	~~T Q R U~~	~~4~~
~~6~~	~~T Q R S~~	~~3~~
~~7~~	~~T Q R S U~~	~~4~~
~~8~~	~~T Q R S~~	~~3~~

After applying constraint, we only have

Tid	items	Range
1	T	0
4	TR	2



The conditional database is

item	
T	{ }
R	T : 1

8. Rule Generation

{F, H, J, L}

① FHJ -> L $P(F,H,J,L) = \frac{5}{10}$, $P(F,H,J) = \frac{5}{10}$, $confidence = \frac{P(FHJL)}{P(FHJ)} = 1 > 60\%$, output

② FH -> JL $P(FHJL) = \frac{5}{10}$, $P(FH) = \frac{6}{10}$, $confidence = \frac{P(FHJL)}{P(FH)} = \frac{5}{6} > 60\%$, output

③ F -> HJL $P(FHJL) = \frac{5}{10}$, $P(F) = \frac{6}{10}$, $confidence = \frac{P(FHJL)}{P(F)} = \frac{5}{6} > 60\%$, output

④ H -> FJL $P(FHJL) = \frac{5}{10}$, $P(H) = \frac{9}{10}$, $confidence = \frac{P(FHJL)}{P(H)} = \frac{5}{9} < 60\%$ ✗

⑤ FJ -> HL $P(FHJL) = \frac{5}{10}$, $P(FJ) = \frac{5}{10}$, $confidence = \frac{P(FHJL)}{P(FJ)} = 1 > 60\%$, output

⑥ F -> HJL already discussed ③

⑦ J -> FHL $P(FHJL) = \frac{5}{10}$, $P(J) = \frac{8}{10}$, $confidence = \frac{P(FHJL)}{P(J)} = \frac{5}{8} > 60\%$, output

⑧ HJ -> FL $P(FHJL) = \frac{5}{10}$, $P(HJ) = \frac{7}{10}$, $confidence = \frac{P(FHJL)}{P(HJ)} = \frac{5}{7} > 60\%$, output

⑨ H -> JFL already discussed ④

⑩ J -> HFL already discussed ⑦

⑪ FHL -> J $P(FHJL) = \frac{5}{10}$, $P(FHL) = \frac{5}{10}$, $confidence = \frac{P(FHJL)}{P(FHL)} = 1 > 60\%$, output

⑫ FH -> JL already discussed ②

⑬ F -> HJL already discussed ⑥

⑭ H -> FJL already discussed ④

⑮ FL -> HJ $P(FHJL) = \frac{5}{10}$, $P(FL) = \frac{5}{10}$, $confidence = \frac{P(FHJL)}{P(FL)} = 1 > 60\%$, output

8. Rule Generation

(16) F->HLJ already discussed (6)

(17) L->FHJ $P(FHJL) = \frac{5}{10}$ $P(L) = \frac{6}{10}$ confidence $= \frac{P(FHJL)}{P(L)} = \frac{5}{6} > 60\%$, output

(18) HL->FJ $P(FHJL) = \frac{5}{10}$ $P(HL) = \frac{6}{10}$ confidence $= \frac{P(FHJL)}{P(HL)} = \frac{5}{6} > 60\%$, output

(19) H->FJL already discussed (4)

(20) L->FHJ already discussed (17)

(21) FJL->H $P(FJLH) = \frac{5}{10}$ $P(FJL) = \frac{5}{10}$ confidence $= \frac{P(FHJL)}{P(FJL)} = 1 > 60\%$, output

(22) FJ->LH already discussed (5)

(23) F->LHJ already discussed (3)

(24) J->FLH already discussed (7)

(25) FL->JH already discussed (15)

(26) F->LJH already discussed (6)

(27) L->FJH already discussed (17)

(28) JL->FH $P(FHJL) = \frac{5}{10}$ $P(JL) = \frac{6}{10}$ confidence $= \frac{P(FHJL)}{P(JL)} = \frac{5}{6} > 60\%$, output

(29) J->FHL already discussed (7)

(30) L->FJH already discussed (17)

(31) HJL->F $P(HJFL) = \frac{5}{10}$ $P(HJL) = \frac{6}{10}$ confidence $= \frac{5}{6} > 60\%$, output

(32) HJ->FL already discussed (12) ; (33) H->JFL already discussed (4) ; (34) J->FHL already discussed (7)

(35) HL->FJ already discussed (18) ; (36) H->FJL already discussed (4) ; (37) L->FJH already discussed (17)

(38) JL->FH $P(FHJL) = \frac{5}{10}$ $P(JL) = \frac{6}{10}$ confidence $= \frac{P(FHJL)}{P(JL)} = \frac{5}{6} > 60\%$, output ; (39) J->FHL already discussed (7) ; (40) L->FJH already discussed (17)

Finally, we will output

FHJ->L	J->FHL	L->FHJ	HJL->F
FH->JL	HJ->FL	HL->FJ	JL->FH
F->HJL	FHL->J	FJL->H	
FJ->HL	FL->HJ	JL->FH	

Maybe we can have fewer items in question 8. :)