a) Apriori P_1

support(X,Y) = P(X,Y) = $\frac{number(XY)}{num(All\ Samples)}$

confidence (A⇒B) = P(B|A) = $\frac{support(A \cup B)}{support(A)}$.

30%×10=3.

Data

1. {a,b,d,e}
2. {b,c,d}
3. {a,b,d,e}
4. {a,c,d,e}
5. {b,c,d,e}
6. {b,d,e}
7. {c,d}
8. {a,b,c}
9. {a,d,e}
10. {b,d}.

Scan D →

C_2

itemset	sup.
{a}	5
{b}	7
{c}	5
{d}	9
{e}	6

cut →

L_2

itemset	sup.
{a}	5
{b}	7
{c}	5
{d}	9
{e}	6.

Scan D →

C_3

itemset	sup.
{a,b}	3
{a,c}	2 ×
{a,d}	4
{a,e}	4
{b,c}	3
{b,d}	6
{b,e}	4
{c,d}	4
{c,e}	2 ×.
{d,e}	6

Cut →

L_3.

itemset	sup.
{a,b}	3
{a,d}	4
{a,e}	4
{b,c}	3
{b,d}	6
{b,e}	4
{c,d}	4
{d,e}.	6.

ScanD →

C4.

itemset.	sup.
{a,b,d}	丅 2 ×
{a,b,e}.	丅 2. ×
{a,d,e}	止 4
{b,c,d}	丅 2. ×
~~{b,c,e}~~.	
{b,d,e}	止 4.

Cut →

L4.

itemset	sup.
{a,d,e}	4
{b,d,e}	4.

综上，频繁项集为 {a,d,e}, {b,d,e}, {a,b}, {a,d}, {a,e}, {b,c}, {b,d}, {b,e}, {c,d}, {d,e}, {a}, {b}, {c}, {d}, {e}.

(2) 由 {a,d,e} 得到关联规则

{a,d} ⇒ e, confidence = $\frac{support(\{a,d,e\})}{support(\{a,d\})} = \frac{4}{4} = 1 > 0.5$

{a,e} ⇒ d, confidence = $\frac{4}{4} = 1 > 0.5$

{d,e} ⇒ a, confidence = $\frac{4}{6} \approx 0.67 > 0.5$

a ⇒ {d,e}, confidence = $\frac{4}{5} = 0.8 > 0.5$

d ⇒ {a,e}, confidence = $\frac{4}{9} \approx 0.44 < 0.5$ 舍弃

e ⇒ {a,d}, confidence = $\frac{4}{6} \approx 0.67 > 0.5$

由 {b,d,e} 得到关联规则

{b,d} ⇒ e, confidence = $\frac{4}{6} = 0.67$

{b,e} ⇒ d, confidence = $\frac{4}{4} = 1$.

{d,e} ⇒ b, confidence = $\frac{4}{6} = 0.67$

b ⇒ {d,e}, confidence = $\frac{4}{7} = 0.57$

d ⇒ {b,e}, confidence = $\frac{4}{9} = 0.44 < 0.5$ 舍弃.

e ⇒ {b,d}, confidence = $\frac{4}{6} = 0.67$

由 {a,b} 得到关联规则

a ⇒ b, confidence = $\frac{3}{5} = 0.6$

b ⇒ a, confidence = $\frac{3}{7} < 0.5$ 舍弃.

由 {a,d} 得到关联规则.

a⇒d, confidence = $\frac{4}{5}$

d⇒a, confidence = $\frac{4}{9} < 0.5$ 舍弃.

由 {a,e} 得到关联规则

a⇒e, confidence = $\frac{4}{5} = 0.8$

e⇒a, confidence = $\frac{4}{6} = 0.67$

由{b,c}得到关联规则

$b \Rightarrow c$, confidence $= \frac{3}{7} < 0.5$ 舍弃.

$c \Rightarrow b$, confidence $= \frac{3}{5} = 0.6$.

由{b,d}得到关联规则.

$b \Rightarrow d$, confidence $= \frac{6}{7}$.

$d \Rightarrow b$, confidence $= \frac{6}{9}$

由{b,e}得到关联规则

$b \Rightarrow e$, confidence $= \frac{4}{7}$

$e \Rightarrow b$, confidence $= \frac{4}{6}$.

由{c,d}得到关联规则

$c \Rightarrow d$, confidence $= \frac{4}{5} = 0.9$

$d \Rightarrow c$, confidence $= \frac{4}{9} < 0.5$ 舍弃.

由{d,e}得到关联规则

$d \Rightarrow e$, confidence $= \frac{6}{9} = 0.67$

$e \Rightarrow d$, confidence $= \frac{6}{6} = 1$

综上,关联规则有 $\{a,d\} \Rightarrow e$, $\{a,e\} \Rightarrow d$, $\{d,e\} \Rightarrow a$. $a \Rightarrow \{d,e\}$, $e \Rightarrow \{a,d\}$.

$\{b,d\} \Rightarrow e$, $\{b,e\} \Rightarrow d$, $\{d,e\} \Rightarrow b$, $b \Rightarrow \{d,e\}$, $e \Rightarrow \{b,d\}$.

$a \Rightarrow b$, $a \Rightarrow e$, $e \Rightarrow a$, $c \Rightarrow b$, $b \Rightarrow d$, $d \Rightarrow b$, $b \Rightarrow e$, $e \Rightarrow b$.

$c \Rightarrow d$, $d \Rightarrow e$, $e \Rightarrow d$. $a \Rightarrow d$.

(2) FP 增长算法

2. 关联 (25 分)

a) 针对图 2 的交易事务数据,采用 FP 增长算法求取频繁项集,

假设最小支持度为≥30%

事务 ID	购买项
1	{a, b, d, e}
2	{b, c, d}
3	{a, b, d, e}
4	{a, c, d, e}
5	{b, c, d, e}
6	{b, d, e}
7	{c, d}
8	{a, b, c}
9	{a, d, e}
10	{b, d}

图 2 交易事务数据

a 5
b 7
c 5
d 9
e 6

d b e

b) 基于上述频繁项集,构造关联规则,要求最小置信度≥50%

a) FP 增长算法求频繁项集

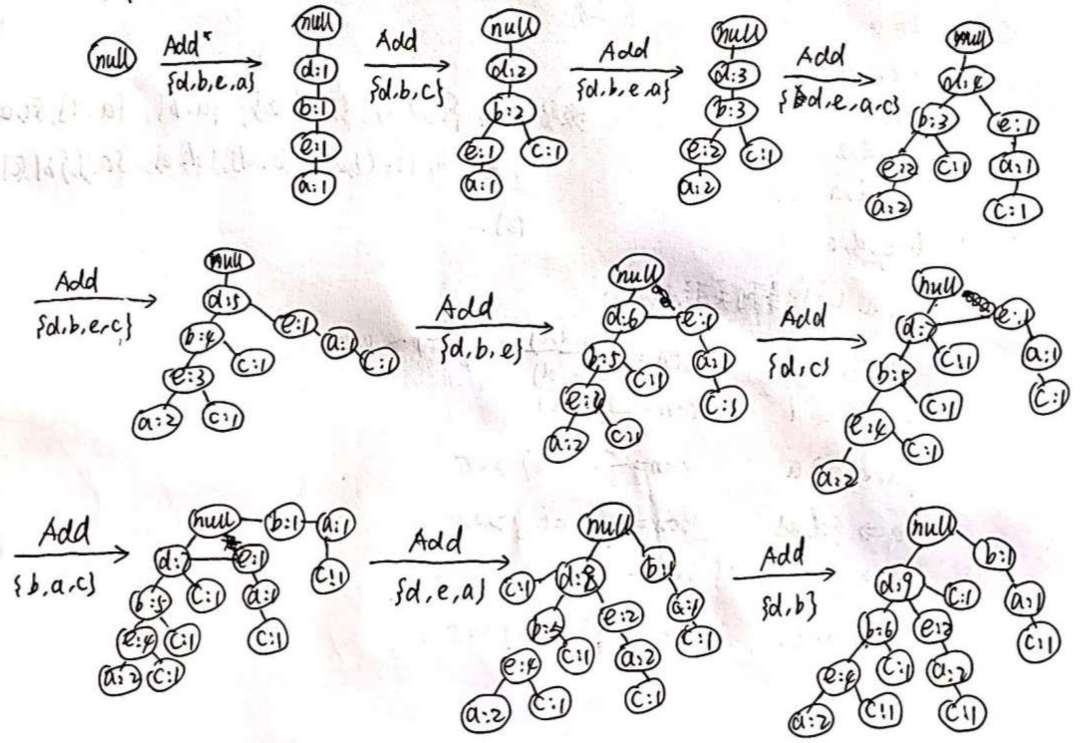
1. 过滤和重排序

事务ID	购买项	过滤和重排序
1	a,b,d,e	d,b,e,a
2	b,c,d	d,b,c
3	a,b,d,e	d,b,e,a
4	a,c,d,e	d,e,a,c
5	b,c,d,e	d,b,e,c
6	b,d,e	d,b,e
7	c,d	d,c
8	a,b,~~d~~c	b,a,c
9	a,d,e	d,e,a
10	b,d	d,b

计数	
a	5
b	7
c	5
d	9
e	6

降序	
d	9
b	7
e	6
a	5
c	5

2. 构造FP树:



3. 抽取条件模式基

表头信息	
d	9
b	7
e	6
a	5
c	5

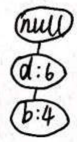
项	条件模式基
b	d:6
e	db:4, d:2
a	dbe:2, de:2, b:1
c	dbe:1, db:1, dea:1, d:1, ba:1

4. 构造条件FP树

b-FP树

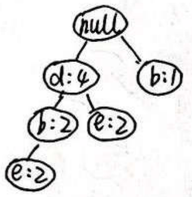
→ {b,d} 6

e-FP树



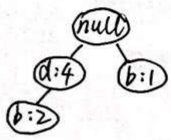
→ {b,e} 4, {d,e} 6, {b,d,e} 4

a-FP树



→ {a,b} 3, {a,d} 4, {a,e} 4, {a,d,e} 4.

c-FP树



→ {b,c} 3, {c,d} 4, ~~{b,c,d}~~

综上，频繁项集为…:

b) 同 Apriori 算法

3. 特征

2. 特征

给定图 2 中的目标集（DOG）和对比集（CAT），使用信息增益计算各个属性与当前概念描述任务之间的相关性。并采用 T=0.1 作为阈值，对属性进行筛选。

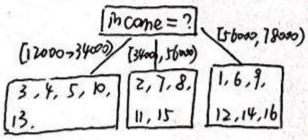
Gender	Tail	Weight	Count
M	Long	5-10	2
M	Middle	0-5	3
F	Long	5-10	3
M	Middle	10-15	1
M	Short	10-15	3
F	Long	15-20	3

Gender	Tail	Weight	Count
M	Long	0-5	2
F	Middle	5-10	1
F	Short	0-5	2
F	Long	5-10	2
M	Middle	0-5	1
F	Short	5-10	2

图 2 目标集 DOG（左）、对比集 CAT（右）

3° 构造决策树

① 由 Gain(income) > Gain(student) > Gain(age)，构造第一级决策树：



② 对 income ∈ [12000, 34000)，其信息 I(3, 2, 0) ≐ 0.971

对于Age：

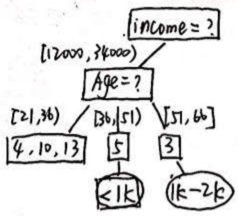
	<1k	1k-2k	>2k	
[21, 36)	2	1	0	I(2,1,0) = 0.918
[36, 51)	1	0	0	0
[51, 66]	0	1	0	0

$E(Age) = \frac{3}{5} \times 0.918 \doteq 0.551$

$Gain(Age) = 0.971 - 0.551 = 0.420$

同理，对student，Gain(student) = 0.322

由于 Gain(Age) > Gain(Student).

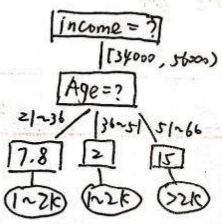


对income ∈ [34000, 56000)，I(0, 4, 1) = 0.722

对Age： E(Age) = 0， Gain(Age) = 0.722 − 0 = 0.722

对Stu： $E(Stu) = \frac{3}{5} \times (-\frac{2}{3}\log_2\frac{2}{3} - \frac{1}{3}\log_2\frac{1}{3})$ Gain(Stu) = 0.722 − E(Stu) = 0.17

由Gain(Age) > Gain(Stu).

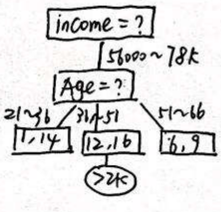


对income ∈ [56000, 78000)，I(1, 1, 4) ≠

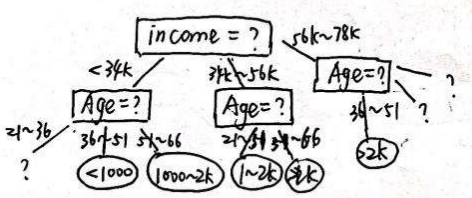
对Age：

对Stu： E(Stu) = I(1, 1, 4)， Gain(Stu) = 0

由Gain(Age) > Gain(Stu).



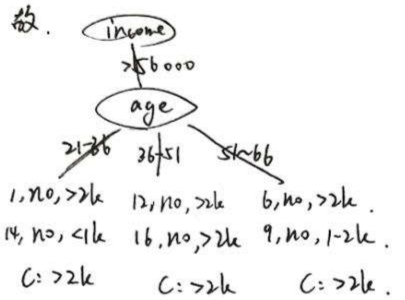
~~对第三层用student分得~~



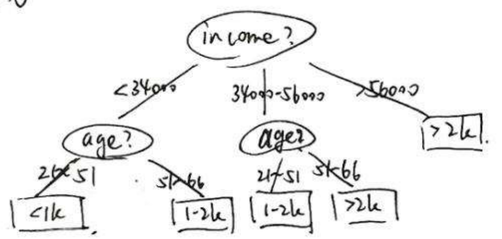
~~同理~~对于 income >56000.

注意到 数据的student属性值均为no，所以 Gain(student)=0.

故.



综上，决策树为



c) 根据 b) 得出的决策树，元组 (24, 75000, yes) 分类到 >2000

当决策树无法再进行分类时，如何确定其类别？

(2) 朴素贝叶斯方法

3. 数据预处理与分类（25分）

a) 针对图3中训练数据集进行离散化处理。要求采用等宽分桶的方式将age和incoming属性离散到3个区间。

b) 依据训练集，采用朴素贝叶斯方法分类未知元组(24, 75000, yes)，对分类属性 Class: buys_MP 进行预测。

ID	age	income	student	Class:buys_MP
1	23	68000	no	>2000
2	49	36000	no	1000..2000
3	55	22000	no	1000..2000
4	34	30000	yes	<1000
5	38	15000	yes	<1000
6	57	75000	no	>2000
7	21	52000	no	1000..2000
8	31	45000	yes	1000..2000
9	66	58000	no	1000..2000
10	34	12000	yes	<1000
11	40	40000	yes	1000..2000
12	50	78000	no	>2000
13	29	20000	yes	1000..2000
14	25	70000	no	<1000
15	61	55000	no	>2000
16	45	65000	no	>2000

图 3 训练数据集

k-means.

a.) Manhattan D: $d(x,y) = |x_1 - y_1| + |x_2 - y_2| + \cdots + |x_p - y_p|$

Euclidean D: $d(x,y) = \sqrt{(x_1-y_1)^2 + (x_2-y_2)^2 + \cdots + (x_3-y_3)^2}$.

$$D = \begin{array}{c|cccccccccccc} & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 \\ \hline 1 & 0 \\ 2 & 2 & 0 \\ 3 & 3 & 3 & 0 \\ 4 & 1 & 3 & 4 & 0 \\ 5 & 6 & 6 & 5 & 7 & 0 \\ 6 & 1 & 3 & 4 & 2 & 5 & 0 \\ 7 & 10 & 12 & 13 & 9 & 8 & 9 & 0 \\ 8 & 6 & 6 & 3 & 7 & 6 & 5 & 14 & 0 \\ 9 & 3 & 1 & 4 & 4 & 7 & 4 & 10 & 7 & 0 \\ 10 & 6 & 6 & 3 & 7 & 2 & 5 & 10 & 4 & 7 & 0 \\ 11 & 5 & 7 & 8 & 4 & 7 & 4 & 5 & 9 & 8 & 7 & 0 \\ 12 & 4 & 6 & 7 & 3 & 8 & 3 & 6 & 8 & 7 & 8 & 10 \end{array}$$

b. 以下为k-means聚类过程.

Loop 1

1-1 分配

center\points	1(3,5)	2(2,6)	3(3,8)	4(3,4)	5(7,7)	6(4,5)	7(9,1)	8(4,10)	9(1,6)	10(6,8)	11(5,2)	12(4,2)
(3,5)	0✓	/	/	1✓	6	1✓	10✓	6	3	6	5✓	4✓
(2,6)	/	0✓	/	3	6	3	12	6	1✓	6	7	6
(3,8)	/	/	0✓	4	5✓	4	13	3✓	4	3✓	8	7

1-2 更新中心点

center 1.

$x_1 = \frac{3+3+4+9+5+4}{6} = 4.667$ $y_1 = \frac{5+4+5+1+2+2}{6} = 3.167$.

center2

$x_2 = \frac{2+1}{2} = 1.5$ $y_2 = \frac{6+6}{2} = 6$.

center3

$x_3 = \frac{3+7+4+6}{4} = 5$ $y_3 = \frac{8+7+10+8}{4} = 8.25$.

Loop 2.

2-1 分配

center\points	1(3,5)	2(2,6)	3(3,8)	4(3,4)	5(7,7)	6(4,5)	7(9,1)	8(4,10)	9(1,6)	10(6,8)	11(5,2)	12(4,2)
(4.667,3.167)	3.5	5.5	6.5	2.5✓	6.166	2.5✓	6.5✓	7.5	6.5	6.17	1.5✓	1.884
(1.5,6)	2.5✓	0.5✓	3.5	3.5	6.5	3.5	12.5	6.5	0.5✓	6.5	7.5	6.5
(5,8.25)	3.25	5.25	2.25✓	6.25	3.25✓	4.25	11.25	2.75✓	6.25	1.25✓	6.25	7.25

2-2 更新中心点. center1.

$x_1 = \frac{3+4+9+5+4}{5} = 5.2$ $y_1 = \frac{4+5+1+2+2}{5} = 2.8$

center2.

$x_2 = \frac{3+2+1}{3} = 2$ $y_2 = \frac{5+6+6}{3} = 5.667$

center3

$x_3 = \frac{3+7+4+6}{4} = 5$ $y_3 = \frac{8+7+10+8}{4} = 8.25$.

Loop3.

3-1 分配

center\points	1(3,5)	2(2,6)	3(3,8)	4(3,4)	5(7,7)	6(4,5)	7(9,1)	8(4,10)	9(1,6)	10(6,8)	11(5,2)	12(4,2)
(5.2, 2.8)	4.4	6.4	7.4	3.4	6.0	3.4	5.6✓	8.4	7.4	6.0	1.0✓	2.0✓
(2, 5.667)	1.67✓	0.33✓	3.33	2.67✓	6.33	2.67✓	11.67	6.33	1.33✓	6.33	6.67	5.67
(5, 8.25)	3.25	5.25	2.25✓	6.25	3.25✓	4.25	11.25	2.75✓	6.25	1.25✓	6.25	7.25

3-2 更新中心点. center1

$x_1 = \frac{9+5+4}{3} = 6$ $y_1 = \frac{1+2+2}{3} = 1.67$

center2 $x_2 = \frac{3+2+3+4+1}{5} = 2.6$ $y_2 = \frac{5+6+4+5+6}{5} = 5.2$

center3 $x_3 = \frac{3+7+4+6}{4} = 5$ $y_3 = \frac{8+7+10+8}{4} = 8.25$

k-means P2.

loop 4.

3-1 分配

center\points.	1(3,5)	2(2,6)	3(3,8)	4(3,4)	5(7,7)	6(4,5)	7(9,1)	8(4,10)	9(1,6)	10(6,8)	11(5,2)	12(4,2)
(6, 1.67)	6.33	8.33	9.33	5.33	6.33	5.33	3.67 ✓	10.33	9.33	6.33	1.33 ✓	2.33 ✓
(2.6, 5.2)	0.6 ✓	1.4 ✓	3.2	1.6 ✓	6.2	1.6 ✓	10.6	6.2	2.4 ✓	6.2	5.6	4.6
(5, 8.25)	5.25	5.25	2.25 ✓	6.25	3.25 ✓	4.25	11.25	2.75 ✓	6.25	1.25 ✓	6.25	7.25

注意到聚类不发生变化,循环终止.

聚类结果为 C_1: 7(9,1) 11(5,2) 12(4,2)

C_2: 1(3,5) 2(2,6) 4(3,4) 6(4,5) 9(1,6)

C_3: 3(3,8) 5(7,7) 8(4,10) 10(6,8)

b) 凝聚式层次式方法——平均曼哈顿距离

4. 聚类 （25 分）

a) 针对图 4 中的数据，采用曼哈顿距离作为距离函数，给出对应的相异矩阵。

b) 采用凝聚式层次式方法对该数据集进行聚类，聚类间的距离使用聚类中数据的平均曼哈顿距离进行度量。即：给定聚类 C_i，C_j，它们之间的平均曼哈顿距离距为：$d(C_i, C_j)=\frac{1}{ni*nj}\sum_{p\in Ci}\sum_{p'\in Cj}|p-p'|$，其中：|p-p'|为对象 p 和 p' 之间的曼哈顿距离，ni 是聚类 C_i 中对象的数目。

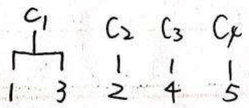
ID	x	y
1	3	8
2	2	7
3	4	8
4	3	4
5	4	5

图 4 聚类数据集

a)

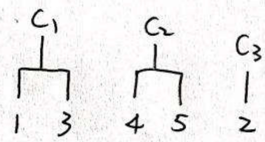
$$D = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{bmatrix} 0 & & & & \\ 2 & 0 & & & \\ 1 & 3 & 0 & & \\ 4 & 4 & 5 & 0 & \\ 4 & 4 & 3 & 2 & 0 \end{bmatrix} \end{array}$$

b) 由D得，第一次聚类结果为：



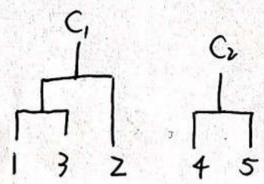
$$D_1 = \begin{array}{c} \\ C_1 \\ C_2 \\ C_3 \\ C_4 \end{array} \begin{array}{c} \begin{array}{cccc} C_1 & C_2 & C_3 & C_4 \end{array} \\ \begin{bmatrix} 0 & & & \\ 1.67 & 0 & & \\ 3 & 2 & 0 & \\ 2.33 & 2 & 1 & 0 \end{bmatrix} \end{array}$$

由D_1得，第二次聚类结果为：



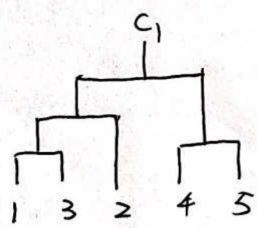
$$D_2 = \begin{array}{c} \\ C_1 \\ C_2 \\ C_3 \end{array} \begin{array}{c} \begin{array}{ccc} C_1 & C_2 & C_3 \end{array} \\ \begin{bmatrix} 0 & & \\ 4 & 0 & \\ 1.67 & 2.67 & 0 \end{bmatrix} \end{array}$$

由D_2得，第三次聚类结果为：



$$D_3 = \begin{array}{c} \\ C_1 \\ C_2 \end{array} \begin{array}{c} \begin{array}{cc} C_1 & C_2 \end{array} \\ \begin{bmatrix} 0 & \\ 4.8 & 0 \end{bmatrix} \end{array}$$

由D_3得，第四次聚类结果为：



c)　凝聚式层次式方法——最大距离

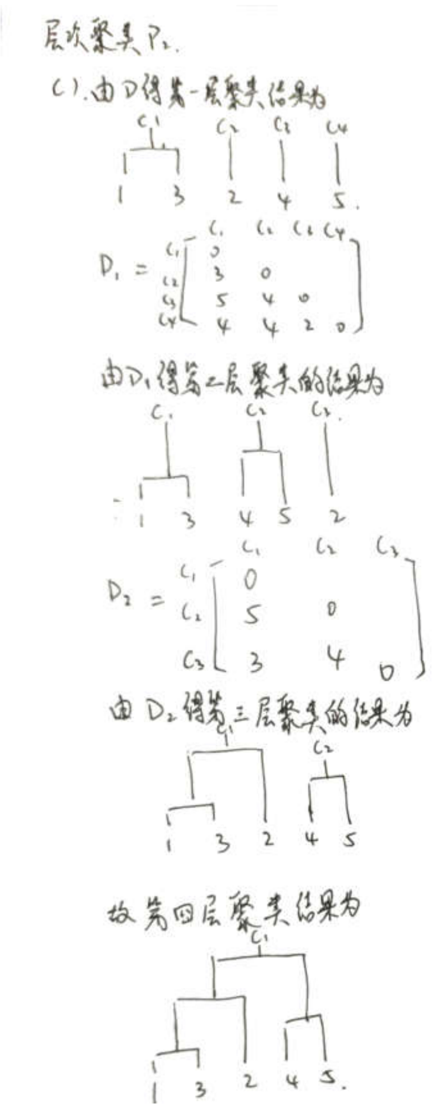
聚类（25 分）

a) 针对图 5 中的数据，采用曼哈顿距离作为距离函数，给出对应的相异矩阵。

b) 采用凝聚式层次式方法对该数据集进行聚类，聚类间的距离使用聚类中数据之间的最大距离进行度量。

ID	x	y
1	3	8
2	2	7
3	4	8
4	3	4
5	4	5

图 5 聚类数据集



d) 基于密度的方法对该数据进行聚类

DBSCAN 算法有两个参数：半径 eps 和密度阈值 MinPts，具体步骤为：

1、以每一个数据点 xi 为圆心，以 eps 为半径画一个圆圈。这个圆圈被称为 xi 的 eps 邻域

2、对这个圆圈内包含的点进行计数。如果一个圆圈里面的点的数目超过了密度阈值 MinPts，那么将该圆圈的圆心记为核心点，又称核心对象。如果某个点的 eps 邻域内点的个数小于密度阈值但是落在核心点的邻域内，则称该点为边界点。既不是核心点也不是边界点的点，就是噪声点。

3、核心点 xi 的 eps 邻域内的所有的点，都是 xi 的直接密度直达。如果 xj 由 xi 密度直达，xk 由 xj 密度直达。。。xn 由 xk 密度直达，那么，xn 由 xi 密度可达。这个性质说明了由密度直达的传递性，可以推导出密度可达。

4、如果对于 xk，使 xi 和 xj 都可以由 xk 密度可达，那么，就称 xi 和 xj 密度相连。将密度相连的点连接在一起，就形成了我们的聚类簇。