

GRADO DE INGENIERO EN INFORMÁTICA
INTRODUCCIÓN A LA MINERÍA DE DATOS
FEBRERO 2014

1. (2.5) La siguiente tabla representa la matriz de similaridad o distancias entre los elementos de un cierto conjunto de datos compuesto por 6 puntos:

	p1	p2	p3	p4	p5	p6
p1	1.00	0.12	0.44	0.98	0.17	0.45
p2	0.12	1.00	0.77	0.88	0.10	0.01
p3	0.44	0.77	1.00	0.67	0.71	0.17
p4	0.98	0.88	0.67	1.00	0.23	0.82
p5	0.17	0.10	0.71	0.23	1.00	0.55
p6	0.45	0.01	0.17	0.82	0.55	1.00

Realiza el dendograma correspondiente al clustering jerárquico mediante el método de *complete link*. ¿Es posible realizar con la información dada el clustering usando el método *average link*?

2. (2.5) Considera la siguiente tabla que incluye la información de 10 transacciones realizadas en un cierto establecimiento:

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

Considera una soporte mínimo del 30%. Construye la rejilla correspondiente a todos los posibles conjuntos de ítems. Marca en la rejilla cada nodo con una *F* si es frecuente, una *I* si es infrecuente y una *N* si es podado por el algoritmo *Apriori*. Adicionalmente marca los nodos maximalmente frecuentes con una *M* y los nodos cerrados y frecuentes con una *C*.

A partir únicamente de los nodos cerrados y frecuentes obtén las 5 reglas con mayor confianza de todas las posibles, sin considerar las reglas triviales.

3. (2.5) Considera el conjunto de datos dónde cada instancia tiene 3 atributos, dos de ellos de tipo lógico y un tercero nominal y puede pertenecer a una de dos clases, c_1 ó c_2 . La siguiente tabla indica el número de instancias para cada una de las dos clases en función de los valores posibles de los atributos:

Complete link

	1	2	3	4	5	6
1	X	0,12	0,44	0,98	0,17	0,45
2		X	0,77	0,89	0,10	0,01
3			X	0,67	0,71	0,17
4				X	0,23	0,92
5					X	0,55
6						X

1^a iteración - C₁ → 1-4

1-4	X	0,12	0,44	0,17	0,45
2	X	0,77	0,10	0,01	
3		X	0,71	0,17	
4			X	0,23	0,92
5				X	0,55
6					X

2^a iteración - C₂ → 2-3

	1-4	2-3	5	6
1-4	X	0,12	0,17	0,45
2-3		X	0,10	0,01
5			X	0,55
6				X

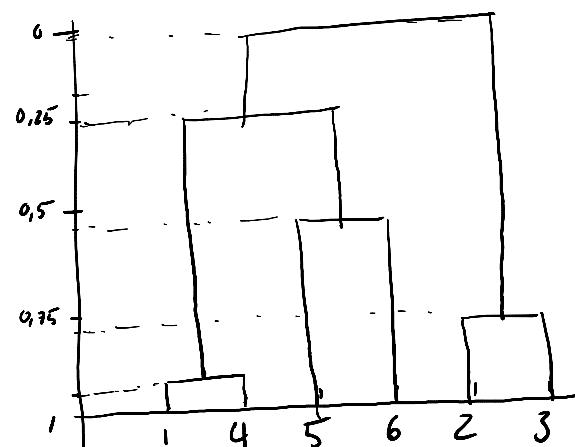
3^a iteración - C₃ → 5-6

	1-4	2-3	5-6
1-4	X	0,12	0,17
2-3		X	0,01
5-6			X

4^a iteración - C₄ → C₁-C₃

	C ₁ -C ₃	2-3
C ₁ -C ₃	X	0,01
2-3		X

→ 5^a iteración C₅ → C₄-C₂



DENDOGRAMA:

2) C_1 - Itemsets de 1 item:

Item	Count
a	5
b	5
c	6
d	7
e	9

Soporte mínimo $\rightarrow 30\%$

Count mínimo $\rightarrow 3$

C_2 - Itemsets de 2 items

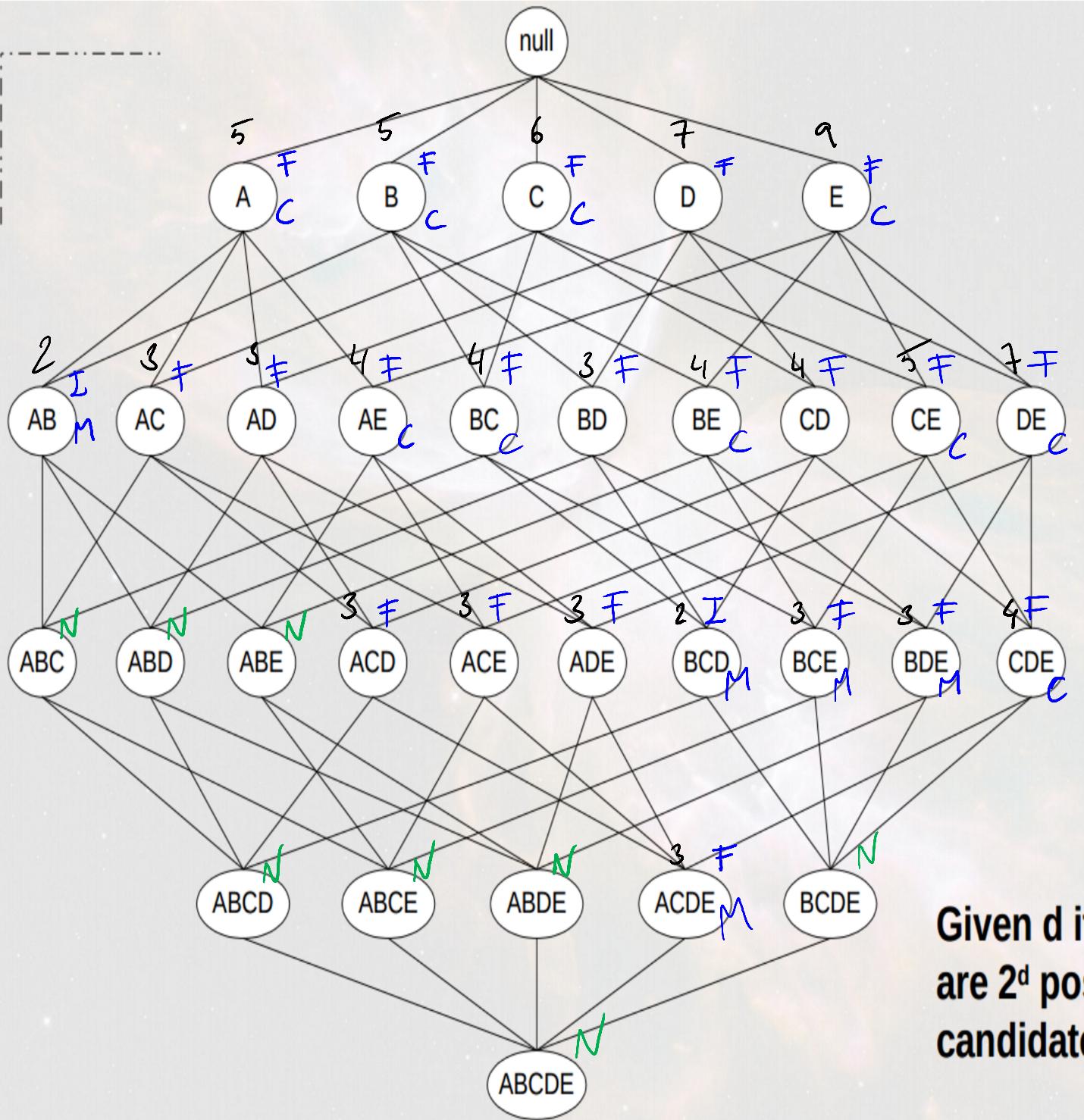
Itemset	Count
ab	2
ac	3
ad	3
ae	4
bc	4
bd	3
be	4
cd	4
ce	5
de	7

C_3 - Itemset de 3 items

Itemset	Count
acd	3
ace	3
ade	3
bcd	2
bce	3
bde	3
cde	4

C_4 - Itemset de 4 items

Itemset	Count
acde	3



Given d items
are 2^d possible
candidate iter

Atributos			Número de instancias	
x_1	x_2	x_3	c_1	c_2
V	V	a	5	32
V	V	b	0	1
V	V	c	40	7
V	F	a	0	0
V	F	b	10	5
V	F	c	12	6
<hr/>			<hr/>	
F	V	a	0	28
F	V	b	0	0
F	V	c	10	3
F	F	a	8	1
F	F	b	1	22
F	F	c	4	17

Construye un árbol de decisión binario utilizando una estrategia voraz y cómo criterio de división de cada nodo el error de clasificación. Para crear una nueva división es necesario que el nuevo subárbol mejore al nodo padre. Calcula el error de entrenamiento del árbol completo.

4. (2.5) La siguiente tabla muestra un conjunto de datos de 12 instancias representadas cada una de ellas por 5 variables de tipo lógico:

	x_1	x_2	x_3	x_4	x_5
p1	V	F	F	V	F
p2	V	F	F	F	F
p3	F	V	V	V	V
p4	F	V	V	F	V
p5	V	F	V	V	V
p6	F	F	F	F	V
p7	V	F	F	V	F
p8	F	V	F	F	F
p9	V	F	V	V	V
p10	V	F	V	F	F
p11	V	F	V	V	V
p12	F	V	V	F	V

Realiza el algoritmo k -medias paso a paso usando la distancia de Hamming y $k = 2$. Selecciona como centros iniciales las instancias p1 y p6. El centroide de cada clúster se construye usando la moda de cada variable que forma el clúster. El algoritmo ha de detenerse si no ha convergido después de 10 iteraciones.

Tiempo de realización: 6 horas. Calificación de cada ejercicio entre paréntesis.

3) Para determinar el nodo padre tenemos que comprobar qué ratio de error tiene cada atributo.

Total instancias $\rightarrow 212$

x_1	C_1	C_2
V	67	51
F	23	71

$$\rightarrow \text{Error } x_1 = (51 + 23) / 212 = 74/212$$

x_2	C_1	C_2
V	55	71
F	35	51

$$\rightarrow \text{Error } x_2 = (55 + 35) / 212 = 90/212$$

x_3	C_1	C_2
a	13	61
b	11	28
c	66	33

(ab)_c
~~abc~~
~~(ac)b~~

Es el que menor error tiene. x_3 nodo padre.

x_3	C_1	C_2
ab	24	89
c	66	33
a	13	61
bc	77	61
ac	79	94
b	11	28

$$\text{Error } x_3 = (24 + 33) / 212 = 57/212$$

$$\text{Error } x_3 = (13 + 61) / 212 = 74/212$$

$$\text{Error } x_3 = (11 + 79) / 212 = 90/212$$

(ab)

x_1	x_2	x_3	c_1	c_2
V	V	a	5	32
V	V	b	0	1.
V	F	a	0	0
V	F	b	10	5.
F	V	a	0	28
F	V	b	0	0.
F	F	a	8	1
F	F	b	1	22
			24	89

x_1	x_2	x_3	c_1	c_2
V	V	c	40	7
V	F	c	12	6
F	V	c	10	3
F	F	c	4	17

Calculamos el error del siguiente nodo, que nos indicará qué atributo activará de padres: Total partición: 113

x_1	c_1	c_2
V	15	39
F	9	51

$$\text{Error } x_1 = (15+9)/113 = 24/113$$

x_2	c_1	c_2
V	5	61
F	19	28

$$\text{Error } x_2 = (5+19)/113 = 24/113$$

x_3	c_1	c_2
a	13	61
b	11	28

$$\text{Error } x_3 = (13+11)/113 = 24/113$$

(Se poden)

Calculamos el error para la rama de $x_3 = c$ para saber quién nodo será el padre a partir de ahí:
 Total partición: 99

x_1	c_1	c_2
V	52	13
F	14	20

$$\text{Error}_{x_1} = (3+14)/99 = 27/99$$

x_2	c_1	c_2
V	50	10
F	16	23

$$\text{Error}_{x_2} = (16+10)/99 = 26/99$$

Se celebra como nuevo nodo el atributo x_2 al tener menor error

x_1	x_2	x_3	c_1	c_2
V	V	c	40	7
F	V	c	10	3

~~50 10 60~~

x_1	x_2	x_3	c_1	c_2
V	F	c	12	6
F	F	c	4	17

~~16 23 39~~

Los nodos hijo deben tener siempre menor error que el padre
 (suma de los errores < error padre)
 Error de un nodo \rightarrow (base unitaria/total)

Ahora calculamos el error de los nodos hijos para los valores $x_1 = V$ y $x_2 = F$.

$$\underline{x_2 = V}$$

x_1	c_1	c_2
V	40	7
F	10	3

$$\text{Error } x_1 = (7+3)/60 = \underline{\underline{10/60}}$$

Sus hijos saldrán de $x_1 = V$ y $x_2 = F$

$x_1 = V$	$x_2 = F$	
c_1	40	10
c_2	7	3

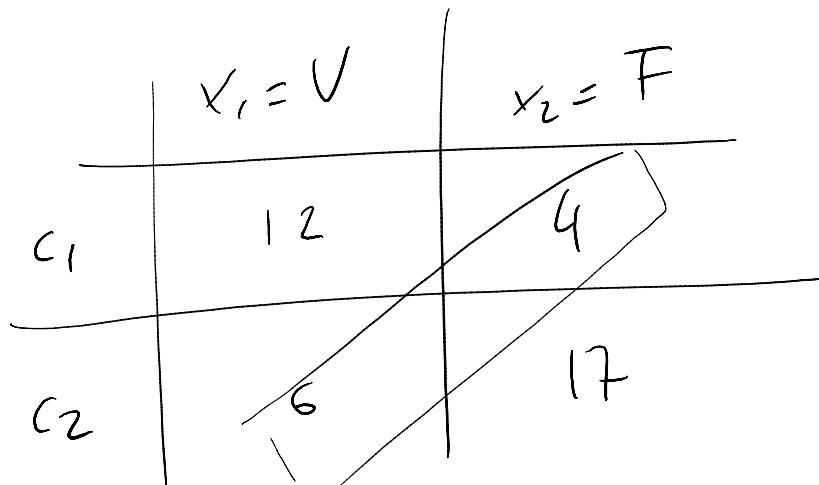
Error de los hijos es $10/60$, por lo que se puede algoritmo que el padre.

$$\underline{x_2 = F}$$

x_1	c_1	c_2
V	12	6
F	4	17

$$\text{Error } x_1 = (12+4)/39 = \underline{\underline{16/39}}$$

Los hijos subirán de $x_1=V$ y $x_2=F$



Error de los hijos es $10/39$, por lo que
se continúa clasificando

Para calcular el error total del grafo hay que coger
las clases descendientes en los nodos hoja que han
sido podados y sumar las instancias de estos.
Al dividirlo entre el total de instancias del problema
tenemos el error (un var árbol)

$$(24 + 10 + 6 + 4) / 212 = 44/212$$

