COMP1942 Exploring and Visualizing Data (Spring Semester 2024)
Homework 1 Solution
Full Mark: 100 Marks

Q1 [20 Marks]

(a)

Item	Count
AngusBeefBurger	4
CrabRoeNoodle	4
LimeJuice	3
MulberryJuice	1
RoastDuck	4
ScallopCongee	4

L_1={{AngusBeefBurger},{CrabRoeNoodle},{LimeJuice},{ RoastDuck},{ ScallopCongee }}
Generate L_2:
Join step/prune step:
C_2={ {AngusBeefBurger, CrabRoeNoodle}, {AngusBeefBurger, LimeJuice},
{AngusBeefBurger, RoastDuck}, {AngusBeefBurger, ScallopCongee},
{CrabRoeNoodle, LimeJuice}, {CrabRoeNoodle, RoastDuck}, {CrabRoeNoodle, ScallopCongee},
{LimeJuice, RoastDuck}, {LimeJuice, ScallopCongee},
{RoastDuck, ScallopCongee} }
Counting step:
L_2={ {AngusBeefBurger, CrabRoeNoodle}, {AngusBeefBurger, LimeJuice}, {AngusBeefBurger, ScallopCongee},
{CrabRoeNoodle, RoastDuck}, {CrabRoeNoodle, ScallopCongee},
{LimeJuice, RoastDuck},
{RoastDuck, ScallopCongee}
}

Generate L_3:
Join step:
C_3={{ AngusBeefBurger, CrabRoeNoodle, LimeJuice }, {AngusBeefBurger, CrabRoeNoodle, ScallopCongee},
{AngusBeefBurger, LimeJuice, ScallopCongee},
{CrabRoeNoodle, RoastDuck, ScallopCongee} }
Prune step:
C_3={{ AngusBeefBurger, CrabRoeNoodle, ScallopCongee}, {CrabRoeNoodle, RoastDuck, ScallopCongee} }

Counting step:
L_3={{ CrabRoeNoodle, RoastDuck, ScallopCongee }}

Large itemsets = $L_1 \cup L_2 \cup L_3$
= {{AngusBeefBurger},{CrabRoeNoodle},{LimeJuice},{ RoastDuck},{ ScallopCongee },
{AngusBeefBurger, CrabRoeNoodle}, {AngusBeefBurger, LimeJuice},
{AngusBeefBurger, ScallopCongee},
{CrabRoeNoodle, RoastDuck}, {CrabRoeNoodle, ScallopCongee},

{LimeJuice, RoastDuck},
{RoastDuck, ScallopCongee},
{ CrabRoeNoodle, RoastDuck, ScallopCongee }}

(b) Association rules

{RoastDuck, ScallopCongee} → CrabRoeNoodle
{CrabRoeNoodle, RoastDuck} → ScallopCongee

Q2 [20 Marks]

(a)

Item	Freq
a	3
b	3
c	5
d	5
e	1
f	6
g	1
h	1
i	1
j	1
k	1
l	1
m	1
n	1
o	1
p	1
q	1

Freq items:

Item	Freq
a	3
b	3
c	5
d	5
f	6

Sorted Freq items:

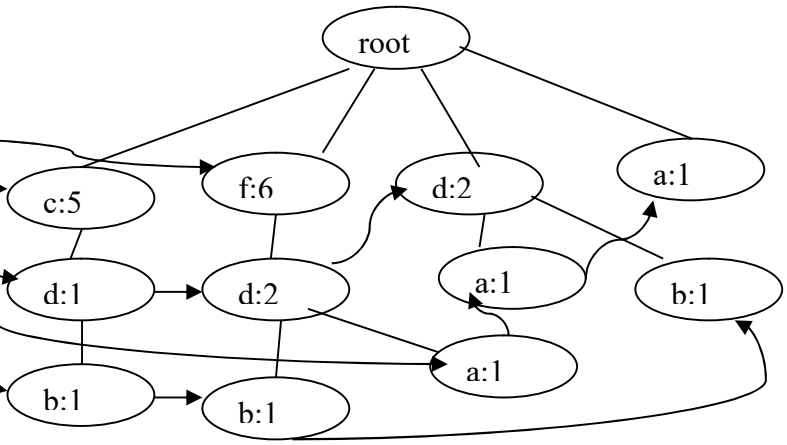
Item	Freq
f	6
c	5
d	5
a	3
b	3

Ordered freq items

TID	Items bought	(ordered) freq items
1	b,c,d,p	c,d,b
2	f,j,q	f
3	c,i	c
4	a,d	d,a
5	c,m	c
6	b,d,f	f,d,b
7	a,d,f	f,d,a
8	a,l	a
9	c,g	c
10	c,k	c
11	f,n,o	F
12	e,f	f
13	f,h	f
14	b,d	d,b

FP-tree

Item	Head of link
f	
c	
d	
a	
b	



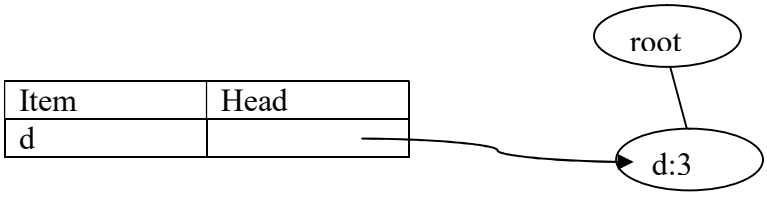
Conditional FP-tree on "b" **count (b)=3**

(c:1,d:1,b:1) (b:1,d:1)
(f:1,d:1,b:1) ⇒ (b:1,d:1)
(d:1,b:1) (b:1,d:1)

Item	Freq
f	1
c	1
d	3
a	0
b	3

↓

Item	Freq
b	3
d	3



{b,d}:3

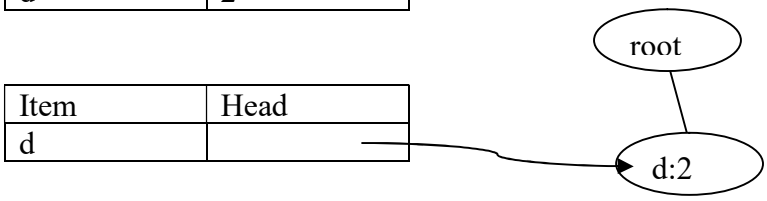
Conditional FP-tree on "a" **count (a)=3**

(f:1,d:1,a:1) (a:1,d:1)
(d:1,a:1) ⇒ (a:1,d:1)
(a:1) (a:1)

Item	Freq
f	1
c	0
d	2
a	3
b	0

↓

Item	Freq
a	3
d	2



{a,d}:2

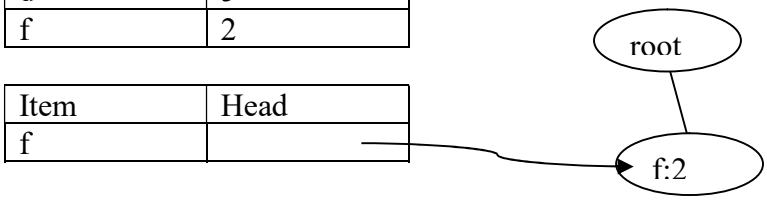
Conditional FP-tree on "d" **count (d)=5**

(c:1,d:1) (d:1)
(f:2,d:2) ⇒ (d:2,f:2)
(d:2) (d:2)

Item	Freq
f	2
c	1
d	5
a	0
b	0

↓

Item	Freq
d	5
f	2



{d,f}:2

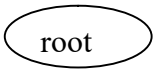
Conditional FP-tree on "c" **count (c)=5**

(c:5) ⇒ (c:5)

Item	freq
c	5

↓

Item	freq
c	5



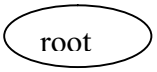
Conditional FP-tree on "f" **count (f)=6**

(f:6) ⇒ (f:6)

Item	freq
f	6

↓

Item	freq
f	6



Freq itemsets

={{b},{b,d},
{a},{a,d},
{d},{d,f},
{c},
{f}}

(b) association rules

{b} → {d}
{d} → {b}
{a} → {d}

Q3 [20 Marks]

(a)

(i)

cluster 1:

mean=(16.5,18.25)

points={x_1,x_3,x_5,x_8}

cluster 2:

mean=(6.25,10)

points={x_2,x_4,x_6,x_7}

(ii)

cluster1:

mean=(4.33,12.33)

points={x_2,x_4,x_6}

cluster2:

mean=(16.5,18.25)

points={x_1,x_3,x_5,x_8}

cluster 3:

mean=(12,3)

points={x_7}

(b)

Advantages:

1. It is easy to implement.
2. It is easy to understand.

Disadvantages:

1. It is difficult to determine the value of k because we do not know the no. of the clusters.
2. K-means is sensitive to the initial guess of the means.

Q4 [20 Marks]

	1	2	3	4	5	6	7	8
1	0							
2	11	0						
3	5	13	0					
4	12	2	14	0				
5	7	7	1	18	0			
6	13	4	15	5	20	0		
7	9	15	12	16	15	19	0	
8	11	20	12	21	17	22	30	0

D(1,*)=13
D(2,*)=20
D(3,*)=15
D(4,*)=21
D(5,*)=20
D(6,*)=22
D(7,*)=30
D(8,*)=30

A={8}
B={1,2,3,4,5,6,7}

D(1,A)=11 D(1,B)=13 Δ_1=2
D(2,A)=20 D(2,B)=17 Δ_2=-3
D(3,A)=12 D(3,B)=15 Δ_3=3
D(4,A)=21 D(4,B)=18 Δ_4=-3
D(5,A)=17 D(5,B)=20 Δ_5=3
D(6,A)=22 D(6,B)=20 Δ_6=-2
D(7,A)=30 D(7,B)=19 Δ_7=-11

A={3,8}
B={1,2,4,5,6,7}

D(1,A)=11 D(1,B)=13 Δ_1=2
D(2,A)=20 D(2,B)=17 Δ_2= -3
D(4,A)=21 D(4,B)=18 Δ_4= -3
D(5,A)=17 D(5,B)=20 Δ_5=3
D(6,A)=22 D(6,B)=20 Δ_6= -2
D(7,A)=30 D(7,B)=19 Δ_7= -11

A={3,5,8}
B={1,2,4,6,7}

D(1,A)=11 D(1,B)=13 Δ_1=2
D(2,A)=20 D(2,B)=15 Δ_2= -5

D(4,A)=21 D(4,B)=16 Δ_4= -5
D(6,A)=22 D(6,B)=19 Δ_6= -3
D(7,A)=30 D(7,B)=19 Δ_7= -11

A={1,3,5,8}
B={2,4,6,7}

D(2,A)=20 D(2,B)=15 Δ_2= -5
D(4,A)=21 D(4,B)=16 Δ_4= -5
D(6,A)=22 D(6,B)=19 Δ_6= -3
D(7,A)=30 D(7,B)=19 Δ_7= -11

Stop!
A={1,3,5,8}
B={2,4,6,7}

So, there are two clusters:
Cluster 1: data points1, 3, 5, 8
Cluster 2: data points 2, 4, 6, 7

The distance between 2 clusters is 30.

Q5 [20 Marks]

(a)(i)

Info(T) = $-\frac{1}{2}\log\frac{1}{2}-\frac{1}{2}\log\frac{1}{2}=1$

For attribute HasMacBook,

Info(T_{no})= $-\frac{1}{2}\log\frac{1}{2}-\frac{1}{2}\log\frac{1}{2}=1$

Info(T_{yes})= $-\frac{1}{2}\log\frac{1}{2}-\frac{1}{2}\log\frac{1}{2}=1$

Info(HasMacBook, T)= $\frac{1}{2}Info(T_{no})+\frac{1}{2}Info(T_{yes})=1$

SplitInfo(HasMacBook)= $-\frac{1}{2}\log\frac{1}{2}-\frac{1}{2}\log\frac{1}{2}=1$

Gain(HasMacBook, T)= $\frac{1-1}{1}=0$

For attribute Income,

Info(T_{high})= $-\frac{1}{2}\log\frac{1}{2}-\frac{1}{2}\log\frac{1}{2}=1$

Info(T_{medium})= $-\frac{1}{2}\log\frac{1}{2}-\frac{1}{2}\log\frac{1}{2}=1$

Info(T_{low})= $-\frac{1}{2}\log\frac{1}{2}-\frac{1}{2}\log\frac{1}{2}=1$

Info(Income, T)= ½ Info(T_{high}) + ¼ Info(T_{medium}) + ¼ Info(T_{low}) = 1

SplitInfo(Income)= $-\frac{1}{2}\log\frac{1}{2}-\frac{1}{4}\log\frac{1}{4}-\frac{1}{4}\log\frac{1}{4}=1.5$

Gain(Income, T)= $\frac{1-1}{1.5}=0$

For attribute Age,

Info(T_{middle})= $-1\log 1-0\log 0=0$

Info(T_{old})= $-\frac{3}{4}\log\frac{3}{4}-\frac{1}{4}\log\frac{1}{4}=0.8113$

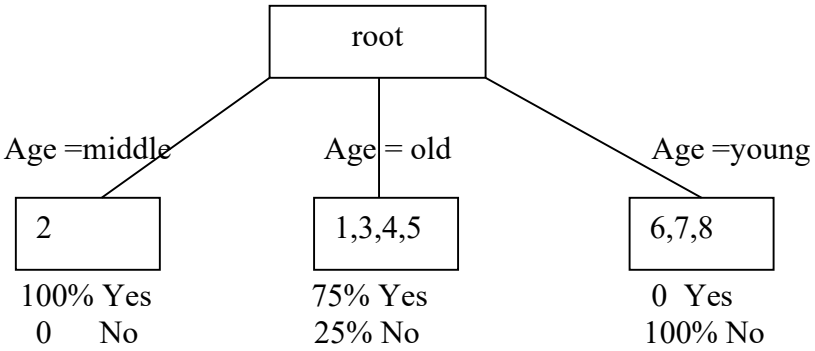
Info(T_{young})= $-0\log 0-1\log 1=0$

Info(Age)= 1/8 Info(T_{middle}) + ½ Info(T_{old}) + 3/8 Info(T_{young}) = 0.405

SplitInfo(Age)= $-\frac{1}{8}\log\frac{1}{8}-\frac{1}{2}\log\frac{1}{2}-\frac{3}{8}\log\frac{3}{8}=1.4056$

Gain(Age, T)= $\frac{1-0.405}{1.4056}=0.4233$

We choose attribute Age for Splitting:



Consider the node for "Age=old"

Info(T)= $-\frac{3}{4}\log\frac{3}{4}-\frac{1}{4}\log\frac{1}{4}=0.8113$

For attribute HasMacBook,

Info(T_{no})= $-\frac{1}{2}\log\frac{1}{2}-\frac{1}{2}\log\frac{1}{2}=1$

Info(T_{yes})= $-1\log 1-0\log 0=0$

Info(HasMacBook, T)= $\frac{1}{2}Info(T_{no})+\frac{1}{2}Info(T_{yes})=0.5$

SplitInfo(HasMacBook)= $-\frac{1}{2}\log\frac{1}{2}-\frac{1}{2}\log\frac{1}{2}=1$

Gain(HasMacBook, T)= $\frac{0.8113-0.5}{1}=0.3113$

For attribute Income,

Info(T_{high})= $-\frac{1}{2}\log\frac{1}{2}-\frac{1}{2}\log\frac{1}{2}=1$

Info(T_{medium})= $-1\log 1-0\log 0=0$

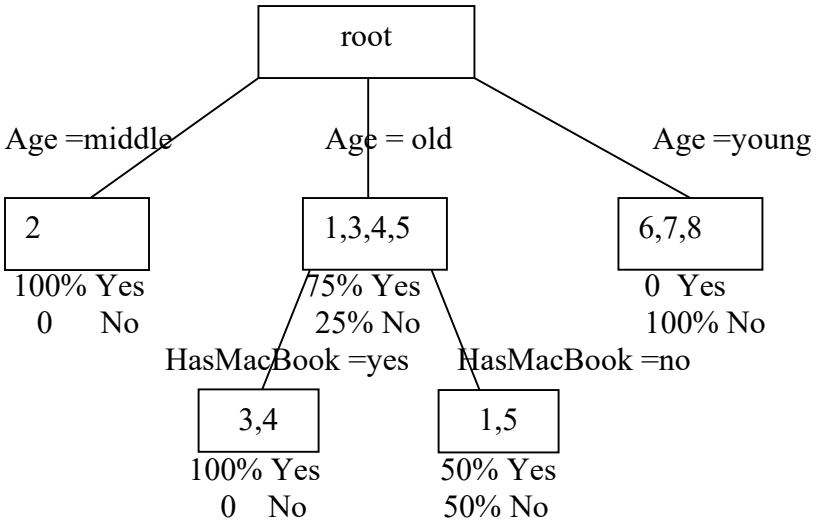
Info(T_{low})= $-1\log 1-0\log 0=0$

Info(Income)= ½ Info(T_{high}) + ¼ Info(T_{medium}) + ¼ Info(T_{low}) = 0.5

SplitInfo(Income)= $-\frac{1}{2}\log\frac{1}{2}-\frac{1}{4}\log\frac{1}{4}-\frac{1}{4}\log\frac{1}{4}=1.5$

Gain(Income, T)= $\frac{0.8113-0.5}{1.5}=0.2075$

We choose attribute HasMacBook for Splitting:



(ii)It is very likely that this customer will buy “Apple Vision Pro”.

(b)

Differences:

The definition of the gain used in C4.5 is different from that used in ID3.
The gain used in C4.5 is equal to the gain used in ID3 divided by SplitInfo.

The reason why there is a difference is described as follows.
In ID3, there is a higher tendency to choose an attribute containing more values (e.g., attribute identifier and attribute HKID). Thus, splitInfo in C4.5 is used to penalize an attribute containing more values. If this value is larger, the penalty is larger.