CHEN Wenda
18410782

1. confusion matrix

Actually class \ predict class	+	−
+	2	3
−	1	4

$$\text{accuracy} = \frac{2+4}{10} = 0.6$$

$$\text{precision} = \frac{2}{2+1} = \frac{2}{3}$$

$$\text{recall} = \frac{2}{2+3} = \frac{2}{5}$$

$$\text{F-measure} = \frac{2 \times \frac{2}{3} \times \frac{2}{5}}{\frac{2}{3}+\frac{2}{5}} = \frac{1}{2}$$

2. (a)

$$P(A=1|+) = \frac{P(A=1,+)}{P(+)} = \frac{3/10}{4/10} = \frac{3}{4}$$

$$P(B=1|+) = \frac{P(B=1,+)}{P(+)} = \frac{1/10}{4/10} = \frac{1}{4}$$

$$P(C=1|+) = \frac{P(C=1,+)}{P(+)} = \frac{4/10}{4/10} = 1$$

$$P(A=1|-) = \frac{P(A=1,-)}{P(-)} = \frac{2/10}{6/10} = \frac{1}{3}$$

$$P(B=1|-) = \frac{P(B=1,-)}{P(-)} = \frac{3/10}{6/10} = \frac{1}{2}$$

$$P(C=1|-) = \frac{P(C=1,-)}{P(-)} = \frac{1/10}{6/10} = \frac{1}{6}$$

(b)

$$P(+|A=1,B=1,C=1) = \frac{P(A=1,B=1,C=1|+)P(+)}{P(A=1,B=1,C=1)} \rightarrow \text{ignore}$$

$$\approx P(A=1,B=1,C=1|+) \cdot P(+)$$

$$= P(A=1|+) \cdot P(B=1|+) \cdot P(C=1|+) \cdot P(+)$$

$$= \frac{3}{4} \times \frac{1}{4} \times 1 \times \frac{4}{10}$$

$$= \frac{3}{40}$$

$$P(-|A=1,B=1,C=1) = \frac{P(A=1,B=1,C=1|-) \cdot P(-)}{P(A=1,B=1,C=1)} \rightarrow \text{ignore}$$

$$\approx P(A=1,B=1,C=1|-) \cdot P(-)$$

$$= P(A=1|-) \cdot P(B=1|-) \cdot P(C=1|-) \cdot P(-)$$

$$= \frac{1}{3} \times \frac{1}{2} \times \frac{1}{6} \times \frac{6}{10}$$

$$= \frac{1}{60}$$

$\because \frac{3}{40} > \frac{1}{60}$

$\therefore$ this instance will be predicted to "+"

(c)

$P(A=1) = 5/10 = 0.5$

$P(B=1) = 4/10 = 0.4$

$P(A=1, B=1) = 2/10 = 0.2$

$P(A=1) \cdot P(B=1) = P(A=1, B=1)$

A and B are independent.

(d) $P(A=1) = 5/10 = 0.5$

$P(B=0) = 6/10 = 0.6$

$P(A=1, B=0) = 3/10 = 0.3$

$P(A=1) \cdot P(B=0) = P(A=1, B=0)$

A and B are independent.

(e) $P(A=1, B=1 \mid +) = 1/4$

$P(A=1 \mid +) = 3/4$

$P(B=1 \mid +) = 1/4$

$\because P(A=1, B=1 \mid +) \neq P(A=1 \mid +) \cdot P(B=1 \mid +)$

$\therefore$ they are not conditionally independent

3.

(a) items = {milk, diaper, beer, butter, bread, cookies}

d = items number = 6

association rule number $= 3^6 - 2^{6+1} + 1 = 602$

(b) 4.

because the number of the longest transaction is 4.

(c). $C_6^3 = \frac{6 \times 5 \times 4}{1 \times 2 \times 3} = 20$

(d)

2-itemset	support
milk, bread	0.3
butter, diapers	0.3
cookies, milk	0.1
cookies, bread	0.1
beer, milk	0.1
milk, diappers	0.4
cookies, diapers	0.2
diapers, bread	0.3
butter, bread	**0.5**
butter, milk	0.3
butter, cookies	0.1
beer, cookies	0.2
beer, diappers	0.3

Since the bigger size of itemsets can not get a larger support than its sub-itemsets, no larger size of itemset can have larger support.

→ largest support {butter, bread}

(e)

	confidence		confidence
Bread=>Milk	0.600	Milk=>Bread	0.600
Diapers=>Butter	0.429	Butter=>Diapers	0.600
Cookies=>Milk	0.250	Milk=>Cookies	0.200
Cookies=>Bread	0.250	Bread=>Cookies	0.200
Beer=>Milk	0.250	Milk=>Beer	0.200
Diapers=>Milk	0.571	Milk=>Diapers	0.800
Cookies=>Diapers	0.500	Diapers=>Cookies	0.286
Diapers=>Bread	0.429	Bread=>Diapers	0.600
Bread=>Butter	1.000	Butter=>Bread	1.000
Butter=>Milk	0.600	Milk=>Butter	0.600
Cookies=>Butter	0.250	Butter=>Cookies	0.200
Cookies=>Beer	0.500	Beer=>Cookies	0.500
Beer=>Diapers	0.750	Diapers=>Beer	0.429

{Bread, milk}
{bread, butter}
{butter, milk}
{cookies, beer}

4 (a).

the support count of all frequent itemsets.

{A} = 3 {C} = 3 {D} = 4

{C, D} = 3.

maximal frequent itemsets

{A}, {C, D}

closed frequent itemsets

{A} {D} {C, D}

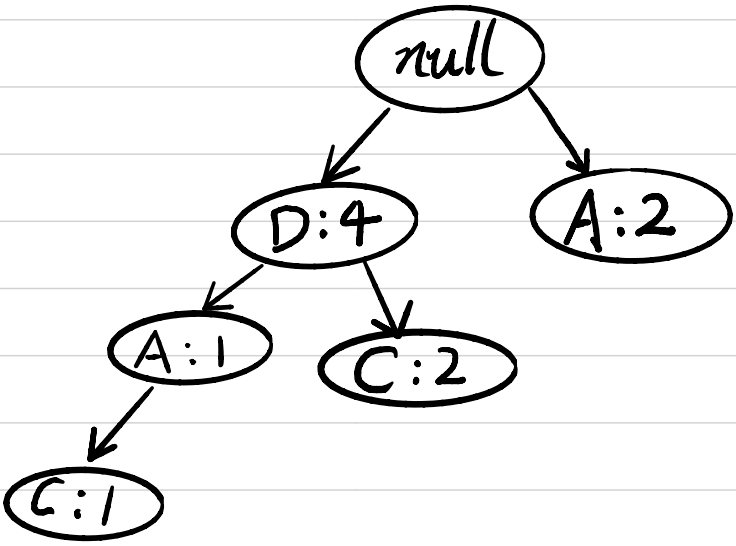
(b) support count of 1 frequent itemsets

D > A > C

Ordered ITEM-SET

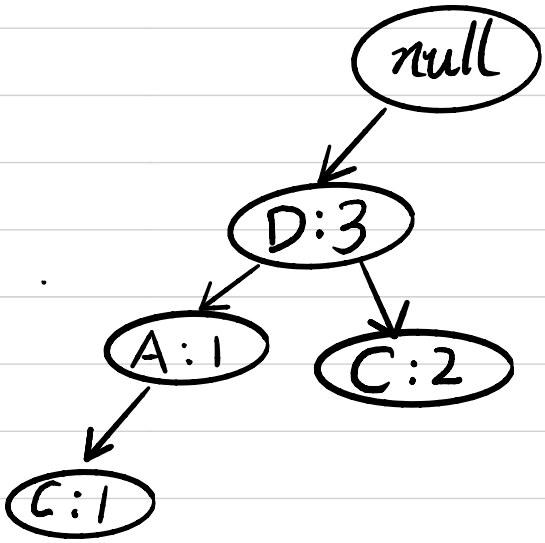
D
A
D A C
D C
A
D C.

FP-TREE



FP-Tree on C

Item	Frequency
D	3
A	1 <2 (ignore)
C	3



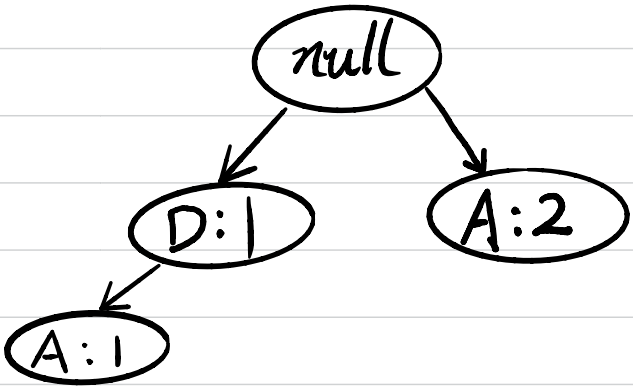
⇓

frequence 2-itemset.

{D, C}

FP-Tree On A.

Item	Frequency
D	1 <2 ignore
A	3

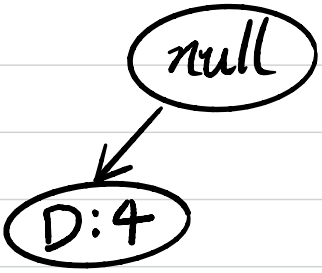


⇓

{A}

FP-Tree On D.

Item	Frequency
D	4



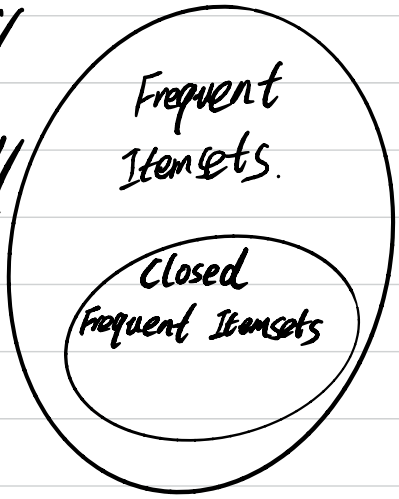
⇓

{D}

All closed. Frequent itemset:

{D}, {A}, {C, D}.

(c) Closed frequent itemsets retain all the information needed to obtain all the frequent itemsets. What's more, mining closed frequent itemsets can reduce the search space and memory consuption.



(d). After mining the all maximum frequent itemsets, we can generate all frequent itemset in a single scan. Because every subset of frequent itemset is frequent. And there is no immediate superset is frequent set.