CSE 5243 Autumn'22 Assignment #4 – Intr Data Mining

Neng Shi

2022/11/15

1 Problem 1.

$P(Class = N) = 4/9$
$P(Class = Y) = 5/9$
$P(a1 = T|Class = N) = 3/4$
$P(a1 = T|Class = Y) = 1/5$
$P(a2 = F|Class = N) = 2/4$
$P(a2 = F|Class = Y) = 2/5$

For $a3$, when $Class = N$, $\mu = 5.00$ and $\sigma^2 = 2.00$,
so $P(a3 = 2.0|Class = N) = \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(2.0-5.0)^2}{2\times 2}} = 0.030$.
And when $Class = Y$, $\mu = 5.00$ and $\sigma^2 = 6.00$,
so $P(a3 = 2.0|Class = Y) = \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(2.0-5.0)^2}{2\times 6}} = 0.077$.

$X = (a1 = T, a2 = F, a3 = 2.0)$
$P(X|Class = N) = P(a1 = T|Class = N) \times P(a2 = F|Class = N) \times P(a3 = 2.0|Class = N) = 0.011$
$P(X|Class = Y) = P(a1 = T|Class = Y) \times P(a2 = F|Class = Y) \times P(a3 = 2.0|Class = Y) = 0.006$

$P(X|Class = N) \times P(Class = N) = 0.005$
$P(X|Class = Y) \times P(Class = Y) = 0.003$,
So the new point $(T, F, 2.0)$ should be classified as N.

2 Problem 2.

$Info(D) = -\frac{4}{8} \times log_2(\frac{4}{8}) - \frac{4}{8} \times log_2(\frac{4}{8}) = 1$

After the partition, $D1$ contains $x1, x2, x4, x5, x7$ and $D2$ contains $x3, x6, x8$.

$$Info(D1) = -\frac{3}{5} \times log_2(\frac{3}{5}) - \frac{2}{5} \times log_2(\frac{2}{5}) = 0.97$$

$$Info(D2) = -\frac{1}{3} \times log_2(\frac{1}{3}) - \frac{2}{3} \times log_2(\frac{2}{3}) = 0.92$$

$$Info_{AB-B^2 \leq 0}(D) = \frac{5}{8} \times Info(D1) + \frac{3}{8} \times Info(D2) = 0.95$$

So $Gain(AB - B^2 \leq 0) = 1 - 0.95 = 0.05$

3 Problem 3.

First, we cluster A and B since they have the minimum distance.

Then the updated distance matrix is

	{A,B}	C	D	E
{A, B}	0	3	2	3
C		0	1	3
D			0	5
E				0

Second, we cluster C and D since they have the minimum distance in the updated matrix.

Then the updated matrix is

	{A,B}	{C, D}	E
{A, B}	0	2	3
C		0	3
D			0

Third, we cluster $\{A, B\}$ and $\{C, D\}$ since they have the minimum distance in the updated matrix. Finally, we cluster $\{A, B, C, D\}$ and E.

4 Problem 4.

Proof. For an item $m \in M$, since it is an max-pattern, it is frequent and there is no frequent super-pattern. Then, obviously, there is no super-pattern with the same support as m, which means $m \in C$. Thus, we prove that $M \subset C$. $\square$

5 Problem 5.

(a)

Proof. If an itemset s appears in one transaction, then any of its non-empty subset would appear in the transaction. Thus, the support of any nonempty subset s' of itemset s must bet at least as great as the support of s. $\square$

(b)

Proof. If an itemset s is frequent, it means that the support of s is larger or equal to a minsup threshold σ. Then, according to the prove in (a), the support of any nonempty subset s' of itemset s is also larger or equal to σ, which means they are also frequent. $\square$

(c)

Proof. $confidence(s' => (l - s')) = \frac{support\_count(l)}{support\_count(s')}$,
and $confidence(s => (l - s)) = \frac{support\_count(l)}{support\_count(s)}$.

Since according to the prove in (a), $support\_count(s') \geq support\_count(s)$, $confidence(s' => (l - s')) \leq confidence(s => (l - s))$. $\square$

(d)

Proof. Prove by contradiction.
Denote the n non-overlapping partitions as $D_i, i \in \{0, ..., n-1\}$, the itemset as s, the support of s in these partitions as $support_i(s), i \in \{0, ..., n-1\}$, and the minsup threshold as σ.
Assume that an itemset s is not frequent in all the partitions, which mean $support_i(s) < \sigma$ for all the i. In that case, the support for s in D would also be less than the minsup threshold σ because it comes from the weighted average of $support_i(s)$, meaning that s is not frequent in D. $\square$

6 Problem 6.

Table 1: C_1

Itemset	sup
{A}	5
{B}	4
{C}	5
{D}	6
{E}	1
{F}	4
{G}	5

Table 2: F_1

Itemset	sup
{A}	5
{B}	4
{C}	5
{D}	6
{F}	4
{G}	5

Table 3: C_2

Itemset	sup
{A, B}	3
{A, C}	3
{A, D}	4
{A, F}	2
{A, G}	2
{B, C}	2
{B, D}	2
{B, F}	1
{B, G}	2
{C, D}	4
{C, F}	2
{C, G}	3
{D, F}	4
{D, G}	3
{F, G}	2

Table 4: F_2

Itemset	sup
{A, B}	3
{A, C}	3
{A, D}	4
{C, D}	4
{C, G}	3
{D, F}	4
{D, G}	3

Table 5: C_3

Itemset	sup
{A, C, D}	3
{C, D, G}	2

Table 6: F_3

Itemset	sup
{A, C, D}	3

All frequent patterns are $F_1 \cup F_2 \cup F_3$.

7 Problem 7.

Table 7: C_1

Itemset	sup
{A}	4
{B}	5
{C}	5
{D}	3
{E}	4

Table 8: F_1

Itemset	sup
{A}	4
{B}	5
{C}	5
{D}	3
{E}	4

Table 9: C_2

Itemset	sup
{A, B}	3
{A, C}	4
{A, D}	2
{A, E}	2
{B, C}	4
{B, D}	2
{B, E}	4
{C, D}	2
{C, E}	3
{D, E}	1

Table 10: F_2

Itemset	sup
{A, B}	3
{A, C}	4
{A, D}	2
{A, E}	2
{B, C}	4
{B, D}	2
{B, E}	4
{C, D}	2
{C, E}	3

Table 11: C_3

Itemset	sup
{A, B, C}	3
{A, B, D}	1
{A, B, E}	2
{A, C, D}	2
{A, C, E}	2
{B, C, D}	1
{B, C, E}	3

Table 12: F_3

Itemset	sup
{A, B, C}	3
{A, B, E}	2
{A, C, D}	2
{A, C, E}	2
{B, C, E}	3

Table 13: C_4

Itemset	sup
{A, B, C, E}	2

Table 14: F_4

Itemset	sup
{A, B, C, E}	2

All frequent itemsets are $F_1 \cup F_2 \cup F_3 \cup F_4$.
All closed itemsets are $\{A, B, C, E\} : 2$, $\{A, B, C\} : 3$, $\{A, C, D\} : 2$, $\{B, C, E\} : 3$, $\{A, C\} : 4$, $\{B, C\} : 4$, $\{B, D\} : 2$, $\{B, E\} : 4$, $\{B\} : 5$, $\{C\} : 5$, and $\{D\} : 3$.
All maximal itemsets are $\{A, B, C, E\} : 2$, $\{A, C, D\} : 2$, and $\{B, D\} : 2$.