Data Mining Exercises

1 Krimp

In the code tables below, only the length of the codes of the itemsets are given (in bits), not the codes themselves (since we are not interested in the codes).

$DB_1 =$

Tid	Transaction
1	ABCD
2	BCD
3	ACD

$\mathrm{CT}_1 =$

Itemset	Code Length
ABC	1.5
BC	2
CD	2.5
C	3.5
A	2.5
D	3

$DB_2 =$

Tid	Transaction
1	ABC
2	ABC
3	ABC
4	BC
5	BCD
6	CD

Question 1 : Compute $\mathrm{Len}(DB_1 \mid CT_1)$. For each transaction, give the itemsets used to compress it.

We assume the input of Krimp is DB_2 and the set of candidate itemsets is $F = \{ABC, BC, CD\}$.

Question 2 : In which order are the candidate itemsets considered by Krimp?

Question 3 : Give the Standard Code Table ST for DB_2 and compute its length $\mathrm{Len}(ST)$.

Question 4 : Compute $\mathrm{TotLen}(DB_2, ST) = \mathrm{Len}(DB_2 \mid ST) + \mathrm{Len}(ST)$

Question 5 : Suppose that a code table CT_2 contains itemsets A, B, C, D, BC, ABC. This code table is used to compress DB_2. In which order do the itemsets appear in the code table? Compute the code lengths of each of these itemsets in CT_2 (if an itemset has a usage of 0, its length is 0). You can use the logarithm table at the end of this subject.

Question 6 : Compute $\mathrm{TotLen}(DB_2, CT_2) = \mathrm{Len}(DB_2 \mid CT_2) + \mathrm{Len}(CT_2)$

2 Sequences

Sid	Sequence
1	abc(ab)bc
2	(abc)baa(abc)
3	abcc(ab)c
4	(abc)(abc)(abc)

Question 7 : Of which sequences of the database are sequences $S1 = (ab)(bc)$ and $S2 = b(ab)c$ subsequences? You have to give the correspondence between the items of $S1$ and $S2$ and the items of the sequences of the database. What are the support of $S1$ and $S2$ in this database?

Question 8 : Given the following set of frequent sequences of size 3, what are the candidates sequences of size 4? How do you compute them? $F_3 = \{a(ab), (ab)c, aac, abc, bcb, (ab)b, (abc)\}$

3 Base 2 Logarithm Table

Recall that $\log_2(ab) = \log_2(a) + \log_2(b)$ and thus $\log_2(a/b) = \log_2(a) - \log_2(b)$.

x	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
$\log_2(x)$	0	1	1.6	2	2.3	2.6	2.8	3	3.2	3.3	3.5	3.6	3.7	3.8	3.9	4

Question 1

1	ABC D	ABC	1.5
2	BC D	BC	2
3	A CD	CD	2.5
		C	3.5
		A	2.5
		D	3

$Len(DB_1 | CT_1) = 1 \times code_{CT}(ABC) + 2 \times code_{CT}(D) + 1 \times code_{CT}(BC) + 1 \times code_{CT}(A) + 1 \times code_{CT}(CD)$

$= 1.5 + 2 \times 3 + 2 + 2.5 + 2.5$

$= 1.5 + 6 + 2 + 5$

$= 14.5$

2. $F = \{ABC, BC, CD\}$ (3, 5, 2)

The order is $\{BC, ABC, CD\}$ (5, 3, 2)

$DB_2 =$

1	ABC
2	ABC
3	ABC
4	BC
5	BCD
6	CD

3.

ST

X	Usage	Code length
A	3	
B	5	
C	6	
D	2	

ST =

X	Usage	Code length	
C	6	$-\log_2 \frac{6}{16}$	$= 4 - \log_2 6 = 1.4$
B	5	$-\log_2 \frac{5}{16}$	$= 4 - \log_2 5 = 1.7$
A	3	$-\log_2 \frac{3}{16}$	$= 4 - \log_2 3 = 2.4$
D	2	$-\log_2 \frac{2}{16}$	$= 3$
Total	16		

$Len(ST) = \sum_{C \in ST} Len(C, code_{ST}(C))$

$= \sum_{C \in ST} Len(C) + Len(code_{ST}(C))$

$= 2 \times (1.4 + 1.7 + 2.4 + 3) = 17$

$Len(DB_2 | ST) = 6 \times 1.4 + 5 \times 1.7 + 3 \times 2.4 + 2 \times 3$

$= 30.1$

$Total Len(DB_2, ST) = 30.1 + 17 = 47.1$

$CT_2 =$

X	Usage	Code length
ABC	3	$-\log \frac{3}{8} = 1.4$
BC	2	$-\log \frac{2}{8} = 2$
C	1	$-\log \frac{1}{8} = 3$
B	0	0
A	0	0
D	2	$-\log \frac{2}{8} = 2$
Total		8.4

$Len(ABC) = Len(code_{ST}(A)) + Len(code_{ST}(B)) + Len(code_{ST}(C))$

$= 2.4 + 1.7 + 1.4 = 5.5$ bits

$Len(BC) = Len(code_{ST}(B)) + Len(code_{ST}(C))$

$= 1.7 + 1.4 = 3.1$ bits

$Len(C) = 1.4$ bits

$Len(D) = 3$ bits

$Len(CT_2) = 8.4 + 3 + 1.4 + 5.5 + 3.1$

$= 8.4 + 6.9 + 3.1 + 3$

$= 8.4 + 10 + 3 = 21.4$ bits

$Len(DB_2 | CT_2) = 3 \times 1.4 + 2 \times 2 + 3 \times 1 + 2 \times 2$

$= 4.2 + 4 + 3 + 4$

$= 15.2$ bits.

$Total Len(DB_2 | CT_2) = 21.4 + 15.2$

$= 36.6$

Database : [ABC, ABC, ABC, BC, BCD, CD]

F : [BC, ABC, CD]
5 3 2

Items: [A, B, C, D]

Candidate is BC

Code Table

	Item	nuage	Code Length	length
	BC	5	1.1	3.1
CT_{bc}	C	1	3.5	1.4
	B	0	0	0
	A	3	1.9	2.4
	D	2	2.5	3.0

Total = (1.1 + 3.5 + 1.9 + 2.5 + 3.1 + 1.4 + 2.4 + 3.0)

= 18.900

Compressed Database

	1	ABC	3.0
DB_2 =	2	ABC	3.0
	3	ABC	3.0
	4	BC	1.1
	5	BC D	3.6
	6	CD	6.0

Total database

length = (18.900 + 19.7) = 38.6

*** candidate OK

Candidate is ABC

$DB_1 =$

TID	
1	ABCD
2	BCD
3	ACD

$CT_1 =$

Itemset	Code Length
ABC	1.5
BC	2
CD	2.5
C	3.5
A	2.5
D	3

1 ABC D

2 BC D

3 A CD

$len(DB_1 | CT_1)$

$= 1.5 + 2 \times 3 + 2 + 2.5 + 2.5$

$= 8.5 + 6 = 14.5$

(2)

$DB_2 =$

Tid	Transaction
1	ABC
2	ABC
3	ABC
4	BC
5	BCD
6	CD

$F = \{ABC, BC, CD\}$

3 5 2

By Krimp's algo, $F' = \{ABC, BC, CD\}$

ST =

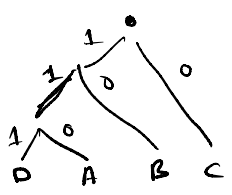
X	Code(X)	occur
A	110	3
B	10	5
C	0	6
D	111	2
Total		16

$Len(ST) = \sum_{X \in ST} Code_{ST}(X)$

$= -\left(\log\frac{3}{16} + \log\frac{5}{16} + \log\frac{6}{16} + \log\frac{2}{16}\right)$

$= -\left(\log \frac{3 \cdot 5 \cdot 6 \cdot 2}{16 \times 16 \times 16 \times 16}\right) = -\log\left(\frac{15}{16^3 \times 2}\right)$

$= 2.74$



$$len(DB_2/ST) = -\left(3 \times \log\frac{3}{16} + 5 \cdot \log\frac{5}{16} + 6 \cdot \log\frac{6}{16} + 2 \cdot \log\frac{2}{16}\right)$$

$$= 9.07$$

$$total(DB_2/ST) = 9.07 + 2.74$$

$$= 11.81$$

A, B, C, D, BC, ABC

3 5 6 2, 5 3

$CT_2 =$

Itemset	Support	$-\ln(Code_{ST}(x))$
ABC	3	$-\ln\frac{3}{8}$
BC	2	$-\ln\frac{2}{8}$
C	1	$-\ln\frac{1}{8}$
B	0	0
A	0	0
D	2	$-\ln\frac{2}{8}$
Total	8	

$$len(CT_2) = -\log\frac{3 \cdot 2 \cdot 1 \cdot 2}{8 \cdot 8 \cdot 8 \cdot 8} + \sum len(code_{ST}(x))$$

$$= 2.55 + \sum_{x \in CT_2} len(code_{ST}(x))$$

$$= 2.55 - \left(\lg\frac{3}{16} + \lg\frac{5}{16} + \ln\frac{6}{16}\right) - \left(\lg\left(\frac{5}{16}\right) + \lg\frac{6}{16}\right)$$

$$+ \left(\ln\frac{6}{16} + \ln\frac{2}{16}\right)$$

$$= 2.55 + 1.66 + 0.93 + 1.33$$

$$= 6.47$$

$DB_2 =$

Tid	Transaction
1	ABC
2	ABC
3	ABC
4	BC
5	BCD
6	CD

$$Len(DB_2 | CT_2) = -\left(3 \times \lg\frac{3}{8} + 2 \times \lg\frac{2}{8} + \ln\frac{1}{8} + 2 \times \ln\frac{2}{8}\right)$$

$$= 4.59$$

$$\text{Total len}(DB_2 | CT_2) = 6.47 + 4.59$$

$$= 11.06$$