Suggested Solution

Data Mining and Data Warehousing

1a)

The choice of minimum support and minimum confidence
The generation of item-set

Let the minimum support = 0.15, minimum confidence = 0.75

A – Egg
B – Potato chips
C – Pasta
D – Cheese
E – Beer
F – Diaper

Itemset	Count	Support
A	12	0.67
B	14	0.78
C	12	0.67
D	5	0.28
E	6	0.33
F	7	0.39

Itemset	Count	Support
AB	8	0.44
AC	8	0.44
AD	3	0.17
AE	4	0.22
AF	5	0.28
BC	8	0.44
BD	4	0.22
BE	5	0.28
BF	4	0.22
CD	1	0.06
CE	5	0.28
CF	6	0.33
DE	0	0.00
DF	1	0.06
EF	1	0.06

Itemset	Count	Support
ABC	4	0.22
ABD	2	0.11
ABE	3	0.17
ABF	2	0.11
ACE	3	0.17
ACF	4	0.22
BCE	4	0.22
BCF	3	0.17

Itemset	Count	Support
ABCE	2	0.11

Interesting association rules (Confidence > minimum confidence)

Rule	Confidence
A->BC	0.33
B->AC	0.28
C->AB	0.33
BC->A	0.50
AC->B	0.50
AB->C	0.50
A->BE	0.25
B->AE	0.22
E->AB	0.52
BE->A	0.61
AE->B	0.77
AB->E	0.39
A->CE	0.25
C->AE	0.25
E->AC	0.52
CE->A	0.61
AE->C	0.77
AC->E	0.39

Rule	Confidence
A->CF	0.33
C->AF	0.33
F->AC	0.56
CF->A	0.67
AF->C	0.79
AC->F	0.50
B->CE	0.28
C->BE	0.33
E->BC	0.67
CE->B	0.79
BE->C	0.79
BC->E	0.50
B->CF	0.22
C->BF	0.25
F->BC	0.43
CF->B	0.52
BF->C	0.77
BC->F	0.39

The interesting rules are:
AE->B, AE->C, AF->C, CE->B, BE->C, BF->C

1b)

This part refers to the interesting rules from 1a). No rules can meet the requirement (Lift ratio>= 1.5), no rules are still interesting.

Rule	Lift ratio
AE->B	0.99
AE->C	1.15
AF->C	1.17
CE->B	1.01
BE->C	1.17
BF->C	1.15

2a)

Let x_1 , x_2 and x_3 be the normalized average monthly payment, normalized average durations of calls, and normalized total calling time respectively.

Let $V = \{v_1, v_2, v_3\}$ be the normalized reference point,
then $v_1 = 0.367$; $v_2 = 0.292$; $v_3 = 0.494$

The distance d is defined as

$$d = \sqrt{\sum_{i=1}^{k} (x_i - v_i)^2}$$

The normalized values and the distance between X and V

	x1	x2	x3	d
1	0.074	0.285	0.000	0.574
2	0.468	0.766	0.401	0.494
3	0.086	0.109	0.599	0.351
4	0.586	0.358	0.104	0.453
5	0.239	0.591	0.294	0.382
6	1.000	0.029	0.869	0.782
7	0.095	0.000	0.203	0.494
8	0.072	0.358	0.398	0.316
9	0.668	0.431	0.582	0.343
10	0.803	0.920	1.000	0.917
11	0.686	1.000	0.302	0.800
12	0.000	0.226	0.700	0.426
13	0.362	0.416	0.500	0.124
14	0.549	0.599	0.394	0.371
15	0.458	0.518	0.610	0.270

The decision of the 5-nn

Customer	Distance	Decision
13	0.124	Stay
15	0.270	Switch
8	0.316	Switch
9	0.343	Undecided
3	0.351	Stay

The decision should be (explanation is required)

Either –

The number of “Stay” is equal to the number of “Switch”, the decision cannot be decided by using k=5, one more data point should be considered.

Customer	Distance	Decision
14	0.371	Stay

The decision should be “Stay”.

Or –

Arbitrary pick one

Or –

The total distance for “Stay” = 0.124 + 0.351 = 0.475
The total distance for “Switch” = 0.270 + 0.316 = 0.586

The total distance for “Stay” is less than the total distance for “Switch”
Therefore, the decision should be “Stay”.

2b)

Let x_1, x_2 be the normalized average monthly payment, and normalized total calling time respectively.

Let $V = \{v_1, v_2\}$ be the normalized reference point,
then $v_1 = 0.285$; $v_2 = 0.521$;

The distance d is defined as

$$d = \sqrt{\sum_{i=1}^{k} (x_i - v_i)^2}$$

The normalized values and the distance between X and V

	X1	X2	d
1	0.074	0.000	0.562
2	0.468	0.401	0.219
3	0.086	0.599	0.214
4	0.586	0.104	0.515
5	0.239	0.294	0.231
6	1.000	0.869	0.795
7	0.095	0.203	0.371
8	0.072	0.398	0.246
9	0.668	0.582	0.388
10	0.803	1.000	0.706
11	0.686	0.302	0.457
12	0.000	0.700	0.337
13	0.362	0.500	0.079
14	0.549	0.394	0.293
15	0.458	0.610	0.195

The average duration of calls of 5-nn

Customer	Distance	Average duration of calls
13	0.079	7.5
15	0.195	8.9
3	0.214	3.3
2	0.219	12.3
5	0.231	9.9

.

The average duration calls expected:
(7.5 + 8.9 + 3.3 + 12.3 + 9.9) / 5 = 8.38

2c) (explanation is required)

If k is free to choose, the choice of k depends on the dataset. k should not be too small. If k is too small (equals to 1 or 2), the k-nn classifier may fit to the noisy data in the dataset.

The distances should be ranked first, and then compare the difference between the current nn and the next nn. If the difference is large, we set a cut between these two.

Ranked distance	Diff	
0.124	0.124	
0.270	0.146	
0.316	0.046	
0.343	0.027	
0.351	0.008	
0.371	0.020	
0.382	0.011	k = 7
0.426	0.044	←
0.453	0.027	
0.494	0.041	
0.494	0.000	
0.574	0.080	
0.782	0.208	
0.800	0.018	
0.917	0.117	

3)

A linear line (x = 1.7) is able to separate two different classes.