Data Mining
Homework 1

Yao Song
301266041

October 7, 2014

1 Question 1

1.1 find the frequent item sets

Due to the independence assumption, the support of an item sets I with k items can be written as $support(I) = \prod_{i=1}^{k} P(i)$,where $support(*)$ is the support of $*$ and $P(*)$ is the probability of $*$. We need to count each item set whose support is no less than 1%. The problem turns out to be finding all groups of numbers from 1 to 10 with a product less than 100. Certainly all the single items will be counted. Then all the pairs will also be counted. Some triples and quadruples will also be counted. The detailed list of groups of items is as follows:
single items:
1 2 3 4 5 6 7 8 9 10
pairs:
(1, 2) (1, 3) (1, 4) (1, 5) (1, 6) (1, 7) (1, 8) (1, 9) (1, 10) (2, 3) (2, 4) (2, 5) (2, 6) (2, 7) (2, 8) (2, 9) (2, 10) (3, 4) (3, 5) (3, 6) (3, 7) (3, 8) (3, 9) (3, 10) (4, 5) (4, 6) (4, 7) (4, 8) (4, 9) (4, 10) (5, 6) (5, 7) (5, 8) (5, 9) (5, 10) (6, 7) (6, 8) (6, 9) (6, 10) (7, 8) (7, 9) (7, 10) (8, 9) (8, 10) (9, 10)
triples:
(1, 2, 3) (1, 2, 4) (1, 2, 5) (1, 2, 6) (1, 2, 7) (1, 2, 8) (1, 2, 9) (1, 2, 10) (1, 3, 4) (1, 3, 5) (1, 3, 6) (1, 3, 7) (1, 3, 8) (1, 3, 9) (1, 3, 10) (1, 4, 5) (1, 4, 6) (1, 4, 7) (1, 4, 8) (1, 4, 9) (1, 4, 10) (1, 5, 6) (1, 5, 7) (1, 5, 8) (1, 5, 9) (1, 5, 10) (1, 6, 7) (1, 6, 8) (1, 6, 9) (1, 6, 10) (1, 7, 8) (1, 7, 9) (1, 7, 10) (1, 8, 9) (1, 8, 10) (1, 9, 10) (2, 3, 4) (2, 3, 5) (2, 3, 6) (2, 3, 7) (2, 3, 8) (2, 3, 9) (2, 3, 10) (2, 4, 5) (2, 4, 6) (2, 4, 7) (2, 4, 8) (2, 4, 9) (2, 4, 10) (2, 5, 6) (2, 5, 7) (2, 5, 8) (2, 5, 9) (2, 6, 7) (2, 6, 8) (3, 4, 5) (3, 4, 6) (3, 4, 7) (3, 4, 8) (3, 5, 6)
quadruples:
(1, 2, 3, 4) (1, 2, 3, 5) (1, 2, 3, 6) (1, 2, 3, 7) (1, 2, 3, 8) (1, 2, 3, 9) (1, 2, 3, 10) (1, 2, 4, 5) (1, 2, 4, 6) (1, 2, 4, 7) (1, 2, 4, 8) (1, 2, 4, 9) (1, 2, 4, 10) (1, 2, 5, 6) (1, 2, 5, 7) (1, 2, 5, 8) (1, 2, 5, 9) (1, 2, 6, 7) (1, 2, 6, 8) (1, 3, 4, 5) (1, 3, 4, 6) (1, 3, 4, 7) (1, 3, 4, 8) (1, 3, 5, 6)

1.2 prove there is no interesting association

Let us assume I is a set of items and j is an item. The association rule is $I \to j$. The confidence of the association rule is $conf(I \to j) = \frac{support(I,j)}{support(I)} = \frac{P(I,j)}{P(I)}$. We know that all the baskets are independent of each other, no sets of items are correlated. So $conf(I \to j) = \frac{P(I,j)}{P(I)} = P(j) = \frac{support(j)}{N}$, where N is the number of baskets in total.

Thus, by the definition of interest: $Interest(j) = conf(I \to j) - \frac{support(j)}{N}$, we know $Interest(j) = 0$ is always true. We conclude that there are no interesting association rules in this data set.

2 Question 2

In pass 1 of the A-Priori algorithm, we need to count each of the one million items. This process needs 4×10^6 bytes of memory.

In the second pass, if we use the triangular matrix method, we need to count all the pairs made up from the N frequent items. That will take $4 \times \frac{N(N-1)}{2} = 2(N^2 - N)$ bytes of memory.

If we use the triples method, we only need to count all the occurring pairs made up from the frequent items, that is $10^6 + M$ pairs according to assumptions 4,5 and 6 in the problem. And we need $12 \times (10^6 + M)$ bytes of memory to count the pairs.

In conclusion, we need a total of $4 \times 10^6 + 2(N^2 - N)$ bytes of main memory if we use the triangular matrix method in the second pass. If we use the triples method, we need a total of $4 \times 10^6 + 12 \times (10^6 + M) = 16 \times 10^6 + 12M$ bytes of main memory. The minimum amount of memory needed is the minimum of the two values $min\{4 \times 10^6 + 2(N^2 - N), 16 \times 10^6 + 12M\}$, which depends on the value of N and M.

3 Question 3

3.1 compute support for each item and each pair of items

The table for items:

item	1	2	3	4	5	6
support	4	6	8	8	6	4

The table for pairs:

pair	1,2	1,3	1,4	1,5	1,6	2,3	2,4	2,5	2,6	3,4	3,5	3,6	4,5	4,6	5,6
support	2	3	2	1	0	3	4	2	1	4	4	2	3	3	2

3.2 which pairs hash to which buckets?

buckets	0	1	2	3	4	5	6	7	8	9	10
pairs	ϕ	(2,6) (3,4)	(1,2) (4,6)	(1,3)	(1,4) (3,5)	(1,5)	(1,6) (2,3)	(3,6)	(2,4) (5,6)	(4,5)	(2,5)
support	0	5	5	3	6	1	3	2	6	3	2

3.3 which buckets are frequent?

From the table above, we can observe buckets 1,2,4,8 are frequent.

3.4 which pairs are counted on the second pass of the PCY algorithm?

All items are frequent, only the pairs in the frequent buckets of pass 1 will be counted in the second pass of PCY algorithm. The pairs are (2,6),(3,4),(1,2),(4,6),(1,4),(3,5),(2,4),(5,6).

4 Question 4

The second pass result is as follows:

buckets	0	1	2	3	4	5	6	7	8
pairs	ϕ	(4,6)	(5,6)	(1,2)	ϕ	(1,4)	(2,4)	(3,4)	(2,6) (3,5)
support	0	3	2	2	0	2	4	4	5

The second pass reduced the set of candidate pairs. Now we only need to count 4 pairs. They are (2,4) (2,6) (3,4) (3,5)

5 Question 5

We know the CF =75%, so z = 1.15.
The computation process is as follows:
adoption of budget resolution =n:
original subtree error: 2.80613948555
replaced leaf error: 2.73057782412
so we **prune** the subtree.

physician fee freeze=n:
original subtree error: 4.61102506017
replaced leaf error: 2.96152268664
so we **prune** the subtree.

duty free exports = n:
original subtree error: 7.69930394053
replaced leaf error: 8.49222452865
so we do **not prune** the subtree

synfuels corporation cutback =y:
original subtree error: 9.06482071828
replaced leaf error: 10.6959586471
so we do **not prune** the subtree

physician freeze=y:
original subtree error: 15.7121952555
replaced leaf error: 15.1970073524
so we **prune** the subtree

water project cost sharing =u:
original subtree error: 2.74613354335
replaced leaf error: 3.22974452245
so we do **not prune** the subtree

physician freeze = u:
original subtree error: 3.74002739023
replaced leaf error: 4.72288283877
so we do **not prune** the subtree

We complete the pruning process now. The final pruned tree is as follows:
physician fee freeze = n: democrat(168/1)
physician fee freeze = y: republican(123/11)
physician fee freeze = u:
..... water project cost sharing = n: democrat(0)
..... water project cost sharing = y: democrat(4)
..... water project cost sharing = u:
..... mx missile = n: republican(0)
..... mx missile = y: democrat(3/1)
..... mx missile = u: republican(2)

6 Question 6

The algorithm will converge in 4 iterations.
For iteration 1:

center	(2,10)	(5,8)	(1,2)	
point	dist1	dist2	dist3	cluster
(2,10)	0	5	9	1
(2,5)	5	6	4	3
(8,4)	12	7	9	2
(5,8)	5	0	10	2
(7,5)	10	5	9	2
(6,4)	10	5	7	2
(1,2)	9	10	0	3
(4,9)	3	2	10	2

For iteration 2:

center	(2,10)	(6,6)	(1.5,3.5)	
point	dist1	dist2	dist3	cluster
(2,10)	0	8	7	1
(2,5)	5	5	2	3
(8,4)	12	4	7	2
(5,8)	5	3	8	2
(7,5)	10	2	7	2
(6,4)	10	2	5	2
(1,2)	9	9	2	3
(4,9)	3	5	8	1

For iteration 3:

center	(3,9.5)	(6.5,5.25)	(1.5,3.5)	
point	dist1	dist2	dist3	cluster
(2,10)	1.5	9.25	7	1
(2,5)	5.5	4.75	2	3
(8,4)	10.5	2.75	7	2
(5,8)	3.5	4.25	8	1
(7,5)	8.5	0.75	7	2
(6,4)	8.5	1.75	5	2
(1,2)	9.5	8.75	2	3
(4,9)	1.5	6.25	8	1

For iteration 4:

center	(3.67,9)	(7,4.33)	(1.5,3.5)	
point	dist1	dist2	dist3	cluster
(2,10)	2.67	10.67	7	1
(2,5)	5.67	5.67	2	3
(8,4)	9.33	1.33	7	2
(5,8)	2.33	5.67	8	1
(7,5)	7.33	0.67	7	2
(6,4)	7.33	1.33	5	2
(1,2)	9.67	8.33	2	3
(4,9)	0.33	7.67	8	1