

Question 2

(c): From results in (a), separate out all strong classification rules, i.e., rules that contain the class attribute (survey answer) on the right-hand-side.

7. distance=2 sex=1 children=1 pets=1 414 ==> answered=1 378 conf:(0.91)

So when customer's distance is in-middle, he is male and he does have pets – then we can say with 91 percent confidence that customer would answer the questionnaire.

(d): Provide a brief summary and interpretation of results.

1. Even though Apriori Algorithm calculates more sets of frequent items (which may include duplicates) but it is beaten by FP-growth Algorithm in that FP Growth has much smaller memory footprint, faster runtime – and is more scalable with huge datasets because of its linear running time.

2. For FP-Growth in part (b), a total of 15 tuples were produced that encompassed all possible combinations.

Customers who did not answer the questionnaire had a support of 87.9 percent which was the highest support of all attributes. And as expected, setting min support any higher (e.g. 0.9) produced no results at all.

Also, all four Boolean attributes being false together had a support of 26.9 percent. Checking the “find min number of itemsets ” meant the min support variable was ignored and it decreases support until specified no. of frequent itemsets is found.

ExampleSet (/Local Repository/Marketing campaign)

ExampleSet (/Local Repository/Marketing campaign_3)

Result History

FrequentItemSets (FP-Growth)

ExampleSet (/Local Repository/Marketing campaign_2)

Data

Annotations

No. of Sets: 15

Total Max. Size: 4

Min. Size:

Max. Size:

Contains Item:

Update View

Size	Support	Item 1	Item 2	Item 3	Item 4
1	0.879	answered			
1	0.636	children			
1	0.631	pets			
1	0.545	sex			
2	0.600	answered	children		
2	0.563	answered	pets		
2	0.515	answered	sex		
2	0.413	children	pets		
2	0.412	children	sex		
2	0.348	pets	sex		
3	0.386	answered	children	pets	
3	0.399	answered	children	sex	
3	0.324	answered	pets	sex	
3	0.283	children	pets	sex	
4	0.269	answered	children	pets	sex

3. For Apriori in part (a), one interesting result was when customer's distance is in-middle, he is male and he does have pets – then we can say with 91 percent confidence that customer would answer the questionnaire.

distance=2 sex=1 children=1 pets=1 414 ==> answered=1 378 conf:(0.91)

There were three instances when confidence was 100 percent e.g. when customer is young, his distance is in-middle and he did answer the questionnaire, then we could say FOR CERTAIN – customer was a male.

age=1 distance=2 answered=1 259 ==> sex=1 259 [conf:\(1\)](#)

Furthermore, when minimum support is 0.01 then the Apriori algorithm has to perform 20 cycles. This is reduced to 18 cycles when when support is increased to 0.1

W-Apriori

Apriori

Minimum support: 0.01 (153 instances)
Minimum metric <confidence>: 0.7
Number of cycles performed: 20

Generated sets of large itemsets:

Size of set of large itemsets L(1): 7
Size of set of large itemsets L(2): 20
Size of set of large itemsets L(3): 27
Size of set of large itemsets L(4): 14
Size of set of large itemsets L(5): 2

Best rules found:

1. age=1 distance=2 answered=1 259 ==> sex=1 259 conf:(1)
2. age=1 answered=1 pets=1 163 ==> children=1 163 conf:(1)
3. age=1 answered=1 sex=1 pets=1 157 ==> children=1 157 conf:(1)
4. age=1 answered=1 pets=1 163 ==> sex=1 157 conf:(0.96)
5. age=1 answered=1 children=1 pets=1 163 ==> sex=1 157 conf:(0.96)
6. age=1 answered=1 pets=1 163 ==> sex=1 children=1 157 conf:(0.96)
7. distance=2 sex=1 children=1 pets=1 414 ==> answered=1 378 conf:(0.91)
8. answered=1 pets=1 813 ==> sex=1 704 conf:(0.87)
9. distance=2 answered=1 pets=1 584 ==> sex=1 492 conf:(0.84)
10. answered=1 children=1 pets=1 672 ==> sex=1 564 conf:(0.84)
11. age=1 answered=1 385 ==> sex=1 321 conf:(0.83)
12. answered=1 pets=1 813 ==> children=1 672 conf:(0.83)
13. distance=2 answered=1 children=1 886 ==> sex=1 725 conf:(0.82)
14. distance=2 answered=1 1264 ==> sex=1 1032 conf:(0.82)
15. age=1 answered=1 sex=1 321 ==> distance=2 259 conf:(0.81)
16. distance=2 answered=1 children=1 pets=1 460 ==> sex=1 378 conf:(0.81)
17. distance=2 answered=1 pets=1 584 ==> children=1 469 conf:(0.8)
18. answered=1 sex=1 pets=1 704 ==> children=1 564 conf:(0.8)
19. answered=1 children=1 1296 ==> sex=1 1037 conf:(0.8)
20. age=1 answered=1 sex=1 children=1 203 ==> pets=1 157 conf:(0.77)
21. distance=2 answered=1 sex=1 pets=1 492 ==> children=1 378 conf:(0.77)
22. age=1 answered=1 children=1 267 ==> sex=1 203 conf:(0.76)
23. answered=1 1853 ==> sex=1 1391 conf:(0.75)
24. answered=1 sex=1 1391 ==> children=1 1037 conf:(0.75)

Repository

DB

Local Repository (zainq)

data (zainq)

processes (zainq)

abc (zainq - v1, 2/19/17 4:15 PM - 4 KB)

communities (zainq - v1, 2/19/17 2:00 PM - 1.6 MB)

communities_data_3 (zainq - v1, 2/19/17 3:13 PM - 1.6 MB)

communities_1 (zainq - v1, 2/19/17 2:15 PM - 1.6 MB)

communities_FINAL (zainq - v1, 2/19/17 4:10 PM - 1.6 MB)

communities_test (zainq - v1, 3/1/17 6:53 PM - 417 KB)

Marketing campaign_2 (zainq - v1, 3/1/17 1:47 PM - 417 KB)

Marketing campaign_3 (zainq - v1, 3/1/17 2:00 PM - 238 KB)

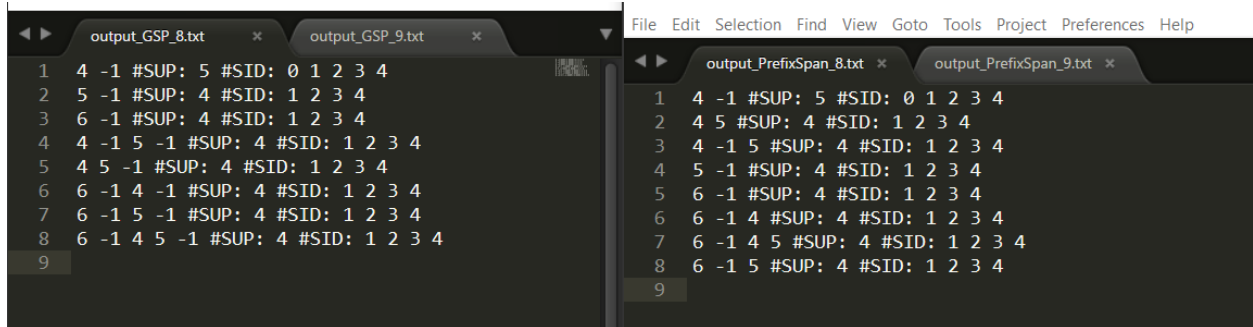
missing_vals_inserted (zainq - v1, 2/19/17 3:00 PM - 4 KB)

Cloud Repository (disconnected)

Question 3

For at least one minimum support value, verify that the the computed sequences are correct:

Output of GSP (first) vs PrefixSpan (second) when confidence is 0.8:




The screenshot shows two side-by-side text editors. The left editor, titled 'output_GSP_8.txt', contains 9 lines of output. The right editor, titled 'output_PrefixSpan_8.txt', contains 9 lines of output. Both outputs list sequences with their support and sequence IDs.

```
1 4 -1 #SUP: 5 #SID: 0 1 2 3 4
2 5 -1 #SUP: 4 #SID: 1 2 3 4
3 6 -1 #SUP: 4 #SID: 1 2 3 4
4 4 -1 5 -1 #SUP: 4 #SID: 1 2 3 4
5 4 5 -1 #SUP: 4 #SID: 1 2 3 4
6 6 -1 4 -1 #SUP: 4 #SID: 1 2 3 4
7 6 -1 5 -1 #SUP: 4 #SID: 1 2 3 4
8 6 -1 4 5 -1 #SUP: 4 #SID: 1 2 3 4
9
```

```
1 4 -1 #SUP: 5 #SID: 0 1 2 3 4
2 4 5 #SUP: 4 #SID: 1 2 3 4
3 4 -1 5 #SUP: 4 #SID: 1 2 3 4
4 5 -1 #SUP: 4 #SID: 1 2 3 4
5 6 -1 #SUP: 4 #SID: 1 2 3 4
6 6 -1 4 #SUP: 4 #SID: 1 2 3 4
7 6 -1 4 5 #SUP: 4 #SID: 1 2 3 4
8 6 -1 5 #SUP: 4 #SID: 1 2 3 4
9
```

Output of GSP (first) vs PrefixSpan (second) when confidence is 0.9:



The screenshot shows two side-by-side text editors. The left editor, titled 'output_GSP_9.txt', contains 2 lines of output. The right editor, titled 'output_PrefixSpan_9.txt', contains 2 lines of output. Both outputs list sequences with their support and sequence IDs.

```
1 4 -1 #SUP: 5 #SID: 0 1 2 3 4
2
```

```
1 4 -1 #SUP: 5 #SID: 0 1 2 3 4
2
```

Sample screenshot while using PrefixSpan:

