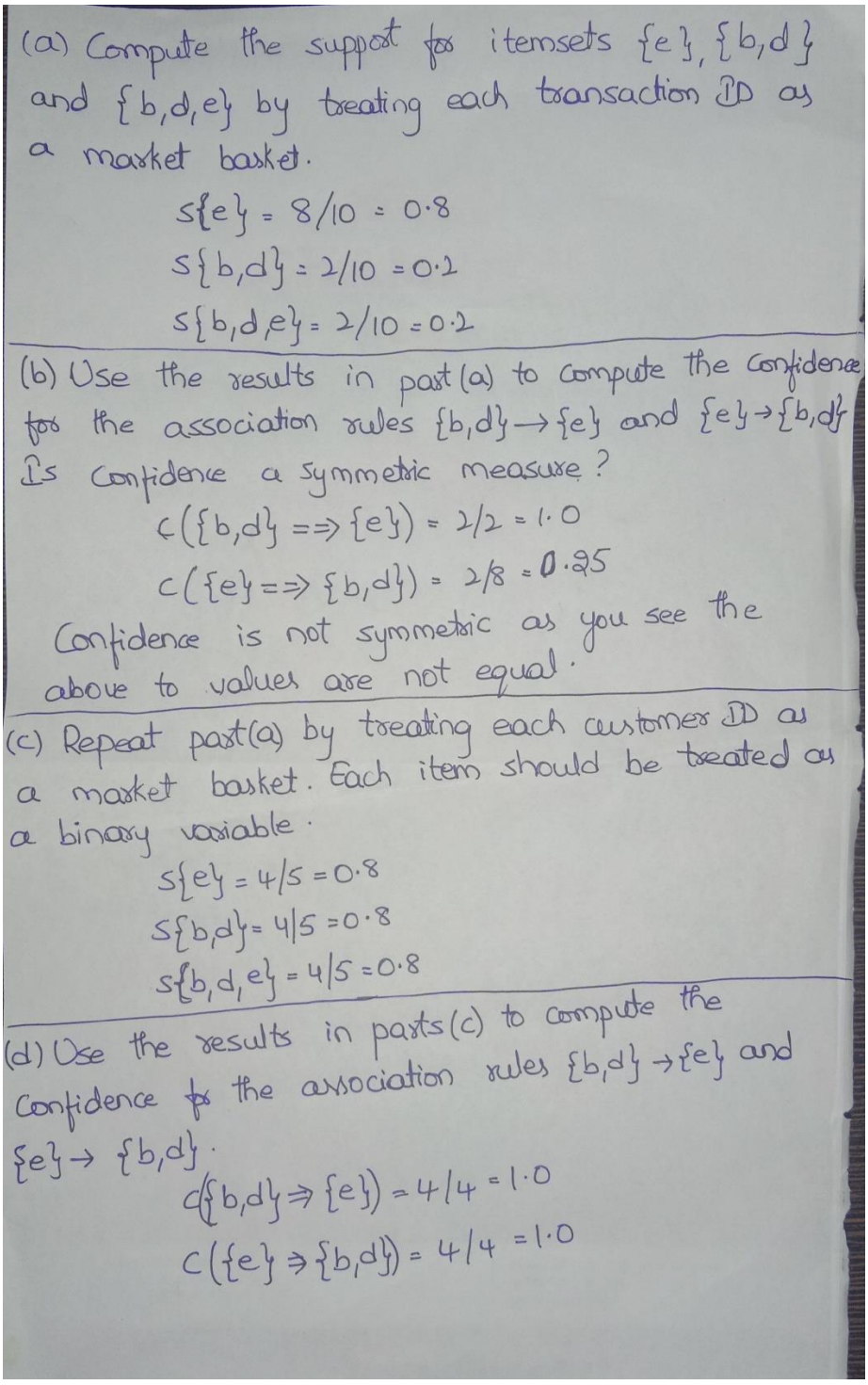
Data Mining Assignment 3

1) Read Chapter 6 (only sections 6.1 and 6.7).

2) Do Chapter 6 textbook problem #2 (parts a, b, c, d only) on page 404.



(a) Compute the support for itemsets {e}, {b,d} and {b,d,e} by treating each transaction ID as a market basket.

$s\{e\} = 8/10 = 0.8$

$s\{b,d\} = 2/10 = 0.2$

$s\{b,d,e\} = 2/10 = 0.2$

(b) Use the results in part (a) to compute the confidence for the association rules {b,d} → {e} and {e} → {b,d}. Is confidence a symmetric measure?

$c(\{b,d\} \Rightarrow \{e\}) = 2/2 = 1.0$

$c(\{e\} \Rightarrow \{b,d\}) = 2/8 = 0.25$

Confidence is not symmetric as you see the above to values are not equal.

(c) Repeat part (a) by treating each customer ID as a market basket. Each item should be treated as a binary variable.

$s\{e\} = 4/5 = 0.8$

$s\{b,d\} = 4/5 = 0.8$

$s\{b,d,e\} = 4/5 = 0.8$

(d) Use the results in parts (c) to compute the confidence for the association rules {b,d} → {e} and {e} → {b,d}.

$c(\{b,d\} \Rightarrow \{e\}) = 4/4 = 1.0$

$c(\{e\} \Rightarrow \{b,d\}) = 4/4 = 1.0$

3) Do Chapter 6 textbook problem #6 (parts d, e only) on page 406.

With one item:

ItemSet	Support
Diapers	7
Milk	5
Bread	5
Butter	5
Beer	4
Cookies	4

With two items:

ItemSet	Support
Diapers, Milk	4
Diapers, Bread	3
Diapers, Butter	3
Diapers, Beer	3
Diapers, Cookies	2
Milk, Bread	3
Milk, Butter	2
Milk, Beer	1
Milk, Cookies	1
Bread, Butter	5
Bread, Beer	0
Bread, Cookies	1

Butter, Bread	0
Butter, Cookies	1
Beer, Cookies	2

(d) Find an itemset (of size 2 or larger) that has the largest support.

Ans: {bread, butter}

(e) Find a pair of items, a and b, such that the rules {a} -> {b} and {b} -> {a} have the same confidence.

Ans: C({bread, butter}) = support({bread, butter})/support({bread}) = 5/5 = 1

C({beer, cookies}) = 1

4) Using the data at www.stats202.com/more_stats202_logs.txt and treating each row as a "market basket" compute the support and confidence for the rule ip=65.57.245.11 → "Mozilla/5.0 (X11; U; Linux i686 (x86_64); en-US; rv:1.8.1.3) Gecko/20070309 Firefox/2.0.0.3".

State what the support and confidence values mean in plain English in this context.

Support for the above rule: Transactions containing all the items in the above rule/Total number of logs

Confidence: Support (entire rule)/Support (IP address)

```
$ python stats.py
Count of A:  3636
Count of B:  234
Count of AB:  1385
Total Transactions:  14809
-------------------------------------
Support of A: 0.2455263690998717
Support of B: 0.015801201971773923
Support of AB: 0.09352420825173881
Confidence of AB: 0.3809130913091309
```

Support: The number of transactions that include the items in the X and Y part of the rule as a percentage of the total number of transactions. It is a measure of how frequently the collection of items occur together as a percentage of all transactions.

X -> Y

Support = Number of transactions containing all the items in X and Y / Total number of transactions

Here X refers to the IP address and Y refers to browser information

Confidence: It is the ratio of the number of transactions that includes all items in {B} as well as the number of transactions that includes all items in {A} to the number of transactions that includes all items in {A}

A->B

Confidence = Support ({A, B})/Support ({A})

Here A refers to the IP address and B refers to browser information