

## 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.

Customer ID	Transaction ID	Items Bought
1	0001	{a, d, e}
1	0024	{a, b, c, e}
2	0012	{a, b, d, e}
2	0031	{a, c, d, e}
3	0015	{b, c, e}
3	0022	{b, d, e}
4	0029	{c, d}
4	0040	{a, b, c}
5	0033	{a, d, e}
5	0038	{a, b, e}

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

10 distinct baskets/transactions.

- {e}:  $s = 8 / 10 = 0.8$
- {b, d}:  $s = 2 / 10 = 0.2$
- {b, d, e}:  $s = 2 / 10 = 0.2$

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

Both rules have support 0.2, (support count is 2):

- $\{b, d\} \rightarrow \{e\}$ :  $c = 0.2 / 0.2 = 1$
- $\{e\} \rightarrow \{b, d\}$ :  $c = 0.2 / 0.8 = 0.25$

Support is a symmetric measure, but confidence is not symmetric

**c) Repeat part (a) by treating each customer ID as a market basket. Each item should be treated as a binary variable (1 if an item appears in at Least one transaction bought by the customer, and 0 otherwise.)**

Now we have 5 baskets in total.

- $\{e\}$ :  $s = 4/5 = 0.8$
- $\{b, d\}$ :  $s = 5/5 = 1$
- $\{b, d, e\}$ :  $s = 4/5 = 0.8$

**d) Use the results in part (c) to compute the confidence for the association rules  $\{b, d\} \rightarrow \{e\}$  and  $\{e\} \rightarrow \{b, d\}$ .**

- $\{b, d\} \rightarrow \{e\}$ :  $c = 0.8/1 = 0.8$
- $\{e\} \rightarrow \{b, d\}$ :  $c = 0.8 / 0.8 = 1$

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

Transaction ID	Items Boueht
1	{Milk, Beer, Diapers}
2	{Bread, Butter, Milk}
3	{Milk, Diapers, Cookies}
4	{Bread, Butter, Cookies}
5	{Beer, Cookies, Diapers}
6	{Milk, Diapers, Bread, Butter}
7	{Bread, Butter, Diapers}

8	{Beer, Diapers}
9	{Milk, Diapers, Bread, Butter}
10	{Beer, Cookies}

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

Itemset	Support
cookies   milk	1
bread   cookies	1
milk	5
beer   cookies	2
beer   diapers	3
bread   butter   milk	3
bread   butter   cookies	1
beer   milk	1
butter   cookies	1
butter   milk	3
butter	5
bread   butter   diapers   milk	2
bread   butter	5
bread	5
butter   diapers   milk	2
bread   diapers	3
cookies	4

beer	4
butter   diapers	3
diapers	7
diapers   milk	4
beer   cookies   diapers	1
beer   diapers   milk	1
bread   diapers   milk	2
bread   butter   diapers	3
bread   milk	3
cookies   diapers   milk	1
cookies   diapers	2
∅	10

The table is having all item sets with non-zero support count Ignoring the 1-itemsets (and ∅ ), the itemset with the largest support is {bread, butter}.

**(e) Find a pair of items, a and b, such that the rules  $\{a\} \rightarrow \{b\}$  and  $\{b\} \rightarrow \{a\}$  have the same confidence.**

Bread and butter have the same support ( $s = 5$ ). This means that the rules  $\{\text{bread}\} \rightarrow \{\text{butter}\}$  and  $\{\text{butter}\} \rightarrow \{\text{bread}\}$  have the same confidence ( $c = 5/5 = 1$ ). The same can be said with beer and cookies ( $s = 4, c = 2/4 = 0.5$ ).

**4) Using the data at [www.stats202.com/more\\_stats202\\_logs.txt](http://www.stats202.com/more_stats202_logs.txt) and treating each row as a "market basket" compute the support and confidence for the rule  $\text{ip}=65.57.245.11 \rightarrow \text{"Mozilla/5.0 (X11; U; Linux i686 (x86_64); en-US; rv:1.8.1.3) Gecko/20070309 Firefox/2.0.0.3"}$ .**

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

The rule for which we have to find the support and confidence of the given Address is {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"} Support for {65.57.245.11} =  $5021/14803 = 0.33$

The support for {"Mozilla/5.0 (X11; U; Linux i686 (x86\_64); en-US; rv:1.8.1.3) Gecko/20070309 Firefox/2.0.0.3"} =  $1619 / 14803 = 0.109$

Confidence for rule {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"} =  $\text{support count}(\{65.57.245.11, \text{"Mozilla/5.0 (X11; U; Linux i686 (x86_64); en-US; rv:1.8.1.3) Gecko/20070309 Firefox/2.0.0.3"}\}) / \text{support count}(\{65.57.245.11\}) = 1619 / 5021 = 0.322$