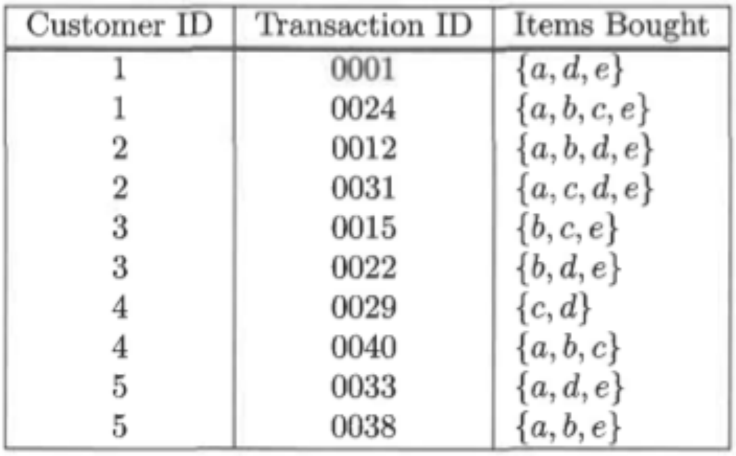
Data Mining Assignment 3

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. Consider the data set shown in Table 6.22

Table 6.22. Example of market basket transactions.



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.

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} -> {e}) = 0.2/0.2 = 100%

c({e} -> {b,d}) = 0.2/0.8 = 25%

No, confidence is not a symmetric measure.

(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.)

s({e}) = 4/5 = 0.8

s({b,d}) = 5/5 = 1

s({b,d,e}) = 4/5 = 0.8

(d)Use the results in part (c) to compute the confidence for the association rules {b,d} -> {e} and {e} -> {b,d}

c({b,d} -> {e}) = 0.8 / 1 = 80%

c({e} -> {b,d}) = 0.8 / 0.8 = 100%

3) Do Chapter 6 textbook problem #6 (parts d,e only) on page 406.
Consider the market basket transactions shown in Table 6.23

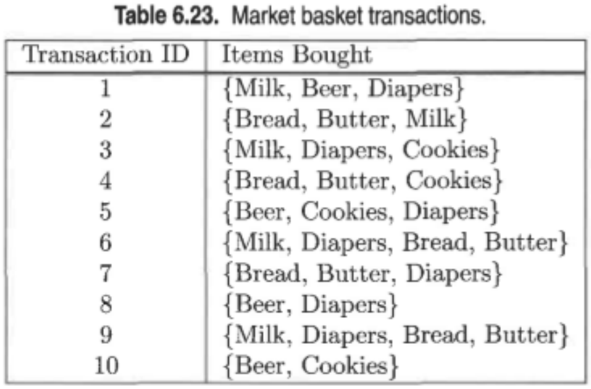


Table 6.23. Market basket transactions.

Transaction ID	Items Bought
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.

{Bread, Butter}

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

{Beer, Cookies} or {Bread, Butter}

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".

Support(S) = 1385 / 14803 = 9%

Confidence(c) = 1385 / 5018 = 27%

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

9% support means there are only 9percent of the records which have this ip=65.57.245.11 and the crawler="Mozilla/5.0 (X11; U; Linux i686 (x86_64); en-US; rv:1.8.1.3) Gecko/20070309 Firefox/2.0.0.3".

Whereas the confidence is 27% times it is that crawler of all the times where it's the ip=65.57.245.11