Assignment-3 Solution

2. Consider the data set shown in Table

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.

support = (Frequency of item set) / (total number of item sets)

The support for **{e}** is **0.8**.

EXPLANATION:In simple words e appeared 8 times out of 10 ie 8/10

The support for **{b, d}** is **0.2**.

EXPLANATION: as mentioned in above explanation

The support for **{b, d, e}** is **0.2.**

EXPLANATION: as mentioned in above explanation

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

Confidence = Measure how frequently Y appears in a transaction that contains X.

Confidence for {b, d} as measured by {e} = support ({b, d, e}) / support ({b, d})

= 0.2 / 0.2 = 1.

Confidence for {e} as measured {b, d} = support ({b, d, e}) / support ({e})

= 0.2 / 0.8 = 0.25

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.)

support = (Frequency of g item set) / (total number of item sets)

The support for **{e}** is **0.8**.

EXPLANATION: ‘e’ appeared 4 times in 5 Customer ID transactions mentioned above.

The support for **{b, d}** is **1.**

EXPLANATION:‘b’ & ‘d’ appeared 5 times together in 5 Customer ID transactions mentioned above.

The support for **{b, d, e}** is **0.8.**

EXPLANATION: ‘b’, ‘d’ & ‘e’ appeared 4 times together in 5 Customer ID transactions mentioned above.

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

Confidence = Measure how often Y appears in a transaction that contains X.

Confidence for {b, d} -> {e} = support ({b, d, e}) / support ({b, d})

= 0.8 / 1 = 0.8 which implies 80%

Confidence for {e} -> {b, d} = support ({b, d, e}) / support ({e})

= 0.8 / 0.8 = 1which implies100%

No, confidence is not an asymmetric measure.

3. 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.

Item	Support Count
Milk	5
Beer	4
Diapers	7
Bread	5
Cookies	4
Butter	5

Support Count for 2 items is as below

Item	Support Count
Milk, Beer	1
Milk, Diaper	4
Milk, Bread	3
Milk, Cookies	1
Milk, Butter	2
Beer, Diaper	3
Beer, Bread	0
Beer, Cookies	2
Beer, Butter	0
Diaper, Bread	3
Diaper, Cookies	2
Diaper, Butter	3
Bread, Cookies	1
Bread, Butter	5
Cookies, Butter	1

As the number of k items sets keep increasing, the support count will be the same or decrease. So, for 2-item set the support will be higher.
So, looking into the above table **{bread, butter}** will have higher support followed by {milk, diaper} data sets.

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

Looking into table 1 of the above problem the support count for Milk, **bread, butter** and Beer, Cookies are the same.

The following pairs will have the same confidence

1. Milk, butter and butter, milk
2. Milk, bread and bread, milk

3. Bread, butter and butter, bread
4. Beer, cookies and cookies, beer

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

The rule for which we have to find the support and confidence 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

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"}

= support count ({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 count ({65.57.245.11})

= 1619 / 5021= 0.322