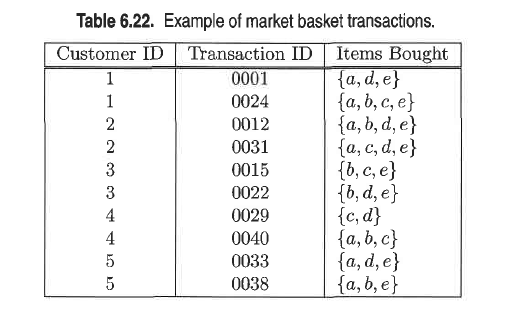
Data Mining Assignment 3

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 6.1.



(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*(*bd 🡪 e*) = 0.2/0.2 = 100%

*c*(*e 🡪 bd*) = 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

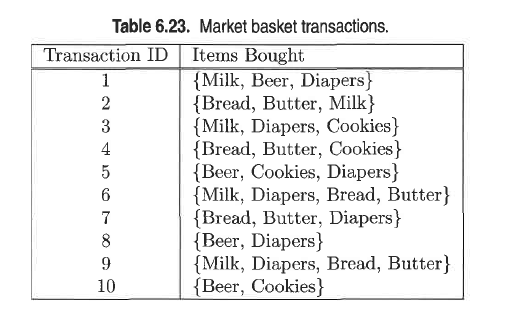
(d) Use the results in part (c) to compute the confidence for the association

rules *{b, d} −→ {e}* and *{e} −→ {b, d}*.

*c*(*bd 🡪 e*) = 0.8/1 = 80%

*c*(*e 🡪 bd*) = 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.



(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.

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

(Beer, Cookies) or (Bread, Butter).  
  
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 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