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

|  |  |  |
| --- | --- | --- |
| 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} |

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

* {e}:  support ---> 8/10 = 80%
* {b, d}: support ---> 2/10 = 20%
* {b, d, e}: support ---> 2/10 = 20%

1. 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?

* {b, d} ---> {e}: confidence --> 2/2 = 100%
* {e} ---> {b, d}: confidence ---> 2/8 = 25%

Confidence is not a symmetric measurement.

1. 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.)

* {e}: support ---> 4/5 = 80%
* {b, d}: support ---> 5/5 = 100%
* {b, d, e}: support ---> 4/5 =80%

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

* {b, d} ---> {e}: confidence ---> 4/5 = 80%
* {e} ---> {b, d}: confidence ---> 4/4 = 100%

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

Consider the market basket transactions shown in Table

|  |  |
| --- | --- |
| 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.

{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