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} |

1. Compute the support for the 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

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?

c(bd 🡪 e) = 0.2/0.2 = 100%

c(e🡪 bd) = 0.2/0.8 = 25%

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 appers 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 = 0.9

1. 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.

|  |  |
| --- | --- |
| Transaction ID | Items Bought |
| 1 | {Milk, Beer, Diapers} |
| 2 | {Milk, Bread, Butter} |
| 3 | {Milk, Diapers, Cookies} |
| 4 | {Bread, Butter, Cookies} |
| 5 | {Beer, Diapers, Cookies} |
| 6 | {Milk, Diapers, Bread, Butter} |
| 7 | {Diapers, Bread, Butter} |
| 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 |
| Milk | Cookies | 1 |
| Bread | Cookies | 1 |
| Milk | 5 |
| Beer | Cookies | 2 |
| Beer | Diapers | 3 |
| Milk | Bread | Butter | 3 |
| Bread | Butter | Cookies | 1 |
| Milk | Beer | 1 |
| Butter | Cookies | 1 |
| Milk | Butter | 3 |
| Butter | 5 |
| Milk | Diapers | Bread | Butter | 2 |
| Bread | Butter | 5 |
| Bread | 5 |
| Milk | Diapers | Butter | 2 |
| Diapers | Bread | 3 |
| Cookies | 4 |
| Beer | 4 |
| Diapers | Butter | 3 |
| Diapers | 7 |
| Milk | Diapers | 4 |
| Cookies | Diapers | 2 |
| ∅ | 10 |

1. Find a pair of items a and b such that the rules {a} 🡪 {b} and {b} 🡪 {a} have the same confidence.

Bread and butter have the same support (s= 5). This means that the rules {bread} 🡪 {butter} and {butter} 🡪 {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 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 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"} = 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