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.
3. 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

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 ({b, d} 🡺 {e}) = 2/2 = 1.0

C({e} 🡺{b, d}) = 2/8 = 0.25

Confidence is not symmetric as you see the above to values are not equal.

1. Repeat part(a) by treating each customer ID as a market basket. Each item should be treated as a binary variable.

S {e} = 4/5 = 0.8

S {b, d} = 4/5 = 0.8

S {b, d, e} = 4/5 = 0.8

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

C ({b, d} 🡺 {e}) = 4/4 = 1.0

C({e} 🡺 {b, d}) = 4/4 = 1.0

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

With one item:

|  |  |
| --- | --- |
| **Item set** | **Support** |
| Diapers | 7 |
| Milk | 5 |
| Bread | 5 |
| Butter | 5 |
| Beer | 4 |
| Cookies | 4 |

With two items:

|  |  |
| --- | --- |
| **Itemset** | **Support** |
| Diapers, milk | 4 |
| Diapers, Bread | 3 |
| Diapers, butter | 3 |
| Diapers, beer | 3 |
| Diapers, cookies | 2 |
| Milk, Bread | 3 |
| Milk, Butter | 2 |
| Milk, Beer | 1 |
| Milk, cookies | 1 |
| Bread, butter | 5 |
| Bread, beer | 0 |
| Bread, cookies | 1 |
| Butter, Bread | 0 |
| Butter, cookies | 1 |
| Beer, cookies | 2 |

d. find an itemset that has the largest support.

{bread, butter}

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

C ({Bread, butter}) = support ({bread, butter})/support({bread}) = 5/5 = 1 C({beer, cookies}) = 1  
  
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".

* Support (for above rule): transactions containing all the items in the total number of logs.
* Confidence: support (entire rule)/support (IP)
* Support: the number of transactions include the items of X and Y part of the rule as the percentage of total number of the transactions. It’s the measure of how frequently the items occur together as the percentage.

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