Mark Gameng

HW 8

Part 1.1

1. Not sure what the whole market basket domain is, so I just went with common grocery items and some items in table 5.2
    a. Milk -> Bread. This has a high support and high confidence as it's a common thing to buy and usually bought together, thus not interesting as well. Can be interesting if the items in the rules aren't so common and known.
    b. Bread -> Water. Water and bread sales are usually high, thus reasonably high support, but not all transactions with bread contain water, so not that high confidence. Due to low confidence, its not really interesting.
    c. Rice -> Oil. Rice and oil aren't bought that often due to their quantity bottle/packaging, thus low support, and also low confidence because rice and oil aren't commonly paired. Also, not interesting.
    d. Diaper -> Beer. Beer and diapers aren't that common, and is only if has family/pregnant, and alcoholic and such. Thus, low support, but diaper and beer can have high confidence because diapers -> baby = stress -> beer. With a low support but high confidence, this rule is quite interesting.
2. s
    a. $S(\{e\}) = 8/10 = 0.8$
       $S(\{b, d\}) = 2/10 = 0.2$
       $S(\{b, d, e\}) = 2/10 = 0.2$
    b. $C(\{b, d\} \rightarrow \{e\}) = S(\{b, d, e\}) / S(\{b, d\}) = 0.2 / 0.2 = 1$
       $C(\{e\} \rightarrow \{b, d\}) = S\{b, d, e\}) / S(\{e\}) = 0.2 / 0.8 = 0.25$
       Thus, confidence is not a symmetric measure
    c. $S(\{e\}) = 4/5 = 0.8$
       $S(\{b, d\}) = 5/5 = 1$
       $S(\{b, d, e\}) = 4/5 = 0.8$
    d. $C(\{b, d\} \rightarrow \{e\}) = S(\{b, d, e\}) / S(\{b, d\}) = 0.8 / 0.1 = 0.8$
       $C(\{e\} \rightarrow \{b, d\}) = S\{b, d, e\}) / S(\{e\}) = 0.8 / 0.8 = 1$
    e. I feel like there is no direct relationship between s1, s2 and c1, c2, other than the fact that s2 and c2 would be higher than s1 and c1 because treating customer ID as a market basket would increase support and confidence.
9.
    a.
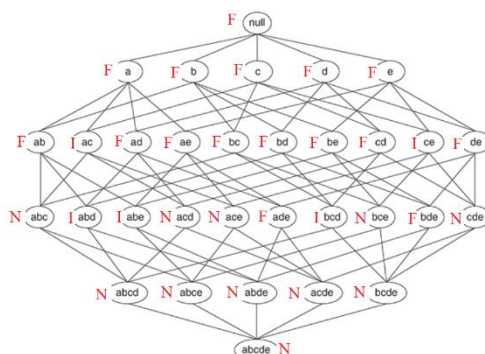
<ol type="b" start="2">
32 item sets with 16 frequent.
Thus, 16/32 = 0.5 -> 50% of frequent item sets

15.

<ol type="a">
Dataset a because each block achieves minsup >= 10%
Dataset d because no items achieves minsup >= 10%
Dataset e because lots of items with minsup >= 10% are also in the same transaction ranges. Like at 2000 to 4000 range. Will be longer than dataset a.
Dataset b because of the long strip, ranging from the lower to upper transaction range.
Dataset e because it has varying length of strips, transactions wise, more so than other datasets.

Part 1.2

1. a – minsup = 3/8

minsup = 3/8, thus minimum support count is 3

Tries all possible item sets and gets the counts of each and if >= 3 then frequent itemset. Prunes whenever possible.

Set	Count
{A}	5
{B}	4
{C}	4
{D}	6
{E}	~~1~~
{G}	5
{F}	4

Set	Count
A,b	3
A,c	3
A,d	4
~~A,f~~	~~2~~
~~A,g~~	~~2~~
~~B,c~~	~~2~~
~~B,d~~	~~2~~
~~B,f~~	~~1~~
~~B,g~~	~~2~~
C,d	4
~~C,f~~	~~2~~
C,g	3
D,f	3
D,g	3
~~F,g~~	~~2~~

Set	Count
~~A,b,c~~	~~1~~
~~A,b,d~~	~~2~~
A,c,d	3
~~C,d,g~~	~~2~~
~~D,f,g~~	~~2~~

Thus, frequent patterns from this dataset are the non-crossed out sets above - {A, B, C, D, G, F, (A,B), (A,C), (A,D), (C,D), (C,G), (D, F), (D, G), (A, C, D)}

4. Possible rules with set ABE: {A} -> {B, E}, {B} -> {A, E}, {E} -> {A, B}, {A, B} -> {E}, {A, E} -> {B}, {B,E} -> {A}, {A} -> {B}, {A} -> {E}, {B} -> {A}, {B} -> {E}, {E} -> {A}, {E} -> {B}

{A} -> {B, E}: Support = 2/6 = 1/3; Confidence = 2/4 = ½

{B} -> {A, E}: Support = 1/3; Confidence = 2/5

{E} -> {A, B}: Support = 1/3; Confidence = 2/4 = ½

{A, B} -> {E}: Support = 1/3; Confidence = 2/3

{A, E} -> {B}: Support = 1/3; Confidence = 2/2 = 1

{B,E} -> {A}: Support = 1/3; Confidence = 2/4 = ½

{A} -> {B}: Support = 3/6 = ½; Confidence = ¾

{A} -> {E}: Support = 2/6 = 1/3; Confidence = 2/4 = 1/2

{B} -> {A}: Support = ½; Confidence = 3/5

{B} -> {E}: Support = 4/6 = 2/3; Confidence = 4/5

{E} -> {A}: Support = 1/3; Confidence = 2/4 = 1/2

{E} -> {B}: Support = 2/3; Confidence = 4/4 = 1