Marc Soda HW 3

1) $Supp(A) = \frac{freq(A)}{num\ tX}$

$conf(A, B) = \frac{supp(A \cup B)}{supp(A)}$

a) bread → milk

Support: 3/5, confidence: $\frac{3/5}{4/5} = 3/4$

Not very interesting because both are things people commonly view as necessary to have.

b) milk → coke

support: 2/5, conf: $\frac{2/5}{4/5} = 2/4$

May be interesting because milk is commonly seen as healthy and soda is not

c) Eggs → Bread

supp: 1/5 conf: $\frac{1/5}{1/5} = 1$

Uninteresting because both foods are very common.

d) Bread - eggs

supp: 1/5, conf: $\frac{1/5}{4/5} = 1/4$

Interesting because one would expect bread and eggs to have a stronger association because both are very common to buy

2) a)

(1, 2, 3) : (1, 2, 3, 4), (1, 2, 3, 5)
(1, 2, 4) : (1, 2, 4, 5)
(1, 2, 5) : none
(1, 3, 4) : (1, 3, 4, 5)
(1, 3, 5) : none
(2, 3, 4) : (2, 3, 4, 5)
(2, 3, 5) : none
(3, 4, 5) : none

Answer: (1,2,3,4), (1,2,3,5), (1,2,4,5), (1,3,4,5), (2,3,4,5)

b) ~~MA-Step=4~~. This step in the algorithm is effectively the same as the procedure for part A

Answer is same as part a:
(1,2,3,4), (1,2,3,5), (1,2,4,5), (1,3,4,5), (2,3,4,5)

c)

- (1, 2, 3, 4) survives because (1, 2, 3), (1, 2, 4), (1, 3, 4), and (2, 3, 4) are frequent
- (1, 2, 3, 5) survives because all of its subsets are frequent
- (1, 2, 4, 5) is pruned because (1, 4, 5) is not frequent
- (1, 3, 4, 5) is pruned because (1, 4, 5) is not frequent
- (2, 3, 4, 5) is pruned because (2, 4, 5) is not frequent.

3) a) $conf(a,b) = \frac{supp(a,b)}{supp(a)}$

$= \frac{1/5}{1/4} = 4/5 = 80\%$

$a \to b$ is interesting because it is above the confidence threshold

b) $int(a,b) = supp(a,b) / supp(a) \cdot supp(b)$ $\quad \frac{1}{4} \quad \frac{9}{10}$

$= 1/5 \;/\; 1/4 \cdot 9/10$

$= 1/5 \;/\; 9/40$

$= 40/45 = .889$

c) a and b are negatively correlated

Just because a rule has high confidence does not mean it is interesting. Other measures sometimes need to be considered when analyzing association.

d) i: Note, I know my notation is strange, but I can't draw curly braces. You can tell what I mean.

$c(a,b) = \frac{s(a,b)}{s(a)} < s(b)$

$\Rightarrow s(a,b) < s(b)\,s(a) \quad (1)$

$s(\bar{a},b) = s(b) - s(a,b)$

$s(\bar{a}) = 1 - s(a)$

$c(\bar{a},b) = \frac{s(\bar{a},b)}{s(\bar{a})} = \frac{s(b) - s(a,b)}{1 - s(a)}$

$c(\bar{a},b) - c(a,b) = \frac{s(a)s(b) - s(a,b)}{(1-s(a))(s(a)}$ ← this is always positive; ← this is always positive

i $\therefore c(\bar{a},b) > c(a,b)$ if $c(a,b) < s(b)$

ii $c(\bar{a},b) - s(b) = \frac{s(a)s(b) - s(a,b)}{1 - s(a)}$ ← this is always positive; ← this is always positive

$\therefore c(\bar{a},b) > s(b)$ if $c(a,b) < s(b)$

4) a) min_sup = 3

items	supp
a	5
b	6
c	5
d	9
e	6

all survive →

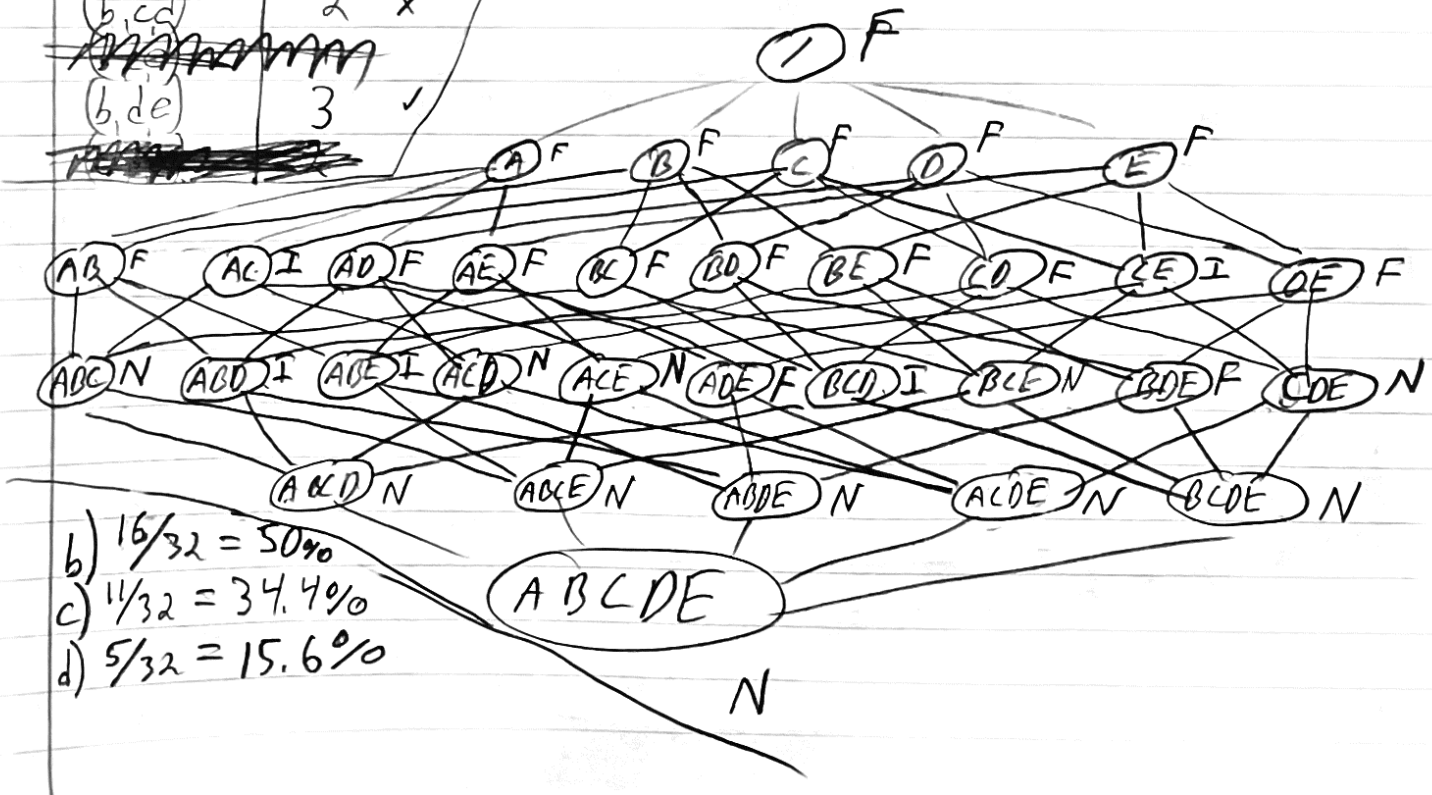
items	supp
(a,b)	3 ✓
~~(a,c)~~	~~2~~
~~(a,e)~~	~~2~~
(a,d)	4 ✓
(a,e)	4 ✓
(b,c)	3 ✓
(b,d)	6 ✓
(b,e)	4 ✓
(c,d)	4 ✓
~~(c,e)~~	~~2~~
(d,e)	6 ✓

next row →

Itemsets that are striked through did not survive

items	supp	
(a,b,d)	2	x
(a,b,e)	2	x
~~(a,c,d)~~	~~2~~	~~x~~
(a,d,e)	4	✓
~~(b,c,d)~~	~~2~~	~~x~~
~~(b,c,e)~~	~~2~~	~~x~~
(b,d,e)	3	✓
~~(c,d,e)~~	~~2~~	~~x~~

items	supp
(a,b,d,e)	2



b) 16/32 = 50%

c) 11/32 = 34.4%

d) 5/32 = 15.6%