Tutorial Solutions Week 5

1. Consider the following supermarket transactions

TID	Items
T1	Milk, Bread, Coffee
T2	Bread, Tea
T3	Bread, Eggs
T4	Milk ,Bread, Tea
T5	Milk, Eggs
T6	Bread, Eggs
T7	Milk, Eggs
T8	Milk, Bread, Eggs, Coffee
T9	Milk, Bread, Eggs

(a)Find the frequent item sets for which the minimum support is 2.

Let: Milk=M, Bread=B, Coffee=C, Tea=T, Egg=E

1-item sets:

Item	{M}	{B}	{C}	{T}	{E}
Support (frequency)	6	7	2	2	6

All have appeared in more than 2 transactions.

2-item sets:

Items	{M,B}	{M,C}	{M,T}	{M,E}	{B,C}	{B,T}	{B,E}	{C,T}	{C,E}	{T,E}
Support (frequency)	4	2	1	3	2	2	4	0	1	0

According to Apriori algorithm, any 2-item set with a lower-than-minimum-support will not be considered, as adding one item to it could only lead to lower or equal support. So we start from the candidate set {MB, MC, ME, BC, BT, BE} (ie. those without the diagonal shading). Any 3-item set that contains a 2-item set not included in our candidate set will not have a support higher than 2. This can be generalized to all higher-level item sets. We consider only the 3-item sets that are crossovers from the candidate 2-item set.

3-item sets:

Items	{M,B,C}	{M,B,E}	{M,B,T}	{M,C,E}
Support (frequency)	2	2	1	1

We do not consider the 2-item sets that include {M,T}, {C,T}, {C,E}, or {T,E} as their support is below 2. That means that the 3-item sets that include any of these item pairs will also not have enough support and will not be considered.

4-item sets:

Similar to the previous steps, we start from the candidate 3-item sets that have a support of 2 or greater (ie. {M,B,C}, and {M,B,E}).

Items	{M,B,C,E}

Support (frequency)	1

We stop at 4-item set, as:

- None of the 4-item sets qualify, so it's impossible to have any qualifying higher-level item set (they always contain 4-item sets), and;
- The longest transaction in our database only has 4 items.

The full set of frequent item sets is:
{{M}, {B}, {C}, {T}, {E}, {M,B}, {M,C}, {M,E}, {B,C}, {B,T}, {B,E}, {M,B,C}, {M,B,E}}

b) Generate association rules from the 3-item set {Milk, Bread, Eggs}, giving the support and confidence of each rule.

Support = (count of the item set) / (number of instances in the dataset)
Confidence = (count of the item set) / (count of the antecedent ie. the left-hand side of the rule)

Rule	Support	Confidence
M->BE	2/9 (0.2)	2/6 (0.33)
B->ME	2/9 (0.2)	2/7 (0.28)
E->MB	2/9 (0.2)	2/6 (0.33)
MB->E	2/9 (0.2)	2/4 (0.5)
ME->B	2/9 (0.2)	2/4 (0.5)
BE->M	2/9 (0.2)	2/4 (0.5)

c) If the minimum confidence threshold is 70% which rules would be output?

None. All six rules have confidence lower than 70% (ie. 0.7).

---

2. Given the transaction database below:

TID	Items bought
100	a , b, c, d, e, f
200	a, c, e, f, g
300	a, d, e
400	b, c, d, f, h
500	a, c, e, f, h

a) Suppose that the support threshold is 40% (or count 2), list the frequent 2-itemsets and their support.

Frequent 1-item sets:

Item set	{a}	{b}	{c}	{d}	{e}	{f}	{g}	{h}
Support	4	2	4	2	4	4	1	2

Frequent 2-item sets:

Item set	{a,b}	{a,c}	{a,d}	{a,e}	{a,f}	{a,h}	{b,c}	{b,d}	{b,e}	{b,f}	{b,h}
Support	1	3	2	4	3	1	2	2	1	2	1
Item set	{c,d}	{c,e}	{c,f}	{c,h}	{d,e}	{d,f}	{d,h}	{e,f}	{e,h}	{f,h}	
Support	2	3	4	2	2	2	1	3	1	2	

b) Compute the support and confidence of the following association rules:

R1: a, b, c => d, e

R2: a, d => f

Rule	**Support**	**Confidence**
a,b,c=>d,e	1/5 (0.4)	1/1 (1.0)
a,d=>f	1/5 (0.4)	½ (0.5)

c) Given the support threshold of 40%, and the confidence threshold of 50%, use the Apriori algorithm to generate association rules.

First find all the frequent item sets with a support of 40% or higher (ie. a count of at least 2):

1-item sets and count (support):

{a}, {b}, {c}, {d}, {e}, {f}, {h}

4 2 4 2 4 4 2

2-item sets and count (support):

{a,c}, {a,d}, {a,e}, {a,f}, {b,c}, {b,d}, {b,f}, {c,d}, {c,e}, {c,f}, {c,h}, {d,e}, {d,f}, {e,f}, {f,h}

3 2 4 3 2 2 2 2 3 4 2 2 2 3 2

3-item sets and count (support):

{a,c,e}, {a,c,f}, {a,d,e}, {a,e,f}, {b,c,d}, {b,c,f}, {b,d,f}, {c,d,f}, {c,e,f}, {c,f,h}

3 3 2 3 2 2 2 2 3 2

4-item sets and count (support):

{a,c,e,f}, {b,c,d,f}

3 2

There are not 5-item sets with a support of 40% or greater.

Combining all frequent item sets above gives the full set of frequent item sets.

Rules and confidence:

a=>c (3/4=0.75) c=>a (3/4=0.75)

a=>d (2/4=0.5) d=>a (2/2=1.0)

a=>e (4/4=1.0) e=>a (4/4=1.0)

a=>f (3/4=0.75) f=>a (3/4=0.75)

b=>c (2/2=1.0) c=>b (2/4=0.5)

b=>d (2/2=1.0) d=>b (2/2=1.0)

b=>f (2/2=1.0)	f=>b (2/4=0.5)		
c=>d (2/4=0.5)	d=>c (2/2=1.0)		
c=>e (3/4=0.75)	e=>c (3/4=0.75)		
c=>f (4/4=1.0)	f=>c (4/4=1.0)		
c=>h (2/4=0.5)	h=>c (2/2=1.0)		
d=>e (2/2=1.0)	e=>d (2/4=0.5)		
d=>f (2/2=1.0)	f=>d (2/4=0.5)		
e=>f (3/4=0.75)	f=>e (3/4=0.75)		
f=>h (2/4=0.5)	h=>f (2/2=1.0)		
a=>c,e (3/4=0.75)	a,c=>e (3/3=1.0)	a,e=>c (3/4=0.75)	c=>a,e (3/4=0.75)
c,e=>a (3/3=1.0)	e=>a,c (3/4=0.75)		
a=>c,f (3/4=0.75)	a,c=>f (3/3=1.0)	a,f=>c (3/3=1.0)	c=>a,f (3/4=0.75)
c,f=>a (3/4=0.75)	f=>a,c (3/4=0.75)		
a=>d,e (2/4=0.5)	a,d=>e (2/2=1.0)	ae=>d (2/4=0.5)	d=>a,e (2/2=1.0)
d,e=>a (2/2=1.0)	e=>a,d (2/4=0.5)		
a=>e,f (3/4=0.75)	a,e=>f (3/4=0.75)	a,f=>e (3/3=1.0)	e=>a,f (3/4=0.75)
e,f=>a (3/3=1.0)	f=>a,e (3/4=0.75)		
b=>c,d (2/2=1.0)	b,c=>d (2/2=1.0)	b,d=>c (2/2=1.0)	c=>b,d (2/4=0.5)
c,d=>b (2/2=1.0)	d=>b,c (2/2=1.0)		
b=>c,f (2/2=1.0)	b,c=>f (2/2=1.0)	b,f=>c (2/2=1.0)	c=>b,f (2/4=0.5)
c,f=>b (2/4=0.5)	f=>b,c (2/4=0.5)		
b=>d,f (2/2=1.0)	b,d=>f (2/2=1.0)	b,f=>d (2/2=1.0)	d=>b,f (2/2=1.0)
d,f=>b (2/2=1.0)	f=>b,d (2/4=0.5)		
c=>d,f (2/4=0.5)	c,d=>f (2/2=1.0)	c,f=>d (2/4=0.5)	d=>c,f (2/2=1.0)
d,f=>c (2/2=1.0)	f=>c,d (2/4=0.5)		
c=>e,f (3/4=0.75)	c,e=>f (3/3=1.0)	c,f=>e (3/4=0.75)	e=>c,f (3/4=0.75)
e,f=>c (3/4=0.75)	f=>c,e (3/4=0.75)		
c=>f,h (2/4=0.5)	cf=>h (2/4=0.5)	ch=>f (2/2=1.0)	f=>c,h (2/4=0.5)
f,h=>c (2/2=1.0)	h=>c,f (2/2=1.0)		
a=>c,e,f (3/4=0.75)	a,c=>e,f (3/3=1.0)	a,c,e=>f (3/3=1.0)	a,e=>c,f (3/4=0.75)
a,e,f=>c(3/3=1.0)	a,f=>c,e (3/3=1.0)	a,c,f=>e (3/3=1.0)	c=>a,e,f (3/4=0.75)
c,e=>a,f (3/3=1.0)	c,e,f=>a (3/3=1.0)	e=>a,c,f (3/4=0.75)	ef=>a,c (3/4=0.75)
f=>a,c,e (3/4=0.75)	c,f=>a,e (3/4=0.75)		
b=>c,d,f (2/2=1.0)	b,c=>d,f (2/2=1.0)	b,c,d=>f (2/2=1.0)	b,d=>c,f (2/2=1.0)
b,d,f=>c (2/2=1.0)	b,f=>c,d (2/2=1.0)		
b,c,f=>d (2/2=1.0)	c=>b,d,f (2/2=1.0)	c,d=>b,f (2/2=1.0)	c,d,f=>b (2/2=1.0)
d=>b,c,f (2/2=1.0)	d,f=>b,c (2/2=1.0)		
f=>b,c,d (2/4=0.5)	c,f=>b,d (2/4=0.5)		

There are no rules with a confidence level of less than 50%. As the smallest item set count is 2 (support of 40%), and the largest is 4 then it is not possible to generate a rule with a confidence lower than 50% for this dataset.

3. Consider the following data:

	Age	Prescription	Astigmatic	Tears	Lenses
1	child	myope	no	reduced	NO
2	child	myope	yes	normal	HARD
3	child	hypermetrope	no	normal	SOFT
4	child	hypermetrope	yes	reduced	NO
5	child	myope	no	normal	SOFT
6	adult	myope	yes	reduced	NO
7	adult	hypermetrope	yes	reduced	NO
8	adult	hypermetrope	yes	normal	NO
9	elderly	myope	no	normal	NO
10	elderly	myope	yes	normal	HARD
11	elderly	hypermetrope	no	reduced	NO
12	elderly	hypermetrope	no	normal	SOFT

(a) Find 2-item sets that have a minimum support of 2.

While there are several different ways to approach this task, the easiest it to enumerate the item sets directly from the data:

1-item sets:

Item sets	**Count**	**Support**
{age=child}	5	5/12≈0.42
{age=adult}	3	3/12=0.25
{age=elderly}	4	4/12≈0.3
{prescription=myope}	6	6/12=0.5
{prescription=hypermetrope}	6	6/12=0.5
{astigmatic=no}	6	6/12=0.5
{astigmatic=yes}	6	6/12=0.5
{tears=reduced}	5	5/12≈0.42
{tears=normal}	7	7/12≈0.58
{lenses=no}	7	7/12≈0.58
{lenses=hard}	2	2/12≈0.17
{lenses=soft}	3	3/12=0.25

All 1-item sets have a count of at least 2 and can therefore all be used to make 2-item sets.

2-item sets:

Item sets	**Count**	**Support**
{age=child, prescription=myope}	3	3/12
{age=child, prescription= hypermetrope}	2	2/12
{age=child, astigmatic=no}	3	3/12
{age=child, astigmatic=yes}	2	2/12
{age=child, tear=reduced}	2	2/12

{age=child, tears=normal}	3	3/12
{age=child, lenses=no}	2	2/12
{age=child, lenses=hard}	1	1/12
{age=child, lenses=soft}	2	2/12
{age=adult, prescription=myope}	1	1/12
{age=adult, prescription= hypermetrope}	2	2/12
{age=adult, astigmatic=no}	0	0/12
{age=adult, astigmatic=yes}	3	3/12
{age=adult, tear=reduced}	2	2/12
{age=adult, tears=normal}	1	1/12
{age=adult, lenses=no}	3	3/12
{age=adult, lenses=hard}	0	0/12
{age=adult, lenses=soft}	0	0/12
{age=elderly, prescription=myope}	2	2/12
{age=elderly, prescription= hypermetrope}	2	2/12
{age=elderly, astigmatic=no}	3	3/12
{age=elderly, astigmatic=yes}	1	1/12
{age=elderly, tear=reduced}	1	1/12
{age=elderly, tears=normal}	3	3/12
{age=elderly, lenses=no}	2	2/12
{age=elderly, lenses=hard}	1	1/12
{age=elderly, lenses=soft}	1	1/12
{prescription=myope, astigmatic=no}	3	3/12
{prescription=myope, astigmatic=yes}	3	3/12
{prescription=myope, tears=reduced}	2	2/12
{prescription=myope, tears=normal}	4	4/12
{prescription=myope, lenses=no}	3	3/12
{prescription=myope, lenses=hard}	2	2/12
{prescription=myope, lenses=soft}	1	1/12
{prescription=hypermetrope, astigmatic=no}	3	3/12
{prescription=hypermetrope, astigmatic=yes}	3	3/12
{prescription=hypermetrope, tears=reduced}	3	3/12
{prescription=hypermetrope, tears=normal}	3	3/12
{prescription=hypermetrope, lenses=no}	4	4/12
{prescription=hypermetrope, lenses=hard}	0	0/12
{prescription=hypermetrope, lenses=soft}	2	2/12
{astigmatic=no, tears=reduced}	2	2/12
{astigmatic=no, tears=normal}	4	4/12
{astigmatic=no, lenses=no}	3	3/12
{astigmatic=no, lenses=hard}	0	0/12
{astigmatic=no, lenses=soft}	3	3/12
{astigmatic=yes, tears=reduced}	3	3/12
{astigmatic=yes, tears=normal}	3	3/12

{astigmatic=yes, lenses=no}	4	4/12
{astigmatic=yes, lenses=hard}	2	2/12
{astigmatic= yes, lenses=soft}	0	0/12
{tears=reduced, lenses=no}	5	5/12
{tears=reduced, lenses=hard}	0	0/12
{tears=reduced, lenses=soft}	0	0/12
{tears=normal, lenses=no}	2	2/12
{tears= normal, lenses=hard}	2	2/12
{tears= normal, lenses=soft}	3	3/12

The shaded 2-item sets do not have the required minimum support are therefore not to be included in the answer.

After such conversion is done, apply the techniques used in question 2 to solve this question.

(b) Generate association rules from your first 2-item set that have a confidence of 100%.

Apply the techniques used in question 2 to solve this question.

---

4. Consider the following rules produced by the Apriori program from the above data.

R1:Tears=reduced 5 ==> lenses=NO 5 <conf:(1)> lift:(1.71) lev:(0.17) [2] conv:(2.08)

R2:Lenses=SOFT 3 ==> Astigmatic=no 3 <conf:(1)> lift:(2) lev:(0.13) [1] conv:(1.5)

R3:Prescription=hypermetrope Astigmatic=no Tears=reduced 1 ==> lenses=NO 1 <conf:(1)> lift:(1.71) lev:(0.03) [0] conv:(0.42)

(a) Which rule is the best, and why?

Generally speaking, higher support and confidence means better rules. These 3 rules all have the same confidence, but rule 1 has the highest support so by these criteria it is the best. However, this system has limitations. For a comparison of these interesting measures go to http://michael.hahsler.net/research/association_rules/measures.html.

5. ***Consider the iris data. All the attributes are numeric. What would be result of applying the apriori algorithm directly to this kind of data. Is there a way to get association rules from this kind of data?***

Discretise the numeric data, then run the Apriori algorithm.