ESSEC BUSINESS SCHOOL

BDA – Practical Sessions

Session 5

Frequent Itemsets

Jen Alchimowicz

Agenda for today

1. Quick recap – support and confidence

2. Exercises 1 and 2 from the book (From Session 7 on moodle)

3. Quick recap – Apriori algorithm

4. Exercise 3 (From Session 7 on moodle)

5. Apriori coding example

Agenda for today

1. Quick recap – support and confidence

2. Exercises 1 and 2 from the book (From Session 7 on moodle)

3. Quick recap – Apriori algorithm

4. Exercise 3 (From Session 7 on moodle)

5. Apriori coding example

Support

If I is a set of items, the support for I is the number of baskets for which I is a subset. We say I is frequent if $support(I) \geq s$, where s is some chosen support threshold.

Example:

$S_1 = \{Bread, Coke, Milk\}$ $S_2 = \{milk, pepsi, juice\}$

$S_3 = \{bread, milk\}$ $S_4 = \{Coke, juice\}$

$S_5 = \{milk, pepsi, bread\}$ $S_6 = \{milk, coke, bread, juice\}$

$S_7 = \{coke, bread, juice\}$ $S_8 = \{bread, coke\}$

$support(\{bread\}) = 6$
$support(\{coke\}) = 5$
$support(\{juice\}) = 4$
$support(\{juice, bread\}) = 2$
$support(\{coke, milk\}) = 2$
$support(\{juice, pepsi, milk\}) = 1$
$support(\{juice, coke, pepsi\}) = 0$

For $s = 5$, the frequent item sets are:

- *{bread}, {coke}, {milk}*

For $s = 4$, the frequent item sets are:

- *{bread}, {coke}, {milk}, {juice}, {bread, milk}, {bread, coke}*

Confidence

Confidence of a rule is the fraction of baskets with all of I that also contain j.

$$conf(I \to j) = \frac{support(I \cup j)}{support(I)}$$

Example:

$S_1 = \{Bread, Coke, Milk\}$ $S_2 = \{milk, pepsi, juice\}$

$S_3 = \{bread, milk\}$ $S_4 = \{Coke, juice\}$

$S_5 = \{milk, pepsi, bread\}$ $S_6 = \{milk, coke, bread, juice\}$

$S_7 = \{coke, bread, juice\}$ $S_8 = \{bread, coke\}$

$confidence(\{bread\} \to milk) = 4/6$

$confidence(\{coke\} \to juice) = 3/5$

$confidence(\{coke\} \to pepsi) = 0/4$

$confidence(\{bread, milk\} \to coke) = 2/4$

$confidence(\{coke, bread, juice\} \to milk) = 1/2$

Agenda for today

1. Quick recap – support and confidence

2. Exercises 1 and 2 from the book (From Session 7 on moodle)

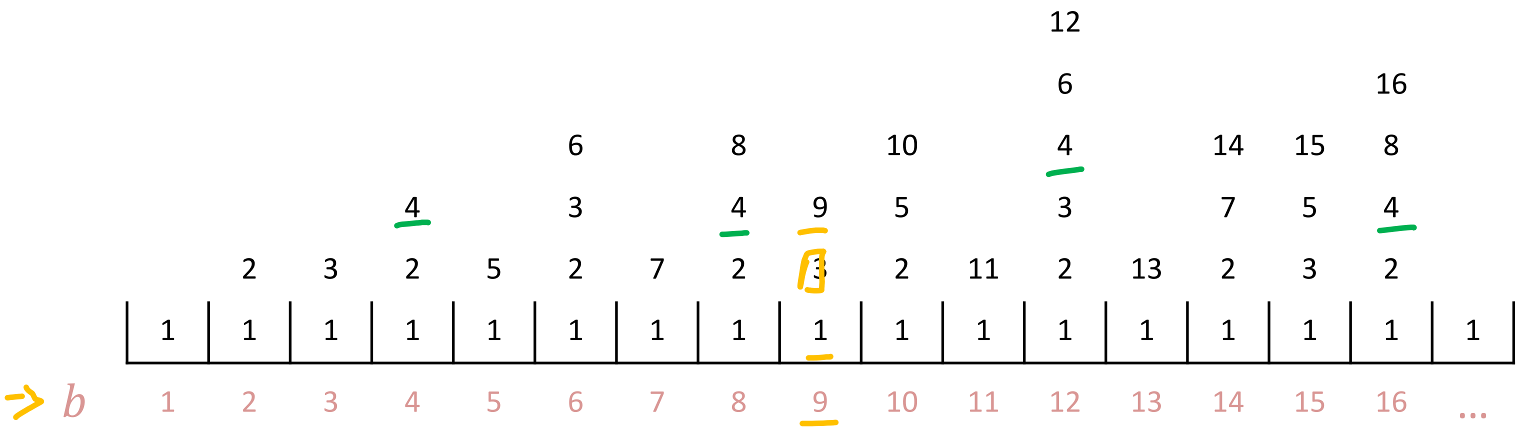
3. Quick recap – Apriori algorithm

4. Exercise 3 (From Session 7 on moodle)

5. Apriori coding example

Exercise 1

1. (**Exercise 6.1.1 MMDS book**) Suppose there are 100 items, numbered 1 to 100, and also 100 baskets, also numbered 1 to 100. Item i is in basket b if and only if i divides b with no remainder. Thus, item 1 is in all the baskets, item 2 is in all fifty of the even-numbered baskets, and so on. Basket 12 consists of items $\{1, 2, 3, 4, 6, 12\}$, since these are all the integers that divide 12. Answer the following questions:



												12					
												6				16	
						6		8		10		4		14	15	8	
				4		3		4	9	5		3		7	5	4	
		2	3	2	5	2	7	2	3	2	11	2	13	2	3	2	
	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
-> b	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	...

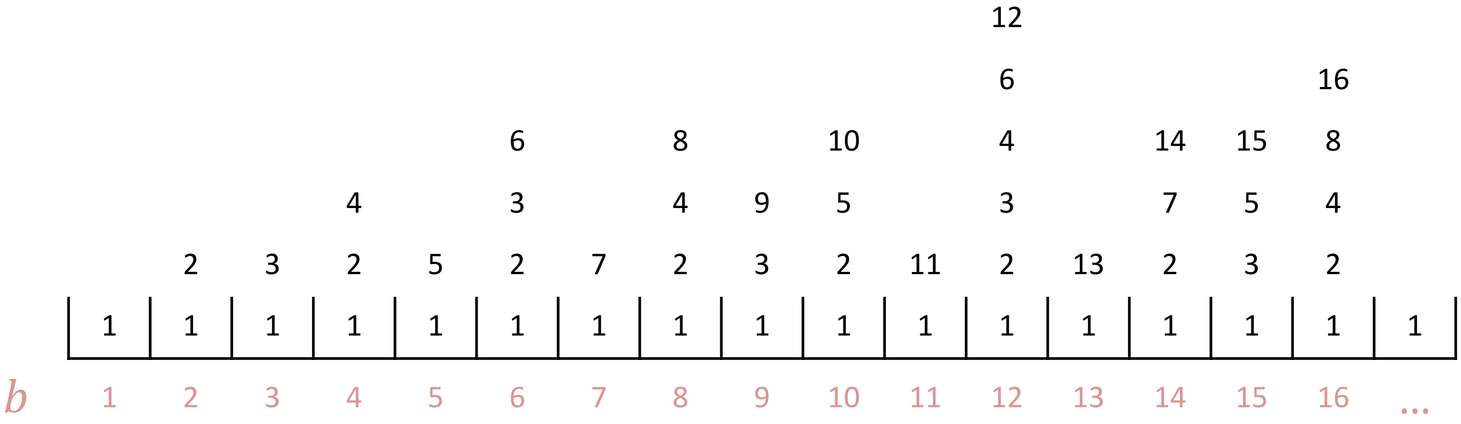
(a) If the support threshold is 5, which items are frequent?

Frequent items: $\{1,2,3,4,5, \dots, 19, 20\}$

(items that have at least 5 multiples that are <= 100)

Exercise 1

1. (**Exercise 6.1.1 MMDS book**) Suppose there are 100 items, numbered 1 to 100, and also 100 baskets, also numbered 1 to 100. Item i is in basket b if and only if i divides b with no remainder. Thus, item 1 is in all the baskets, item 2 is in all fifty of the even-numbered baskets, and so on. Basket 12 consists of items $\{1, 2, 3, 4, 6, 12\}$, since these are all the integers that divide 12. Answer the following questions:



												12					
												6				16	
						6		8		10		4		14	15	8	
				4		3		4	9	5		3		7	5	4	
		2	3	2	5	2	7	2	3	2	11	2	13	2	3	2	
	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
b	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	...

(b) what is the confidence of the following association rules?

$$confidence(\{5,7\} \rightarrow 2) = \frac{support(I \cup j)}{support(I)} = \frac{support(\{5,7\} \cup \{2\})}{support(\{5,7\})} = \frac{1}{2}$$

{5,7,2} will appear in basket 70

{5,7} will appear in baskets 35 and 70

$$confidence(\{2,3,4\} \rightarrow 5) = \frac{support(\{2,3,4\} \cup \{5\})}{support(\{2,3,4\})} = \frac{1}{8}$$

Lowest common multiple of {2,3,4} is 12.
So {2,3,4} will appear in baskets
$\{12, 24, 36, 48, \boxed{60}, 72, 84, 96\}$

Exercise 2

2. (**Exercise 6.1.3 MMDS book**) Suppose there are 100 items, numbered 1 to 100, and also 100 baskets, also numbered 1 to 100. Item i is in basket b if and only if b divides i with no remainder. For example, basket 12 consists of items $\{12, 24, 36, 48, 60, 72, 84, 96\}$

	...	...	...	...	...	...	...		...					
	5	10	15	22	25	30	35		60					
	4	8	12	16	20	24	28		48					
	3	6	9	12	15	18	21		36					
	2	4	6	8	10	12	14		24		100			
	1	2	3	4	5	6	7		12		50	51	52	
b	1	2	3	4	5	6	7	...	12	...	50	51	52	...

(a) If the support threshold is 5, which items are frequent?

Answer: every item that has at least 5 dividers (including 1 and itself)

Examples:

- 10 has 4 divisors: {1,2,5,10} and thus will be in 4 baskets: {1,2,5,10}
- 12 has 6 divisiors: {1,2,3,4,6,12} and thus will be in 6 baskets: {1,2,3,4,6,12}
- 50 has 6 divisors: {1,2,5,10,25,50}

Exercise 2

2. **(Exercise 6.1.3 MMDS book)** Suppose there are 100 items, numbered 1 to 100, and also 100 baskets, also numbered 1 to 100. Item i is in basket b if and only if b divides i with no remainder. For example, basket 12 consists of items {12, 24, 36, 48, 60, 72, 84, 96}

	...	...	...	...	...	...	...		...				
	5	10	15	22	25	30	35		60				
	4	8	12	16	20	24	28		48				
	3	6	9	12	15	18	21		36				
	2	4	6	8	10	12	14		24		100		
	1	2	3	4	5	6	7		12		50	51	52	1
b	1	2	3	4	5	6	7	...	12	...	50	51	52	...

(b) what is the confidence of the following association rules?

$$confidence(\underline{\{24,60\}} \rightarrow \underline{8}) = \frac{support(\{24,60\} \cup 8)}{support(\{24,60\})} = \frac{3}{6} = \frac{1}{2}$$

$$support(\{60\}) \colon \{\underline{1},\underline{2},\underline{3},\underline{4},5,\underline{6},\underline{\overset{10}{12}},15,20,30,60\}$$

$$support(\{24\}) \colon \{\underline{1},\underline{2},\underline{3},\underline{4},\underline{6},\underline{12},24\}$$

$$support(\{8\}) \colon \{\underline{1},\underline{2},\underline{4},8\} \leftarrow$$

$$support(\{24,60\}) \colon \{\underline{1},\underline{2},3,\underline{4},6,12\}$$

$$support(\{24,60,8\}) \colon \{1,2,4\}$$

Agenda for today

1. Quick recap – support and confidence

2. Exercises 1 and 2 from the book (From Session 7 on moodle)

3. Quick recap – Apriori algorithm

4. Exercise 3 (From Session 7 on moodle)

5. Apriori coding example

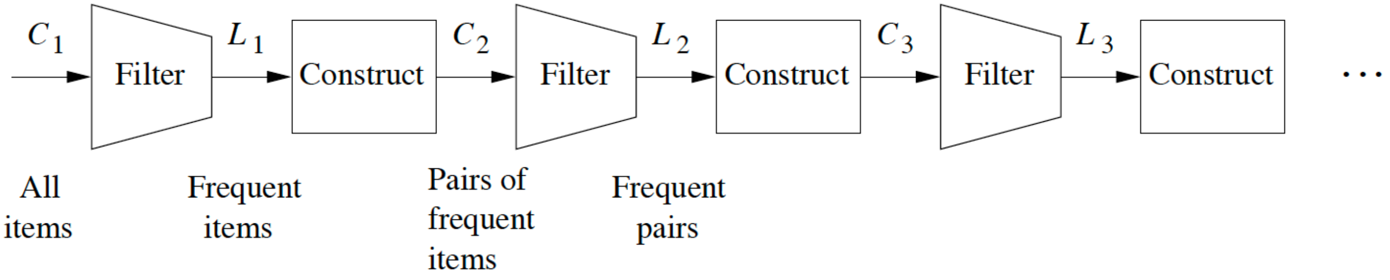
Apriori algorithm

Apriori algorithm reduces the number of counts we need to keep in main memory.
Monotonicity: ***If a set I of items is frequent, then so is every subset of I.***
For association rules to be actionable on we need to generate only few of them.

The algorithm:
Repeat until no frequent sets are found of size k:

1. Take frequent items from step k-1
2. Construct candidate sets of size k
3. Calculate the supports for candidate sets
4. Filter out sets below the support threshold

Monotonicity:
{bread, milk, coke}



Agenda for today

1. Quick recap – support and confidence

2. Exercises 1 and 2 from the book (From Session 7 on moodle)

3. Quick recap – Apriori algorithm

4. Exercise 3 (From Session 7 on moodle)

5. Apriori coding example

Exercise 3

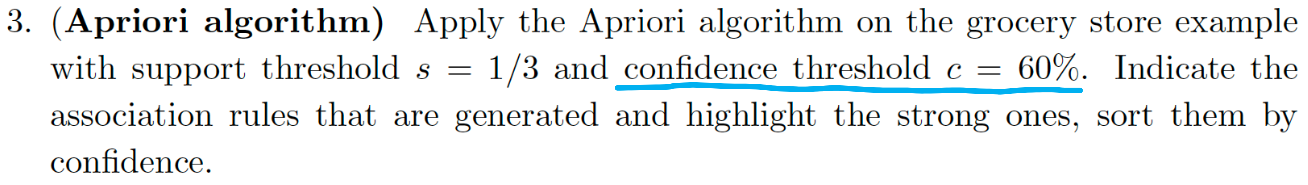
3. (**Apriori algorithm**) Apply the Apriori algorithm on the grocery store example with support threshold $s = 1/3$ and confidence threshold $c = 60\%$. Indicate the association rules that are generated and highlight the strong ones, sort them by confidence.

Transaction ID	Items
1	Milk, Bread, Juice
2	Milk, Bread
3	Milk, Coke, Chips
4	Chips, Coke
5	Chips, Juice
6	Milk, Coke, Chips

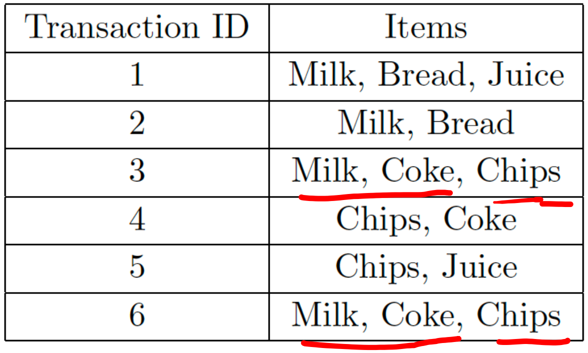
Frequent itemsets: itemsets with support >= 2

Pass (k)	Candidate k-sets and their support	Frequent k-sets
k=1	{milk} (4), {bread} (2), {juice} (2), {coke} (3), {chips} (4)	{milk}, {bread}, {juice}, {coke}, {chips}
k=2	{milk, bread} (2), {milk, juice} (1), {milk, coke} (2), {milk, chips} (2), {bread, juice} (1), {bread, coke} (0), {bread, chips} (0), {juice, chips} (1), {juice, coke} (0), {coke, chips} (3)	{milk, bread}, {milk, coke}, {milk, chips}, {coke, chips}
k=3	{milk, coke, chips} (2)	{milk, coke, chips}
k=4		

Exercise 3



3. (**Apriori algorithm**) Apply the Apriori algorithm on the grocery store example with support threshold $s = 1/3$ and confidence threshold $c = 60\%$. Indicate the association rules that are generated and highlight the strong ones, sort them by confidence.



Transaction ID	Items
1	Milk, Bread, Juice
2	Milk, Bread
3	Milk, Coke, Chips
4	Chips, Coke
5	Chips, Juice
6	Milk, Coke, Chips

Frequent k-sets: {milk, bread}, {milk, coke}, {milk, chips}, {coke, chips}, {milk, coke, chips}

Association rules:

{milk, bread}: $confidence(\{milk\} \rightarrow \{bread\}) = \frac{support(\{milk\}\cup\{bread\})}{support(\{milk\})} = \frac{2}{4} = \frac{1}{2}$

$confidence(\{bread\} \rightarrow \{milk\}) = \frac{support(\{bread\}\cup\{milk\})}{support(\{bread\})} = \frac{2}{2} = 1$

{milk, coke}: $confidence(\{milk\} \rightarrow \{coke\}) = \frac{2}{4} = \frac{1}{2}$, $confidence(\{coke\} \rightarrow \{milk\}) = \frac{2}{3}$

{milk, chips}: $confidence(\{milk\} \rightarrow \{chips\}) = \frac{2}{4} = \frac{1}{2}$, $confidence(\{chips\} \rightarrow \{milk\}) = \frac{2}{4} = \frac{1}{2}$

{coke, chips}: $confidence(\{coke\} \rightarrow \{chips\}) = \frac{3}{3} = 1$, $confidence(\{chips\} \rightarrow \{coke\}) = \frac{3}{4}$

{milk, coke, chips}:

$confidence(\{milk, coke\} \rightarrow \{chips\}) = 2/2 = 1$

$confidence(\{milk, chips\} \rightarrow \{coke\}) = 2/2 = 1$

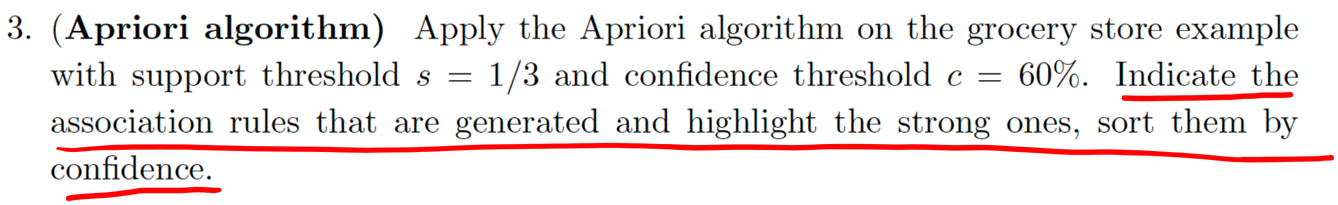
$confidence(\{coke, chips\} \rightarrow \{milk\}) = 2/3$

$confidence(\{milk\} \rightarrow \{coke, chips\}) = 2/4 = 1/2$

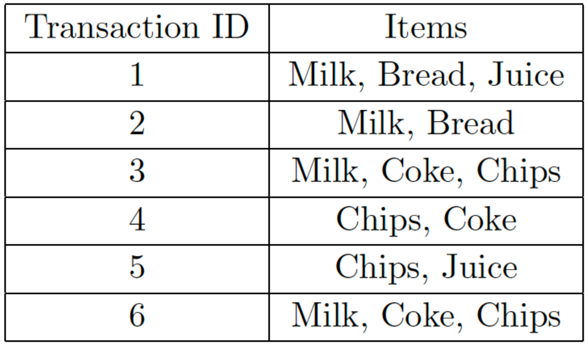
$confidence(\{chips\} \rightarrow \{coke, milk\}) = 2/4 = 1/2$

$confidence(\{coke\} \rightarrow \{milk, chips\}) = 2/3$

Exercise 3



3. (**Apriori algorithm**) Apply the Apriori algorithm on the grocery store example with support threshold $s = 1/3$ and confidence threshold $c = 60\%$. Indicate the association rules that are generated and highlight the strong ones, sort them by confidence.



Transaction ID	Items
1	Milk, Bread, Juice
2	Milk, Bread
3	Milk, Coke, Chips
4	Chips, Coke
5	Chips, Juice
6	Milk, Coke, Chips

We take association rules with confidence >= 0.6

Sorted association rules with $\boldsymbol{support \geq 0.33}$ and $\boldsymbol{confidence \geq 0.6}$:

1. {coke} -> {chips}, support=0.5, confidence=1

2. {bread} -> {milk}, support=0.33, confidence=1

3. {milk, coke} -> {chips}, support=0.33, confidence=1

4. {milk, chips} -> {coke}, support=0.33, confidence=1

5. {chips} -> {coke}, support=0.5, confidence=0.75

6. {coke} -> {milk}, support=0.33, confidence=0.66

7. {coke} -> {chips, milk}, support=0.33, confidence=0.66

8. {coke, chips} -> {milk}, support=0.33, confidence=0.66

Agenda for today

1. Quick recap – support and confidence

2. Exercises 1 and 2 from the book (From Session 7 on moodle)

3. Quick recap – Apriori algorithm

4. Exercise 3 (From Session 7 on moodle)

5. Apriori coding example