ANSWERS

Exercise 1 ANSWER

Note that an exam question of this type, while similar, would have a considerably smaller data set.

	Barley	Corn	Gram	Millet	Rice	Wheat
1	1	1	1	1	1	1
2		1		1	1	1
3	1		1		1	1
4		1	1		1	1
5	1		1	1	1	
6			1		1	1
7	1			1	1	1
8				1	1	1
9	1	1	1	1		
10	1	1	1		1	1
11		1		1	1	1
12	1	1		1	1	1

Checking support

The general formula for support for the rule is the estimated probability of antecedent and consequent appearing together:

$$P(antecedent\ AND\ consequent)$$

The support for any rule involving items A, B and C, regardless of which items are in the antecedent or consequent, is calculated as:

$$suppport(A, B, C) = P(A \wedge B \wedge C) = \frac{f(A \wedge B \wedge C)}{n}$$

where n is the number of instances in the data set and $f(A)$ is the frequency of A in the data set (the number of times it appears in the data set). The symbol $\wedge$ is the logical 'and' operator.

With the Apriori algorithm we start by identifying the **individual items** that feature at or above the required support level. In this case the required support level is 33% and we are looking for attributes that have the value 1 4/12 times or more frequently in the data set. This includes all the types of cereal in the table.

Next we look for groups of **two items** that feature in the set 4/12 or more times. The list includes most (14) of the 15 (see note below) pairs. The excluded pair (Gram, Millet) have value 1 in the same instance only 3 times. Any group of three that contains this pair is discarded automatically, according to the apriori algorithm.

There are 20 possible **groups of three**, of which 4 are excluded immediately because they include the pair (Gram, Millet). Let's write them out and check them one by one to see if they have a support greater than 33%:

$support$(BCG) $= 3/12$	**X** $(3/12 < 33\%)$
$support$(BC M) $= 3/12$	**X** $(3/12 < 33\%)$
$support$(BC R) $= 3/12$	**X** $(3/12 < 33\%)$
$support$(BC W) $= 3/12$	**X** $(3/12 < 33\%)$
$support$(B GM)	**X (Gram and Millet)**
$support$(B G R) $= 4/12$	✓ $(4/12 > 33\%)$
$support$(B G W) $= 3/12$	**X** $(3/12 < 33\%)$
$support$(B MR) $= 4/12$	✓ $(4/12 > 33\%)$
$support$(B M W) $= 3/12$	**X** $(3/12 < 33\%)$
$support$(B RW) $= 5/12$	✓ $(5/12 > 33\%)$
$support$(CGM)	**X (Gram and Millet)**
$support$(CG R) $= 3/12$	**X** $(3/12 < 33\%)$
$support$(CG W) $= 3/12$	**X** $(3/12 < 33\%)$
$support$(C MR) $= 4/12$	✓ $(4/12 > 33\%)$
$support$(C M W) $= 4/12$	✓ $(4/12 > 33\%)$
$support$(C RW) $= 6/12$	✓ $(6/12 > 33\%)$
$support$(GMR)	**X (Gram and Millet)**
$support$(GM W)	**X (Gram and Millet)**
$support$(G RW) $= 5/12$	✓ $(5/12 > 33\%)$
$support$(MRW) $= 6/12$	✓ $(6/12 > 33\%)$

Eight of the groups of three products fulfil the support requirement (ticked in the list).

Checking confidence

Now for each of the 8 groups of three that meet the support requirement check if the confidence requirement can be met. We can use any of the three items in the group as the consequent (as we are at a stage where we are *searching* for rules.

The general formula for confidence is the ratio between estimated probability of antecedent and consequent occurring together and the estimated probability of the antecedent occurring:

$$\frac{P(antecedent\ AND\ consequent)}{P(antecedent)}$$

The confidence for group A, B and C with antecedent (A *and* B) and consequent $C$is calculated as:

$$confidence((A \wedge B) \rightarrow C) = \frac{P(A \wedge B \wedge C)}{P(A \wedge B)} = \frac{\frac{f(A \wedge B \wedge C)}{n}}{\frac{f(A \wedge B)}{n}} = \frac{f(A \wedge B \wedge C)}{f(A \wedge B)}$$

where n is the number of instances in the data set and $f(A)$ is the frequency of A in the data set (the number of times it appears in the data set).

The confidence values are:

$confidence((G,R) \rightarrow B) = 4/6$	✓ $(4/6 > 66\%)$
$confidence((R,B) \rightarrow G) = 4/6$	✓ $(4/6 > 66\%)$
$confidence((B,G) \rightarrow R) = 4/5$	✓ $(4/5 > 66\%)$
$confidence((M,R) \rightarrow B) = 4/7$	**X** $(4/7 < 66\%)$
$confidence((R,B) \rightarrow M) = 4/6$	✓ $(4/6 > 66\%)$
$confidence((B,M) \rightarrow R) = 4/5$	✓ $(4/5 > 66\%)$
$confidence((R,W) \rightarrow B) = 5/10$	**X** $(5/10 < 66\%)$
$confidence((W,B) \rightarrow R) = 5/5$	✓ $(5/5 > 66\%)$
$confidence((B,R) \rightarrow W) = 5/6$	✓ $(5/6 > 66\%)$
$confidence((M,R) \rightarrow C) = 4/7$	**X** $(4/7 < 66\%)$
$confidence((R,C) \rightarrow M) = 4/6$	✓ $(4/6 > 66\%)$
$confidence((C,M) \rightarrow R) = 4/5$	✓ $(4/5 > 66\%)$
$confidence((M,W) \rightarrow C) = 4/6$	✓ $(4/6 > 66\%)$
$confidence((W,C) \rightarrow M) = 4/6$	✓ $(4/6 > 66\%)$
$confidence((C,M) \rightarrow W) = 4/5$	✓ $(4/5 > 66\%)$
$confidence((R,W) \rightarrow C) = 6/10$	**X** $(6/10 < 66\%)$
$confidence((W,C) \rightarrow R) = 6/6$	✓ $(6/6 > 66\%)$
$confidence((C,R) \rightarrow W) = 6/6$	✓ $(6/6 > 66\%)$
$confidence((R,W) \rightarrow G) = 5/10$	**X** $(5/10 < 66\%)$
$confidence((W,G) \rightarrow R) = 5/5$	✓ $(5/5 > 66\%)$
$confidence((G,R) \rightarrow W) = 5/6$	✓ $(5/6 > 66\%)$
$confidence((R,W) \rightarrow M) = 6/10$	**X** $(6/10 < 66\%)$
$confidence((W,M) \rightarrow R) = 6/6$	✓ $(6/6 > 66\%)$
$confidence((M,R) \rightarrow W) = 6/7$	✓ $(6/7 > 66\%)$

Calculating lift

The general formula for lift is the ratio between estimated probability of antecedent and consequent occurring together and the product of estimated probabilities of the antecedent occurring and of the consequent occurring:

$$\frac{P(antecedent\ AND\ consequent)}{P(antecedent)P(consequent)}$$

The lift for group A, B and C with antecedent ($A\ and\ B$) and consequent $C$is calculated as:

$$lift((A,B) \rightarrow C) = \frac{P(A \wedge B \wedge C)}{P(A \wedge B)P(C)} = \frac{\frac{f(A \wedge B \wedge C)}{n}}{\frac{f(A \wedge B)}{n}\frac{f(C)}{n}} = \frac{f(A \wedge B \wedge C)n}{f(A \wedge B)f(C)}$$

where n is the number of instances in the data set and $f(A)$ is the frequency of A in the data set (the number of times it appears in the data set).

$lift((G,R) \to B) = (4*12)/(6*7) = 1.14$
$lift((G,R) \to B) = (4*12)/(6*7) = 1.14$
$lift((G,R) \to B) = (4*12)/(5*11) = 0.87$
$lift((M,R) \to B) = (4*12)/(6*8) = 1$
$lift((M,R) \to B) = (4*12)/(5*11) = 0.87$
$lift((R,W) \to B) = (5*12)/(5*11) = 1.09$
$lift((R,W) \to B) = (5*12)/(6*10) = 1$
$lift((M,R) \to C) = (4*12)/(6*8) = 1$
$lift((M,R) \to C) = (4*12)/(5*11) = 0.87$
$lift((M,W) \to C) = (4*12)/(6*7) = 1.14$
$lift((M,W) \to C) = (4*12)/(6*8) = 1$
$lift((M,W) \to C) = (4*12)/(5*10) = 0.96$
$lift((R,W) \to C) = (6*12)/(6*11) = 1.09$
$lift((R,W) \to C) = (6*12)/(6*10) = 1.2$
$lift((R,W) \to G) = (5*12)/(5*11) = 1.09$
$lift((R,W) \to G) = (5*12)/(6*10) = 1$
$lift((R,W) \to M) = (6*12)/(6*11) = 1.09$
$lift((R,W) \to M) = (6*12)/(7*10) = 1.03$

All the values of lift are close to 1 and indicate that there is no significant association in any of these rules, as the probability that the antecedent and consequent will appear together is close to what it would be if the appearance of one were independent of the appearance of the other.

NOTE: The number of groups of 2 out of 6 elements is $\frac{6 \times 5}{1 \times 2}$, the number of groups of 3 out of 6 elements is $\frac{6 \times 5 \times 4}{1 \times 2 \times 3}$ etc.

Exercise 2 ANSWER

Let's introduce some short symbols for the values: **Exhaustion** to E, **Stuffy nose** to S and **Diagnosis=cold** to C.

Now, for the support, we find the relative frequency of all three items in the rule occurring together:

$$support(E \wedge S \wedge C) = \frac{f(E \wedge S \wedge C)}{n} = \frac{7}{17}$$

The confidence is the ratio between the support for the rule and the relative frequency of the antecedent

$$confidence((E \wedge S) \to C) = \frac{\frac{f(E \wedge S \wedge C)}{n}}{\frac{f(E \wedge S)}{n}} = \frac{f(E \wedge S \wedge C)}{f(E \wedge S)} = \frac{7}{7} = 1$$

The lift is the ration between the support for the rule and the product of the relative frequencies of antecedent and consequent:

$$lift((E \wedge S) \to C) = \frac{\frac{f(E \wedge S \wedge C)}{n}}{\frac{f(E \wedge S)}{n}\frac{f(C)}{n}} = \frac{f(E \wedge S \wedge C)n}{f(E \wedge S)f(C)} = \frac{17 \times 7}{7 \times 9} = 1.89$$

Exercise 3 ANSWER

a) Finding 'interesting' rules

With the Apriori algorithm we use **support** levels of progressively larger groups of items to eliminate any combinations that cannot be part of a rule because of low support. This reduces the number of data instances that need to be dealt with.

We start by looking at the incidence of the different grocery items. As the required support level is 33%(4/12), any items that occur less than 4 times will not be examined (if an item does not appear 4 times then no combination of items including tht item can appear 4 times or more).

Grocery	**Occurrences**	**To be examined further?**
Milk	9	yes
Egg	3	no
Bread	10	yes
Butter	10	yes
Ketchup	3	no
Cookies	5	yes

Now we list the two-item combinations including the items that are 'still in the game'. Again, any combinations that occur les than 4 times are not examined further.

Groceries	**Occurrences**	**To be examined further?**
Milk, Bread	7	yes
Milk, Butter	7	yes
Milk, Cookies	3	no
Bread, Butter	9	yes
Bread, Cookies	4	yes
Butter, Cookies	3	no

Based on the results above there is only one three-item combination that can be considered (Cookies can't be combined either with Butter or Milk, which means we can't make up a three-item combination including Cookies). It meets the support level, so we have a candidate combination of items.

Groceries	**Occurrences**	**To be examined further?**
Milk, Bread, Butter	6	yes

****

The next step is to investigate the candidate item combinations (of which we have only one in this example) to see which antecedent/consequent arrangements fulfill the minimal **confidence** level requirement. The different arrangements are listed in the table, together with their confidence values.

Antecedent	**Consequent**	**Confidence**
Milk, Bread	Butter	$\frac{P(Milk, Break, Butter)}{P(Milk, Bread)} = \frac{6}{7} = 0.86$
Milk, Butter	Bread	$\frac{P(Milk, Break, Butter)}{P(Milk, Butter)} = \frac{6}{7} = 0.86$
Butter, Bread	Milk	$\frac{P(Milk, Break, Butter)}{P(Butter, Bread)} = \frac{6}{9} = 0.67$

The confidence level for all the rules is higher than the required 50%, hence the three rules listed above are all 'interesting' based on the given criteria. The rules are:

1. **IF** Milk and Bread **THEN** Butter
2. **IF** Milk and Butter **THEN** Bread
3. **IF** Butter and Bread **THEN** Milk

b) Calculating the lift

If A is the antecedent and C the consequent, the lift is calculated as $\frac{P(A,C)}{P(A)P(C)}$.

Lift for rule 1: $\frac{P(Milk, Break, Butter)}{P(Milk, Bread)P(Butter)} = \frac{\frac{6}{12}\frac{7}{12}}{\frac{10}{12}} = 1.03$

Lift for rule 2: $\frac{P(Milk, Break, Butter)}{P(Milk, Butter)P(Bread)} = \frac{\frac{6}{12}\frac{7}{12}}{\frac{10}{12}} = 1.03$

Lift for rule 3: $\frac{P(Milk, Break, Butter)}{P(Butter, Bread)P(Milk)} = \frac{\frac{6}{12}\frac{9}{12}}{\frac{9}{12}} = 0.89$

The lift values are below or close to 1, indicating that there is no positive association between the antecedents and consequents in the rules and that the high confidence values stem simply from high occurrence in the data sets.