Tarea 03

Kevin Hernández

Ejercicio 6.1.1: 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:

a. If the support threshold is 5, which items are frequent?
b. If the support threshold is 5, which pairs of items are frequent?
c. What is the sum of the sizes of all the baskets?

Solución 6.1.1

a. El artículo i está en la canasta b si i es un factor de b. En otras palabras, i está en la canasta b si y solo si existe un entero constante $k >= 1$ tal que $b = k * i$. Como resultado, el artículo i se encuentra en 5 o más cestas si $100/i >= 5$. Por lo tanto, los elementos {1}, {2}, ..., {20} representan los singleton frecuentes.

b. Se puede obtener el conjunto de pares frecuentes, contando explícitamente el conjunto de cada par y devolviendo aquellos cuyos recuentos son superiores a 5.

(1, 2) (1, 3) (1, 4) (1, 5) (1, 6) (1, 7)
(1, 8) (1, 9) (1, 10) (1, 11) (1, 12) (1, 13) (1, 14)
(1, 15) (1, 16) (1, 17) (1, 18) (1, 19) (1, 20)
(2, 3) (2, 4) (2, 5) (2, 6) (2, 7) (2, 8) (2, 9)
(2, 10) (2, 12) (2, 14) (2, 16) (2, 18) (2, 20)
(3, 4) (3, 5) (3, 6) (3, 9) (3, 12) (3, 15) (3, 18)
(4, 5) (4, 6) (4, 8) (4, 10) (4, 12) (4, 16) (4, 20)
(5, 10) (5, 15) (5, 20)
(6, 9) (6, 12) (6, 18)
(7, 14)
(8, 16)

(9, 18)

(10, 20)

c. Podemos definir num_factors(b) como el número de factores que tiene b. Luego suma de los tamaños todos los canastos = suma (num_factors(b), $b = 1, 2, , 20$). Los factores de un número, por ejemplo 12, sería $2^2 \cdot 3$ y serían 6 factores en total. Entonces la suma de los tamaños de todas los canastos sería: **482**

Exercise 6.1.3: 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}

Solución 6.1.3

a. La canasta b consta de elementos que son múltiplos de b. Alternativamente, el elemento i está en la canasta b si b es un factor de i. Así, el ítem i es frecuente si tiene al menos 5 factores <= 100. Estos serían:

12 16 18 20 24 28 30 32 36 40 42 44 45 48 50 52 54 56 60 63 64 66 68 70 72 75 76 78 80 81 84 88 90 92 96 98 99 100

b. Claramente, (i, j) representan un par frecuente si i y j comparten al menos 5 factores comunes. Por lo que el número de factores comunes sería: **6**.

c. El tamaño de la canasta b es parte_entera($100/b$). Es decir: **482**.

Exercise 6.1.5:For the data of Exercise 6.1.1, what is the confidence of the following association rules?

Solución 6.1.5

(a) $\{5, 7\} \rightarrow 2$.

Las cestas que contienen tanto el artículo 5 como el artículo 7 son la cesta 35 y la cesta 70, en las que solo la cesta 70 también contiene el artículo 2. Por lo tanto, la confianza de esta regla de asociación es **1/2**.

(b) $\{2, 3, 4\} \rightarrow 5$.

Las cestas cuyos números son múltiplos de 12 contienen el conjunto de artículos {2,3,4} como un subconjunto, hay 8 cestas de este tipo, mientras que solo aquellas cuyos números son múltiplos de 60 contienen el conjunto de artículos {2,3,4,5 } como subconjunto, hay 1 cesta de este tipo. Por tanto, la confianza de la regla de asociación $\{2, 3, 4\} \rightarrow 5$ es **1/8**.

Exercise 6.2.1: If we use a triangular matrix to count pairs, and n, the number of items, is 20, what pair's count is in a[100]?

Solución 6.2.1

Para cualquier par {i, j} en la matriz triangular, el índice correspondiente k es:

$$(i-1)\ (20-\tfrac{i}{2})+j-i$$

Resolviendo la ecuación:

$$100=(i-1)\ (20-\tfrac{i}{2})+j-i,\ \text{con}\ 1\leqslant i<j\leqslant 20$$

Podemos obtener el par {7,8}.

Exercise 6.2.6: Apply the A-Priori Algorithm with support threshold 5 to the data of:

(a) Exercise 6.1.1.

(b) Exercise 6.1.3.

Solución 6.2.6

(a)

Conjuntos de elementos candidatos de tamaño 1 (C_1):

Elementos con índice $1\ldots 100$

Conjuntos de elementos verdaderamente frecuentes de tamaño 1 (L_1):

Elementos con índice $1\ldots 20$

Conjuntos de elementos candidatos de tamaño 2 (C2):

$$\forall 1\leqslant i<j\leqslant 20\{i,j\}$$

Conjuntos de elementos verdaderamente frecuentes de tamaño 2 (L_2):

Lo mismo que Solución 6.1.1 (b)

Conjuntos de elementos candidatos de tamaño 3 (C_3):

$$\forall c,|c|=3\ \text{y}\ \exists c_i,c_j\in L_2, c=c_i\cup c_j$$

Conjuntos de elementos verdaderamente frecuentes de tamaño 3 (L_3):

{1, 2, 3}, {1, 2, 4}, {1, 2, 5}, {1, 2, 6}, {1, 2, 7}, {1, 2, 8}, {1, 2, 9}, {1, 2, 10}, {1, 2, 12}, {1, 2, 14}, {1, 2, 16}, {1, 2, 18}, {1, 2, 20},

{1, 3, 4}, {1, 3, 5}, {1, 3, 6}, {1, 3, 9}, {1, 3, 12}, {1, 3, 15}, {1, 3, 18},

{1, 4, 5}, {1, 4, 6}, {1, 4, 8}, {1, 4, 10}, {1, 4, 12}, {1, 4, 16}, {1, 4, 20},

{1, 5, 10}, {1, 5, 15}, {1, 5, 20},

{1, 6, 9}, {1, 6, 12}, {1, 6, 18},

{1, 7, 14}, {1, 8, 16}, {1, 9, 18}, {1, 10, 20},

{2, 3, 4}, {2, 3, 6}, {2, 3, 9}, {2, 3, 12}, {2, 3, 18},

{2, 4, 5}, {2, 4, 6}, {2, 4, 8}, {2, 4, 10}, {2, 4, 12}, {2, 4, 16}, {2, 4, 20},

{2, 5, 10}, {2, 5, 20},

{2, 6, 9}, {2, 6, 12}, {2, 6, 18},

{2, 7, 14}, {2, 8, 16}, {2, 9, 18}, {2, 10, 20},

{3, 4, 12}, {3, 5, 15},

{4, 5, 10}, {4, 5, 20}, {4, 6, 12},

{5, 10, 20},

{6, 9, 18}

Conjuntos de elementos candidatos de tamaño 4 (C_4):

$$\forall c, |c| = 4 \text{ y } \exists c_i, c_j \in L_3, c = c_i \cup c_j$$

Conjuntos de elementos verdaderamente frecuentes de tamaño 4 (L_4):

{1, 2, 3, 4}, {1, 2, 3, 6}, {1, 2, 3, 9}, {1, 2, 3, 12}, {1, 2, 3, 18},

{1, 2, 4, 5}, {1, 2, 4, 6}, {1, 2, 4, 8}, {1, 2, 4, 10}, {1, 2, 4, 12}, {1, 2, 4, 16}, {1, 2, 4, 20},

{1, 2, 5, 10}, {1, 2, 5, 20}, {1, 2, 6, 9}, {1, 2, 6, 12}, {1, 2, 6, 18},

{1, 2, 7, 14}, {1, 2, 8, 16}, {1, 2, 9, 18}, {1, 2, 10, 20},

{1, 3, 4, 12}, {1, 3, 5, 15},

{1, 4, 5, 10}, {1, 4, 5, 20}, {1, 4, 6, 12},

{1, 5, 10, 20}, {1, 6, 9, 18}

{2, 3, 4, 12}, {2, 4, 5, 10}, {2, 4, 5, 20}, {2, 4, 6, 12},

{2, 5, 10, 20}, {2, 6, 9, 18}

{4, 5, 10, 20}

Conjuntos de elementos candidatos de tamaño 4 (C_4):

$$\forall c, |c| = 5 \text{ y } \exists c_i, c_j \in L_4, c = c_i \cup c_j$$

Conjuntos de elementos verdaderamente frecuentes de tamaño 5 (L_5):

{1, 2, 3, 4, 12}, {1, 2, 4, 5, 10}, {1, 2, 4, 5, 20}, {1, 2, 4, 6, 12}, {1, 2, 5, 10, 20}, {1, 2, 6, 9, 18},

{1, 4, 5, 10, 20}, {2, 4, 5, 10, 20}

Conjuntos de elementos candidatos de tamaño 6 (C_6):

$$\forall c, |c| = 6 \text{ y } \exists c_i, c_j \in L_5, c = c_i \cup c_j$$

Conjuntos de elementos verdaderamente frecuentes de tamaño 6 (L_6):

{1, 2, 4, 5, 10, 20}

Dado que solo hay un conjunto de elementos, no habrá conjuntos de elementos candidatos de tamaño 7, **el algoritmo se detiene aquí**.

(b)

Mismo procedimiento que el paso anterior.

Exercise 6.3.1:Here is a collection of twelve baskets. Each contains three of the six items 1 through 6.

{1,2,3} {2,3,4} {3,4,5} {4,5,6}

{1,3,5} {2,4,6} {1,3,4} {2,4,5}

{3,5,6} {1,2,4} {2,3,5} {3,4,6}

Suppose the support threshold is 4. On the first pass of the PCY Algorithm we use a hash table with 11 buckets, and the set {i, j} is hashed to bucket $i \times j$ *mod* 11.

(a) By any method, compute the support for each item and each pair of items.

(b) Which pairs hash to which buckets?

(c) Which buckets are frequent?

(d) Which pairs are counted on the second pass of the PCY Algorithm?

Solución 6.3.1

(a)

item	1	2	3	4	5	6
support	4	6	8	8	6	4

pair	{1, 2}	{1, 3}	{1, 4}	{1, 5}	{1, 6}	{2, 3}	{2, 4}	{2, 5}
support	2	3	2	1	0	3	4	2

pair	{2, 6}	{3, 4}	{3, 5}	{3, 6}	{4, 5}	{4, 6}	{5, 6}
support	1	4	4	2	3	3	2

(b)

pair	{1, 2}	{1, 3}	{1, 4}	{1, 5}	{1, 6}	{2, 3}	{2, 4}	{2, 5}
bucket	2	3	4	5	6	6	8	10
pair	{2, 6}	{3, 4}	{3, 5}	{3, 6}	{4, 5}	{4, 6}	{5, 6}	
bucket	1	1	4	7	9	2	8	

(c)

bucket	0	1	2	3	4	5	6	7	8	9	10
support	0	5	5	3	6	1	3	2	6	3	2

Los cubos frecuentes son aquellos con soporte por encima de 4, en este caso serían: 1, 2, 4 y 8.

(d)

Como solo los pares en cubos frecuentes se contarán en la segunda pasada de PCY, serían:

{1, 2}, {1, 4}, {2, 4}, {2, 6}, {3, 4}, {3, 5}, {4, 6}, {5, 6}

Exercise 6.3.2: Suppose we run the Multistage Algorithm on the data of Exercise 6.3.1, with the same support threshold of 4. The first pass is the same as in that exercise, and for the second pass, we hash pairs to nine buckets, using the hash function that hashes {i, j} to bucket i + j mod 9. Determine the counts of the buckets on the second pass. Does the second pass reduce the set of candidate pairs? Note that all items are frequent, so the only reason a pair would not be hashed on the second pass is if it hashed to an infrequent bucket on the first pass.

Exercise 6.3.3: Suppose we run the Multihash Algorithm on the data of Exercise 6.3.1. We shall use two hash tables with five buckets each. For one, the set {i,j}, is hashed to bucket 2i+3j+4 mod 5, and for the other, the set is hashed to i + 4j mod 5. Since these hash functions are not symmetric in i and j, order the items so that i < j when evaluating each hash function. Determine the counts of each of the 10 buckets. How large does the support threshold have to be for the Multistage Algorithm to eliminate more pairs than the PCY Algorithm would, using the hash table and function described in Exercise 6.3.1?