Esercises on Association Analysis

Exercise. Argue rigorously that given a family F of itemsets of the same length, represented as sorted arrays of items, function APRIORI-GEN(F) does not generate the same itemset twice.

Solution. Consider an itemset $Z = Z[1]Z[2]\cdots Z[k]$ generated by APRIORI-GEN(F), and assume the items are sorted. During the candidate generation phase of APRIORI-GEN(F), Z can be generated only by the pair of itemsets $X = Z[1]Z[2]\cdots Z[k-2]Z[k-1]$ and $Y = Z[1]Z[2]\cdots Z[k-2]Z[k]$. □

Exercise. Consider a dataset T of transactions over a set I of d items and suppose that there exist M frequent itemsets w.r.t. some support threshold minsup. Show that A-Priori explicitly computes the support of at most $d + \min\{M^2, dM\}$ itemsets.

Solution. The claim is a consequence of the following observations:

- In order to compute F_1, A-Priori computes the support of all of the d items
- In order to compute F_k, with $k > 1$, A-Priori computes the support of the candidates C_k generated by invoking APRIORI-GEN(F_{k-1}). By construction, APRIORI-GEN(F_{k-1}) generates at most $\frac{|F_{k-1}|^2}{2} \leq |F_{k-1}|^2$ candidates. Moreover, each itemset $X = X[1]X[2]\cdots X[k-1]X[k] \in C_k$ can be seen as $X' \cup \{X[k]\}$, with $X' = X[1]X[2]\cdots X[k-1] \in F_{k-1}$, and for each $X' \in F_{k-1}$ there must be less than d candidates $X' \cup \{a\}$, with $a \in I$, in C_k. Hence,

$$\begin{aligned} |C_k| &< |F_{k-1}|^2 \\ |C_k| &< d|F_{k-1}|, \end{aligned}$$

  which implies that in order to produce the frequent itemsets of length greater than 1, A-Priori computes the support of at most

$$\sum_{k>1} \min\{F_{k-1}|^2, dF_{k-1}|\} \leq \min\{M^2, dM\}$$

□

Exercise. Consider two association rules $r_1 : A \to B$, and $r_2 : B \to C$, and suppose that both satisfy support and confidence requirements. Is it true that also $r_3 : A \to C$ satisfies the requirements? If so, prove it, otherwise show a counterexample.

Solution. The answer is no. Here is a counterexample. Consider the following dataset T:

TID	Items
1	ABC
2	AB
3	BC

Fix minsup = 1/2 and minconf = 2/3. We have that:

Rule	Support	Confidence
$r_1 : A \to B$	2/3	1
$r_2 : B \to C$	2/3	2/3
$r_3 : A \to C$	1/3	1/2

Clearly, rules r_1 and r_2 satisfy the support and confidence requirements, while rule r_3 satisfies neither of them. □

Exercise. Let

$$\begin{aligned} c_1 &= \text{Conf}(A \to B) \\ c_2 &= \text{Conf}(A \to BC) \\ c_3 &= \text{Conf}(AC \to B) \end{aligned}$$

What relationships do exist among the $c_i's$?

Solution. By the anti-monotonicity of support, we have that

$$c_2 = \frac{\text{Supp}(ABC)}{\text{Supp}(A)} \leq \frac{\text{Supp}(AB)}{\text{Supp}(A)} = c_1$$

and

$$c_2 = \frac{\text{Supp}(ABC)}{\text{Supp}(A)} \leq \frac{\text{Supp}(ABC)}{\text{Supp}(AC)} = c_3.$$

Instead, there is no fixed relationship between c_1 and c_3. As an exercise, think of an example where $c_1 < c_3$, and one where $c_3 < c_1$. □

Exercise. For a given itemset $X = \{x_1, x_2, \ldots, x_k\}$, define the measure:

$$\zeta(X) = \min\{\text{Conf}(x_i \to X - \{x_i\}) \ : \ 1 \leq i \leq k\}.$$

Say whether ζ is *monotone*, *anti-monotone* or neither one. Justify your answer.

Solution. Fix an arbitrary itemset $X = \{x_1, x_2, \ldots, x_k\}$ and let i be the index, between 1 and k, such that $\zeta(X) = \text{Conf}(x_i \to X - \{x_i\})$. Let X' be an itemset that strictly contains X (i.e., $X' \supset X$). We have that:

$$\zeta(X) = \text{Conf}(x_i \to X - \{x_i\}) = \frac{\text{Supp}(X)}{\text{Supp}(\{i\})} \geq \frac{\text{Supp}(X')}{\text{Supp}(\{i\})} \geq \zeta(X').$$

Hence, ζ is anti-monotone. □

Exercise. Consider the following alternative implementation of procedure APRIORI-GEN(F_{k-1}) (regard an itemset $X \in F_{k-1}$ as an array of items $X[1], X[2], \ldots, X[k-1]$ sorted according to some specified ordering of the items):

```
C_k ← ∅;
for each X ∈ F_{k-1} do
  for each (i ∈ F_1) do
    if (i > X[k-1]) then add X ∪ {i} to C_k
remove from C_k every itemset containing at least
  one subset of length k-1 not in F_{k-1}
return C_k
```

Show that the set C_k returned by the above procedure contains all frequent itemsets of length k.

Solution. Consider an arbitrary frequent itemset Z of length k, sorted by increasing item, and let $X = Z[1 \div k-1]$ and $i = Z[k]$. For the anti-monotonicity of support we have that $X \in F_{k-1}$, $i \in F_1$, and any subset of Z of length $k-1$ is in F_{k-1}. Note also that $i > X[k-1]$, since Z is assumed to be sorted. Hence $Z = X \cup \{i\}$ is added to C_k by the two nested for-each loops, and cannot be subsequently removed. □

Exercise. Let T be a dataset of transactions over I. Recall that the *closure* of an itemset $X \subseteq I$ is defined as $\text{Closure}(X) = \bigcap_{t \in T_X} t$, where T_X is the set of transactions that contain X. Recall also that X and $\text{Closure}(X)$ have the same support.

1. Show that if X is a closed itemset then $X = \text{Closure}(X)$.
2. Let $X, Y \subseteq I$ be two closed itemsets and define $Z = X \cap Y$.
   (a) Find a relation among T_X, T_Y and T_Z (i.e., the sets of transactions containing X, Y, and Z, respectively). Justify your answer.
   (b) Show that Z is also closed.

Solution.

1. Since $X \subset t$ for every $t \in T_X$, we have that $X \subseteq \text{Closure}(X)$. Since X and $\text{Closure}(X)$ have the same support, and X is closed, $\text{Closure}(X)$ cannot be larger than X.
2. (a) Since Z is contained in every transaction of T_X and in every transaction of T_Y, we have that $T_X \cup T_Y \subseteq T_Z$.
   (b) If Z were not closed, there would exist an itemset $V = Z \cup \{a\}$, for some $a \notin Z$, with the same support as Z. This itemset would be contained in every transaction $t \in T_Z$. Hence, a would be contained in every transaction $t \in T_X$ and in every transaction $t \in T_Y$, and since $X = \bigcap_{t \in T_X} t$ and $Y = \bigcap_{t \in T_Y} t$ (from the Point (1)), this would imply that $a \in X$ and $a \in Y$, hence $a \in X \cap Y = Z$, which is a contradiction.

□

Exercise. Let $I = \{a_1, a_2, \ldots, a_n\} \cup \{b_1, b_2, \ldots, b_n\}$ be a set of $2n$ item, e let $T = \{t_1, t_2, \ldots, t_n\}$ be a set of n transactions over I, where

$$t_i = \{a_1, a_2, \ldots a_n, b_i\} \qquad \text{per } 1 \leq i \leq n.$$

For minsup $= 1/n$, determine the number of frequent closed itemsets and the number of maximal itemsets.

Solution. Sia $A = \{a_1, a_2, \ldots, a_n\}$ e $B = \{b_1, b_2, \ldots, b_n\}$. Ogni sottoinsieme di A ha supporto 1, mentre ogni itemset formato da un sottoinsieme di A e un item di B ha supporto $1/n$. Tutti gli altri itemset hanno supporto 0. In questo caso gli itemset chiusi frequenti sono $n+1$, ovvero, l'itemset A e tutti gli itemset del tipo $A \cup \{b_i\}$, per $1 \leq i \leq n$. Tutti questi itemset, tranne A sono anche massimali, quindi il numero di itemset massimali è n. □

Exercise. Let d be an even integer, and define T as the set of the following $N = (3/2)d$ transactions over $I = \{1, 2, \ldots, d\}$

$$\begin{aligned} t_i &= \{i\} \quad 1 \leq i \leq d \\ t_{d+i} &= I - \{i\} \quad 1 \leq i \leq d/2. \end{aligned}$$

1. Identify the itemsets of support $> 1/3$ and the itemsets of support $= 1/3$.
2. Using the result of the previous point, show that the number of Top-K frequent itemsets, with $K = d$, is exponential in d.

Solution.

1. Gli itemset X con supporto $> 1/3$ sono tutti e soli gli 1-itemset $\{i\}$ con $d/2 < i \leq d$. In totale sono $d/2$. Gli itemset X con supporto $= 1/3$ sono tutti e soli gli 1-itemset $\{i\}$ con $1 \leq i \leq d/2$, e gli itemset X con $|X| > 1$, tali che $X \subseteq \{i : d/2 < i \leq d\}$. In totale sono $d/2 + 2^{d/2} - 1 - d/2 = 2^{d/2} - 1$. Per qualsiasi altro itemset Y, diverso da quelli sopra citati, il supporto è inferiore a $1/3$.
2. Per $K = d$ si ha che $s(k) = 1/3$, e quindi i top-K frequent itemset sono tutti e soli gli itemset X con supporto $\geq 1/3$, quindi in totale $d/2 + 2^{d/2} - 1$ itemset.

□

Exercise. Consider the mining di association rules from a dataset T of transactions. Call *standard* the rules extracted with the classical framework. We say that a standard rule $r : X \to Y$ is also *essential* if $|X| = 1$ or for each non-empty subset $X' \subset X$, $\text{Conf}(X' \to Y \cup (X - X')) < \text{Conf}(r)$.

1. Let T consists of the following 5 transactions: $(ABCD)$, $(ABCE)$, (ABC), (ABE), (BCD). Using minsup=0.5 and minconf=0.5, identify a standard rule $X \to Y$ with $|X| > 1$ which is not essential.

2. Each essential rule can be regarded as *representative* of a set of non-essential standard rule. Which subset? Justify your answer.

Solution.

1. L'itemset ABC ha supporto $3/5 > 0.5$. Le regole $A \to BC$ e $AB \to C$ hanno entrambe confidenza $3/4 > 0.5$, quindi la seconda di esse è standard ma non essenziale.

2. Una regola essenziale $r: \;\; X \to Y$ con confidenza c può essere considerata rappresentante di tutte le regole $r': \;\; X \cup Y' \to Y - Y'$, con $\emptyset \subseteq Y' \subset Y$ che hanno confidenza(r') = confidenza(r). Infatti, relativamente a queste regole X è l'itemset minimale la cui presenza in una transazione implica la presenza di $X \cup Y$ con confidenza c.

$\square$