Exercises 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 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 of the two. 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 $X[1], X[2], \ldots, X[k-1]$ in increasing order):

```
$C_k \leftarrow \emptyset$;
for each $X \in F_{k-1}$ do
  for each ($i \in F_1$) do
    if ($i > X[k-1]$) then add $X \cup \{i\}$ 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 $i = Z[k]$. Since Z is frequent, 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 a set of items I, and let CLO_T be the family of closed itemsets with respect to T. Prove that

$$\text{Supp}(X) = \max\{\text{Supp}(Y) \ : \ Y \supseteq X \wedge Y \in \text{CLO}_T\}.$$

Solution. For every itemset X there exists a closed superset $\bar{Y} \supseteq X$ such that $\text{Supp}(\bar{Y}) = \text{Supp}(X)$. The argument is as follows: if X is closed then $\bar{Y} = X$, otherwise we may keep adding items to X without decreasing its support until we reach a closed itemset $\bar{Y} \supset X$. Since, by anti-monotonicity of support, for every closed superset $Y \supseteq X$ we have $\text{Supp}(Y) \leq \text{Supp}(X) = \text{Supp}(\bar{Y})$, we conclude that

$$\text{Supp}(X) = \text{Supp}(\bar{Y}) = \max\{\text{Supp}(Y) \ : \ Y \supseteq X \wedge Y \in \text{CLO}_T\}.$$

□

Exercise. Let $I = \{1, 2, \ldots, 2n\}$ for some integer $n > 0$, and let $T = \{t_1, t_2, \ldots, t_{2n}\}$ be a set of $2n$ transactions, where $t_j = I - \{j\}$, for every $1 \leq j \leq 2n$.

1. For every $X \subseteq I$ determine $\text{Supp}(X)$ as a function of n and of the length of X.
2. Using the result of the first point, determine the number of maximal itemsets w.r.t. minsup=1/2 and argue that it is exponential in the input size.
3. What can you say about frequent closed itemsets?

Solution.

1. Consider an itemset $X \subseteq I$ of $k \leq 2n$ items. It is easy to see that X is contained in every transaction of T except those transactions t_i with $i \in X$. Therefore, X is contained in $2n - k$ transactions, hence $\text{Supp}(X) = 1 - k/(2n)$.
2. From the first point we can immediately observe that the frequent itemsets w.r.t. minsup=1/2 are all itemsets of size $\leq n$. Among these itemsets, the maximal ones are those of size n, since they are frequent and any of their supersets has support $< 1 - n/(2n) = 1/2$. Therefore, the number of maximal itemsets is $\binom{2n}{n} \geq 2^n$, while the dataset size is polynomial in n.
3. The frequent closed itemsets include all maxima itemsets, so by the previous point, the number of frequent closed itemsets is also exponential in the input size[1].

□

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] We can give a precise characterization of the frequent closed itemsets in the given dataset. Since, by the first point, every itemset is closed, the family of frequent closed itemsets coincides with the family of all frequent itemsets, that is all itemsets of size $\leq n$.

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. The itemsets of support $> 1/3$ are exactly the singleton itemsets $\{i\}$, with $d/2 < i \leq d$. Hence, there are $d/2$ such itemsets. The itemsets of support $= 1/3$ are exactly the itemsets $\{i\}$, with $1 \leq i \leq d/2$, and the itemsets X with $|X| > 1$ such that $X \subseteq \{i \; : \; d/2 < i \leq d\}$. Hence, there are $d/2 + 2^{d/2} - 1 - d/2 = 2^{d/2} - 1$ such itemsets. Any other itemset has support less than $1/3$.
2. For $K = d$ we have that $s(K) = 1/3$, hence the top-K frequent itemsets are exactly the $d/2 + 2^{d/2} - 1$ itemsets of support $\geq 1/3$.

□

Exercise. Let T be a dataset of transactions over I. Let $X, Y \subseteq I$ be two closed itemsets and define $Z = X \cap Y$.

1. 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.
2. Show that Z is also closed.

Solution.

1. Since Z is contained in every transaction of T_X and in every transaction of T_Y, we have that $T_X \cup T_Y \subseteq T_Z$.
2. By contradiction, suppose that Z is not closed. Then, some strictly larger superset of Z must have the same support, which implies that there exists $a \notin Z$ such that the itemset $V = Z \cup \{a\}$ has the same support as Z. Therefore, V must be contained in every transaction $t \in T_Z$ and, in particular, a must be contained in every transaction $t \in T_X$ and in every transaction $t \in T_Y$. Since X and Y are closed, a must belong to both X and Y, otherwise, adding a to either itemset would yield a larger itemset with the same support. This implies that $a \in X \cap Y = Z$, which contradicts the initial hypothesis.

□

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. Let $A = \{a_1, a_2, \ldots, a_n\}$ and $B = \{b_1, b_2, \ldots, b_n\}$. Every itemset subset of A has support equal to 1, while every itemset consisting of a subset of A and one item of B has support $1/n$. Every other itemset has support 0. In this case, there are $n+1$ frequent closed itemsets, namely A and every itemset of $A \cup \{b_i\}$, with $1 \leq i \leq n$. These itemsets, except for A, are also maximal, hence the number of maximal itemsets is n. □

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 $r: \; X \rightarrow Y$ is also *essential* if $|X| = 1$ or for each non-empty subset $X' \subset X$, $\text{Conf}(X' \rightarrow Y \cup (X - X')) < \text{Conf}(r)$.

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 $X \rightarrow Y$ 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. The itemset ABC has support $3/5 > 0.5$. Both the rules $A \rightarrow BC$ and $AB \rightarrow C$ have cofindence $3/4 > 0.5$, hence the latter is standard but not essential.

2. An essential rule $r: \; X \rightarrow Y$ of confidence c can be regarded to represent all rules $r': \; X \cup Y' \rightarrow Y - Y'$, with $\emptyset \subseteq Y' \subset Y$, for which $\text{Conf}(r') = \text{Conf}(r)$. Indeed, relatively to these rules X is the minimal antecedent whose presence in a transaction implies the existence of $X \cup Y$ in the transaction with confidence c.

□

Exercise. Let T be a dataset of N transactions over a set of items I, and let $\epsilon > 0$ be a parameter. For each itemset X, let s_X be an approximation of its true support $\text{Supp}_T(X)$ such that

$$\text{Supp}_T(X) - \epsilon \leq s_X \leq \text{Supp}_T(X) + \epsilon$$

Consider an ordering of all itemsets $X_1, X_2, X_3, \ldots$ such that $s_{X_1} \geq s_{X_2} \geq s_{X_3} \ldots$, and let $K < K'$ be two positive indices for which $s_{X_K} > s_{X_{K'}} + 2\epsilon$.

1. Show that for each pair of indices $i, j \geq 1$, with $i \leq K < K' \leq j$, we have $\text{Supp}_T(X_i) > \text{Supp}_T(X_j)$

2. Based on the previous points, show that the set $\{X_1, X_2, \ldots, X_{K'}\}$ contains the Top-K frequent itemsets with respect to the true support.

Solution.

1. From the property of the support approximation and from the chosen ordering we conclude that

$$\begin{aligned}\text{Supp}_D(X_i) &\geq s_{X_i} - \epsilon \\ &\geq s_{X_K} - \epsilon \\ &> s_{X_{K'}} + \epsilon \\ &\geq \text{Supp}_D(X_j).\end{aligned}$$

2. From the previous point we know that for each X_j with $j > K'$ there are at least K distinct itemsets whose support is strictly greater than $\text{Supp}(X_j)$, namely $X_1, X_2, \ldots, X_K$. Therefore, X_j cannot belong to the Top-K frequent itemsets.

□

Exercise. Consider a dataset T of N transactions over a set of items I, where each transaction contains $O(1)$ items. Show how to draw a sample S of K transactions from T, uniformly at random with replacement, in one MapReduce round. How much local space is needed by your method? Assume that T is provided in input as a set of key-value pairs (i, t_i) where the key is a TID i, with $0 \leq i < N$, and the value is a transaction t_i.

Solution. We can extract the sample in one round by first choosing the indices of the transactions to be sampled in the map phase, and then extracting the transactions in the reduce phase. The details of the algorithm are as follows.

- *Map phase*: map each pair (i, t_i) into the pair $(i \bmod (N/K), (i, t_i))$, where the key is $i \bmod (N/K)$. Moreover, if $i < K$ generate a random value, $x_i \in [0, N)$, producing the N/K pairs $(0, x_i), (1, x_i), \ldots, (N/K - 1, x_i)$.

- *Map phase*: For each key j independently, with $0 \leq j < N/K$, gather the set T_j of all intermediate pairs $(j, (i, t_i))$ (i.e., with $j = i \bmod (N/K)$), and the set H_j of K, intermediate pairs (j, x_i) with $0 \leq i < K$. Note that T_j and H_j contain K pairs each. Extract from T_j the multiset S_j of all pairs $(j, (i, t_i))$ such that i corresponds to some index x_ℓ with $\ell \in [0, K)$. Multiple copies of $(j, (i, t_i))$ will be included in S_j in case more than one of the indices $x_1, x_2, \ldots, x_\ell$ (which are chosen with replacement) are equal to i.

The final sample S will be the aggregation of all S_j's. It is immediate to see that the algorithm requires $M_L = \Theta(K)$ and aggregate space $M_A = \Theta(N)$. □