University of Southern Denmark
IMADA

DM566: Data Mining and Machine Learning

Spring term 2022

Exercise 2: Apriori, Confidence, Itemsets and Association Rules

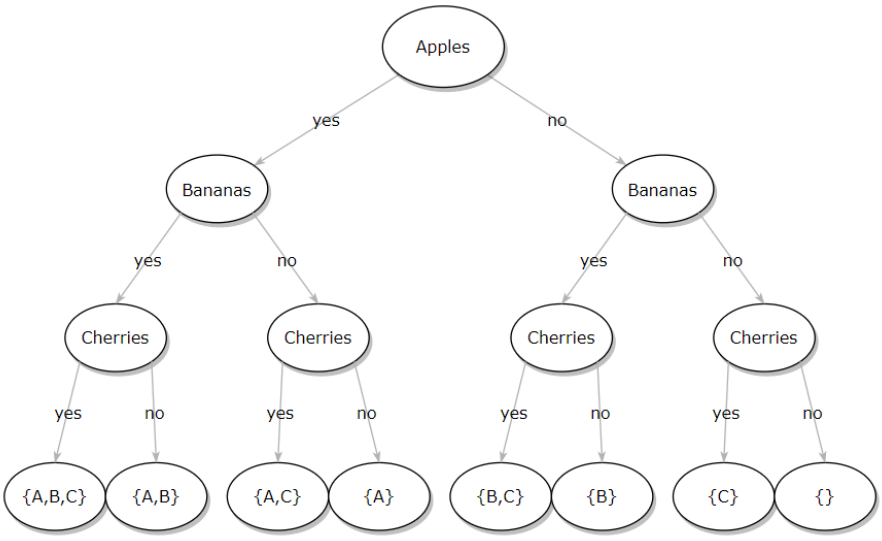
Exercise 2-1 Combinatoric Explosion (1 point)

1. A database contains transactions over the following items: *apples*, *bananas*, and *cherries*. How many different combinations of these items can exist (i.e., how many different transactions could possibly occur in the database)?

   (We do not distinguish whether a transaction contains a fruit once or several times, e.g., regardless if someone bough one apple or several apples would just result in the transaction containing *apples*).

   Suggested solution:
   A transaction can either contain apples or not. We have 2 possibilities here.
   Each of these possiblities can either contain bananas or not. That is, for each of the 2 previous possibilities, we have 2 possibilities. Therefore we have four overall.
   Each of these four possibilities can either contain cranberries or not. Eight possibilities.

   For illustration, you could sketch this at the blackboard as a branching pattern (like a binary tree: first layer: apples: yes/no, second layer at each branch: bananas: yes/no, etc.).



2. The database now also contains the items *dates*, *eggplants*, *figs*, and *guavas*. How many possible transactions do we have now?

   Suggested solution:
   It becomes clear that sketching a tree is not convenient anymore. We can explain this in another way, using a table.
   Fill in the TID column later, after discussion about the numbers and their relationship to the code we have in each row (binary vs. decimal):

TID	A	B	C	D	E	F	G
0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	1
2	0	0	0	0	0	1	0
3	0	0	0	0	0	1	1
4	0	0	0	0	1	0	0
5	0	0	0	0	1	0	1
6	0	0	0	0	1	1	0
7	0	0	0	0	1	1	1
8	0	0	0	1	0	0	0
⋮							
16	0	0	1	0	0	0	0
⋮							
127	1	1	1	1	1	1	1

3. How many combinations (possible different transactions) do we have with n items?

   Suggested solution:
   From the previous discussion it should have become clear now that the set of all possible combinations is the powerset over the set of items, where each item can be either in or out. This property (in or out) can be represented as a binary code, i.e., each element of the powerset can be uniquely mapped to exactly one number in binary representation, and each number x, $0 \leq x < 2^n$, can be uniquely mapped to exactly one element of the powerset.

   So we have overall 2^n possible combinations (i.e., different transactions), where n is the number of items.

   We say, the number of possibilities grows exponentially. And this growth rate is quite fast. For $n = 10$ we have 1024, for $n = 20$ we have 1,048,576, for $n = 30$ we have 1,073,741,824.

4. How many transactions with exactly two items (i.e., 2-itemsets) can we have when the database contains 3 items? When it contains 5 items? How many k-itemsets do we have when the database contains n items?

   Suggested solution:
   Use the database with 3 items as example: we can have two of the three elements A, B, C: $\{A, B\}$, $\{A, C\}$, $\{B, C\}$.

   To answer the question with complete enumeration of all possibilities for 5 elements already becomes tiresome, so we will derive the answer from the general solution.

   You can view this question as having a collection of n elements and drawing k of them sequentially without putting an element back.

   Let us first assume, we care for the order of drawing, i.e., we distinguish $\{A, B\}$ from $\{B, A\}$. Then we have n possibilities to draw the first element, $n-1$ possibilities to draw the second, and so on until we have $n-k+1$ possibilities to draw the k-th element. Altogether:

$$\begin{aligned} n \cdot (n-1) \cdot (n-2) \cdot \ldots \cdot (n-k+1) &= \prod_{i=0}^{k-1} (n-i) \\ &= \frac{n!}{(n-k)!} \end{aligned}$$

   Now we do actually not care for the order, i.e., we do not distinguish $\{A, B\}$ from $\{B, A\}$. Therefore we have to divide the result by the number of possible orderings. A set of k elements can be ordered/permuted in $k!$ different ways.

   The number of k-itemsets out of n different items is therefore the expression from before, divided by $k!$:

$$\frac{n!}{k!(n-k)!}$$

   This is also written with the expression

$$\binom{n}{k}$$

   and called the binomial coefficient.

Exercise 2-2 Itemsets and Association Rules (1 point)

Given a set of transactions T according to the following table:

Transaction ID	Items in basket
1	{ Milk, Eggs, Pasta }
2	{ Bread, Butter, Milk }
3	{ Milk, Pasta, Onions }
4	{ Bread, Butter, Onions }
5	{ Eggs, Onions, Pasta }
6	{ Milk, Pasta, Bread, Butter }
7	{ Bread, Butter, Pasta }
8	{ Eggs, Pasta }
9	{ Milk, Pasta, Bread, Butter }
10	{ Eggs, Onions }

1. What are the support and confidence of { Pasta } $\rightarrow$ { Milk }?

   Suggested solution:
   Support is the amount of times they appear together in the table.
   Support is 4.

   Confidence is the support, divided by the times pasta appears alone in the table.
   Confidence is $\frac{4}{7} = 57\%$.

2. What is the maximum number of size-3 itemsets that can be derived from this data set?

   Suggested solution:
   First we need to know the number of items:

   {Milk, Eggs, Pasta, Bread, Butter, Onions}

   To choose any 3 of 6, the mathematical term is

   $$\binom{6}{3} = \frac{6!}{3!(6-3)!}$$
   $$= 20$$

3. What is the maximum number of association rules that can be extracted from this dataset (including rules which have zero support)?

   Suggested solution:
   From six items, we can generate association rules by having 1 or 2 or ... or 5 items in the antecedent and include all or some of the remaining items in the consequent (we exclude the case of an empty consequent, hence we subtract 1 from the number of elements in the powerset of the remaining items).

   Mathematically:

   $$\binom{6}{1} \cdot (2^5 - 1) + \binom{6}{2} \cdot (2^4 - 1) + \binom{6}{3} \cdot (2^3 - 1) + \binom{6}{4} \cdot (2^2 - 1) + \binom{6}{5} \cdot (2^1 - 1),$$

that is for d items:

$$\sum_{i=1}^{d-1} \binom{d}{i} \cdot (2^{d-i} - 1)$$

The actual number is therefore:

$$\sum_{i=1}^{5} \binom{6}{i} \cdot (2^{6-i} - 1) = 602.$$

4. What is the maximum size of frequent itemsets that can be extracted (assuming $\sigma > 0$)?

   Suggested solution:
   The maximum frequent itemset occuring in the database has size 4. We can therefore not find any larger itemset with support > 0.

5. Find an itemset (of size 2 or larger) that has the largest support.

   Suggested solution:
   $s(\{Bread, Butter\}) = 5$

6. Find a pair of items a, b, s.t. the rules $\{a\} \rightarrow \{b\}$ and $\{b\} \rightarrow \{a\}$ have the same confidence.

   Suggested solution:

$$\begin{aligned}
\text{conf}(\{Bread\} \Rightarrow \{Butter\}) &= \frac{s(\{Bread, Butter\})}{s(\{Bread\})} \\
&= \frac{5}{5} \\
&= 1 \\
\text{conf}(\{Butter\} \Rightarrow \{Bread\}) &= \frac{s(\{Bread, Butter\})}{s(\{Butter\})} \\
&= \frac{5}{5} \\
&= 1
\end{aligned}$$

Exercise 2-3 Apriori Candidate Generation (1 point)

Given the frequent 3-itemsets:

$$\{a,b,c\}, \{a,b,d\}, \{a,b,e\}, \{a,c,d\}, \{a,c,e\}, \{b,c,d\}, \{b,c,e\}, \{c,d,e\},$$

list all candidate 4-itemsets following the Apriori joining and pruning procedure.

Suggested solution:

Joining:

Two frequent $(k-1)$-itemsets are joined if they are identical in the first $k-2$ items.

$$\{a,b,c\}+\{a,b,d\} \rightarrow \{a,b,c,d\}$$
$$\{a,b,c\}+\{a,b,e\} \rightarrow \{a,b,c,e\}$$
$$\{a,b,d\}+\{a,b,e\} \rightarrow \{a,b,d,e\}$$
$$\{a,c,d\}+\{a,c,e\} \rightarrow \{a,c,d,e\}$$
$$\{b,c,d\}+\{b,c,e\} \rightarrow \{b,c,d,e\}$$
$$\{c,d,e\} \text{ no joining partner}$$

Pruning:

Remove any non-frequent itemsets.

$\{a,b,d,e\}$ cannot be frequent as $\{a,d,e\}$ and $\{b,d,e\}$ is not frequent.

$\{a,c,d,e\}$ cannot be frequent as $\{a,d,e\}$ is not frequent.

$\{b,c,d,e\}$ cannot be frequent as $\{b,d,e\}$ is not frequent.

Result:

After joining and pruning we have the following itemsets.

$$\{a,b,c,d\} \text{ and } \{a,b,c,e\}$$

Exercise 2-4 The monotonicity of confidence (1 point)

Theorem 2.1 (lecture) states:

Given:

- itemset X
- $Y \subset X, Y \neq \emptyset$,

If $\text{conf}(Y \rightarrow (X \setminus Y)) < c$, then $\forall Y' \subset Y$: $\text{conf}(Y' \rightarrow (X \setminus Y')) < c$.

1. Prove the theorem.

   Suggested solution:

   Consider the following two rules:

   $$Y' \Rightarrow X \backslash Y'$$

   and

   $$Y \Rightarrow X \backslash Y$$

   where $Y' \subset Y$.

   The confidence of the rules are: $\frac{s(X)}{s(Y')}$ and $\frac{s(X)}{s(Y)}$, respectively.

   Since $Y' \subset Y$, we have: $s(Y') \geq s(Y)$.

   Therefore the former rule cannot have a higher confidence than the latter rule.

2. Sketch an algorithm (pseudo code) that generates all association rules with support σ or above and a minimum confidence of c, provided the set F of all frequent itemsets (w.r.t. σ) with their support, efficiently using the pruning power of the given theorem. What is the asymptotic time complexity of the algorithm?

   Suggested solution:

   AssociationRules(F, c):

   ```
   foreach Z ∈ F, |Z| ≥ 2 do:
     A = {X|X ⊂ Z, X ≠ ∅}
     while A ≠ ∅ do:
       X = maximal element in A
       A = A\{X}
       c_tmp = s(Z)\s(X)
       if c_tmp ≥ c then
         print X ⇒ (Z\X), s(Z), c_tmp
       else
         A = A\{W|W ⊂ X}
       end if
     end while
   end foreach
   ```