UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
School of Professional & Executive Development

Algorithms for association rules

Lluís A. Belanche
Computer Science Department
belanche@cs.upc.edu

Big Data Management and Analytics

February 8, 2017

Introduction to association rules

The **Market Basket Analysis** (MBA) problem assumes that we have a big set of **items** (like *products*: milk, bread, ...) with which we can fill **transactions** (like *market baskets*).

The goal pursued is twofold:

1. know which items appear often in the same transactions (which products are often bought together)
2. use this information to increase some kind of benefit (e.g. alter prices, offer pack discounts, redeploy products in corridors, do stock previsions)

Introduction to association rules

The applications are countless:

- **transactions = documents; items = words** "Which words appear often together in documents". May indicate the existence of complex concepts
- **transactions = documents; items = sentences** "Which sentences appear often together in documents". May indicate the existence of plagiarism
- **transactions = web pages; items = html tags** May indicate web mirroring (needs research for the web 2.0, much more interesting to analyze)
- **transactions = courses; items = enrolments** "Which courses are taken often together by students". We do that at the FIB (useful for making enrolment forecasting)

Introduction to association rules

We are particularly interested in:

Frequent itemsets

Set of items ("itemset") that appear together quite often (above a given level of frequency) (just occasionally we will be interested in products that are seldom sold)

Association rules

LHS ⇒ RHS, LHS is an itemset, RHS is a singleton. Useful for making predictions on the probability of the item in the RHS (given the itemset in the LHS), we call it the confidence of the rule

Association rules: Causality

The implication in association rules is not equivalent to "causality". Examples:

- $\{milk, butter\} \Rightarrow bread$. This rule may have a high confidence just because many people buy bread
- $\{pasta\ sauce\} \Rightarrow pasta$. This rule could well imply causality. How could we check it? What about raising the price of the pasta and lower the price of the pasta sauce?

1. If there *really* is a causality, people who buy pasta will continue buying the sauce (and it is cheaper now!), because the only reason to buy the sauce is having bought the pasta first ...
2. What if we make a "special pack" (pasta + pasta sauce)? Should we price it above, equal or below the separate price?

Association rules: Causality

The implication in association rules is not equivalent to "causality". Examples:

- $\{milk, butter\} \Rightarrow bread$. This rule may have a high confidence just because many people buy bread
- $\{pasta\ sauce\} \Rightarrow pasta$. This rule could well imply causality. How could we check it? What about raising the price of the pasta and lower the price of the pasta sauce?

1. If there *really* is a causality, people who buy pasta will continue buying the sauce (and it is cheaper now!), because the only reason to buy the sauce is having bought the pasta first ...
2. What if we make a "special pack" (pasta + pasta sauce)? Should we price it above, equal or below the separate price?

Association rules: Causality

The implication in association rules is not equivalent to "causality". Examples:

- $\{milk, butter\} \Rightarrow bread$. This rule may have a high confidence just because many people buy bread
- $\{pasta\ sauce\} \Rightarrow pasta$. This rule could well imply causality. How could we check it? What about raising the price of the pasta and lower the price of the pasta sauce?

1. If there *really* is a causality, people who buy pasta will continue buying the sauce (and it is cheaper now!), because the only reason to buy the sauce is having bought the pasta first ...
2. What if we make a "special pack" (pasta + pasta sauce)? Should we price it above, equal or below the separate price?

Association rules: Causality

The implication in association rules is not equivalent to "causality". Examples:

- $\{milk, butter\} \Rightarrow bread$. This rule may have a high confidence just because many people buy bread
- $\{pasta\ sauce\} \Rightarrow pasta$. This rule could well imply causality. How could we check it? What about raising the price of the pasta and lower the price of the pasta sauce?

  1. If there *really* is a causality, people who buy pasta will continue buying the sauce (and it is cheaper now!), because the only reason to buy the sauce is having bought the pasta first ...
  2. What if we make a "special pack" (pasta + pasta sauce)? Should we price it above, equal or below the separate price?

Association rules: Causality

The implication in association rules is not equivalent to "causality". Examples:

- $\{milk, butter\} \Rightarrow bread$. This rule may have a high confidence just because many people buy bread
- $\{pasta\ sauce\} \Rightarrow pasta$. This rule could well imply causality. How could we check it? What about raising the price of the pasta and lower the price of the pasta sauce?

1. If there *really* is a causality, people who buy pasta will continue buying the sauce (and it is cheaper now!), because the only reason to buy the sauce is having bought the pasta first ...
2. What if we make a "special pack" (pasta + pasta sauce)? Should we price it above, equal or below the separate price?

Association rules: Causality

The implication in association rules is not equivalent to "causality". Examples:

- $\{\mathit{milk}, \mathit{butter}\} \Rightarrow \mathit{bread}$. This rule may have a high confidence just because many people buy bread
- $\{\mathit{pasta\ sauce}\} \Rightarrow \mathit{pasta}$. This rule could well imply causality. How could we check it? What about raising the price of the pasta and lower the price of the pasta sauce?

1. If there *really* is a causality, people who buy pasta will continue buying the sauce (and it is cheaper now!), because the only reason to buy the sauce is having bought the pasta first ...
2. What if we make a "special pack" (pasta + pasta sauce)? Should we price it above, equal or below the separate price?

Association rules: Causality

The implication in association rules is not equivalent to "causality". Examples:

- $\{milk, butter\} \Rightarrow bread$. This rule may have a high confidence just because many people buy bread
- $\{pasta\ sauce\} \Rightarrow pasta$. This rule could well imply causality. How could we check it? What about raising the price of the pasta and lower the price of the pasta sauce?

    1. If there *really* is a causality, people who buy pasta will continue buying the sauce (and it is cheaper now!), because the only reason to buy the sauce is having bought the pasta first ...
    2. What if we make a "special pack" (pasta + pasta sauce)? Should we price it above, equal or below the separate price?

Formalization

We have a binary data base (transactions × items):

	items (R)
transactions (T)	$01000\cdots11000$
	$00100\cdots10010$
	$\vdots$
	$01100\cdots10010$

Let R be the total set of items and T the set of transactions. We call **itemset** to any $X \subseteq R$. Let $X = \{X_1, \ldots, X_n\}$ an itemset. Define the **support** (a.k.a. frequency) of X as:

$$s(X) := \frac{|t \in T / X_1 = X_2 = \ldots = X_n = 1|}{|T|}$$

Let $\alpha_f \in (0, 1)$; we say an itemset X is **frequent** if $s(X) \geq \alpha_f$

Formalization

Our first goal is to find **all the frequent itemsets** in the data base Consider again $X = \{X_1, \dots, X_n\}$ an itemset, $Y \in R$ an item, where $Y \notin X$. A rule is of the form $\{X_1, \dots, X_n\} \Rightarrow Y$. Define the **frequency** of the rule as:

$$freq(X \Rightarrow Y) := s(X \cup \{Y\})$$

and the **confidence** of the rule (aka rule cover) as:

$$conf(X \Rightarrow Y) := \frac{s(X \cup \{Y\})}{s(X)}$$

The confidence can be seen as a frequentist estimation of $Pr(Y|X)$, given the database.

Formalization

Notice $freq(X \Rightarrow Y) = freq(Y \Rightarrow X)$ but $conf(X \Rightarrow Y) \neq conf(Y \Rightarrow X)$

Let $\alpha_f, \alpha_c \in (0, 1)$ fixed. We say a rule $X \Rightarrow Y$ is a **basic rule** if:

1. $freq(X \Rightarrow Y) \geq \alpha_f$
2. $conf(X \Rightarrow Y) \geq \alpha_c$

Our second (and main) goal is thus an **algorithm** to find all the basic rules in the data base (for fixed α_f, α_c) in an efficient way.

This algorithm can be used with different values of α_f, α_c to create sets of **interesting** rules (basic, novel, non-trivial and useful).

Example

id	A	B	C	D	E	F	G	H	I	J	K
t_1	1	1	1	1	0	0	1	0	0	0	0
t_2	1	1	0	0	1	1	0	0	0	0	0
t_3	0	1	0	0	0	0	0	0	1	0	1
t_4	1	1	0	0	0	0	0	1	0	0	0
t_5	0	0	0	0	1	0	1	0	0	1	0

Note the *sparsity* of the matrix, something to be expected.
Exercise: compute the basic rules if $\alpha_f = 0.3$ and $\alpha_c = 0.9$.

Example

Solution: $\mathit{freq}(\{A\} \Rightarrow B) = \frac{3}{5}, \mathit{conf}(\{A\} \Rightarrow B) = 1, \ldots$

We arrive at the set of frequent itemsets:
$\mathcal{F} = \{\{A\}, \{B\}, \{E\}, \{G\}, \{A, B\}\}$.

The only possible rules are therefore $\{A\} \Rightarrow B$ and $\{B\} \Rightarrow A$.

Since $\mathit{freq}(\{B\} \Rightarrow A) = \frac{3}{5}, \mathit{conf}(\{B\} \Rightarrow A) = \frac{3}{4}$, the answer is the sole rule $\{A\} \Rightarrow B$.

Association rules: an Algorithm

The previous example suggests this way of proceeding:

1. Find all the frequent itemsets $\mathcal{F}$
2. Generate all the rules of the form $F \setminus \{Z\} \Rightarrow Z$, where $Z \in F \in \mathcal{F}$.
3. Check if these rules are basic

And do it in an efficient way ... what is the computational cost of "generate & test" all the rules by brute force? $O(|R|^2 \cdot 2^{|R|-1} \cdot |T|)$

Association rules: an Algorithm

What ??????? How on Earth can we make this efficient?

Observation:

$\boxed{F \text{ is frequent} \Longrightarrow \text{all the subsets of } F \text{ are frequent}}$

(the opposite is not true: the best we can say is that we have no evidence to discard F as frequent)

Proof. Consider $X \subseteq Y \Rightarrow s(X) \geq s(Y)$. If X is not frequent, then neither is Y. Since this is valid for all $Y \supseteq X$, if an $X \subseteq Y$ is not frequent, none of its supersets will be. So, if Y is frequent, all of its subsets must be.

We have the **inspiring idea** of starting with all the frequent itemsets of size 1, use them to discard itemsets of size 2, check if the rest are, use these to discard itemsets of size 3, and so forth ...

Association rules: an Algorithm

What ??????? How on Earth can we make this efficient?

Observation:

$\boxed{F \text{ is frequent} \Longrightarrow \text{all the subsets of } F \text{ are frequent}}$

(the opposite is not true: the best we can say is that we have no evidence to discard F as frequent)

Proof. Consider $X \subseteq Y \Rightarrow s(X) \geq s(Y)$. If X is not frequent, then neither is Y. Since this is valid for all $Y \supseteq X$, if an $X \subseteq Y$ is not frequent, none of its supersets will be. So, if Y is frequent, all of its subsets must be.

We have the **inspiring idea** of starting with all the frequent itemsets of size 1, use them to discard itemsets of size 2, check if the rest are, use these to discard itemsets of size 3, and so forth …

Association rules: an Algorithm

What ??????? How on Earth can we make this efficient?

Observation:

$\boxed{F \text{ is frequent} \Longrightarrow \text{all the subsets of } F \text{ are frequent}}$

(the opposite is not true: the best we can say is that we have no evidence to discard F as frequent)

Proof. Consider $X \subseteq Y \Rightarrow s(X) \geq s(Y)$. If X is not frequent, then neither is Y. Since this is valid for all $Y \supseteq X$, if an $X \subseteq Y$ is not frequent, none of its supersets will be. So, if Y is frequent, all of its subsets must be.

We have the **inspiring idea** of starting with all the frequent itemsets of size 1, use them to discard itemsets of size 2, check if the rest are, use these to discard itemsets of size 3, and so forth ...

Association rules: an Algorithm

What ??????? How on Earth can we make this efficient?

Observation:

$\boxed{F \text{ is frequent} \Longrightarrow \text{all the subsets of } F \text{ are frequent}}$

(the opposite is not true: the best we can say is that we have no evidence to discard F as frequent)

Proof. Consider $X \subseteq Y \Rightarrow s(X) \geq s(Y)$. If X is not frequent, then neither is Y. Since this is valid for all $Y \supseteq X$, if an $X \subseteq Y$ is not frequent, none of its supersets will be. So, if Y is frequent, all of its subsets must be.

We have the **inspiring idea** of starting with all the frequent itemsets of size 1, use them to discard itemsets of size 2, check if the rest are, use these to discard itemsets of size 3, and so forth ...

Association rules: an Algorithm

Algorithm 1: APriori

```
input  : $T$ set of transactions, $R$ set of items, $\alpha_f$
output: $\mathcal{F}$ set of all the frequent itemsets
$i \leftarrow 1$
$C_i \leftarrow \{\{A\} | A \in R\}$
while $C_i \neq \emptyset$ do
    $\mathcal{L}_i \leftarrow \emptyset$
    for $c \in C_i$ do
        if $s(c) \geq \alpha_f$ then $\mathcal{L}_i \leftarrow \mathcal{L}_i \cup \{c\}$
    end
    $C_{i+1} \leftarrow \{$sets of size $i+1$ having all of their subsets of size $i$ within $\mathcal{L}_i\}$
    $i \leftarrow i+1$
end
$\mathcal{F} \leftarrow \cup_{i \geq 2} \mathcal{L}_i$
```

Questions, questions ...

1. How do we form the set C_{i+1} from the set $\mathcal{L}_i$?
2. How efficiently can we compute $s(c) \geq \alpha_f$ for all $c \in C_i$?
3. How do we generate the rules once we have $\mathcal{F}$? Where do we get their support and confidences from?

Answers, answers ...

How do we form the set C_{i+1} from the set $\mathcal{L}_i$?

join : find all the set pairs $\{A, B\}$, where $A, B \in \mathcal{L}_i$, such that $|A \cup B| = i + 1$ ("potential candidates")

prune : check whether each potential candidate is really a candidate: all of its subsets of size i must belong to $\mathcal{L}_i$

Answers, answers ...

How efficiently can we compute $s(c) \geq \alpha_f$ for all $c \in C_i$?

- We go transaction after transaction, keeping a counter γ_c (initialized to 0) for every $c \in C_i$.
- The counter γ_c is incremented by one when a transaction covers c
  (a transaction **covers** a set if all the items in the set have value 1 in the transaction)
- So, the whole inner **for** loop (lines 5–7) can be done in just one pass through the database

Answers, answers ...

How do we generate the rules once we have $\mathcal{F}$? Where do we get their support and confidences from?
First we simplify the set, by removing every $F \in \mathcal{F}$ for which $\exists G \in \mathcal{F}$ such that $G \supset F$. Now letting $F \in \mathcal{F}$ such that $|F| \geq 2$; do for each $Z \in F$:

1. generate the rule $F \setminus \{Z\} \Rightarrow Z$
2. the frequency of the rule is $\mathit{freq}(F \setminus \{Z\} \Rightarrow Z) = s((F \setminus \{Z\}) \cup \{Z\})$. Since $Z \in F$, this is equal to $s(F)$
3. the confidence of the rule is $\mathit{conf}(F \setminus \{Z\} \Rightarrow Z) = \frac{s(F)}{s(F \setminus \{Z\})}$

We already have these quantities, since both F and $F \setminus \{Z\}$ are frequent sets, and thus the algorithm computed their frequencies

Example

	t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9	t_{10}	t_{11}	t_{12}	t_{13}	t_{14}	t_{15}	t_{16}	t_{17}	t_{18}	t_{19}	t_{20}
A	1	1	0	1	0	1	1	1	1	1	1	1	1	0	1	1	0	0	1	1
B	1	0	1	1	1	0	1	1	0	1	1	1	1	1	1	1	1	1	0	1
C	0	1	1	1	0	1	0	0	1	1	0	0	1	1	0	0	0	1	0	0
D	0	0	1	0	0	1	0	0	1	0	0	1	1	0	1	0	0	0	0	0
E	1	1	0	0	1	0	1	0	0	1	1	1	1	0	0	1	1	1	1	1
F	0	1	1	0	1	0	0	0	1	1	0	0	1	0	0	0	0	0	0	0

Execute APriori to obtain $\mathcal{F}$; use this set to obtain all the basic rules if $\alpha_f = \frac{1}{4}, \alpha_c = \frac{3}{4}$.

Solution to the example

$s(\{A\}) = \frac{3}{4}, s(\{B\}) = \frac{4}{5}, s(\{C\}) = \frac{9}{20}, s(\{D\}) = \frac{3}{10}, s(\{E\}) = \frac{13}{20}, s(\{F\}) = \frac{3}{10}$

$C_1 = \{\{A\}, \{B\}, \{C\}, \{D\}, \{E\}, \{F\}\}, \mathcal{L}_1 = \{\{A\}, \{B\}, \{C\}, \{D\}, \{E\}, \{F\}\}$

$C_2 = \{\{A, B\}, \{A, C\}, \dots, \{E, F\}\}$ $s(\{A, B\}) = \frac{11}{20}, s(\{A, C\}) = \frac{3}{10}, \dots, s(\{E, F\}) = \frac{1}{5}$

$\mathcal{L}_2 = \{\{A, B\}, \{A, C\}, \{A, D\}, \{A, E\}, \{B, C\}, \{B, E\}, \{C, F\}\}$

$C_3 = \{\{A, B, C\}, \{A, B, E\}\}$ why $\{A, C, F\} \notin C_3$? because $\{A, F\} \notin \mathcal{L}_2$

$s(\{A, B, C\}) = \frac{3}{20}, s(\{A, B, E\}) = \frac{8}{20}$

$\mathcal{L}_3 = \{\{A, B, E\}\}$

$C_4 = \emptyset$

Solution to the example

$s(\{A\}) = \frac{3}{4}, s(\{B\}) = \frac{4}{5}, s(\{C\}) = \frac{9}{20}, s(\{D\}) = \frac{3}{10}, s(\{E\}) = \frac{13}{20}, s(\{F\}) = \frac{3}{10}$

$C_1 = \{\{A\}, \{B\}, \{C\}, \{D\}, \{E\}, \{F\}\}, \mathcal{L}_1 = \{\{A\}, \{B\}, \{C\}, \{D\}, \{E\}, \{F\}\}$

$C_2 = \{\{A, B\}, \{A, C\}, \ldots, \{E, F\}\}$ $s(\{A, B\}) = \frac{11}{20}, s(\{A, C\}) = \frac{3}{10}, \ldots, s(\{E, F\}) = \frac{1}{5}$

$\mathcal{L}_2 = \{\{A, B\}, \{A, C\}, \{A, D\}, \{A, E\}, \{B, C\}, \{B, E\}, \{C, F\}\}$

$C_3 = \{\{A, B, C\}, \{A, B, E\}\}$ why $\{A, C, F\} \notin C_3$? because $\{A, F\} \notin \mathcal{L}_2$

$s(\{A, B, C\}) = \frac{3}{20}, s(\{A, B, E\}) = \frac{8}{20}$

$\mathcal{L}_3 = \{\{A, B, E\}\}$

$C_4 = \emptyset$

Solution to the example

$s(\{A\}) = \frac{3}{4}, s(\{B\}) = \frac{4}{5}, s(\{C\}) = \frac{9}{20}, s(\{D\}) = \frac{3}{10}, s(\{E\}) = \frac{13}{20}, s(\{F\}) = \frac{3}{10}$

$C_1 = \{\{A\}, \{B\}, \{C\}, \{D\}, \{E\}, \{F\}\}, \mathcal{L}_1 = \{\{A\}, \{B\}, \{C\}, \{D\}, \{E\}, \{F\}\}$

$C_2 = \{\{A, B\}, \{A, C\}, \ldots, \{E, F\}\}$ $s(\{A, B\}) = \frac{11}{20}, s(\{A, C\}) = \frac{3}{10}, \ldots, s(\{E, F\}) = \frac{1}{5}$

$\mathcal{L}_2 = \{\{A, B\}, \{A, C\}, \{A, D\}, \{A, E\}, \{B, C\}, \{B, E\}, \{C, F\}\}$

$C_3 = \{\{A, B, C\}, \{A, B, E\}\}$ why $\{A, C, F\} \notin C_3$? because $\{A, F\} \notin \mathcal{L}_2$

$s(\{A, B, C\}) = \frac{3}{20}, s(\{A, B, E\}) = \frac{8}{20}$

$\mathcal{L}_3 = \{\{A, B, E\}\}$

$C_4 = \emptyset$

Solution to the example

$s(\{A\}) = \frac{3}{4}, s(\{B\}) = \frac{4}{5}, s(\{C\}) = \frac{9}{20}, s(\{D\}) = \frac{3}{10}, s(\{E\}) = \frac{13}{20}, s(\{F\}) = \frac{3}{10}$

$C_1 = \{\{A\}, \{B\}, \{C\}, \{D\}, \{E\}, \{F\}\}, \mathcal{L}_1 = \{\{A\}, \{B\}, \{C\}, \{D\}, \{E\}, \{F\}\}$

$C_2 = \{\{A, B\}, \{A, C\}, \ldots, \{E, F\}\}$ $s(\{A, B\}) = \frac{11}{20}, s(\{A, C\}) = \frac{3}{10}, \ldots, s(\{E, F\}) = \frac{1}{5}$

$\mathcal{L}_2 = \{\{A, B\}, \{A, C\}, \{A, D\}, \{A, E\}, \{B, C\}, \{B, E\}, \{C, F\}\}$

$C_3 = \{\{A, B, C\}, \{A, B, E\}\}$ why $\{A, C, F\} \notin C_3$? because $\{A, F\} \notin \mathcal{L}_2$

$s(\{A, B, C\}) = \frac{3}{20}, s(\{A, B, E\}) = \frac{8}{20}$

$\mathcal{L}_3 = \{\{A, B, E\}\}$

$C_4 = \emptyset$

Solution to the example

$s(\{A\}) = \frac{3}{4}, s(\{B\}) = \frac{4}{5}, s(\{C\}) = \frac{9}{20}, s(\{D\}) = \frac{3}{10}, s(\{E\}) = \frac{13}{20}, s(\{F\}) = \frac{3}{10}$

$C_1 = \{\{A\}, \{B\}, \{C\}, \{D\}, \{E\}, \{F\}\}, \mathcal{L}_1 = \{\{A\}, \{B\}, \{C\}, \{D\}, \{E\}, \{F\}\}$

$C_2 = \{\{A, B\}, \{A, C\}, \dots, \{E, F\}\}$ $s(\{A, B\}) = \frac{11}{20}, s(\{A, C\}) = \frac{3}{10}, \dots, s(\{E, F\}) = \frac{1}{5}$

$\mathcal{L}_2 = \{\{A, B\}, \{A, C\}, \{A, D\}, \{A, E\}, \{B, C\}, \{B, E\}, \{C, F\}\}$

$C_3 = \{\{A, B, C\}, \{A, B, E\}\}$ why $\{A, C, F\} \notin C_3$? because $\{A, F\} \notin \mathcal{L}_2$

$s(\{A, B, C\}) = \frac{3}{20}, s(\{A, B, E\}) = \frac{8}{20}$

$\mathcal{L}_3 = \{\{A, B, E\}\}$

$C_4 = \emptyset$

Solution to the example

$s(\{A\}) = \frac{3}{4}, s(\{B\}) = \frac{4}{5}, s(\{C\}) = \frac{9}{20}, s(\{D\}) = \frac{3}{10}, s(\{E\}) = \frac{13}{20}, s(\{F\}) = \frac{3}{10}$

$C_1 = \{\{A\}, \{B\}, \{C\}, \{D\}, \{E\}, \{F\}\}, \mathcal{L}_1 = \{\{A\}, \{B\}, \{C\}, \{D\}, \{E\}, \{F\}\}$

$C_2 = \{\{A, B\}, \{A, C\}, \ldots, \{E, F\}\}$ $s(\{A, B\}) = \frac{11}{20}, s(\{A, C\}) = \frac{3}{10}, \ldots, s(\{E, F\}) = \frac{1}{5}$

$\mathcal{L}_2 = \{\{A, B\}, \{A, C\}, \{A, D\}, \{A, E\}, \{B, C\}, \{B, E\}, \{C, F\}\}$

$C_3 = \{\{A, B, C\}, \{A, B, E\}\}$ why $\{A, C, F\} \notin C_3$? because $\{A, F\} \notin \mathcal{L}_2$

$s(\{A, B, C\}) = \frac{3}{20}, s(\{A, B, E\}) = \frac{8}{20}$

$\mathcal{L}_3 = \{\{A, B, E\}\}$

$C_4 = \emptyset$

Solution to the example

$s(\{A\}) = \frac{3}{4}, s(\{B\}) = \frac{4}{5}, s(\{C\}) = \frac{9}{20}, s(\{D\}) = \frac{3}{10}, s(\{E\}) = \frac{13}{20}, s(\{F\}) = \frac{3}{10}$

$C_1 = \{\{A\}, \{B\}, \{C\}, \{D\}, \{E\}, \{F\}\}, \mathcal{L}_1 = \{\{A\}, \{B\}, \{C\}, \{D\}, \{E\}, \{F\}\}$

$C_2 = \{\{A, B\}, \{A, C\}, \ldots, \{E, F\}\}$ $s(\{A, B\}) = \frac{11}{20}, s(\{A, C\}) = \frac{3}{10}, \ldots, s(\{E, F\}) = \frac{1}{5}$

$\mathcal{L}_2 = \{\{A, B\}, \{A, C\}, \{A, D\}, \{A, E\}, \{B, C\}, \{B, E\}, \{C, F\}\}$

$C_3 = \{\{A, B, C\}, \{A, B, E\}\}$ why $\{A, C, F\} \notin C_3$? because $\{A, F\} \notin \mathcal{L}_2$

$s(\{A, B, C\}) = \frac{3}{20}, s(\{A, B, E\}) = \frac{8}{20}$

$\mathcal{L}_3 = \{\{A, B, E\}\}$

$C_4 = \emptyset$

Solution to the example

$s(\{A\}) = \frac{3}{4}, s(\{B\}) = \frac{4}{5}, s(\{C\}) = \frac{9}{20}, s(\{D\}) = \frac{3}{10}, s(\{E\}) = \frac{13}{20}, s(\{F\}) = \frac{3}{10}$

$C_1 = \{\{A\}, \{B\}, \{C\}, \{D\}, \{E\}, \{F\}\}, \mathcal{L}_1 = \{\{A\}, \{B\}, \{C\}, \{D\}, \{E\}, \{F\}\}$

$C_2 = \{\{A, B\}, \{A, C\}, \ldots, \{E, F\}\}$ $s(\{A, B\}) = \frac{11}{20}, s(\{A, C\}) = \frac{3}{10}, \ldots, s(\{E, F\}) = \frac{1}{5}$

$\mathcal{L}_2 = \{\{A, B\}, \{A, C\}, \{A, D\}, \{A, E\}, \{B, C\}, \{B, E\}, \{C, F\}\}$

$C_3 = \{\{A, B, C\}, \{A, B, E\}\}$ why $\{A, C, F\} \notin C_3$? because $\{A, F\} \notin \mathcal{L}_2$

$s(\{A, B, C\}) = \frac{3}{20}, s(\{A, B, E\}) = \frac{8}{20}$

$\mathcal{L}_3 = \{\{A, B, E\}\}$

$C_4 = \emptyset$

Solution to the example

$\mathcal{F} = \{\{A, B\}, \{A, C\}, \{A, D\}, \{A, E\}, \{B, C\}, \{B, E\}, \{C, F\}, \{A, B, E\}\}$

If we now generate the rules and compute their confidences:

- $conf(\{E\} \Rightarrow A) = \frac{10}{13}$
- $conf(\{E\} \Rightarrow B) = \frac{11}{13}$
- $conf(\{A, E\} \Rightarrow B) = \frac{8}{10}$
- $conf(\{D\} \Rightarrow A) = \frac{5}{6}$

are the only basic rules. Note that $\{E\} \Rightarrow B$ is a more general rule than $\{A, E\} \Rightarrow B$ (and the latter has lower confidence). Therefore we could eliminate $\{A, E\} \Rightarrow B$.

Measuring how good the rules are

Once we have the (polished) basic rules we are interested in "sorting'em out", from "best" to "worst". Unfortunately, confidence is not enough in many situations ...

To begin with, confidence does not obey augmentation:

$$conf(\{A\} \Rightarrow C) \not\leq conf(\{A, B\} \Rightarrow C)$$

Why confidence does not tell the big picture? See an example: Consider the rule $R : \{milk, butter\} \Rightarrow bread$ with $conf(R) = 0.9$ (very high, indeed). May be simply a lot of people buy bread ...

Measuring how good the rules are

Once we have the (polished) basic rules we are interested in "sorting'em out", from "best" to "worst". Unfortunately, confidence is not enough in many situations ...
To begin with, confidence does not obey augmentation:

$$conf(\{A\} \Rightarrow C) \not\leq conf(\{A, B\} \Rightarrow C)$$

Why confidence does not tell the big picture? See an example: Consider the rule $R: \{milk, butter\} \Rightarrow bread$ with $conf(R) = 0.9$ (very high, indeed). May be simply a lot of people buy bread ...

Measuring how good the rules are

Once we have the (polished) basic rules we are interested in “sorting'em out”, from “best” to “worst”. Unfortunately, confidence is not enough in many situations ...
To begin with, confidence does not obey augmentation:

$$conf(\{A\} \Rightarrow C) \not\leq conf(\{A, B\} \Rightarrow C)$$

Why confidence does not tell the big picture? See an example: Consider the rule $R : \{milk, butter\} \Rightarrow bread$ with $conf(R) = 0.9$ (very high, indeed). May be simply a lot of people buy bread ...

Measuring how good the rules are

Define the **lift** of a rule as

$$lift(X \Rightarrow Y) := \frac{conf(X \Rightarrow Y)}{s(Y)} \in [0, \infty)$$

Example: suppose

$$\begin{aligned} s(\{milk, butter\}) &= 0.7 \\ s(\{bread\}) &= 0.9 \\ freq(\{milk, butter, bread\}) &= 0.63 \end{aligned}$$

Then $conf(R) = \frac{0.63}{0.7} = 0.9$ and $lift(R) = \frac{0.9}{0.9} = 1$.

Measuring how good the rules are

Define the **lift** of a rule as

$$lift(X \Rightarrow Y) := \frac{conf(X \Rightarrow Y)}{s(Y)} \in [0, \infty)$$

Example: suppose

$$\begin{aligned} s(\{milk, butter\}) &= 0.7 \\ s(\{bread\}) &= 0.9 \\ freq(\{milk, butter, bread\}) &= 0.63 \end{aligned}$$

Then $conf(R) = \frac{0.63}{0.7} = 0.9$ and $lift(R) = \frac{0.9}{0.9} = 1$.

Measuring how good the rules are

$$\mathit{lift}(X \Rightarrow Y) \begin{cases} < 1 & \text{the rule is worse than } P(Y) \text{ in predicting } Y \\ = 1 & \text{the rule is equal than } P(Y) \text{ in predicting } Y \\ > 1 & \text{the rule is better than } P(Y) \text{ in predicting } Y \end{cases}$$

Hint: lift measures “deviation from independence”, i.e. deviation from $P(X, Y) = P(X)P(Y)$

In our **example** ($X = \{\mathit{milk}, \mathit{butter}\}$ and $Y = \{\mathit{bread}\}$) the conclusion is that the buying of milk and butter does not condition the buying of bread, because the proportions are the same

An example on using lift

Problem: Adult Data Set. Prediction task is to determine whether a person makes over 50K a year, based on census data (age, workclass, education, marital status, ...)

- Number of transactions: 48,842 (the variables need to be binarized)
- We find the rule:
  $\{marital\_status = Married\_civ\_spouse\} \Rightarrow native\_country = United\_States\}$
  (obviously there is no causality!)
- The reason is that the data come from the US Census (though there are some nationals from Mexico, the Philippines and many other countries).

$$s(\{native\_country = United\_States\}) = 0.8974$$
$$s(\{marital\_status = Married\_civ\_spouse\}) = 0.4582$$

An example on using lift

- Now, if being married (civil wedding) and having US nationality are independent, we can expect that: $s(\{native\_country = United\_States, marital\_status = Married\_civ\_spouse\}) \approx 0.41$
- Therefore $freq(\{marital\_status = Married\_civ\_spouse\} \Rightarrow native\_country = United\_States) \approx 0.41$ and $conf(\{marital\_status = Married\_civ\_spouse\} \Rightarrow native\_country = United\_States) \approx \frac{0.41}{0.4582} \approx 0.895$ (quite high!)
- ... but $lift(\{marital\_status = Married\_civ\_spouse\} \Rightarrow native\_country = United\_States) \approx \frac{0.895}{0.8974} \approx 0.997$

Implementation issues

- A particularly important operation is that of quickly checking whether a **transaction covers an itemset** (e.g, via short-cut low-level boolean operators)
- Another one is fetching the **frequency of an itemset** (e.g., via a hash table)
- It is convenient to store the elements of $\mathcal{L}_i$ in lexicographical order, as in

$$\mathcal{L}_3 = \{\{A, B, C\}, \{A, B, D\}, \{A, C, D\}, \{A, C, E\}, \{B, C, D\}\}$$

(e.g., we store $\{A, B, C\}$ and not $\{B, A, C\}$)

- Now it is a matter of merging only those pairs of subsets with the same prefix of size $i - 1$ (these sets are consecutive), e.g.

$$\{A, B, C\} \cup \{A, B, D\} = \{A, B, C, D\}$$

Pros & Cons of Association Rules

Pros

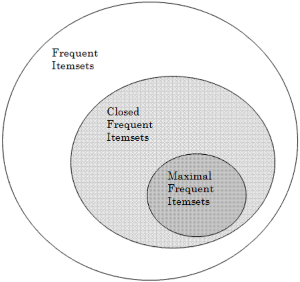
1. Algorithms are reasonably efficient
2. Data preprocessing is not immediate but it's pretty straightforward
3. Very general purpose

Cons

1. Users are overwhelmed by the number of rules identified
2. The mining algorithm assumes sparsity, which is not always true (or may be "inversely true")
3. An "optimal" value for minimum support can not be known beforehand (confidence is more intuitive)
4. What if we are interested in negative associations?

Closed and Maximal Itemsets

- An itemset X is **closed** if there exists no superset of X with the same support count as X
- A **maximal** frequent itemset is a frequent itemset which is not contained in another frequent itemset
- A **maximal itemset** is closed but a closed itemset is not necessarily maximal.



Closed and Maximal Itemsets

Frequent maximal itemsets determine all frequent itemsets! The aim of any association rule algorithm is to find these sets.

Example. Suppose $R = \{B, C, D\}, |T| = 100, \alpha_f = 0.08$ and

$$\begin{aligned} s(\{B, D\}) &= 8/100 \\ s(\{B, C, D\}) &= 8/100 \\ s(\{C, D\}) &= 9/100 \end{aligned}$$

Then $\{C, D\}$ is closed but not maximal.

Other methods

- There exist other algorithms besides Apriori to find frequent itemsets (also closed and maximal) : Eclat, FP-growth, CHARM and CHARM-L, ...
  - It is believed that Apriori is generally better than other algorithms if the support required is high
  - If the support is very low, none of the algorithms is able to handle large frequent sets smoothly
  - In any other case (intermediate supports), Apriori may not be the best option
  - If you are unsure which one to choose, it is recommended to use either Apriori or Eclat as a second option
- The interested practitioner can find the excellent work by C. Borgelt here:

  `http://www.borgelt.net/fpm.html`

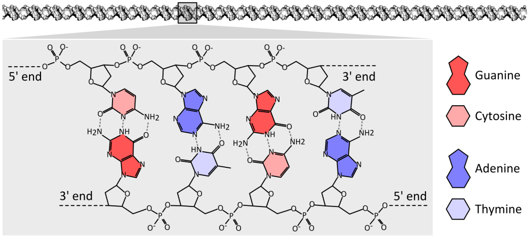
Worked example 1: Document downloads

The **Epub data set** contains the download history of documents from the electronic publication platform of the Vienna University of Economics and Business Administration. The data was recorded between Jan 2003 and Dec 2008.

Worked example 2: DNA sequences

The **Promoter Gene data set** contains DNA sequences of promoter and non-promoters.

- A *promoter* is a region of DNA that initiates or facilitates transcription of a particular gene.
- The dataset consists of 106 observations and 57 categorical variables describing the DNA sequence, represented as the nucleotide at each position: [A] adenine, [C] cytosine, [G] guanine, [T] thymine.
- There is also a response variable: "+" for a promoter gene and "−" for a non-promoter gene.



Exercise: Adult Data Set

- Play with the dataset, trying to find a set of about 10 interesting rules:
    1. Changing the pre-processing
    2. Removing or keeping the NAs (inputs and response)
    3. Changing support and confidence
    4. Focus on specific RHSs
- Deliver a **two-page** (max.) pdf with what you finally did and what you got