Association rules and market basket analysis
What are association rules (AR) and what are they used for:
- The paradigmatic application: Market Basket Analysis
- The single dimensional AR (intra-attribute)

How to compute AR
- Basic Apriori Algorithm and its optimizations
- Multi-Dimension AR (inter-attribute)
- Quantitative AR
Market Basket Analysis: the context

Customer buying habits by finding associations and correlations between the different items that customers place in their "shopping basket"
Market Basket Analysis: the context

Given: a database of customer transactions, where each transaction is a set of items

- Find groups of items which are frequently purchased together
Goal of MBA

- Extract information on purchasing behavior
- Actionable information: can suggest
  - new store layouts
  - new product assortments
  - which products to put on promotion
- MBA applicable whenever a customer purchases multiple things in proximity
  - credit cards
  - services of telecommunication companies
  - banking services
  - medical treatments
MBA: applicable to many other contexts

Telecommunication:
Each customer is a transaction containing the set of customer’s phone calls

Atmospheric phenomena:
Each time interval (e.g. a day) is a transaction containing the set of observed event (rains, wind, etc.)

Etc.
Association Rules

- Express how product/services relate to each other, and tend to group together
- “if a customer purchases three-way calling, then will also purchase call-waiting”
- Simple to understand
- Actionable information: bundle three-way calling and call-waiting in a single package

Examples.
- Rule form: “Body → Head [support, confidence]”.
- buys(x, “diapers”) → buys(x, “beers”) [0.5%, 60%]
- major(x, “CS”) and takes(x, “DB”) → grade(x, “A”) [1%, 75%]
Useful, trivial, unexplicable

**Useful:** “On Thursdays, grocery store consumers often purchase diapers and beer together”.

**Trivial:** “Customers who purchase maintenance agreements are very likely to purchase large appliances”.

**Unexplicable:** “When a new hardware store opens, one of the most sold items is toilet rings.”
Association Rules Road Map

- **Single dimension vs. multiple dimensional AR**
  - E.g., association on items bought vs. linking on different attributes.
  - Intra-Attribute vs. Inter-Attribute

- **Qualitative vs. quantitative AR**
  - Association on categorical vs. numerical attributes

- **Simple vs. constraint-based AR**
  - E.g., small sales (sum < 100) trigger big buys (sum > 1,000)?

- **Single level vs. multiple-level AR**
  - E.g., what brands of beers are associated with what brands of diapers?

- **Association vs. correlation analysis**
  - Association does not necessarily imply correlation.
Association Rule Mining

Given a set of transactions, find rules that will predict the occurrence of an item based on the occurrences of other items in the transaction.

Market-Basket transactions

<table>
<thead>
<tr>
<th>TID</th>
<th>Items</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Bread, Milk</td>
</tr>
<tr>
<td>2</td>
<td>Bread, Diaper, Beer, Eggs</td>
</tr>
<tr>
<td>3</td>
<td>Milk, Diaper, Beer, Coke</td>
</tr>
<tr>
<td>4</td>
<td>Bread, Milk, Diaper, Beer</td>
</tr>
<tr>
<td>5</td>
<td>Bread, Milk, Diaper, Coke</td>
</tr>
</tbody>
</table>

Example of Association Rules

\{\text{Diaper}\} \rightarrow \{\text{Beer}\}, \\
\{\text{Milk}, \text{Bread}\} \rightarrow \{\text{Eggs}, \text{Coke}\}, \\
\{\text{Beer}, \text{Bread}\} \rightarrow \{\text{Milk}\},

Implication means co-occurrence, not causality!
### Basic Concepts: Frequent Patterns and Association Rules

- **Itemset** $X = \{x_1, \ldots, x_k\}$
- Find all the rules $X \rightarrow Y$ with minimum support and confidence
  - **support**, $s$, probability that a transaction contains $X \cup Y$
  - **confidence**, $c$, conditional probability that a transaction having $X$ also contains $Y$

<table>
<thead>
<tr>
<th>Transaction-id</th>
<th>Items bought</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>A, B, D</td>
</tr>
<tr>
<td>20</td>
<td>A, C, D</td>
</tr>
<tr>
<td>30</td>
<td>A, D, E</td>
</tr>
<tr>
<td>40</td>
<td>B, E, F</td>
</tr>
<tr>
<td>50</td>
<td>B, C, D, E, F</td>
</tr>
</tbody>
</table>

Let $sup_{\text{min}} = 50\%$, $conf_{\text{min}} = 50\%$

Freq. Pat.: $\{A:3, B:3, D:4, E:3, AD:3\}$

Association rules:
- $A \rightarrow D$ (60%, 100%)
- $D \rightarrow A$ (60%, 75%)
Definition: Frequent Itemset

- **Itemset**
  - A collection of one or more items
  - ✓ Example: \{Milk, Bread, Diaper\}

- **k-itemset**
  - ✓ An itemset that contains \( k \) items

- **Support count** (\( \sigma \))
  - Frequency of occurrence of an itemset
  - E.g. \( \sigma(\{\text{Milk, Bread, Diaper}\}) = 2 \)

- **Support**
  - Fraction of transactions that contain an itemset
  - E.g. \( s(\{\text{Milk, Bread, Diaper}\}) = \frac{2}{5} \)

- **Frequent Itemset**
  - An itemset whose support is greater than or equal to a \textit{minsup} threshold

<table>
<thead>
<tr>
<th>TID</th>
<th>Items</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Bread, Milk</td>
</tr>
<tr>
<td>2</td>
<td>Bread, Diaper, Beer, Eggs</td>
</tr>
<tr>
<td>3</td>
<td>Milk, Diaper, Beer, Coke</td>
</tr>
<tr>
<td>4</td>
<td>Bread, Milk, Diaper, Beer</td>
</tr>
<tr>
<td>5</td>
<td>Bread, Milk, Diaper, Coke</td>
</tr>
</tbody>
</table>
Definition: Association Rule

Association Rule
- An implication expression of the form $X \rightarrow Y$, where $X$ and $Y$ are itemsets
- Example: 
  $\{\text{Milk, Diaper}\} \rightarrow \{\text{Beer}\}$

Rule Evaluation Metrics
- **Support ($s$)**
  - Fraction of transactions that contain both $X$ and $Y$
- **Confidence ($c$)**
  - Measures how often items in $Y$ appear in transactions that contain $X$

Example:
$\{\text{Milk, Diaper}\} \Rightarrow \text{Beer}$

<table>
<thead>
<tr>
<th>TID</th>
<th>Items</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Bread, Milk</td>
</tr>
<tr>
<td>2</td>
<td>Bread, Diaper, Beer, Eggs</td>
</tr>
<tr>
<td>3</td>
<td>Milk, Diaper, Beer, Coke</td>
</tr>
<tr>
<td>4</td>
<td>Bread, Milk, Diaper, Beer</td>
</tr>
<tr>
<td>5</td>
<td>Bread, Milk, Diaper, Coke</td>
</tr>
</tbody>
</table>

$$s = \frac{\sigma(\text{Milk, Diaper, Beer})}{|T|} = \frac{2}{5} = 0.4$$

$$c = \frac{\sigma(\text{Milk, Diaper, Beer})}{\sigma(\text{Milk, Diaper})} = \frac{2}{3} = 0.67$$
Association rules - module outline

- What are association rules (AR) and what are they used for:
  - The paradigmatic application: Market Basket Analysis
  - The single dimensional AR (intra-attribute)

- How to compute AR
  - Basic Apriori Algorithm and its optimizations
  - Multi-Dimension AR (inter-attribute)
  - Quantitative AR
Given a set of transactions T, the goal of association rule mining is to find all rules having

- support ≥ minsup threshold
- confidence ≥ minconf threshold

**Brute-force approach:**

- List all possible association rules
- Compute the support and confidence for each rule
- Prune rules that fail the minsup and minconf thresholds

⇒ **Computationally prohibitive!**
Mining Association Rules

Example of Rules:

\[
\begin{align*}
\{\text{Milk,Diaper}\} & \rightarrow \{\text{Beer}\} \quad (s=0.4, \ c=0.67) \\
\{\text{Milk,Beer}\} & \rightarrow \{\text{Diaper}\} \quad (s=0.4, \ c=1.0) \\
\{\text{Diaper,Beer}\} & \rightarrow \{\text{Milk}\} \quad (s=0.4, \ c=0.67) \\
\{\text{Diaper}\} & \rightarrow \{\text{Milk,Beer}\} \quad (s=0.4, \ c=0.5) \\
\{\text{Milk}\} & \rightarrow \{\text{Diaper,Beer}\} \quad (s=0.4, \ c=0.5)
\end{align*}
\]

Observations:

- All the above rules are binary partitions of the same itemset: \{\text{Milk, Diaper, Beer}\}
- Rules originating from the same itemset have identical support but can have different confidence
- Thus, we may decouple the support and confidence requirements
Mining Association Rules

Two-step approach:

1. Frequent Itemset Generation
   - Generate all itemsets whose support $\geq \minsup$

2. Rule Generation
   - Generate high confidence rules from each frequent itemset, where each rule is a binary partitioning of a frequent itemset

Frequent itemset generation is still computationally expensive
Basic Apriori Algorithm

Problem Decomposition

1. **Find the frequent itemsets:** the sets of items that satisfy the support constraint
   - A subset of a frequent itemset is also a frequent itemset, i.e., if \{A, B\} is a frequent itemset, both \{A\} and \{B\} should be a frequent itemset
   - Iteratively find frequent itemsets with cardinality from 1 to \(k\) (\(k\)-itemset)

2. **Use the frequent itemsets to generate association rules.**
Given $d$ items, there are $2^d$ possible candidate itemsets.
Frequent Itemset Generation

- Brute-force approach:
  - Each itemset in the lattice is a candidate frequent itemset
  - Count the support of each candidate by scanning the database

```
TID | Items
---|------
1   | Bread, Milk
2   | Bread, Diaper, Beer, Eggs
3   | Milk, Diaper, Beer, Coke
4   | Bread, Milk, Diaper, Beer
5   | Bread, Milk, Diaper, Coke
```

- Match each transaction against every candidate
- Complexity $\sim O(NMw) \Rightarrow$ Expensive since $M = 2^d$ !!!
Frequent Itemset Generation Strategies

- **Reduce the number of candidates (M)**
  - Complete search: $M = 2^d$
  - Use pruning techniques to reduce $M$

- **Reduce the number of transactions (N)**
  - Reduce size of $N$ as the size of itemset increases
  - Used by DHP and vertical-based mining algorithms

- **Reduce the number of comparisons (NM)**
  - Use efficient data structures to store the candidates or transactions
  - No need to match every candidate against every transaction
Reducing Number of Candidates

Apriori principle:
- If an itemset is frequent, then all of its subsets must also be frequent

Apriori principle holds due to the following property of the support measure:

\[ \forall X, Y : (X \subseteq Y) \Rightarrow s(X) \geq s(Y) \]

- Support of an itemset never exceeds the support of its subsets
- This is known as the anti-monotone property of support
Illustrating Apriori Principle

Found to be Infrequent

Pruned supersets
Apriori Execution Example \((min\_sup = 2)\)

**Database TDB**

<table>
<thead>
<tr>
<th>TID</th>
<th>Items</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>1 3 4</td>
</tr>
<tr>
<td>200</td>
<td>2 3 5</td>
</tr>
<tr>
<td>300</td>
<td>1 2 3 5</td>
</tr>
<tr>
<td>400</td>
<td>2 5</td>
</tr>
</tbody>
</table>

**Scan TDB**

**C₁**

<table>
<thead>
<tr>
<th>itemset</th>
<th>sup.</th>
</tr>
</thead>
<tbody>
<tr>
<td>{1}</td>
<td>2</td>
</tr>
<tr>
<td>{2}</td>
<td>3</td>
</tr>
<tr>
<td>{3}</td>
<td>3</td>
</tr>
<tr>
<td>{4}</td>
<td>1</td>
</tr>
<tr>
<td>{5}</td>
<td>3</td>
</tr>
</tbody>
</table>

**L₁**

<table>
<thead>
<tr>
<th>itemset</th>
<th>sup.</th>
</tr>
</thead>
<tbody>
<tr>
<td>{1}</td>
<td>2</td>
</tr>
<tr>
<td>{2}</td>
<td>3</td>
</tr>
<tr>
<td>{3}</td>
<td>3</td>
</tr>
<tr>
<td>{5}</td>
<td>3</td>
</tr>
</tbody>
</table>

**C₂**

<table>
<thead>
<tr>
<th>itemset</th>
<th>sup.</th>
</tr>
</thead>
<tbody>
<tr>
<td>{1 2}</td>
<td>1</td>
</tr>
<tr>
<td>{1 3}</td>
<td>2</td>
</tr>
<tr>
<td>{1 5}</td>
<td>1</td>
</tr>
<tr>
<td>{2 3}</td>
<td>2</td>
</tr>
<tr>
<td>{2 5}</td>
<td>3</td>
</tr>
<tr>
<td>{3 5}</td>
<td>2</td>
</tr>
</tbody>
</table>

**L₂**

**C₃**

<table>
<thead>
<tr>
<th>itemset</th>
<th>sup.</th>
</tr>
</thead>
<tbody>
<tr>
<td>{2 3 5}</td>
<td>2</td>
</tr>
</tbody>
</table>

**C₃**

**Scan TDB**

Giannotti & Pedreschi
The Apriori Algorithm

**Join Step:** $C_k$ is generated by joining $L_{k-1}$ with itself

**Prune Step:** Any $(k-1)$-itemset that is not frequent cannot be a subset of a frequent $k$-itemset

**Pseudo-code:**

- $C_k$: Candidate itemset of size $k$
- $L_k$: frequent itemset of size $k$

$L_1 = \{\text{frequent items}\}$

\[
\text{for } (k = 1; L_k \neq \emptyset; k++) \text{ do begin} \\
\quad C_{k+1} = \text{candidates generated from } L_k; \\
\quad \text{for each transaction } t \text{ in database do} \\
\quad \quad \text{increment the count of all candidates in } C_{k+1} \text{ that are contained in } t \\
\quad L_{k+1} = \text{candidates in } C_{k+1} \text{ with min-support} \\
\text{end} \\
\text{return } \bigcup_k L_k.
\]
Example of Generating Candidates

- $L_3 = \{abc, abd, acd, ace, bcd\}$

- Self-joining: $L_3 \times L_3$
  - $abcd$ from $abc$ and $abd$
  - $acde$ from $acd$ and $ace$

- Pruning:
  - $acde$ is removed because $ade$ is not in $L_3$

- $C_4 = \{abcd\}$
Factors Affecting Complexity

- **Choice of minimum support threshold**
  - lowering support threshold results in more frequent itemsets
  - this may increase number of candidates and max length of frequent itemsets

- **Dimensionality (number of items) of the data set**
  - more space is needed to store support count of each item
  - if number of frequent items also increases, both computation and I/O costs may also increase

- **Size of database**
  - since Apriori makes multiple passes, run time of algorithm may increase with number of transactions

- **Average transaction width**
  - transaction width increases with denser data sets
  - This may increase max length of frequent itemsets and traversals of hash tree (number of subsets in a transaction increases with its width)
The KDD process

Data Mining

Selection and Preprocessing

Data Consolidation

Interpretation and Evaluation

Knowledge

\[ p(x) \leq 0.02 \]
Generating Association Rules from Frequent Itemsets

- Only strong association rules are generated
- Frequent itemsets satisfy minimum support threshold
- Strong rules are those that satisfy minimum confidence threshold

\[
\text{confidence}(A \implies B) = \Pr(B \mid A) = \frac{\text{support}(A \cup B)}{\text{support}(A)}
\]

For each frequent itemset, \( f \), generate all non-empty subsets of \( f \)
For every non-empty subset \( s \) of \( f \) do
  if \( \frac{\text{support}(f)}{\text{support}(s)} \geq \text{min\_confidence} \) then
    output rule \( s \implies (f-s) \)
end
Computational Complexity

- Given $d$ unique items:
  - Total number of itemsets = $2^d$
  - Total number of possible association rules:

$$R = \sum_{k=1}^{d-1} \binom{d}{k} \times \sum_{j=1}^{d-k} \binom{d-k}{j}$$

$$= 3^d - 2^{d+1} + 1$$

If $d=6$, $R = 602$ rules
Rule Generation

Given a frequent itemset $L$, find all non-empty subsets $f \subset L$ such that $f \rightarrow L - f$ satisfies the minimum confidence requirement

If \{A,B,C,D\} is a frequent itemset, candidate rules:

- $ABC \rightarrow D,$
- $ABD \rightarrow C,$
- $ACD \rightarrow B,$
- $BCD \rightarrow A,$
- $A \rightarrow BCD,$
- $B \rightarrow ACD,$
- $C \rightarrow ABD,$
- $D \rightarrow ABC,$
- $AB \rightarrow CD,$
- $AC \rightarrow BD,$
- $AD \rightarrow BC,$
- $BC \rightarrow AD,$
- $BD \rightarrow AC,$
- $CD \rightarrow AB,$

If $|L| = k$, then there are $2^k - 2$ candidate association rules (ignoring $L \rightarrow \emptyset$ and $\emptyset \rightarrow L$)
Rule Generation

- How to efficiently generate rules from frequent itemsets?
  - In general, confidence does not have an anti-monotone property
    \[ c(ABC \rightarrow D) \text{ can be larger or smaller than } c(AB \rightarrow D) \]
  - But confidence of rules generated from the same itemset has an anti-monotone property
  - e.g., \( L = \{A, B, C, D\} \):
    \[
    c(ABC \rightarrow D) \geq c(AB \rightarrow CD) \geq c(A \rightarrow BCD)
    \]
    ✓ Confidence is anti-monotone w.r.t. number of items on the RHS of the rule

Reg. Ass.
Rule Generation for Apriori Algorithm

- Candidate rule is generated by merging two rules that share the same prefix in the rule consequent.

- \( \text{join}(CD \Rightarrow AB, BD \Rightarrow AC) \) would produce the candidate rule \( D \Rightarrow ABC \).

- Prune rule \( D \Rightarrow ABC \) if its subset \( AD \Rightarrow BC \) does not have high confidence.

Reg. Ass.
How to choose intervals?

1. Interval with a fixed “reasonable” granularity
   Ex. intervals of 10 cm for height.

2. Interval size is defined by some domain dependent criterion
   Ex.: 0-20ML, 21-22ML, 23-24ML, 25-26ML, >26ML

3. Interval size determined by analyzing data, studying the distribution or using clustering

![Weight distribution graph]

- 50 - 58 kg
- 59-67 kg
- > 68 kg
Discretization of quantitative attributes

1. Quantitative attributes are **statically** discretized by using predefined concept hierarchies:
   - elementary use of background knowledge

   **Loose interaction between Apriori and discretizer**

2. Quantitative attributes are **dynamically** discretized
   - into “bins” based on the distribution of the data.
   - considering the distance between data points.

   **Tighter interaction between Apriori and discretizer**
Reasoning with AR

**Significance:**

Example:  
\[ <1, \{a, b\}> \]
\[ <2, \{a\}> \]
\[ <3, \{a, b, c\}> \]
\[ <4, \{b, d\}> \]

\([b] \Rightarrow [a]\) has confidence (66%), but is not significant as \(support([a]) = 75\%\).
Beyond Support and Confidence

Example 1: (Aggarwal & Yu, PODS98)

<table>
<thead>
<tr>
<th></th>
<th>coffee</th>
<th>not coffee</th>
<th>sum(row)</th>
</tr>
</thead>
<tbody>
<tr>
<td>tea</td>
<td>20</td>
<td>5</td>
<td>25</td>
</tr>
<tr>
<td>not tea</td>
<td>70</td>
<td>5</td>
<td>75</td>
</tr>
<tr>
<td>sum(col.)</td>
<td>90</td>
<td>10</td>
<td>100</td>
</tr>
</tbody>
</table>

\{tea\} => \{coffee\} has high support (20%) and confidence (80%)

However, a priori probability that a customer buys coffee is 90%

- A customer who is known to buy tea is less likely to buy coffee (by 10%)
- There is a negative correlation between buying tea and buying coffee
- \{~tea\} => \{coffee\} has higher confidence (93%)
Correlation and Interest

- Two events are independent if $P(A \wedge B) = P(A) \times P(B)$, otherwise are correlated.

- Interest = $P(A \wedge B) / P(B) \times P(A)$

- Interest expresses measure of correlation
  
  - $= 1$ $\Rightarrow$ A and B are independent events
  
  - less than 1 $\Rightarrow$ A and B negatively correlated,
  
  - greater than 1 $\Rightarrow$ A and B positively correlated.

- In our example, $I(\text{buy tea } \wedge \text{buy coffee }) = 0.89$ i.e. they are negatively correlated.
Computing Interestingness Measure

Given a rule $X \rightarrow Y$, information needed to compute rule interestingness can be obtained from a contingency table.

Contingency table for $X \rightarrow Y$

<table>
<thead>
<tr>
<th></th>
<th>$Y$</th>
<th>$\bar{Y}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X$</td>
<td>$f_{11}$</td>
<td>$f_{10}$</td>
</tr>
<tr>
<td>$\bar{X}$</td>
<td>$f_{01}$</td>
<td>$f_{00}$</td>
</tr>
<tr>
<td>$+$</td>
<td>$f_{+1}$</td>
<td>$f_{+0}$</td>
</tr>
</tbody>
</table>

- $f_{11}$: support of $X$ and $Y$
- $f_{10}$: support of $X$ and $\bar{Y}$
- $f_{01}$: support of $\bar{X}$ and $Y$
- $f_{00}$: support of $\bar{X}$ and $\bar{Y}$

Used to define various measures:

- support, confidence, lift, Gini, J-measure, etc.
Statistical-based Measures

Measures that take into account statistical dependence

\[
\text{Lift} = \frac{P(Y | X)}{P(Y)}
\]

\[
\text{Interest} = \frac{P(X, Y)}{P(X)P(Y)}
\]

\[
PS = P(X, Y) - P(X)P(Y)
\]

\[
\phi - \text{coefficient} = \frac{P(X, Y) - P(X)P(Y)}{\sqrt{P(X)[1 - P(X)]P(Y)[1 - P(Y)]}}
\]
Example: Lift/Interest

<table>
<thead>
<tr>
<th></th>
<th>Coffee</th>
<th>Coffee</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tea</td>
<td>15</td>
<td>5</td>
</tr>
<tr>
<td>Tea</td>
<td>75</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>90</td>
<td>10</td>
</tr>
</tbody>
</table>

Association Rule: Tea $\rightarrow$ Coffee

Confidence = $P(\text{Coffee}|\text{Tea}) = 0.75$

but $P(\text{Coffee}) = 0.9$

$\Rightarrow$ Lift = $0.75/0.9 = 0.8333$ (< 1, therefore is negatively associated)
Drawback of Lift & Interest

<table>
<thead>
<tr>
<th></th>
<th>Y</th>
<th>\bar{Y}</th>
</tr>
</thead>
<tbody>
<tr>
<td>X</td>
<td>10</td>
<td>0</td>
</tr>
<tr>
<td>\bar{X}</td>
<td>0</td>
<td>90</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>90</td>
</tr>
</tbody>
</table>

\[
Lift = \frac{0.1}{(0.1)(0.1)} = 10
\]

<table>
<thead>
<tr>
<th></th>
<th>Y</th>
<th>\bar{Y}</th>
</tr>
</thead>
<tbody>
<tr>
<td>X</td>
<td>90</td>
<td>0</td>
</tr>
<tr>
<td>\bar{X}</td>
<td>0</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>90</td>
<td>10</td>
</tr>
</tbody>
</table>

\[
Lift = \frac{0.9}{(0.9)(0.9)} = 1.11
\]

Statistical independence:
If \( P(X,Y) = P(X)P(Y) \) => Lift = 1
Association rules - module outline

What are association rules (AR) and what are they used for:
- The paradigmatic application: Market Basket Analysis
- The single dimensional AR (intra-attribute)

How to compute AR
- Basic Apriori Algorithm and its optimizations
- Multi-Dimension AR (inter-attribute)
- Quantitative AR
## Multidimensional AR

Associations between values of different attributes:

<table>
<thead>
<tr>
<th>CID</th>
<th>nationality</th>
<th>age</th>
<th>income</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Italian</td>
<td>50</td>
<td>low</td>
</tr>
<tr>
<td>2</td>
<td>French</td>
<td>40</td>
<td>high</td>
</tr>
<tr>
<td>3</td>
<td>French</td>
<td>30</td>
<td>high</td>
</tr>
<tr>
<td>4</td>
<td>Italian</td>
<td>50</td>
<td>medium</td>
</tr>
<tr>
<td>5</td>
<td>Italian</td>
<td>45</td>
<td>high</td>
</tr>
<tr>
<td>6</td>
<td>French</td>
<td>35</td>
<td>high</td>
</tr>
</tbody>
</table>

**RULES:**

- **nationality** = French $\Rightarrow$ **income** = high $[50\%, 100\%]$  
- **income** = high $\Rightarrow$ **nationality** = French $[50\%, 75\%]$  
- **age** = 50 $\Rightarrow$ **nationality** = Italian $[33\%, 100\%]$
### Single-dimensional vs Multi-dimensional AR

#### Multi-dimensional

1. <1, Italian, 50, low>
2. <2, French, 45, high>

Schema: <ID, a?, b?, c?, d?>

1. <1, yes, yes, no, no>
2. <2, yes, no, yes, no>

#### Single-dimensional

1. <1, {nat/Ita, age/50, inc/low}>
2. <2, {nat/Fre, age/45, inc/high}>

1. <1, {a, b}>
2. <2, {a, c}>

Giannotti & Pedreschi
Quantitative Attributes

- Quantitative attributes (e.g. age, income)
- Categorical attributes (e.g. color of car)

<table>
<thead>
<tr>
<th>CID</th>
<th>height</th>
<th>weight</th>
<th>income</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>168</td>
<td>75,4</td>
<td>30,5</td>
</tr>
<tr>
<td>2</td>
<td>175</td>
<td>80,0</td>
<td>20,3</td>
</tr>
<tr>
<td>3</td>
<td>174</td>
<td>70,3</td>
<td>25,8</td>
</tr>
<tr>
<td>4</td>
<td>170</td>
<td>65,2</td>
<td>27,0</td>
</tr>
</tbody>
</table>

**Problem:** too many distinct values

**Solution:** transform quantitative attributes in categorical ones via discretization.
Quantitative Association Rules

<table>
<thead>
<tr>
<th>CID</th>
<th>Age</th>
<th>Married</th>
<th>NumCars</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>23</td>
<td>No</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>25</td>
<td>Yes</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>29</td>
<td>No</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>34</td>
<td>Yes</td>
<td>2</td>
</tr>
<tr>
<td>5</td>
<td>38</td>
<td>Yes</td>
<td>2</td>
</tr>
</tbody>
</table>

[Age: 30..39] and [Married: Yes] ⇒ [NumCars: 2]

support = 40%
confidence = 100%
Discretization of quantitative attributes

Solution: each value is replaced by the interval to which it belongs.

height: 0-150cm, 151-170cm, 171-180cm, >180cm
weight: 0-40kg, 41-60kg, 60-80kg, >80kg
income: 0-10ML, 11-20ML, 20-25ML, 25-30ML, >30ML

<table>
<thead>
<tr>
<th>CID</th>
<th>height</th>
<th>weight</th>
<th>income</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>151-171</td>
<td>60-80</td>
<td>&gt;30</td>
</tr>
<tr>
<td>2</td>
<td>171-180</td>
<td>60-80</td>
<td>20-25</td>
</tr>
<tr>
<td>3</td>
<td>171-180</td>
<td>60-80</td>
<td>25-30</td>
</tr>
<tr>
<td>4</td>
<td>151-170</td>
<td>60-80</td>
<td>25-30</td>
</tr>
</tbody>
</table>

Problem: the discretization may be useless (see weight).