Data Mining

Part 3. Associations Rules

3.1 Introduction

Spring 2010

Instructor: Dr. Masoud Yaghini

Outline

- Basic Definitions
- Market Basket Analysis
- Weather Problem
- Frequent Itemsets and Association Rules
- Kinds of Frequent Pattern Mining
- References

Basic Definitions

Frequent Patterns

- **Frequent patterns:**
  - patterns that occur frequently in data
- The kinds of frequent patterns:
  - **A frequent itemset pattern**: a set of items that frequently appear together in a transactional data set, such as milk and bread.
  - **A frequent sequential pattern**: such as the pattern that customers tend to purchase first a PC, followed by a digital camera, and then a memory card, is a frequent sequential pattern.

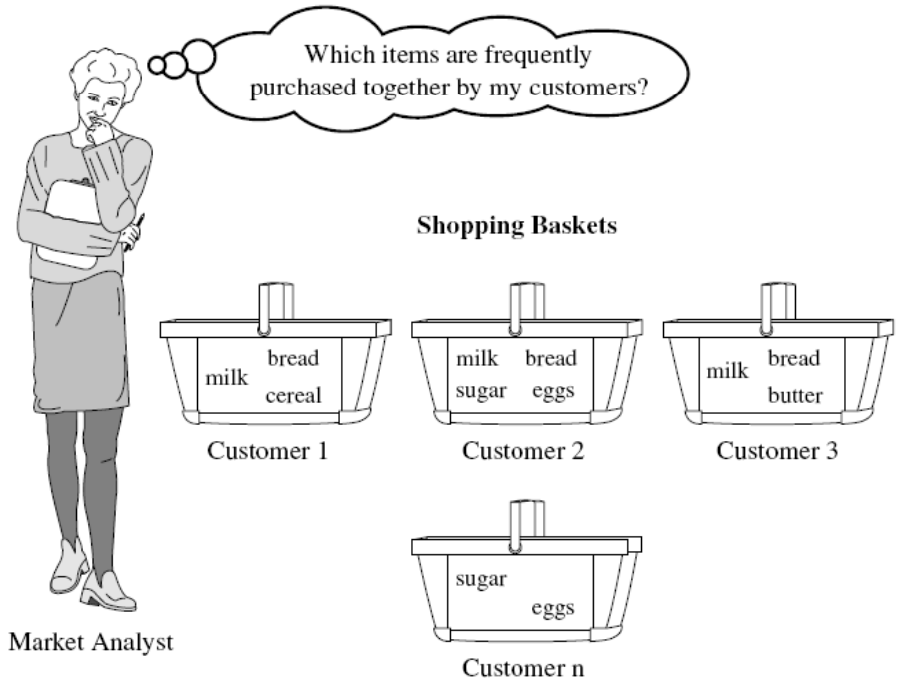
Frequent Patterns

- Mining frequent patterns leads to the discovery of interesting **associations** and **correlations** within data.
- Applications:
  - Basket data analysis,
  - Cross-marketing,
  - Catalog design,
  - Sale campaign analysis
  - Web log (click stream) analysis,
  - DNA sequence analysis,
  - etc.

Market Basket Analysis

Market Basket Analysis

- **Market basket analysis**
  - A typical example of **frequent itemset mining**
- Finding associations between the different items that customers place in their "shopping baskets"



Market Basket Analysis

- The discovery of such associations can help retailers
  - Develop marketing strategies
    - If customers tend to purchase computers and printers together, then having a sale on printers may encourage the sale of printers as well as computers.
  - Design different store layouts
    - If customers who purchase computers also tend to buy antivirus software at the same time, then placing the hardware display close to the software display may help increase the sales of both items.

Market Basket Analysis

- Market basket data can be represented in a binary format.

TID	Bread	Milk	Diapers	Beer	Eggs	Cola
1	1	1	0	0	0	0
2	1	0	1	1	1	0
3	0	1	1	1	0	1
4	1	1	1	1	0	0
5	1	1	1	0	0	1

- The Boolean vectors can be analyzed for buying patterns that reflect items that are frequently associated or purchased together.
- These patterns can be represented in the form of **association rules**.

Example

- The information that customers who purchase computers also tend to buy antivirus software at the same time is represented in **Association Rule**

$$computer \Rightarrow antivirus\_software \;\; [support = 2\%, confidence = 60\%]$$

- Measures of rule interestingness:
  - A support of 2% means that 2% of all the transactions under analysis show that computer and antivirus software are purchased together.
  - A confidence of 60% means that 60% of the customers who purchased a computer also bought the antivirus software.

Support & Confidence Threshold

- Typically, association rules are considered interesting if they satisfy both
  - a minimum support threshold
  - a minimum confidence threshold
- Such thresholds can be set by users or domain experts.
- Additional analysis can be performed to uncover interesting statistical correlations between associated items.

Weather Problem

Weather data

Outlook	Temperature	Humidity	Windy	Play
sunny	hot	high	false	no
sunny	hot	high	true	no
overcast	hot	high	false	yes
rainy	mild	high	false	yes
rainy	cool	normal	false	yes
rainy	cool	normal	true	no
overcast	cool	normal	true	yes
sunny	mild	high	false	no
sunny	cool	normal	false	yes
rainy	mild	normal	false	yes
sunny	mild	normal	true	yes
overcast	mild	high	true	yes
overcast	hot	normal	false	yes
rainy	mild	high	true	no

Item sets for weather data

One-item sets	Two-item sets	Three-item sets	Four-item sets
outlook = sunny (5)	outlook = sunny temperature = mild (2)	outlook = sunny temperature = hot humidity = high (2)	outlook = sunny temperature = hot humidity = high play = no (2)
outlook = overcast (4)	outlook = sunny temperature = hot (2)	outlook = sunny temperature = hot play = no (2)	outlook = sunny humidity = high windy = false play = no (2)
outlook = rainy (5)	outlook = sunny humidity = normal (2)	outlook = sunny humidity = normal play = yes (2)	outlook = overcast temperature = hot windy = false play = yes (2)
.....	.....	.....	.....

- In total: 12 one-item sets, 47 two-item sets, 39 Three-item sets, 6 four-item sets and 0 five-item sets (with minimum support of 2)

Generating rules from an item set

- Once all item sets with minimum support have been generated, we can turn them into rules

```
humidity = normal, windy = false, play = yes
```

- Seven potential rules:

```
If humidity = normal and windy = false then play = yes          4/4
If humidity = normal and play = yes then windy = false          4/6
If windy = false and play = yes then humidity = normal          4/6
If humidity = normal then windy = false and play = yes          4/7
If windy = false then humidity = normal and play = yes          4/8
If play = yes then humidity = normal and windy = false          4/9
If - then humidity = normal and windy = false and play = yes    4/12
```

Rules for weather data

- Rules with support > 1 and confidence = 100%:

	Association rule			Coverage	Accuracy
1	humidity = normal windy = false	⇒	play = yes	4	100%
2	temperature = cool	⇒	humidity = normal	4	100%
3	outlook = overcast	⇒	play = yes	4	100%
4	temperature = cool play = yes	⇒	humidity = normal	3	100%
5	outlook = rainy windy = false	⇒	play = yes	3	100%
6	outlook = rainy play = yes	⇒	windy = false	3	100%
7	outlook = sunny humidity = high	⇒	play = no	3	100%
8	outlook = sunny play = no	⇒	humidity = high	3	100%
9	temperature = cool windy = false	⇒	humidity = normal play = yes	2	100%

- In total: 3 rules with support four, 5 with support three, 50 with support two

Example rules from the same set

- Item set:

```
temperature = cool, humidity = normal, windy = false, play = yes
```

- Resulting rules (all with 100% confidence):

temperature = cool windy = false	⇒	humidity = normal play = yes
temperature = cool humidity = normal windy = false	⇒	play = yes
temperature = cool windy = false play = yes	⇒	humidity = normal

- Three subsets of this item set also have coverage 2:

```
temperature = cool, windy = false
temperature = cool, humidity = normal, windy = false
temperature = cool, windy = false, play = yes
```

Generating rules efficiently

- We are looking for all high-confidence rules
  - But: rough method is (2^N-1)
- Better way: building $(c + 1)$ consequent rules from c consequent ones
  - Observation: $(c + 1)$ consequent rule can only hold if all corresponding c consequent rules also hold
- Resulting algorithm similar to procedure for large item sets

Example

- 1 consequent rules:

```
If humidity = high and windy = false and play = no
   then outlook = sunny
If outlook = sunny and windy = false and play = no
   then humidity = high
```

- Corresponding 2 consequent rule:

```
If windy = false and play = no then outlook = sunny
                                    and humidity = high
```

Frequent Itemsets and Association Rules

Frequent Itemsets

- Let
  - I = $\{I_1, I_2, \ldots, I_m\}$ *be a set of items.*
  - *D:* a set of database transactions
  - T : a transaction which is a set of items such that $T \subseteq I$.
  - TID: transaction identifier
  - *A : a set of items*
- A transaction T is said to contain A if and only if $A \subseteq T$.
- An association rule is the form:

$$A \Rightarrow B, \text{ where } A \subset I,\ B \subset I, \text{ and } A \cap B = \emptyset$$

Support

- **Support**
  - The rule $A \Rightarrow B$ holds in the transaction set D with **support *s***
  - where *s* is the percentage of transactions in D that contain $A \cup B$ (i.e., the union of sets A and B, or say, both A and B).
  - This is taken to be the probability, $P(A \mid B)$.

Confidence

- **Confidence**
  - The rule A ⇒ B has **confidence c** in the transaction set D,
  - where c is the percentage of transactions in D containing A that also contain B.
  - This is taken to be the conditional probability, P(B | A).
- That is,

$$\begin{aligned} support(A \Rightarrow B) &= P(A \cup B) \\ confidence(A \Rightarrow B) &= P(B|A). \end{aligned}$$

Association Rules

Transaction-id	Items bought
10	A, B, D
20	A, C, D
30	A, D, E
40	B, E, F
50	B, C, D, E, F

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

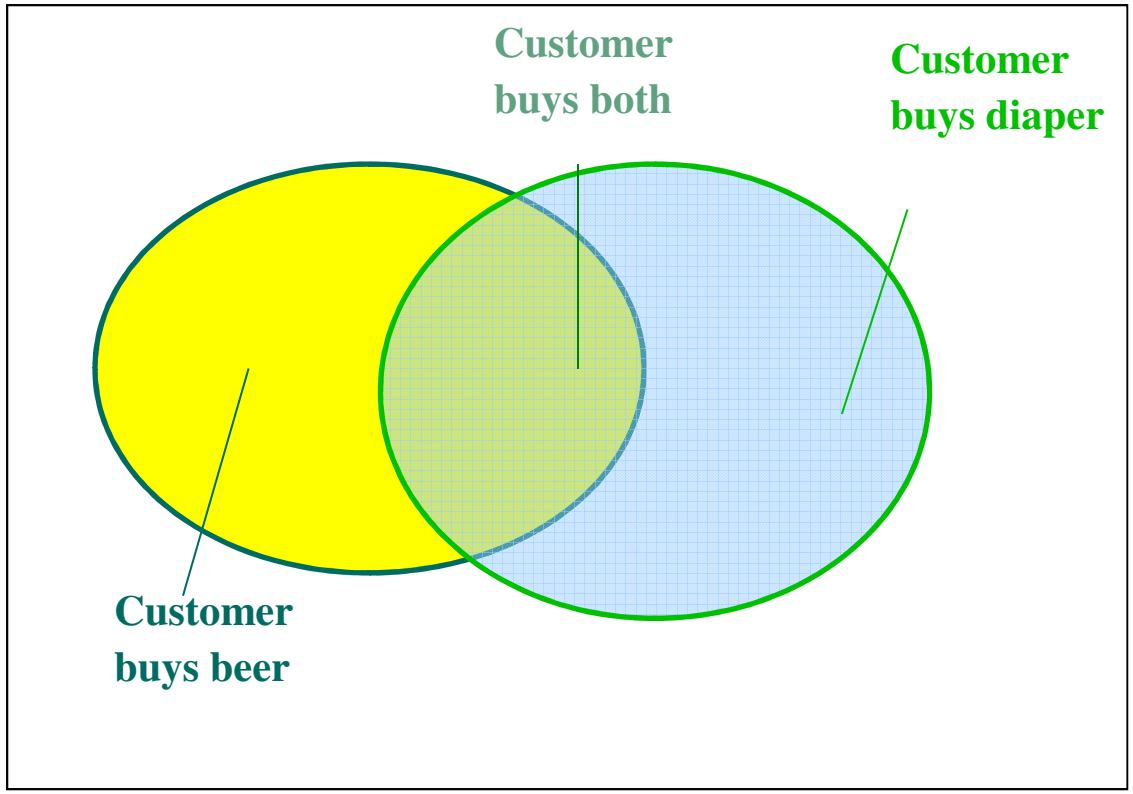
Freq. Pat.: {*A:3, B:3, D:4, E:3, AD:3*}

Association rules:

$A \rightarrow D$ (60%, 100%)

$D \rightarrow A$ (60%, 75%)

Association Rules



Minimum Support and Confidence

- Rules that satisfy both the following thresholds are called strong:
  - a minimum support threshold (min_sup)
  - a minimum confidence threshold (min_conf)
- By convention, we write support and confidence values so as to occur between 0% and 100%, rather than 0 to 1.0.

Definitions

- **Item**
  - one attribute-value (e.g. Milk, Bread)
- **Itemset**
  - A collection of one or more items
- **k-itemset**
  - An itemset that contains k items
  - The set {*computer*, *antivirus_software*} is a 2-itemset.
- **Support count (or frequency)**
  - The number of transactions that contain the itemset.

Frequent Itemsets

- **Frequent Itemset**
  - If the relative support of an itemset I satisfies a prespecified minimum support threshold then I is a frequent itemset

- The confidence of rule A⇒B can be easily derived from the support counts of A and A∪B.

$$confidence(A \Rightarrow B) = P(B|A) = \frac{support(A \cup B)}{support(A)} = \frac{support\_count(A \cup B)}{support\_count(A)}$$

Frequent Itemsets

- That is, once the support counts of A, B, and A∪B are found
  - it is straightforward to derive the corresponding association rules $A \Rightarrow B$ and $B \Rightarrow A$ and check whether they are strong.
- Mining association rules
  - Thus the problem of mining association rules can be reduced to that of mining frequent itemsets.

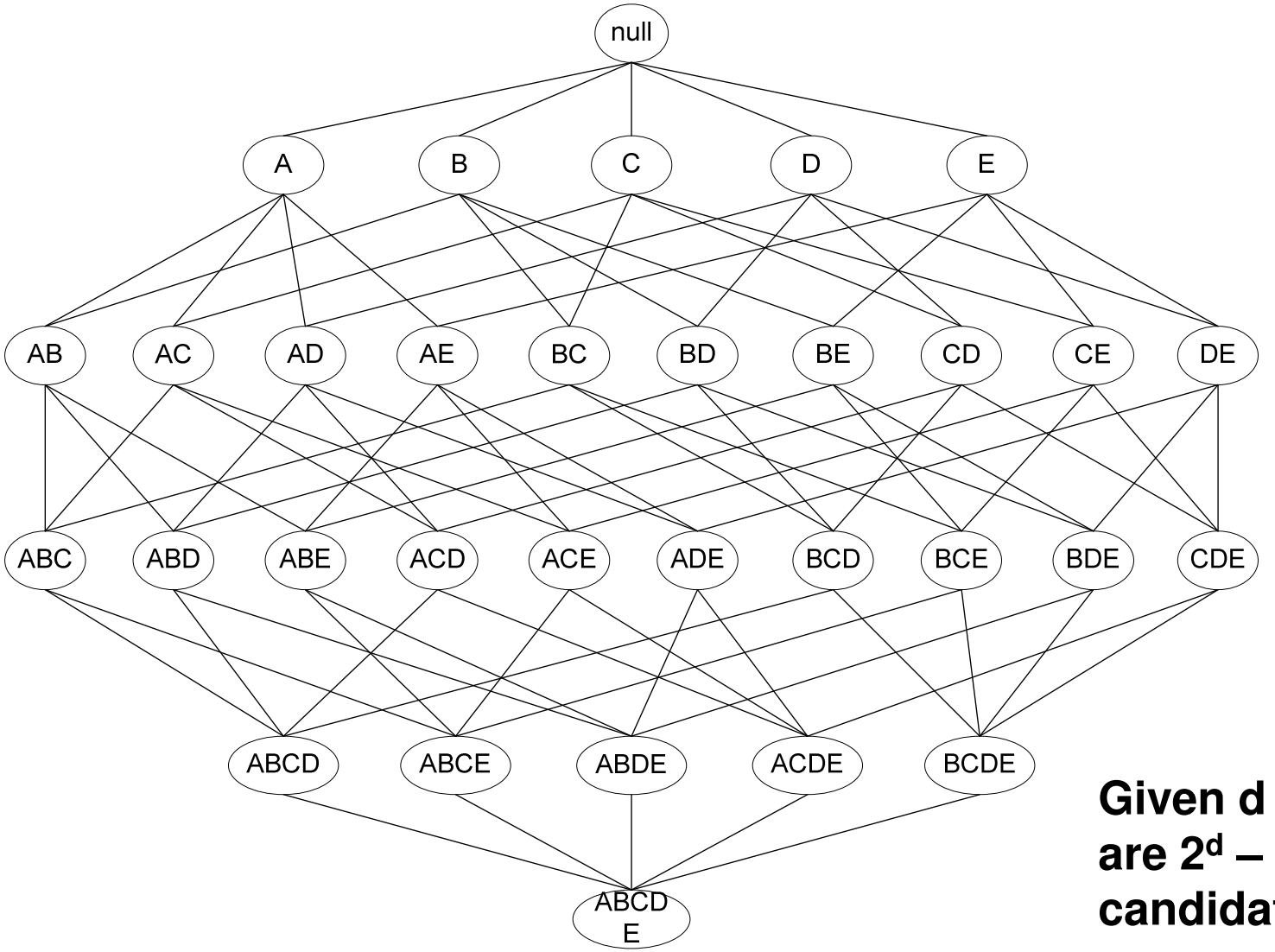
Mining Association Rules

- Two-step approach for mining association rules:
  1. Frequent Itemset Generation
     - Generate all itemsets whose support ≥ min_sup
  2. Rule Generation
     - Generate rules that satisfy minimum support and minimum confidence.
- Frequent itemset generation is computationally expensive
- The second step is much less costly

Frequent Itemset Generation

- Mining frequent itemsets often generates a huge number of itemsets satisfying the minimum support (*min_sup)* threshold, especially when *min_sup* is set low.
- This is because if an itemset is frequent, each of its subsets is frequent as well.

Frequent Itemset Generation



Given d items, there are $2^d - 1$ possible candidate itemsets

Long Pattern

- **Long pattern**
  - A long pattern will contain a combinatorial number of shorter, frequent sub-pattern.
- For example, a frequent itemset of length 100, such as $\{a_1, a_2, \ldots, a_{100}\}$, *contains*
  - $\binom{100}{1}$ frequent 1-itemsets: $a_1, a_2, \ldots, a_{100}$
  - $\binom{100}{2}$ frequent 2-itemsets: $(a_1, a_2), (a_1, a_3), \vdots, (a_{99}, a_{100})$, *and so on.*
  - The total number of frequent itemset:

$$\binom{100}{1} + \binom{100}{2} + \cdots + \binom{100}{100} = 2^{100} - 1 \approx 1.27 \times 10^{30}$$

Long Pattern

- This is too huge a number of itemsets for any computer to compute or store.
- Solution: Mine **closed patterns** and max-patterns instead

Closed and Maximal Frequent Itemsets

- **Closed itemset**
  - An itemset X is **closed** in a data set S if there exists no proper super-itemset Y, $X \subset Y$, such that Y has the **same support count** as X in S.
- **Closed frequent itemset**
  - An itemset X is a **closed frequent itemset** in set S if X is both **closed and frequent** in S.
- **Maximal frequent itemset**
  - An itemset X is a **maximal frequent itemset** (or **max-itemset**) in set S if X is frequent, and there exists no super-itemset Y such that $X \subset Y$ and Y is frequent in S.

Example

- Suppose that a transaction database has only two transactions:
  - $\{<a_1, a_2, \ldots, a_{100}>; <a_1, a_2, \ldots, a_{50}>\}$.
- Let the minimum support count threshold be *min_sup* = 1.
- We find two closed frequent itemsets and their support counts, that is,

  $C = \{ \{a_1, a_2, \ldots, a_{100}\} : 1; \{a_1, a_2, \ldots, a_{50}\} :2 \}$.
- There is one maximal frequent itemset:

  $M = \{\{a_1, a_2, \ldots, a_{100}\} : 1\}$.

Example

- For example, from C, we can derive, say,
  - (1) $\{a_2, a_{45} : 2\}$ since $\{a2, a45\}$ *is a sub-itemset of the itemset* $\{a_1, a_2, \ldots, a_{50} : 2\}$
  - (2) $\{a_8, a_{55} : 1\}$ *since* $\{a_8, a_{55}\}$ *is not a sub-itemset of the previous itemset but of the itemset* $\{a_1, a_2, \ldots, a_{100} : 1\}$.
- However, from the maximal frequent itemset, we can only assert that both itemsets ($\{a_2, a_{45}\}$ *and* $\{a_8, a_{55}\}$) are frequent, but we cannot assert their actual support counts.

Kinds of Frequent Pattern Mining

Kinds of Frequent Pattern Mining

- There are many kinds of frequent patterns, association rules, and correlation relationships.
- Frequent pattern mining can be classified in various ways, based on:
  - the completeness of patterns to be mined
  - the levels of abstraction involved in the rule set
  - the number of data dimensions involved in the rule
  - the types of values handled in the rule
  - the kinds of rules to be mined
  - the kinds of patterns to be mined

Kinds of Frequent Pattern Mining

- Based on the completeness of patterns to be mined:
  - Complete set of frequent itemsets
  - Closed frequent itemsets
  - Maximal frequent itemsets
  - Constrained frequent itemsets
    - those that satisfy a set of user-defined constraints
  - Top-k frequent itemsets
    - the k most frequent itemsets for a user-specified value

Kinds of Frequent Pattern Mining

- Based on the levels of abstraction involved in the rule set:
  - Single-level association rules
    - the rules within a given set do reference items or attributes at single level of abstraction, e.g.

$$buys(X, \text{“}computer\text{”}) \Rightarrow buys(X, \text{“}HP\_printer\text{”})$$

  - Multilevel association rules
    - methods for association rule mining can find rules at differing levels of abstraction, e.g.

$$buys(X, \text{“}computer\text{”}) \Rightarrow buys(X, \text{“}HP\_printer\text{”})$$

$$buys(X, \text{“}laptop\_computer\text{”}) \Rightarrow buys(X, \text{“}HP\_printer\text{”})$$

    - “computer” is a higher-level abstraction of “laptop computer”

Kinds of Frequent Pattern Mining

- Based on the number of data dimensions involved in the rule:
  - Single-dimensional association rule
    - If an association rule references only one dimension, e.g.

      $$buys(X, \text{“}computer\text{”}) \Rightarrow buys(X, \text{“}antivirus\_software\text{”})$$
    - refer to only one dimension, *buys*
  - Multidimensional association rule
    - If a rule references two or more dimensions, e.g.

$$age(X, \text{“}30\ldots39\text{”}) \wedge income(X, \text{“}42K\ldots48K\text{”}) \Rightarrow buys(X, \text{“}high\ resolution\ TV\text{”})$$

    - the dimensions are *age*, *income*, and *buys*

Kinds of Frequent Pattern Mining

- Based on the types of values handled in the rule
  - Boolean association rule
    - If a rule involves associations between the presence or absence of items. e.g.

$$buys(X, \text{“}computer\text{”}) \Rightarrow buys(X, \text{“}antivirus\_software\text{”})$$

  - Quantitative association rule
    - If a rule describes associations between quantitative items or attributes, e.g.

$$age(X, \text{“}30\ldots39\text{”}) \wedge income(X, \text{“}42K\ldots48K\text{”}) \Rightarrow buys(X, \text{“}high\ resolution\ TV\text{”})$$

Kinds of Frequent Pattern Mining

- Based on the kinds of rules to be mined:
  - Association rules
    - the most popular kind of rules generated from frequent patterns.
    - such mining can generate a large number of rules, many of which are redundant or do not indicate a correlation relationship among itemsets.
  - Correlation rules
    - the discovered associations can be further analyzed to uncover statistical correlations

Kinds of Frequent Pattern Mining

- Based on the kinds of patterns to be mined
  - Frequent itemset mining
    - the mining of frequent itemsets (sets of items) from transactional or relational data sets.
  - Sequential pattern mining
    - searches for frequent subsequences in a sequence data set, where a sequence records an ordering of events.
    - For example, we can study the order in which items are frequently purchased.
      - For instance, customers may tend to first buy a PC, followed by a digital camera, and then a memory card.

References

References

- J. Han, M. Kamber, **Data Mining: Concepts and Techniques**, Elsevier Inc. (2006). (Chapter 5)

The end