IN5526 - Web Intelligence

Lecture 2

Juan Domingo Velasquez Silva
Cristobal Gaspar Ignacio Pizarro Venegas

Departamento de Ingeniería Industrial
Universidad de Chile

October 15, 2016

Contents

Beer and diapers

Case large US supermarket

- Customer purchase behaviour:
- Product linked with another
  - Bread $\rightarrow$ butter,
  - Beer $\rightarrow$ diapers

Beer and diapers

Case large US supermarket

- Customer purchase behaviour:
- Product linked with another
  - Bread $\rightarrow$ butter,
  - Beer $\rightarrow$ diapers *wait, what?*
- Market segment
  - Young men married in the last three years with small children.

Based on this information, we deduce:

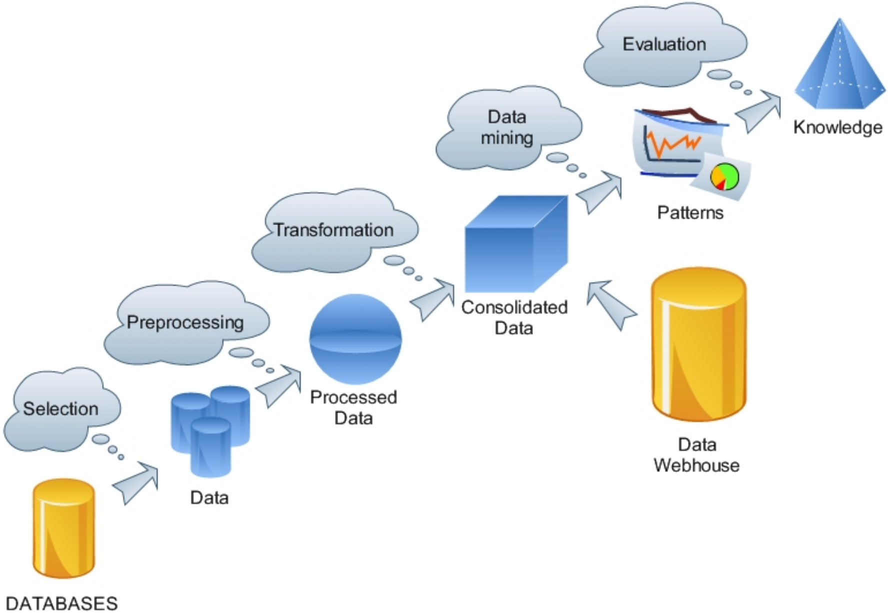
- Place diaper and beer on the same place on Friday afternoons.

A definition

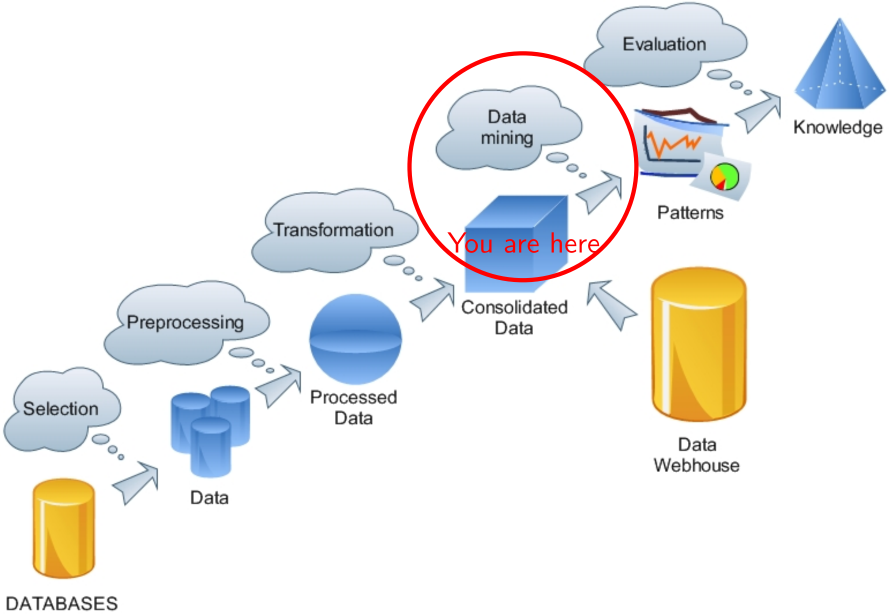
Data Mining

The non-trivial extraction of implicit, previously unknown and potentially useful information from data.

Data Mining in the KDD process



Data Mining in the KDD process



Common Data Mining tasks

Predictive mining

Predict attributes of unknown data based on attributes of known data.

Descriptive mining

Find human-readable structure in data.

Common Data Mining tasks

Classification

Generalizing known structure to apply to new data.

Regression

Attempts to find a function which models the data with the least error.

Clustering

The task of discovering groups and structures in the data.

Association rules

Searches for relationships between variables

Machine Learning

- One of the tools used in Data Mining
- Build models from data

Common Machine Learning tasks

Supervised learning

Build models to predict a variable/class using data for other variables/features. There is a "teacher" who tells what is the right class of any given example in the training set (direct feedback).

Unsupervised learning

Build models to describe a set of variables (or relations). Given a population of unclassified examples, invent reasonable concepts (clusters), and find definitions/meanings of those concepts. No teacher exists during training (no feedback).

Reinforcement learning

Indirect feedback after many examples, an agent that evolve according to its environment (Robotic Movement).

Machine learning

Example data

Can we play golf?

Day	**Outlook**	**Temperature**	**Humidity**	**Wind**	**PlayGolf**
1	Sunny	Hot	High	Weak	**No**
2	Sunny	Hot	High	Strong	**No**
3	Overcast	Hot	High	Weak	**Yes**
4	Rain	Cool	Normal	Weak	**Yes**
5	Sunny	Mild	Normal	Weak	**Yes**
6	Rain	Mild	High	Strong	**No**
7	Overcast	Hot	Normal	Weak	**Yes**

Machine learning

Example data

Can we play golf?

Day	Outlook	Temperature	Humidity	Wind	PlayGolf
1	Sunny	Hot	High	Weak	**No**
2	Sunny	Hot	High	Strong	**No**
3	Overcast	Hot	High	Weak	**Yes**
4	Rain	Cool	Normal	Weak	**Yes**
5	Sunny	Mild	Normal	Weak	**Yes**
6	Rain	Mild	High	Strong	**No**
7	Overcast	Hot	Normal	Weak	**Yes**

Day	Outlook	Temperature	Humidity	Wind	PlayGolf
8	Rain	Hot	Normal	Strong	**?**

Machine learning

Example data

What class of Iris flower is this?

Sepal-length	Sepal-width	Petal-length	Petal-width	Class
6.8	3	6.3	2.3	**Versicolour**
7	3.9	2.4	1.1	**Setosa**
2	3	2.3	1.7	**Verginica**
3	3.4	1.5	1.5	**Verginica**
5.5	3.6	6.8	2.4	**Versicolour**
7.7	4.1	1.2	1.4	**Setosa**
6.3	4.3	1.6	1.2	**Setosa**
1	3.7	2.2	2	**Verginica**

Machine learning

Example data

What class of Iris flower is this?

Sepal-length	Sepal-width	Petal-length	Petal-width	Class
6.8	3	6.3	2.3	**Versicolour**
7	3.9	2.4	1.1	**Setosa**
2	3	2.3	1.7	**Verginica**
3	3.4	1.5	1.5	**Verginica**
5.5	3.6	6.8	2.4	**Versicolour**
7.7	4.1	1.2	1.4	**Setosa**
6.3	4.3	1.6	1.2	**Setosa**
1	3.7	2.2	2	**Verginica**

Sepal-length	Sepal-width	Petal-length	Petal-width	Class
4.5	3.0	6.3	1.5	**?**

Machine Learning

Some applications

- Medical diagnosis
- Industrial fault diagnosis
- **Text categorization**
- Speech Recognition
- **Natural Language Processing**
- Signal and Image Processing
- Industrial control/automation
- **Data Mining**

Contents

Association rules

People at a supermarket

Transaction	Items
t_1	Beef, Chicken, Milk
t_2	Beef, Cheese
t_3	Cheese, Boots
t_4	Beef, Chicken, Cheese
t_5	Beef, Chicken, Clothes, Cheese, Milk
t_6	Chicken, Clothes, Milk
t_7	Chicken, Milk, Clothes

Association rules

- Item set $I = \{\text{Beer}, \text{Chicken}, \text{Clothes}, \text{Chesse}, \text{Milk}, \text{Boots}\}$
- Transaction set $T = \{t_1, t_2, \cdots\}$

An association rule is of the form

$$X \Rightarrow Y$$

Where $X \subset I$, $Y \subset I$, $X \cap Y = \emptyset$. X and Y are called *itemsets*.
Example: $\{\text{Beer}, \text{Chicken}\} \Rightarrow \{\text{Cheese}\}$

Association rule strength

Support

$$\frac{|X \cup Y|}{|T|}$$

It can be interpreted as the probability of occurrence of X and Y together in a transaction.

Confidence

$$\frac{|X \cup Y|}{|X|}$$

It can be interpreted as the conditional probability of Y given X.

The objective is to find rules with some minimum support and confidence.

Apriori algorithm

Two steps

- Generate all frequent itemsets, with support greater than some minimum
- Generate all association rules, with confidence greater than some minimum

Apriori algorithm

Two steps

- Generate all frequent itemsets, with support greater than some minimum
- Generate all association rules, with confidence greater than some minimum

Downward closure property

If an itemset has minimum support, every non-empty subset of it has minimum support

Apriori algorithm

Generate frequent itemsets

Algorithm: Frequent itemset generation

Data: $T = \{t_1, t_2, \cdots, t_n\}$, tx
$I = \{i_1, i_2, \cdots, i_m\}$, items
S^*, minimum support

Result: $F \subset 2^I$, itemsets with *support* $\geq S^*$

$F_1 = \{i | support(\{i\}) \geq S^*\}$;
$k \leftarrow 2$;
while $F_{k-1} \neq \emptyset$ **do**
    $C_k \leftarrow candidate\_gen(F_{k-1})$;
    $F_k \leftarrow \{c \in C_k | support(c) \geq S^*\}$
    $k \leftarrow k + 1$
end
return $\bigcup_k F_k$

Algorithm: Candidate generation

Data: $F_k = \{\{i_1, i_2, \cdots, i_k\}, \cdots\}$,
k-sized itemsets

Result: $C_{k+1} = \{\{i_1, i_2, \cdots, i_{k+1}\}, \cdots\}$,
k+1-sized itemsets

$C_{k+1} \leftarrow \emptyset$;
forall the $f_1, f_2 | f_1$ *differs from* f_2 *only in the last element* **do**
    if *every k-subset of* $f_1 \cup f_2$ *is in* F_k **then**
        $C_{k+1} \leftarrow C_{k+1} \cup \{f_1 \cup f_2\}$;
    end
end
return C_{k+1}

Apriori algorithm
A little example

Transaction	Items
t_1	1, 2, 3
t_2	1, 4
t_3	4, 5
t_4	1, 2, 4
t_5	1, 2, 3, 4, 6
t_6	2, 3, 6
t_7	2, 3, 6

$S^* = 30\%$ (At least 3 examples (because $3/7 > 0.3$**?**))