Notes 01-4: What Can Data Mining Learn?

Four levels of learning can be differentiated by the nature of the learning task:

- *Fact*: A simple statement of truth.
- *Concept*: A set of objects, symbols, or events grouped together because they share certain characteristics.
- *Procedure*: A step-by-step course of action developed to achieve some goal.
- *Principle*: A general truth or law basic to other truths or laws.

Data mining is good at learning *concepts* (i.e., the output of a data mining task). For example, trees, rules, and networks.

Concepts can be viewed from *three* different perspectives: classical, probabilistic, and exemplar.

Classical: Assumes that all concepts have crisply defined properties to determine whether an individual example is representative of the concept.

Example:

```
if Annual Income >= $30,000
   and Years at Current Position >= 5
   and Owns a Home = True
then Good Credit Risk = True
```

In a nutshell: All three rule conditions must be satisfied.

Probabilistic: Assumes that concepts are represented by properties that are probable for individual examples that are representative of the concept.

Example:

The mean annual income for individuals who usually make loans payments on time is $30,000.

The mean number of years at the current employer for individuals who are usually good credit risks is five years.

Individuals who are usually good credit risks tend to own their own home.

In a nutshell: The three rules are used as a general guideline and a probability is usually associated with how well an individual satisfies membership in the concept.

Exemplar: Assumes that a given individual is an example of the concept if the individual is similar enough to a set of one or more known examples of the concept.

Example

Exemplar #1:

```
Annual Income = $32,000
Years at Current Position = 6
Owns a Home = True
```

Exemplar #2:

```
Annual Income = $56,000
Years at Current Position = 13
Owns a Home = False
```

Exemplar #3:

```
Annual Income = $21,000
Years at Current Position = 17
Owns a Home = True
```

In a nutshell: A probability is usually associated with each exemplar indicating the likelihood of concept membership.

Four general types of knowledge can be used as a guide to determine when data mining should be considered: shallow knowledge, multidimensional knowledge, hidden knowledge, and deep knowledge.

Shallow knowledge: Factual in nature and easily stored and manipulated in a database using SQL (data mining not required).

Multidimensional knowledge: Also factual in nature, but stored in a multidimensional format requiring OLAP tools for manipulation (data mining not required).

Hidden knowledge: Represents patterns or regularities in data that cannot be easily found using SQL or OLAP tools (ideal for data mining).

Deep knowledge: Knowledge that can only be found if some direction is provided about what to look for (beyond the capabilities of current data mining tools).