# Notes 01-3: Primary Tasks of Data Mining

Data mining techniques are generally divided into *two* main categories.

*Predictive*: The objective is to predict the value of a particular attribute based upon the values of other attributes.

* The *target* (or dependent) variable is the attribute whose value is to be predicted.
* The *explanatory* (or independent) variables are the attributes used to make the prediction.

*Descriptive*: The objective is to derive patterns (e.g., correlations, trends, clusters, trajectories, and anomalies) that describe the fundamental relationships in the data.

## Core Data Mining Techniques

***Predictive modeling***: Used to build a model for the target variable as a function of the explanatory variables. There are *two* types: *classification* and *regression*.

*Classification*: Used for discrete target variables.

Example - A simple decision tree for mammal classification

DIAGRAM = Introduction.F.2.a1 – **TO BE DONE**

*Regression*: Used for continuous target variables.

Example - Predicting salary based upon years of service

|  |  |
| --- | --- |
| Salary | Years of Service |
| 30 |  3 |
| 57 |  8 |
| 64 |  9 |
| 72 | 13 |
| 36 |  3 |
| 43 |  6 |
| 59 | 11 |
| 90 | 21 |
| 20 |  1 |
| 83 | 16 |

Plot the points on a graph and find a line that best represents the relationship between salary and years of service.

***Association analysis***: Used to discover patterns that describe strongly associated features in the data.

Example - Profiling sales

Assume some store sells the following products: milk, cheese, bread, eggs, diapers, and beer.

Store keeps track of when items are sold (i.e., AM or PM).

Customers can buy any combination of products.

Associations describe products that are purchased together.

milk → bread (AM and PM)

eggs → cheese (AM)

diapers → beer (PM)

***Cluster analysis***: Used to find groups of closely related objects so that objects in the same cluster are more similar to each other than to objects in other clusters.

Example – Clustering based upon customer profiles

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Income | Age | Children | Marital Status | Education |
| 25K | 35 | y | *single* | high school |
| 15K | 25 | n | *married* | B.Sc. |
| 20K | 40 | y | *divorced* | M.Sc. |
| 30K | 20 | y | *married* | Ph.D. |
| 20K | 25 | n | *married* | high school |
| 70K | 60 | n | *single* | B.Sc. |
| 90K | 30 | y | *divorced* | B.Sc. |

Depending upon the attribute chosen, the clusters will be different.

***Summarization*** involves methods for finding a compact description for a subset of data.

***Dependency Modeling*** consists of finding a model which describes significant dependencies between variables. Dependency models exist at two levels:

1. The *structural* level of the model specifies (often graphically) which variables are locally dependent on each other, and
2. The *quantitative* level of the model specifies the strengths of the dependencies using some numerical scale.

***Change and Deviation Detection*** focuses on discovering the most significant changes in the data from previously measured or normative values.