Definition: A computer program is said to **learn** from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.

Machine learning is programming computers to optimize a performance criterion using example data or past experience. We have a model defined up to some parameters, and learning is the execution of a computer program to optimize the parameters of the model using the training data or past experience. The model may be *predictive* to make predictions in the future, or *descriptive* to gain knowledge from data, or both.

Machine learning can be broadly defined as computational methods using experience to improve performance or to make accurate predictions. Here, *experience* refers to the past information available to the learner, which typically takes the form of electronic data collected and made available for analysis.

Learning from data:

In a typical scenario, we have an outcome measurement, usually quantitative or categorical, that we wish to predict based on a set of features. In other words, we have a training set of data, in which we observe the outcome and feature measurements for a set of objects (such as people).

Using this data we build a prediction model, or learner, which will enable us to predict the outcome for new unseen objects. A good learner is one that accurately predicts such an outcome.

Machine learning is defined as a set of methods that can automatically detect patterns in data, and then use the uncovered patterns to predict future data, or to perform other kinds of decision making under uncertainty.

1.1 GENERAL MODEL OF LEARNING FROM EXAMPLES

Consider the following model of searching for functional dependency, which we call the *model of learning from examples*.

The model contains three elements (Fig 1.1):

1. The generator of the data (examples), G.

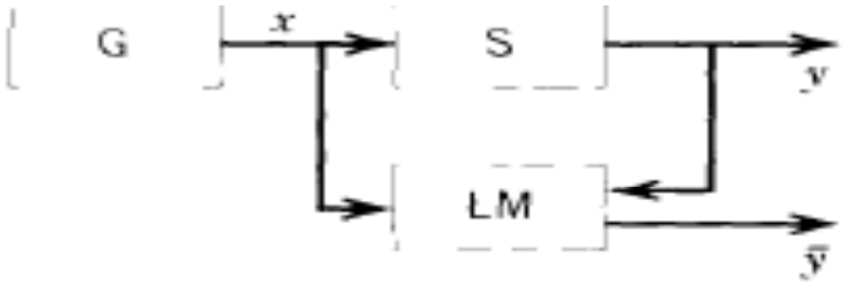

FIGURE 1.1. A model of learning from examples. During the learning process, the learning machine observes the pairs (x, y) (the training set). After training, the machine must on any given x return a value $\bar{y}$. The goal is to return a value $\bar{y}$ which is close to the supervisor's response y.

2. The target operator (sometimes called *supervisor's operator* or, for simplicity, *supervisor*), S.
3. The learning machine, LM.

The learning machine observes ℓ pairs

$$(x_1, y_1), \ldots, (x_\ell, y_\ell)$$

(the *training set*) which contain input vectors x and the supervisor's response y. During this period, the learning machine constructs some operator which will be used for prediction of the supervisor's answer y_i on any specific vector x_i generated by the generator G. The goal of the learning machine is to construct an appropriate approximation.

To be a mathematical statement, this general scheme of learning from examples needs some clarification. First of all, we have to describe what kind of operators are used by the supervisor. In this book, we suppose that the supervisor returns the output y on the vector x according to a conditional distribution function $F(y|x)$ (this includes the case when the supervisor uses some function $y = f(x)$).

1.1 FUNCTION ESTIMATION MODEL

We describe the general model of learning from examples through three components (Fig.1.1):

(i) A generator (G) of random vectors $x \in R^n$, drawn independently from a fixed but unknown probability distribution function $F(x)$.

(ii) A supervisor (S) who returns an output value y to every input vector x, according to a conditional distrihution function[1] $F(y|x)$, also fixed but unknown.

(iii) A learning machine (LM) capable of implementing a set of functions $f(x, \alpha)$, $\alpha \in \Lambda$, where Λ is a set of parameters.[2]

The problem of learning is that of choosing from the given set of functions $f(x, \alpha)$, $\alpha \in \Lambda$, the one that best approximates the supervisor's response.

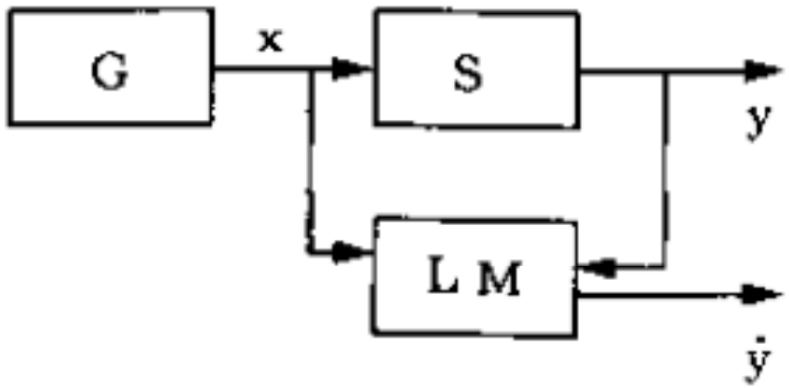

FIGURE 1.1. A model of learning from examples. During the learning process, the learning machine observes the pairs (x, y) (the training set). After training, the machine must on any given x return a value $\hat{y}$. The goal is to return a value $\hat{y}$ that is close to the supervisor's response y.

The selection of the desired function is based on a training set of ℓ independent and identically distributed (i.i.d.) observations drawn according to $F(x, y) = F(x)F(y|x)$:

$$(x_1, y_1), \ldots, (x_\ell, y_\ell). \tag{1.1}$$

More generally, suppose that we observe a quantitative response Y and p different predictors, $X_1, X_2, \ldots, X_p$. We assume that there is some relationship between Y and $X = (X_1, X_2, \ldots, X_p)$, which can be written in the very general form

$$Y = f(X) + \epsilon. \tag{2.1}$$

Here f is some fixed but unknown function of $X_1, \ldots, X_p$, and ϵ is a random *error term*, which is independent of X and has mean zero. In this formulation, f represents the *systematic* information that X provides about Y.

In essence, statistical learning refers to a set of approaches for estimating f.