# Chapter 1

# Introduction to machine learning

In this chapter, we consider different definitions of the term "machine learning" and explain what is meant by "learning" in the context of machine learning. We also discuss the various components of the machine learning process. There are also brief discussions about different types learning like supervised learning, unsupervised learning and reinforcement learning.

## 1.1 Introduction

### 1.1.1 Definition of machine learning

Arthur Samuel, an early American leader in the field of computer gaming and artificial intelligence, coined the term "Machine Learning" in 1959 while at IBM. He defined machine learning as "the field of study that gives computers the ability to learn without being explicitly programmed." However, there is no universally accepted definition for machine learning. Different authors define the term differently. We give below two more definitions.

1. 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 (see [2] p.3).

2. The field of study known as machine learning is concerned with the question of how to construct computer programs that automatically improve with experience (see [4], Preface.).

#### Remarks

In the above definitions we have used the term "model" and we will be using this term at several contexts later in this book. It appears that there is no universally accepted one sentence definition of this term. Loosely, it may be understood as some mathematical expression or equation, or some mathematical structures such as graphs and trees, or a division of sets into disjoint subsets, or a set of logical "if ... then ... else ..." rules, or some such thing. It may be noted that this is not an exhaustive list.

### 1.1.2 Definition of learning

#### 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 $T$, as measured by $P$, improves with experience $E$.

**Examples**

i) Handwriting recognition learning problem

- Task $T$: Recognising and classifying handwritten words within images
- Performance $P$: Percent of words correctly classified
- Training experience $E$: A dataset of handwritten words with given classifications

ii) A robot driving learning problem

- Task $T$: Driving on highways using vision sensors
- Performance measure $P$: Average distance traveled before an error
- training experience: A sequence of images and steering commands recorded while observing a human driver

iii) A chess learning problem

- Task $T$: Playing chess
- Performance measure $P$: Percent of games won against opponents
- Training experience $E$: Playing practice games against itself

**Definition**

A computer program which learns from experience is called a *machine learning program* or simply a *learning program*. Such a program is sometimes also referred to as a *learner*.

## 1.2 How machines learn

### 1.2.1 Basic components of learning process

The learning process, whether by a human or a machine, can be divided into four components, namely, data storage, abstraction, generalization and evaluation. Figure 1.1 illustrates the various components and the steps involved in the learning process.

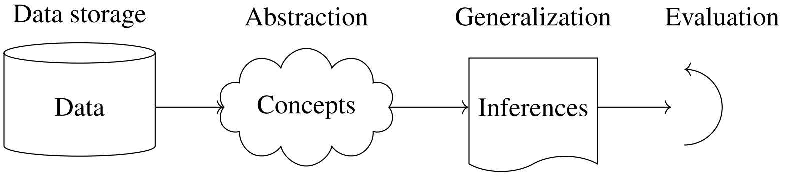


Figure 1.1: Components of learning process

1. **Data storage**

   Facilities for storing and retrieving huge amounts of data are an important component of the learning process. Humans and computers alike utilize data storage as a foundation for advanced reasoning.

   - In a human being, the data is stored in the brain and data is retrieved using electrochemical signals.
   - Computers use hard disk drives, flash memory, random access memory and similar devices to store data and use cables and other technology to retrieve data.

2. **Abstraction**

   The second component of the learning process is known as *abstraction*.

   Abstraction is the process of extracting knowledge about stored data. This involves creating general concepts about the data as a whole. The creation of knowledge involves application of known models and creation of new models.

   The process of fitting a model to a dataset is known as *training*. When the model has been trained, the data is transformed into an abstract form that summarizes the original information.

3. **Generalization**

   The third component of the learning process is known as *generalisation.*

   The term generalization describes the process of turning the knowledge about stored data into a form that can be utilized for future action. These actions are to be carried out on tasks that are similar, but not identical, to those what have been seen before. In generalization, the goal is to discover those properties of the data that will be most relevant to future tasks.

4. **Evaluation**

   *Evaluation* is the last component of the learning process.

   It is the process of giving feedback to the user to measure the utility of the learned knowledge. This feedback is then utilised to effect improvements in the whole learning process.

## 1.3 Applications of machine learning

Application of machine learning methods to large databases is called data mining. In data mining, a large volume of data is processed to construct a simple model with valuable use, for example, having high predictive accuracy.

The following is a list of some of the typical applications of machine learning.

1. In retail business, machine learning is used to study consumer behaviour.
2. In finance, banks analyze their past data to build models to use in credit applications, fraud detection, and the stock market.
3. In manufacturing, learning models are used for optimization, control, and troubleshooting.
4. In medicine, learning programs are used for medical diagnosis.
5. In telecommunications, call patterns are analyzed for network optimization and maximizing the quality of service.
6. In science, large amounts of data in physics, astronomy, and biology can only be analyzed fast enough by computers. The World Wide Web is huge; it is constantly growing and searching for relevant information cannot be done manually.
7. In artificial intelligence, it is used to teach a system to learn and adapt to changes so that the system designer need not foresee and provide solutions for all possible situations.
8. It is used to find solutions to many problems in vision, speech recognition, and robotics.
9. Machine learning methods are applied in the design of computer-controlled vehicles to steer correctly when driving on a variety of roads.
10. Machine learning methods have been used to develop programmes for playing games such as chess, backgammon and Go.

## 1.4 Understanding data

Since an important component of the machine learning process is data storage, we briefly consider in this section the different types and forms of data that are encountered in the machine learning process.

### 1.4.1 Unit of observation

By a *unit of observation* we mean the smallest entity with measured properties of interest for a study.

**Examples**

- A person, an object or a thing
- A time point
- A geographic region
- A measurement

Sometimes, units of observation are combined to form units such as person-years.

### 1.4.2 Examples and features

Datasets that store the units of observation and their properties can be imagined as collections of data consisting of the following:

- **Examples**

  An "example" is an instance of the unit of observation for which properties have been recorded. An "example" is also referred to as an "instance", or "case" or "record." (It may be noted that the word "example" has been used here in a technical sense.)

- **Features**

  A "feature" is a recorded property or a characteristic of examples. It is also referred to as "attribute", or "variable" or "feature."

**Examples for "examples" and "features"**

1. **Cancer detection**

   Consider the problem of developing an algorithm for detecting cancer. In this study we note the following.

   (a) The units of observation are the patients.

   (b) The examples are members of a sample of cancer patients.

   (c) The following attributes of the patients may be chosen as the features:

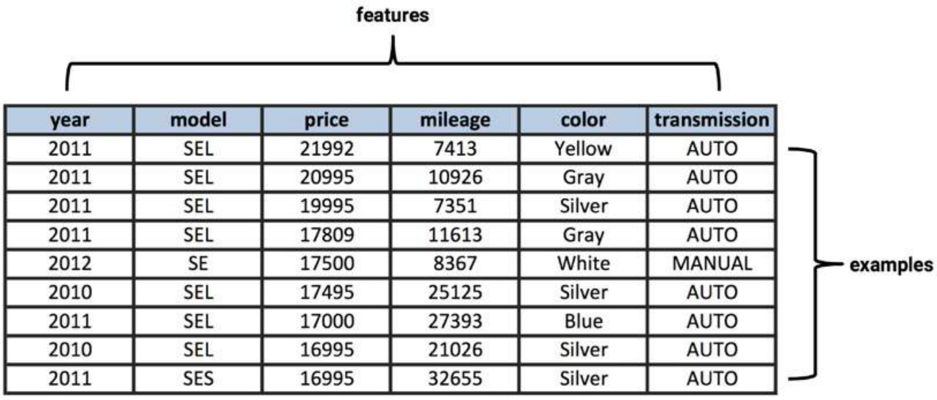
   - gender
   - age
   - blood pressure
   - the findings of the pathology report after a biopsy

2. **Pet selection**

   Suppose we want to predict the type of pet a person will choose.

   (a) The units are the persons.

   (b) The examples are members of a sample of persons who own pets.


features
examples

| year | model | price | mileage | color | transmission |
|---|---|---|---|---|---|
| 2011 | SEL | 21992 | 7413 | Yellow | AUTO |
| 2011 | SEL | 20995 | 10926 | Gray | AUTO |
| 2011 | SEL | 19995 | 7351 | Silver | AUTO |
| 2011 | SEL | 17809 | 11613 | Gray | AUTO |
| 2012 | SE | 17500 | 8367 | White | MANUAL |
| 2010 | SEL | 17495 | 25125 | Silver | AUTO |
| 2011 | SEL | 17000 | 27393 | Blue | AUTO |
| 2010 | SEL | 16995 | 21026 | Silver | AUTO |
| 2011 | SES | 16995 | 32655 | Silver | AUTO |

Figure 1.2: Example for “examples” and “features” collected in a matrix format (data relates to automobiles and their features)

(c) The features might include age, home region, family income, etc. of persons who own pets.

3. **Spam e-mail**

   Let it be required to build a learning algorithm to identify spam e-mail.

   (a) The unit of observation could be an e-mail messages.

   (b) The examples would be specific messages.

   (c) The features might consist of the words used in the messages.

Examples and features are generally collected in a “matrix format”. Fig. 1.2 shows such a data set.

### 1.4.3 Different forms of data

1. **Numeric data**

   If a feature represents a characteristic measured in numbers, it is called a numeric feature.

2. **Categorical or nominal**

   A categorical feature is an attribute that can take on one of a limited, and usually fixed, number of possible values on the basis of some qualitative property. A categorical feature is also called a nominal feature.

3. **Ordinal data**

   This denotes a nominal variable with categories falling in an ordered list. Examples include clothing sizes such as small, medium, and large, or a measurement of customer satisfaction on a scale from “not at all happy” to “very happy.”

**Examples**

In the data given in Fig.1.2, the features “year”, “price” and “mileage” are numeric and the features “model”, “color” and “transmission” are categorical.

## 1.5 General classes of machine learning problems

### 1.5.1 Learning associations

**1. Association rule learning**

*Association rule learning* is a machine learning method for discovering interesting relations, called "association rules", between variables in large databases using some measures of "interestingness".

**2. Example**

Consider a supermarket chain. The management of the chain is interested in knowing whether there are any patterns in the purchases of products by customers like the following:

> "If a customer buys onions and potatoes together, then he/she is likely to also buy hamburger."

From the standpoint of customer behaviour, this defines an association between the set of products {onion, potato} and the set {burger}. This association is represented in the form of a rule as follows:

$$\{\text{onion, potato}\} \Rightarrow \{\text{burger}\}$$

The measure of how likely a customer, who has bought onion and potato, to buy burger also is given by the conditional probability

$$P(\{\text{onion, potato}\}|\{\text{burger}\}).$$

If this conditional probability is 0.8, then the rule may be stated more precisely as follows:

> "80% of customers who buy onion and potato also buy burger."

**3. How association rules are made use of**

Consider an association rule of the form

$$X \Rightarrow Y,$$

that is, if people buy $X$ then they are also likely to buy $Y$.

Suppose there is a customer who buys $X$ and does not buy $Y$. Then that customer is a potential $Y$ customer. Once we find such customers, we can target them for cross-selling. A knowledge of such rules can be used for promotional pricing or product placements.

**4. General case**

In finding an association rule $X \Rightarrow Y$, we are interested in learning a conditional probability of the form $P(Y|X)$ where $Y$ is the product the customer may buy and $X$ is the product or the set of products the customer has already purchased.

If we may want to make a distinction among customers, we may estimate $P(Y|X, D)$ where $D$ is a set of customer attributes, like gender, age, marital status, and so on, assuming that we have access to this information.

**5. Algorithms**

There are several algorithms for generating association rules. Some of the well-known algorithms are listed below:

a) Apriori algorithm

b) Eclat algorithm

c) FP-Growth Algorithm (FP stands for Frequency Pattern)

### 1.5.2 Classification

**1. Definition**

In machine learning, *classification* is the problem of identifying to which of a set of categories a new observation belongs, on the basis of a training set of data containing observations (or instances) whose category membership is known.

**2. Example**

Consider the following data:

| Score1 | 29 | 22 | 10 | 31 | 17 | 33 | 32 | 20 |
|---|---|---|---|---|---|---|---|---|
| Score2 | 43 | 29 | 47 | 55 | 18 | 54 | 40 | 41 |
| Result | Pass | Fail | Fail | Pass | Fail | Pass | Pass | Pass |

Table 1.1: Example data for a classification problem

Data in Table 1.1 is the training set of data. There are two attributes "Score1" and "Score2". The class label is called "Result". The class label has two possible values "Pass" and "Fail". The data can be divided into two categories or classes: The set of data for which the class label is "Pass" and the set of data for which the class label is"Fail".

Let us assume that we have no knowledge about the data other than what is given in the table. Now, the problem can be posed as follows: If we have some new data, say "Score1 = 25" and "Score2 = 36", what value should be assigned to "Result" corresponding to the new data; in other words, to which of the two categories or classes the new observation should be assigned? See Figure 1.3 for a graphical representation of the problem.

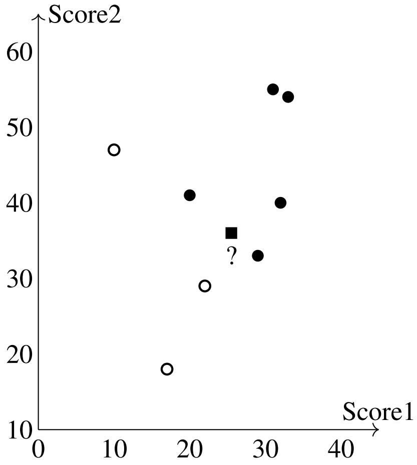


Figure 1.3: Graphical representation of data in Table 1.1. Solid dots represent data in "Pass" class and hollow dots data in "Fail" class. The class label of the square dot is to be determined.

To answer this question, using the given data alone we need to find the rule, or the formula, or the method that has been used in assigning the values to the class label "Result". The problem of finding this rule or formula or the method is the classification problem. In general, even the general form of the rule or function or method will not be known. So several different rules, etc. may have to be tested to obtain the correct rule or function or method.

**3. Real life examples**

i) **Optical character recognition**

*Optical character recognition* problem, which is the problem of recognizing character codes from their images, is an example of classification problem. This is an example where there are multiple classes, as many as there are characters we would like to recognize. Especially interesting is the case when the characters are handwritten. People have different handwriting styles; characters may be written small or large, slanted, with a pen or pencil, and there are many possible images corresponding to the same character.

ii) **Face recognition**

In the case of *face recognition*, the input is an image, the classes are people to be recognized, and the learning program should learn to associate the face images to identities. This problem is more difficult than optical character recognition because there are more classes, input image is larger, and a face is three-dimensional and differences in pose and lighting cause significant changes in the image.

iii) **Speech recognition**

In *speech recognition*, the input is acoustic and the classes are words that can be uttered.

iv) **Medical diagnosis**

In *medical diagnosis*, the inputs are the relevant information we have about the patient and the classes are the illnesses. The inputs contain the patient's age, gender, past medical history, and current symptoms. Some tests may not have been applied to the patient, and thus these inputs would be missing.

v) **Knowledge extraction**

Classification rules can also be used for *knowledge extraction*. The rule is a simple model that explains the data, and looking at this model we have an explanation about the process underlying the data.

vi) **Compression**

Classification rules can be used for *compression*. By fitting a rule to the data, we get an explanation that is simpler than the data, requiring less memory to store and less computation to process.

vii) **More examples**

Here are some further examples of classification problems.

(a) An emergency room in a hospital measures 17 variables like blood pressure, age, etc. of newly admitted patients. A decision has to be made whether to put the patient in an ICU. Due to the high cost of ICU, only patients who may survive a month or more are given higher priority. Such patients are labeled as "low-risk patients" and others are labeled "high-risk patients". The problem is to device a rule to classify a patient as a "low-risk patient" or a "high-risk patient".

(b) A credit card company receives hundreds of thousands of applications for new cards. The applications contain information regarding several attributes like annual salary, age, etc. The problem is to devise a rule to classify the applicants to those who are credit-worthy, who are not credit-worthy or to those who require further analysis.

(c) Astronomers have been cataloguing distant objects in the sky using digital images created using special devices. The objects are to be labeled as star, galaxy, nebula, etc. The data is highly noisy and are very faint. The problem is to device a rule using which a distant object can be correctly labeled.

### 4. Discriminant

A *discriminant* of a classification problem is a rule or a function that is used to assign labels to new observations.

#### Examples

i) Consider the data given in Table 1.1 and the associated classification problem. We may consider the following rules for the classification of the new data:

   IF Score1 + Score2 ≥ 60, THEN "Pass" ELSE "Fail".
   IF Score1 ≥ 20 AND Score2 ≥ 40 THEN "Pass" ELSE "Fail".

   Or, we may consider the following rules with unspecified values for $M, m_1, m_2$ and then by some method estimate their values.

   IF Score1 + Score2 ≥ $M$, THEN "Pass" ELSE "Fail".
   IF Score1 ≥ $m_1$ AND Score2 ≥ $m_2$ THEN "Pass" ELSE "Fail".

ii) Consider a finance company which lends money to customers. Before lending money, the company would like to assess the risk associated with the loan. For simplicity, let us assume that the company assesses the risk based on two variables, namely, the annual income and the annual savings of the customers.

   Let $x_1$ be the annual income and $x_2$ be the annual savings of a customer.

   - After using the past data, a rule of the following form with suitable values for $\theta_1$ and $\theta_2$ may be formulated:

     IF $x_1 > \theta_1$ AND $x_2 > \theta_2$ THEN "low-risk" ELSE "high-risk".

     This rule is an example of a discriminant.

   - Based on the past data, a rule of the following form may also be formulated:

     IF $x_2 - 0.2x_1 > 0$ THEN "low-risk" ELSE "high-risk".

     In this case the rule may be thought of as the discriminant. The function $f(x_1, x_2) = x_2 - 0,2x_1$ can also be considered as the discriminant.

### 5. Algorithms

There are several machine learning algorithms for classification. The following are some of the well-known algorithms.

a) Logistic regression

b) Naive Bayes algorithm

c) $k$-NN algorithm

d) Decision tree algorithm

e) Support vector machine algorithm

f) Random forest algorithm

**Remarks**

- A classification problem requires that examples be classified into one of two or more classes.
- A classification can have real-valued or discrete input variables.
- A problem with two classes is often called a two-class or binary classification problem.
- A problem with more than two classes is often called a multi-class classification problem.
- A problem where an example is assigned multiple classes is called a multi-label classification problem.

### 1.5.3 Regression

**1. Definition**

In machine learning, a *regression problem* is the problem of predicting the value of a numeric variable based on observed values of the variable. The value of the output variable may be a number, such as an integer or a floating point value. These are often quantities, such as amounts and sizes. The input variables may be discrete or real-valued.

**2. Example**

Consider the data on car prices given in Table 1.2.

| Price | Age | Distance | Weight |
|---|---|---|---|
| (US$) | (years) | (KM) | (pounds) |
| 13500 | 23 | 46986 | 1165 |
| 13750 | 23 | 72937 | 1165 |
| 13950 | 24 | 41711 | 1165 |
| 14950 | 26 | 48000 | 1165 |
| 13750 | 30 | 38500 | 1170 |
| 12950 | 32 | 61000 | 1170 |
| 16900 | 27 | 94612 | 1245 |
| 18600 | 30 | 75889 | 1245 |
| 21500 | 27 | 19700 | 1185 |
| 12950 | 23 | 71138 | 1105 |

Table 1.2: Prices of used cars: example data for regression

Suppose we are required to estimate the price of a car aged 25 years with distance 53240 KM and weight 1200 pounds. This is an example of a regression problem beause we have to predict the value of the numeric variable "Price".

**3. General approach**

Let $x$ denote the set of input variables and $y$ the output variable. In machine learning, the general approach to regression is to assume a model, that is, some mathematical relation between $x$ and $y$, involving some parameters say, $\theta$, in the following form:

$$y = f(x, \theta)$$

The function $f(x, \theta)$ is called the *regression function*. The machine learning algorithm optimizes the parameters in the set $\theta$ such that the approximation error is minimized; that is, the estimates of the values of the dependent variable $y$ are as close as possible to the correct values given in the training set.

**Example**

For example, if the input variables are "Age", "Distance" and "Weight" and the output variable is "Price", the model may be

$$\begin{aligned} y &= f(x, \theta) \\ \text{Price} &= a_0 + a_1 \times (\text{Age}) + a_2 \times (\text{Distance}) + a_3 \times (\text{Weight}) \end{aligned}$$

where $x = (\text{Age, Distance, Weight})$ denotes the the set of input variables and $\theta = (a_0, a_1, a_2, a_3)$ denotes the set of parameters of the model.

**4. Different regression models**

There are various types of regression techniques available to make predictions. These techniques mostly differ in three aspects, namely, the number and type of independent variables, the type of dependent variables and the shape of regression line. Some of these are listed below.

- *Simple linear regression*: There is only one continuous independent variable $x$ and the assumed relation between the independent variable and the dependent variable $y$ is

$$y = a + bx.$$

- *Multivariate linear regression*: There are more than one independent variable, say $x_1, \ldots, x_n$, and the assumed relation between the independent variables and the dependent variable is

$$y = a_0 + a_1 x_1 + \cdots + a_n x_n.$$

- *Polynomial regression*: There is only one continuous independent variable $x$ and the assumed model is

$$y = a_0 + a_1 x + \cdots + a_n x^n.$$

- *Logistic regression*: The dependent variable is binary, that is, a variable which takes only the values 0 and 1. The assumed model involves certain probability distributions.

## 1.6 Different types of learning

In general, machine learning algorithms can be classified into three types.

### 1.6.1 Supervised learning

*Supervised learning* is the machine learning task of learning a function that maps an input to an output based on example input-output pairs.

In supervised learning, each example in the training set is a pair consisting of an input object (typically a vector) and an output value. A supervised learning algorithm analyzes the training data and produces a function, which can be used for mapping new examples. In the optimal case, the function will correctly determine the class labels for unseen instances. Both classification and regression problems are supervised learning problems.

A wide range of supervised learning algorithms are available, each with its strengths and weaknesses. There is no single learning algorithm that works best on all supervised learning problems.

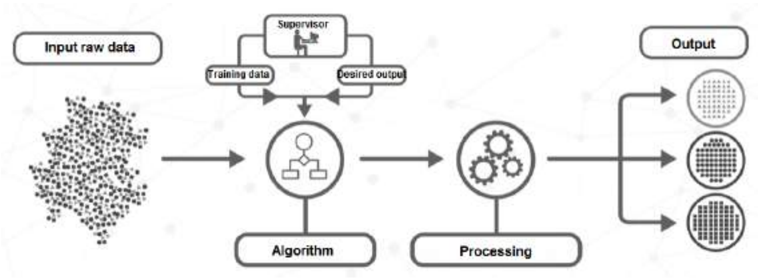


Figure 1.4: Supervised learning

**Remarks**

A “supervised learning” is so called because the process of an algorithm learning from the training dataset can be thought of as a teacher supervising the learning process. We know the correct answers (that is, the correct outputs), the algorithm iteratively makes predictions on the training data and is corrected by the teacher. Learning stops when the algorithm achieves an acceptable level of performance.

**Example**

Consider the following data regarding patients entering a clinic. The data consists of the gender and age of the patients and each patient is labeled as “healthy” or “sick”.

| gender | age | label |
|---|---|---|
| M | 48 | sick |
| M | 67 | sick |
| F | 53 | healthy |
| M | 49 | healthy |
| F | 34 | sick |
| M | 21 | healthy |

Based on this data, when a new patient enters the clinic, how can one predict whether he/she is healthy or sick?

### 1.6.2 Unsupervised learning

*Unsupervised learning* is a type of machine learning algorithm used to draw inferences from datasets consisting of input data without labeled responses.

In unsupervised learning algorithms, a classification or categorization is not included in the observations. There are no output values and so there is no estimation of functions. Since the examples given to the learner are unlabeled, the accuracy of the structure that is output by the algorithm cannot be evaluated.

The most common unsupervised learning method is cluster analysis, which is used for exploratory data analysis to find hidden patterns or grouping in data.

**Example**

Consider the following data regarding patients entering a clinic. The data consists of the gender and age of the patients.

| gender | age |
|---|---|
| M | 48 |
| M | 67 |
| F | 53 |
| M | 49 |
| F | 34 |
| M | 21 |

Based on this data, can we infer anything regarding the patients entering the clinic?

### 1.6.3 Reinforcement learning

*Reinforcement learning* is the problem of getting an agent to act in the world so as to maximize its rewards.

A learner (the program) is not told what actions to take as in most forms of machine learning, but instead must discover which actions yield the most reward by trying them. In the most interesting and challenging cases, actions may affect not only the immediate reward but also the next situations and, through that, all subsequent rewards.

For example, consider teaching a dog a new trick: we cannot tell it what to do, but we can reward/punish it if it does the right/wrong thing. It has to find out what it did that made it get the reward/punishment. We can use a similar method to train computers to do many tasks, such as playing backgammon or chess, scheduling jobs, and controlling robot limbs.

Reinforcement learning is different from supervised learning. Supervised learning is learning from examples provided by a knowledgeable expert.

## 1.7 Sample questions

**(a) Short answer questions**

1. What is meant by "learning" in the context of machine learning?
2. List out the types of machine learning.
3. Distinguish between classification and regression.
4. What are the differences between supervised and unsupervised learning?
5. What is meant by supervised classification?
6. Explain supervised learning with an example.
7. What do you mean by reinforcement learning?
8. What is an association rule?
9. Explain the concept of Association rule learning. Give the names of two algorithms for generating association rules.
10. What is a classification problem in machine learning. Illustrate with an example.
11. Give three examples of classification problems from real life situations.
12. What is a discriminant in a classification problem?
13. List three machine learning algorithms for solving classification problems.
14. What is a binary classification problem? Explain with an example. Give also an example for a classification problem which is not binary.
15. What is regression problem. What are the different types of regression?

**(b) Long answer questions**

1. Give a definition of the term “machine learning”. Explain with an example the concept of learning in the context of machine learning.
2. Describe the basic components of the machine learning process.
3. Describe in detail applications of machine learning in any three different knowledge domains.
4. Describe with an example the concept of association rule learning. Explain how it is made use of in real life situations.
5. What is the classification problem in machine learning? Describe three real life situations in different domains where such problems arise.
6. What is meant by a discriminant of a classification problem? Illustrate the idea with examples.
7. Describe in detail with examples the different types of learning like the supervised learning, etc.