Chapter 1

Introduction to Machine Learning

OBJECTIVE OF THE CHAPTER:

The objective of this chapter is to venture into the arena of machine learning. New comers struggle a lot to understand the philosophy of machine learning. Also, they do not know where to start from and which problem could be and should be solved using machine learning tools and techniques. This chapter intends to give the new comers a starting point to the journey in machine learning. It starts from a historical journey in this field and takes it forward to give a glimpse of the modern day applications.

1.1 INTRODUCTION

It has been more than 20 years since a computer program defeated the reigning world champion in a game which is considered to need a lot of intelligence to play. The computer program was IBM's Deep Blue and it defeated world chess champion, Gary Kasparov. That was the time, probably, when the most number of people gave serious attention to a fast-evolving field in computer science or more specifically artificial intelligence – i.e. machine learning (ML).

As of today, machine learning is a mature technology area finding its application in almost every sphere of life. It can recommend toys to toddlers much in the same way as it can suggest a technology book to a geek or a rich title in literature to a writer. It predicts the future market to help amateur traders compete with seasoned stock traders. It helps an oncologist find whether a tumour is malignant or benign. It helps

in optimizing energy consumption thus helping the cause of Green Earth. Google has become one of the front-runners focusing a lot of its research on machine learning and artificial intelligence – Google self-driving car and Google Brain being two most ambitious projects of Google in its journey of innovation in the field of machine learning. In a nutshell, machine learning has become a way of life, no matter whichever sphere of life we closely look at. But where did it all start from?

The foundation of machine learning started in the 18th and 19th centuries. The first related work dates back to 1763. In that year, Thomas Bayes's work 'An Essay towards solving a Problem in the Doctrine of Chances' was published two years after his death. This is the work underlying Bayes Theorem, a fundamental work on which a number of algorithms of machine learning is based upon. In 1812, the Bayes theorem was actually formalized by the French mathematician Pierre-Simon Laplace. The method of least squares, which is the foundational concept to solve regression problems, was formalized in 1805. In 1913, Andrey Markov came up with the concept of Markov chains.

However, the real start of focused work in the field of machine learning is considered to be Alan Turing's seminal work in 1950. In his paper 'Computing Machinery and Intelligence' (*Mind*, New Series, Vol. 59, No. 236, Oct., 1950, pp. 433–460), Turing posed the question 'Can machines think?' or in other words, 'Do machines have intelligence?'. He was the first to propose that machines can 'learn' and become artificially intelligent. In 1952, Arthur Samuel of IBM laboratory started working on machine learning programs, and first developed programs that could play Checkers. In 1957, Frank Rosenblatt designed the first neural network program simulating the human brain. From then on, for the next 50 years, the journey of machine learning has been fascinating. A number of machine learning algorithms were formulated by different researchers, e.g. the nearest neighbour algorithm in 1969, recurrent neural network in 1982, support vector machines and random forest algorithms in 1995. The latest feather in the cap of machine learning development has been Google's AlphaGo program, which has beaten professional human Go player using machine learning techniques.

Points to Ponder

- While Deep Blue was searching some 200 million positions per second, Kasparov was searching not more than 5–10 positions probably, per second. Yet he played almost at the same level. This clearly shows that humans have some trick up their sleeve that computers could not master yet.
- **Go** is a board game which can be played by two players. It was invented in China almost 2500 years ago and is considered to be the oldest board game. Though it has relatively simple rules, Go is a very complex game (more complex than chess) because of its larger board size and more number of possible moves.

■

The evolution of machine learning from 1950 is depicted in Figure 1.1.

The rapid development in the area of machine learning has triggered a question in everyone's mind – can machines learn better than human? To find its answer, the first step would be to understand what learning is from a human perspective. Then, more light can be shed on what machine learning is. In the end, we need to know whether machine learning has already surpassed or has the potential to surpass human learning in every facet of life.

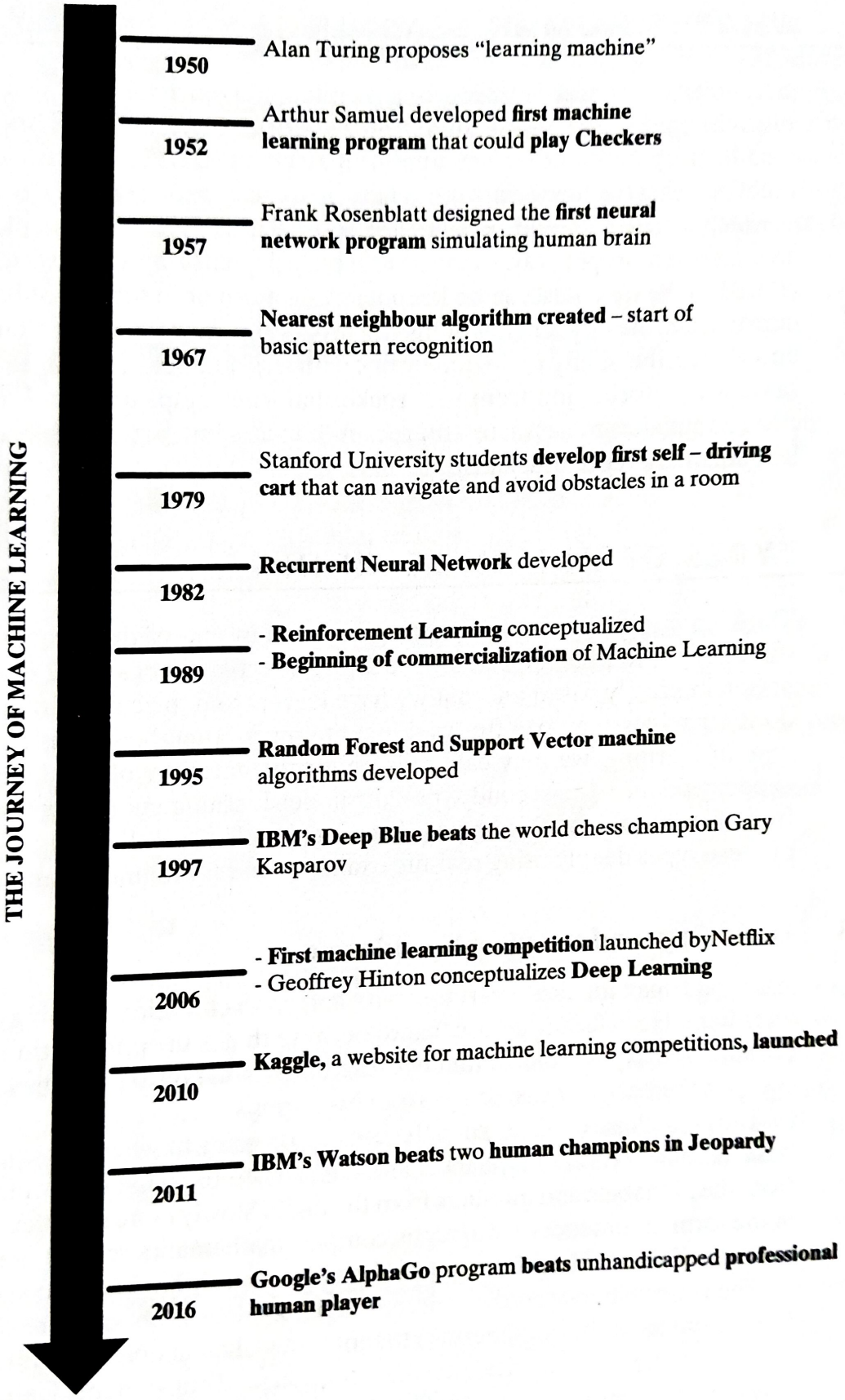

FIG. 1.1
Evolution of machine learning

1.2 WHAT IS HUMAN LEARNING?

In cognitive science, learning is typically referred to as the process of gaining information through observation. And why do we need to learn? In our daily life, we need to carry out multiple activities. It may be a task as simple as walking down the street or doing the homework. Or it may be some complex task like deciding the angle in which a rocket should be launched so that it can have a particular trajectory. To do a task in a proper way, we need to have prior information on one or more things related to the task. Also, as we keep learning more or in other words acquiring more information, the efficiency in doing the tasks keep improving. For example, with more knowledge, the ability to do homework with less number of mistakes increases. In the same way, information from past rocket launches helps in taking the right precautions and makes more successful rocket launch. Thus, with more learning, tasks can be performed more efficiently.

1.3 TYPES OF HUMAN LEARNING

Thinking intuitively, human learning happens in one of the three ways – (1) either somebody who is an expert in the subject directly teaches us, (2) we build our own notion indirectly based on what we have learnt from the expert in the past, or (3) we do it ourselves, may be after multiple attempts, some being unsuccessful. The first type of learning, we may call, falls under the category of learning directly under expert guidance, the second type falls under learning guided by knowledge gained from experts and the third type is learning by self or self-learning. Let's look at each of these types deeply using real-life examples and try to understand what they mean.

1.3.1 Learning under expert guidance

An infant may inculcate certain traits and characteristics, learning straight from its guardians. He calls his hand, a 'hand', because that is the information he gets from his parents. The sky is 'blue' to him because that is what his parents have taught him. We say that the baby 'learns' things from his parents.

The next phase of life is when the baby starts going to school. In school, he starts with basic familiarization of alphabets and digits. Then the baby learns how to form words from the alphabets and numbers from the digits. Slowly more complex learning happens in the form of sentences, paragraphs, complex mathematics, science, etc. The baby is able to learn all these things from his teacher who already has knowledge on these areas.

Then starts higher studies where the person learns about more complex, application-oriented skills. Engineering students get skilled in one of the disciplines like civil, computer science, electrical, mechanical, etc. medical students learn about anatomy, physiology, pharmacology, etc. There are some experts, in general the teachers, in the respective field who have in-depth subject matter knowledge, who help the students in learning these skills.

Then the person starts working as a professional in some field. Though he might have gone through enough theoretical learning in the respective field, he still needs

to learn more about the hands-on application of the knowledge that he has acquired. The professional mentors, by virtue of the knowledge that they have gained through years of hands-on experience, help all new comers in the field to learn on-job.

In all phases of life of a human being, there is an element of guided learning. This learning is imparted by someone, purely because of the fact that he/she has already gathered the knowledge by virtue of his/her experience in that field. So guided learning is the process of gaining information from a person having sufficient knowledge due to the past experience.

1.3.2 Learning guided by knowledge gained from experts

An essential part of learning also happens with the knowledge which has been imparted by teacher or mentor at some point of time in some other form/context. For example, a baby can group together all objects of same colour even if his parents have not specifically taught him to do so. He is able to do so because at some point of time or other his parents have told him which colour is blue, which is red, which is green, etc. A grown-up kid can select one odd word from a set of words because it is a verb and other words being all nouns. He could do this because of his ability to label the words as verbs or nouns, taught by his English teacher long back. In a professional role, a person is able to make out to which customers he should market a campaign from the knowledge about preference that was given by his boss long back.

In all these situations, there is no direct learning. It is some past information shared on some different context, which is used as a learning to make decisions.

1.3.3 Learning by self

In many situations, humans are left to learn on their own. A classic example is a baby learning to walk through obstacles. He bumps on to obstacles and falls down multiple times till he learns that whenever there is an obstacle, he needs to cross over it. He faces the same challenge while learning to ride a cycle as a kid or drive a car as an adult. Not all things are taught by others. A lot of things need to be learnt only from mistakes made in the past. We tend to form a check list on things that we should do, and things that we should not do, based on our experiences.

1.4 WHAT IS MACHINE LEARNING?

Before answering the question 'What is machine learning?' more fundamental questions that peep into one's mind are

- Do machines really learn?
- If so, how do they learn?
- Which problem can we consider as a well-posed learning problem? What are the important features that are required to well-define a learning problem?

At the onset, it is important to formalize the definition of machine learning. This will itself address the first question, i.e. if machines really learn. There are multiple ways to define machine learning. But the one which is perhaps most relevant, concise and accepted universally is the one stated by Tom M. Mitchell, Professor of Machine Learning Department, School of Computer Science, Carnegie Mellon University. Tom M. Mitchell has defined machine learning as

> **'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.'**

What this essentially means is that a machine can be considered to learn if it is able to gather experience by doing a certain task and improve its performance in doing the similar tasks in the future. When we talk about past experience, it means past data related to the task. This data is an input to the machine from some source.

In the context of the learning to play checkers, E represents the experience of playing the game, T represents the task of playing checkers and P is the performance measure indicated by the percentage of games won by the player. The same mapping can be applied for any other machine learning problem, for example, image classification problem. In context of image classification, E represents the past data with images having labels or assigned classes (for example whether the image is of a class cat or a class dog or a class elephant etc.), T is the task of assigning class to new, unlabelled images and P is the performance measure indicated by the percentage of images correctly classified.

The first step in any project is defining your problem. Even if the most powerful algorithm is used, the results will be meaningless if the wrong problem is solved.

1.4.1 How do machines learn?

The basic machine learning process can be divided into three parts.

1. **Data Input:** Past data or information is utilized as a basis for future decision-making
2. **Abstraction:** The input data is represented in a broader way through the underlying algorithm
3. **Generalization:** The abstracted representation is generalized to form a framework for making decisions

Figure 1.2 is a schematic representation of the machine learning process.

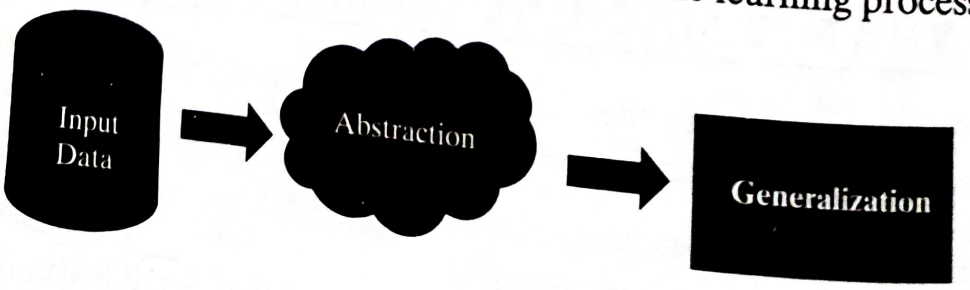

FIG. 1.2
Process of machine learning

Let's put the things in perspective of the human learning process and try to understand the machine learning process more clearly. Reason is, in some sense, machine learning process tries to emulate the process in which humans learn to a large extent.

Let's consider the situation of typical process of learning from classroom and books and preparing for the examination. It is a tendency of many students to try and memorize (we often call it 'learn by heart') as many things as possible. This may work well when the scope of learning is not so vast. Also, the kinds of questions which are asked in the examination are pretty much simple and straightforward. The questions can be answered by simply writing the same things which have been memorized. However, as the scope gets broader and the questions asked in the examination gets more complex, the strategy of memorizing doesn't work well. The number of topics may get too vast for a student to memorize. Also, the capability of memorizing varies from student to student. Together with that, since the questions get more complex, a direct reproduction of the things memorized may not help. The situation continues to get worse as the student graduates to higher classes.

So, what we see in the case of human learning is that just by great memorizing and perfect recall, i.e. just based on knowledge input, students can do well in the examinations only till a certain stage. Beyond that, a better learning strategy needs to be adopted:

1. to be able to deal with the vastness of the subject matter and the related issues in memorizing it
2. to be able to answer questions where a direct answer has not been learnt

A good option is to figure out the key points or ideas amongst a vast pool of knowledge. This helps in creating an outline of topics and a conceptual mapping of those outlined topics with the entire knowledge pool. For example, a broad pool of knowledge may consist of all living animals and their characteristics such as whether they live in land or water, whether they lay eggs, whether they have scales or fur or none, etc. It is a difficult task for any student to memorize the characteristics of all living animals – no matter how much photographic memory he/she may possess. It is better to draw a notion about the basic groups that all living animals belong to and the characteristics which define each of the basic groups. The basic groups of animals are invertebrates and vertebrates. Vertebrates are further grouped as mammals, reptiles, amphibians, fishes, and birds. Here, we have mapped animal groups and their salient characteristics.

1. Invertebrate: Do not have backbones and skeletons
2. Vertebrate
   - **(a)** Fishes: Always live in water and lay eggs
   - **(b)** Amphibians: Semi-aquatic i.e. may live in water or land; smooth skin; lay eggs
   - **(c)** Reptiles: Semi-aquatic like amphibians; scaly skin; lay eggs; cold-blooded
   - **(d)** Birds: Can fly; lay eggs; warm-blooded
   - **(e)** Mammals: Have hair or fur; have milk to feed their young; warm-blooded

This makes it easier to memorize as the scope now reduces to know the animal groups that the animals belong to. Rest of the answers about the characteristics of the animals may be derived from the concept of mapping animal groups and their characteristics.

Moving to the machine learning paradigm, the vast pool of knowledge is available from the data input. However, rather than using it in entirety, a concept map, much in line with the animal group to characteristic mapping explained above, is drawn from the input data. This is nothing but knowledge abstraction as performed by the machine. In the end, the abstracted mapping from the input data can be applied to make critical conclusions. For example, if the group of an animal is given, understanding of the characteristics can be automatically made. Reversely, if the characteristic of an unknown animal is given, a definite conclusion can be made about the animal group it belongs to. This is generalization in context of machine learning.

1.4.1.1 Abstraction

During the machine learning process, knowledge is fed in the form of input data. However, the data cannot be used in the original shape and form. As we saw in the example above, abstraction helps in deriving a conceptual map based on the input data. This map, or a **model** as it is known in the machine learning paradigm, is summarized knowledge representation of the raw data. The model may be in any one of the following forms

- Computational blocks like if/else rules
- Mathematical equations
- Specific data structures like trees or graphs
- Logical groupings of similar observations

The choice of the model used to solve a specific learning problem is a human task. The decision related to the choice of model is taken based on multiple aspects, some of which are listed below:

- The type of problem to be solved: Whether the problem is related to forecast or prediction, analysis of trend, understanding the different segments or groups of objects, etc.
- Nature of the input data: How exhaustive the input data is, whether the data has no values for many fields, the data types, etc.
- Domain of the problem: If the problem is in a business critical domain with a high rate of data input and need for immediate inference, e.g. fraud detection problem in banking domain.

Once the model is chosen, the next task is to fit the model based on the input data. Let's understand this with an example. In a case where the model is represented by a mathematical equation, say '$y = c_1 + c_2x$' (the model is known as simple linear regression which we will study in a later chapter), based on the input data, we have to find out the values of c_1 and c_2. Otherwise, the equation (or the model) is of no use. So, fitting the model, in this case, means finding the values of the unknown coefficients or constants of the equation or the model. This process of fitting the model based on the input data is known as training. Also, the input data based on which the model is being finalized is known as training data.

1.4.1.2 Generalization

The first part of machine learning process is abstraction i.e. abstract the knowledge which comes as input data in the form of a model. However, this abstraction process, or more popularly training the model, is just one part of machine learning. The other key part is to tune up the abstracted knowledge to a form which can be used to take future decisions. This is achieved as a part of generalization. This part is quite difficult to achieve. This is because the model is trained based on a finite set of data, which may possess a limited set of characteristics. But when we want to apply the model to take decision on a set of unknown data, usually termed as test data, we may encounter two problems:

1. The trained model is aligned with the training data too much, hence may not portray the actual trend.
2. The test data possess certain characteristics apparently unknown to the training data.

Hence, a precise approach of decision-making will not work. An approximate or heuristic approach, much like gut-feeling-based decision-making in human beings, has to be adopted. This approach has the risk of not making a correct decision – quite obviously because certain assumptions that are made may not be true in reality. But just like machines, same mistakes can be made by humans too when a decision is made based on intuition or gut-feeling – in a situation where exact reason-based decision-making is not possible.

1.4.2 Well-posed learning problem

For defining a new problem, which can be solved using machine learning, a simple framework, highlighted below, can be used. This framework also helps in deciding whether the problem is a right candidate to be solved using machine learning. The framework involves answering three questions:

1. What is the problem?
2. Why does the problem need to be solved?
3. How to solve the problem?

Step 1: What is the Problem?

A number of information should be collected to know what is the problem.

Informal description of the problem, e.g. I need a program that will prompt the next word as and when I type a word.

Formalism

Use Tom Mitchell's machine learning formalism stated above to define the T, P, and E for the problem.
For example:

- Task (T): Prompt the next word when I type a word.
- Experience (E): A corpus of commonly used English words and phrases.
- Performance (P): The number of correct words prompted considered as a percentage (which in machine learning paradigm is known as learning accuracy).

Assumptions - Create a list of assumptions about the problem.

Similar problems

What other problems have you seen or can you think of that are similar to the problem that you are trying to solve?

Step 2: Why does the problem need to be solved?

Motivation

What is the motivation for solving the problem? What requirement will it fulfil?

For example, does this problem solve any long-standing business issue like finding out potentially fraudulent transactions? Or the purpose is more trivial like trying to suggest some movies for upcoming weekend.

Solution benefits

Consider the benefits of solving the problem. What capabilities does it enable?

It is important to clearly understand the benefits of solving the problem. These benefits can be articulated to sell the project.

Solution use

How will the solution to the problem be used and the life time of the solution is expected to have?

Step 3: How would I solve the problem?

Try to explore how to solve the problem manually.

Detail out step-by-step data collection, data preparation, and program design to solve the problem. Collect all these details and update the previous sections of the problem definition, especially the assumptions.

Summary

- **Step 1: What is the problem?** Describe the problem informally and formally and list assumptions and similar problems.
- **Step 2: Why does the problem need to be solved?** List the motivation for solving the problem, the benefits that the solution will provide and how the solution will be used.
- **Step 3: How would I solve the problem?** Describe how the problem would be solved manually to flush domain knowledge.

Did you know?

Sony created a series of robotic pets called Aibo. It was built in 1998. Although most models sold were dog-like, other inspirations included lion-cubs. It could express emotions and could also recognize its owner. In 2006, Aibo was added to the Carnegie Mellon University's 'Robot Hall of Fame'. A new generation of Aibo was launched in Japan in January 2018.

1.5 TYPES OF MACHINE LEARNING

As highlighted in Figure 1.3, Machine learning can be classified into three broad categories:

1. Supervised learning – Also called predictive learning. A machine predicts the class of unknown objects based on prior class-related information of similar objects.
2. Unsupervised learning – Also called descriptive learning. A machine finds patterns in unknown objects by grouping similar objects together.
3. Reinforcement learning – A machine learns to act on its own to achieve the given goals.

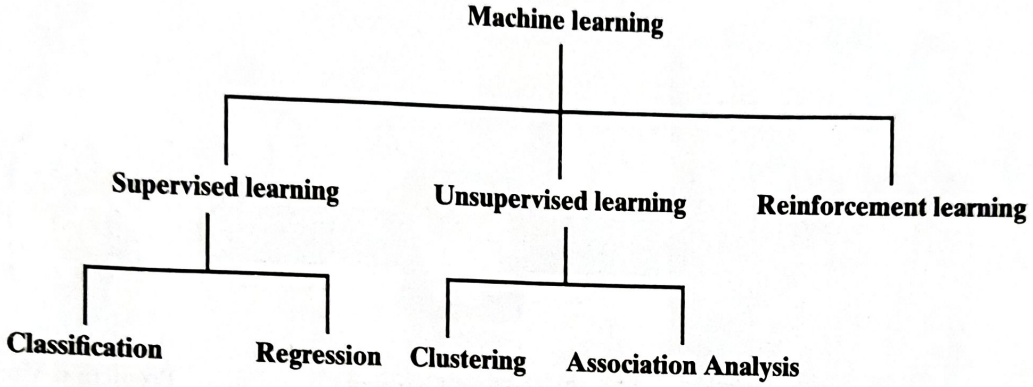

FIG. 1.3
Types of machine learning

Did you know?

Many video games are based on artificial intelligence technique called Expert System. This technique can imitate areas of human behaviour, with a goal to mimic the human ability of senses, perception, and reasoning.

1.5.1 Supervised learning

The major motivation of supervised learning is to learn from past information. So what kind of past information does the machine need for supervised learning? It is the information about the task which the machine has to execute. In context of the definition of machine learning, this past information is the experience. Let's try to understand it with an example.

Say a machine is getting images of different objects as input and the task is to segregate the images by either shape or colour of the object. If it is by shape, the images which are of round-shaped objects need to be separated from images of triangular-shaped objects, etc. If the segregation needs to happen based on colour, images of blue objects need to be separated from images of green objects. But how can the machine know what is round shape, or triangular shape? Same way, how can the machine distinguish image of an object based on whether it is blue or green in colour? A machine is very much like a little child whose parents or adults need to guide him with the basic information on shape and colour before he can start doing the task. A machine needs the basic

information to be provided to it. This basic input, or the experience in the paradigm of machine learning, is given in the form of **training data**. Training data is the past information on a specific task. In context of the image segregation problem, training data will have past data on different aspects or features on a number of images, along with a tag on whether the image is round or triangular, or blue or green in colour. The tag is called '**label**' and we say that the training data is labelled in case of supervised learning.

Figure 1.4 is a simple depiction of the supervised learning process. Labelled training data containing past information comes as an input. Based on the training data, the machine builds a predictive model that can be used on test data to assign a label for each record in the test data.

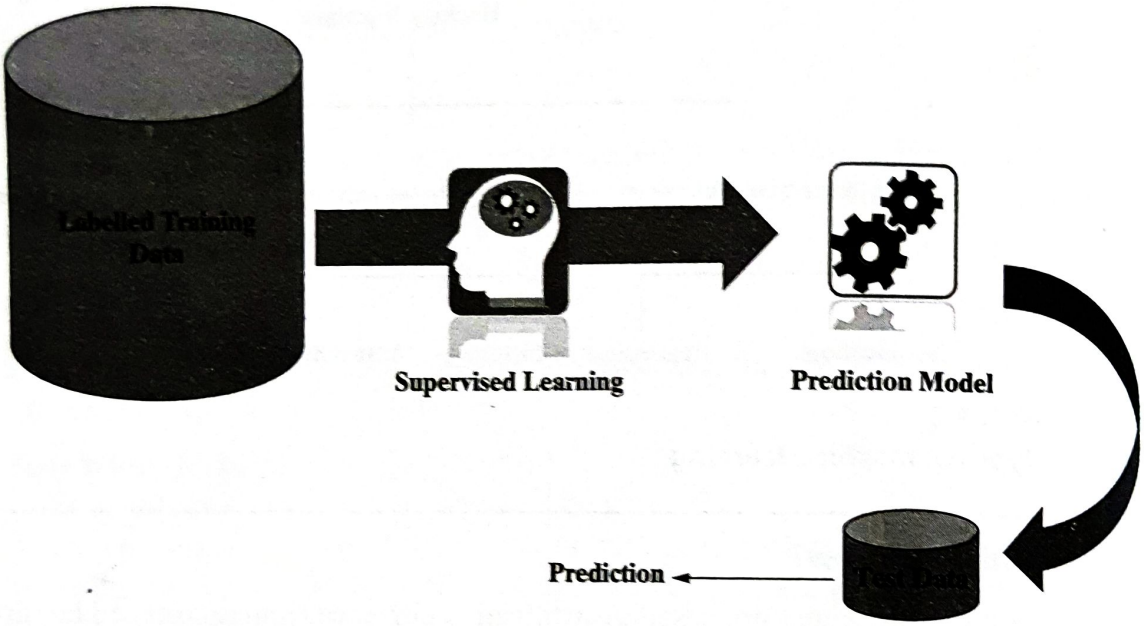

FIG. 1.4
Supervised learning

Some examples of supervised learning are

- Predicting the results of a game
- Predicting whether a tumour is malignant or benign
- Predicting the price of domains like real estate, stocks, etc.
- Classifying texts such as classifying a set of emails as spam or non-spam

Now, let's consider two of the above examples, say 'predicting whether a tumour is malignant or benign' and 'predicting price of domains such as real estate'. Are these two problems same in nature? The answer is 'no'. Though both of them are prediction problems, in one case we are trying to predict which category or class an unknown data belongs to whereas in the other case we are trying to predict an absolute value and not a class. When we are trying to predict a categorical or nominal variable, the problem is known as a **classification** problem. Whereas when we are trying to predict a real-valued variable, the problem falls under the category of **regression**.

Note:

Supervised machine learning is as good as the data used to train it. If the training data is of poor quality, the prediction will also be far from being precise.

Let's try to understand these two areas of supervised learning, i.e. classification and regression in more details.

1.5.1.1 Classification

Let's discuss how to segregate the images of objects based on the shape. If the image is of a round object, it is put under one category, while if the image is of a triangular object, it is put under another category. In which category the machine should put an image of unknown category, also called a **test data** in machine learning parlance, depends on the information it gets from the past data, which we have called as training data. Since the training data has a label or category defined for each and every image, the machine has to map a new image or test data to a set of images to which it is similar to and assign the same label or category to the test data.

So we observe that the whole problem revolves around assigning a label or category or class to a test data based on the label or category or class information that is imparted by the training data. Since the target objective is to assign a class label, this type of problem as classification problem. Figure 1.5 depicts the typical process of classification.

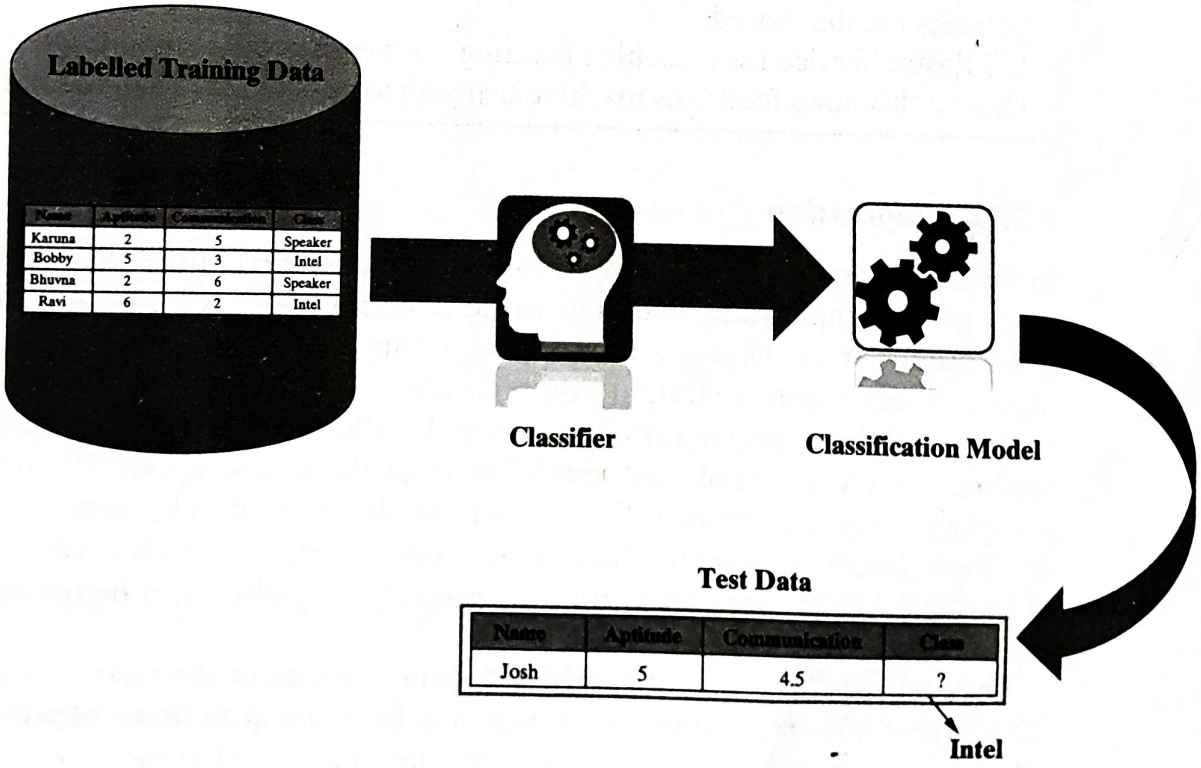

FIG. 1.5
Classification

There are number of popular machine learning algorithms which help in solving classification problems. To name a few, Naïve Bayes, Decision tree, and k-Nearest Neighbour algorithms are adopted by many machine learning practitioners.

A critical classification problem in context of banking domain is identifying potential fraudulent transactions. Since there are millions of transactions which have to be scrutinized and assured whether it might be a fraud transaction, it is not possible for any human being to carry out this task. Machine learning is effectively leveraged to do this task and this is a classic case of classification. Based on the past transaction data, specifically the ones labelled as fraudulent, all new incoming transactions are marked or labelled as normal or suspicious. The suspicious transactions are subsequently segregated for a closer review.

In summary, classification is a type of supervised learning where a target feature, which is of type categorical, is predicted for test data based on the information imparted by training data. The target categorical feature is known as **class**.

Some typical classification problems include:

- Image classification
- Prediction of disease
- Win–loss prediction of games
- Prediction of natural calamity like earthquake, flood, etc.
- Recognition of handwriting

Did you know?

- Machine learning saves life – ML can spot 52% of breast cancer cells, a year before patients are diagnosed.
- US Postal Service uses machine learning for handwriting recognition.
- Facebook's news feed uses machine learning to personalize each member's feed.

1.5.1.2. Regression

In linear regression, the objective is to predict numerical features like real estate or stock price, temperature, marks in an examination, sales revenue, etc. The underlying predictor variable and the target variable are continuous in nature. In case of linear regression, a straight line relationship is 'fitted' between the predictor variables and the target variables, using the statistical concept of least squares method. As in the case of least squares method, the sum of square of error between actual and predicted values of the target variable is tried to be minimized. In case of simple linear regression, there is only one predictor variable whereas in case of multiple linear regression, multiple predictor variables can be included in the model.

Let's take the example of yearly budgeting exercise of the sales managers. They have to give sales prediction for the next year based on sales figure of previous years vis-à-vis investment being put in. Obviously, the data related to past as well as the data to be predicted are continuous in nature. In a basic approach, a simple linear regression model can be applied with investment as predictor variable and sales revenue as the target variable.

Figure 1.6 shows a typical simple regression model, where regression line is fitted based on values of target variable with respect to different values of predictor variable. A typical linear regression model can be represented in the form –

$$y = \alpha + \beta x$$

where 'x' is the predictor variable and 'y' is the target variable.

The input data come from a famous multivariate data set named Iris introduced by the British statistician and biologist Ronald Fisher. The data set consists of 50 samples from each of three species of Iris – *Iris setosa*, *Iris virginica,* and *Iris versicolor*. Four features were measured for each sample – sepal length, sepal width, petal length, and petal width. These features can uniquely discriminate the different species of the flower.

The Iris data set is typically used as a training data for solving the classification problem of predicting the flower species based on feature values. However, we can also demonstrate regression using this data set, by predicting the value of one feature using another feature as predictor. In Figure 1.6, petal length is a predictor variable which, when fitted in the simple linear regression model, helps in predicting the value of the target variable sepal length.

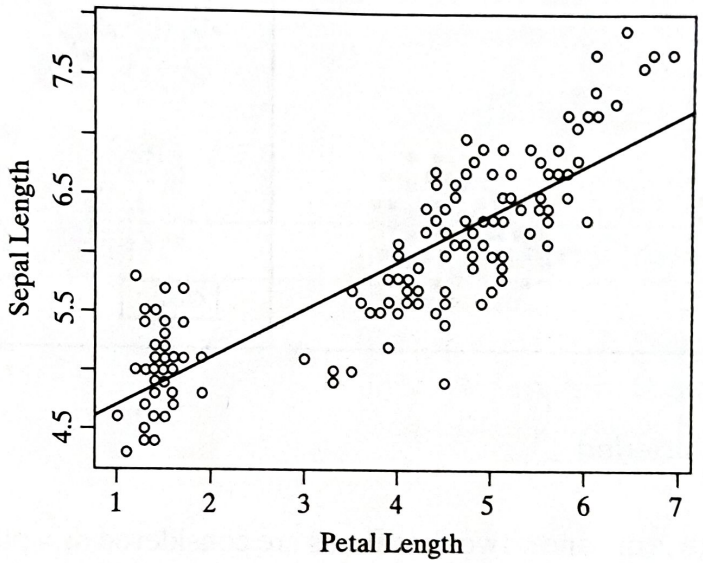

FIG. 1.6
Regression

Typical applications of regression can be seen in

- Demand forecasting in retails
- Sales prediction for managers
- Price prediction in real estate
- Weather forecast
- Skill demand forecast in job market

1.5.2 Unsupervised learning

Unlike supervised learning, in unsupervised learning, there is no labelled training data to learn from and no prediction to be made. In unsupervised learning, the objective is to take a dataset as input and try to find natural groupings or **patterns** within the data

elements or records. Therefore, unsupervised learning is often termed as **descriptive model** and the process of unsupervised learning is referred as **pattern discovery** or **knowledge discovery**. One critical application of unsupervised learning is customer segmentation.

Clustering is the main type of unsupervised learning. It intends to group or organize similar objects together. For that reason, objects belonging to the same cluster are quite similar to each other while objects belonging to different clusters are quite dissimilar. Hence, the objective of clustering to discover the intrinsic grouping of unlabelled data and form clusters, as depicted in Figure 1.7. Different measures of similarity can be applied for clustering. One of the most commonly adopted similarity

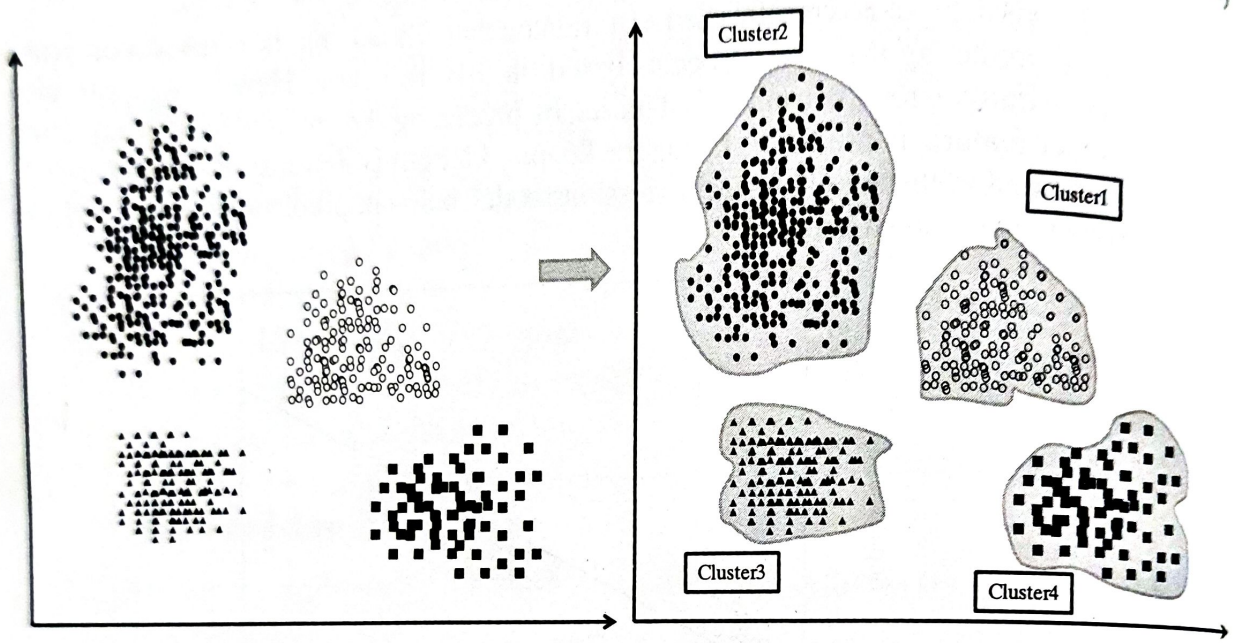

FIG. 1.7
Distance-based clustering

measure is distance. Two data items are considered as a part of the same cluster if the distance between them is less. In the same way, if the distance between the data items is high, the items do not generally belong to the same cluster. This is also known as distance-based clustering. Figure 1.8 depicts the process of clustering at a high level.

Other than clustering of data and getting a summarized view from it, one more variant of unsupervised learning is **association analysis**. As a part of association analysis, the association between data elements is identified. Let's try to understand the approach of association analysis in context of one of the most common examples, i.e. market basket analysis as shown in Figure 1.9. From past transaction data in a grocery store, it may be observed that most of the customers who have bought item A, have also bought item B and item C or at least one of them. This means that there is a strong association of the event 'purchase of item A' with the event 'purchase of item B', or 'purchase of item C'. Identifying these sorts of associations is the goal of association analysis. This helps in boosting up sales pipeline, hence a critical input for the sales group. Critical applications of association analysis include market basket analysis and recommender systems.

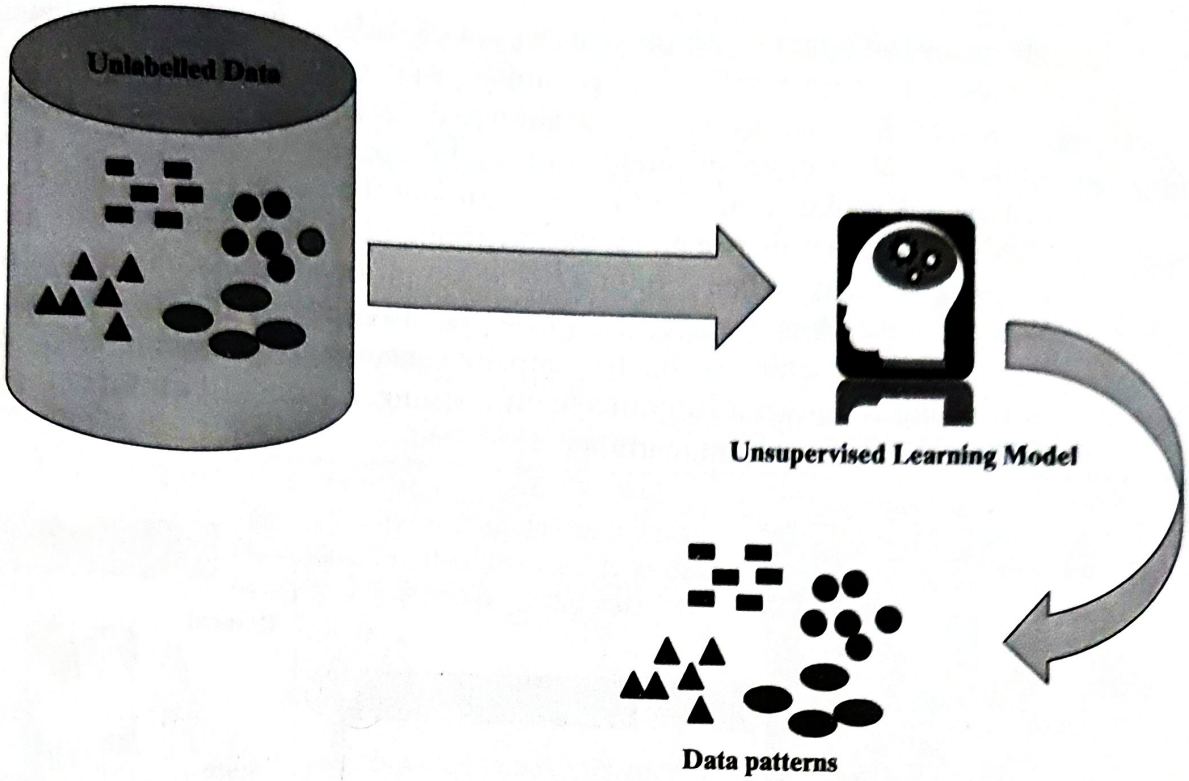

FIG. 1.8
Unsupervised learning

TransID	Items Bought
1	{Butter, Bread}
2	{Diaper, Bread, Milk, Beer}
3	{Milk, Chicken, Beer, Diaper}
4	{Bread, Diaper, Chicken, Beer}
5	{Diaper, Beer, Cookies, Ice cream}
...	...

Market Basket transactions
Frequent itemsets → (Diaper, Beer)
Possible association: Diaper → Beer

FIG. 1.9
Market basket analysis

1.5.3 Reinforcement learning

We have seen babies learn to walk without any prior knowledge of how to do it. Often we wonder how they really do it. They do it in a relatively simple way.

First they notice somebody else walking around, for example parents or anyone living around. They understand that legs have to be used, one at a time, to take a step. While walking, sometimes they fall down hitting an obstacle, whereas other times they are able to walk smoothly avoiding bumpy obstacles. When they are able to walk overcoming the obstacle, their parents are elated and appreciate the baby with loud claps / or may be a chocolates. When they fall down while circumventing an obstacle,

obviously their parents do not give claps or chocolates. Slowly a time comes when the babies learn from mistakes and are able to walk with much ease.

In the same way, machines often learn to do tasks autonomously. Let's try to understand in context of the example of the child learning to walk. The action tried to be achieved is walking, the child is the agent and the place with hurdles on which the child is trying to walk resembles the environment. It tries to improve its performance of doing the task. When a sub-task is accomplished successfully, a reward is given. When a sub-task is not executed correctly, obviously no reward is given. This continues till the machine is able to complete execution of the whole task. This process of learning is known as reinforcement learning. Figure 1.10 captures the high-level process of reinforcement learning.

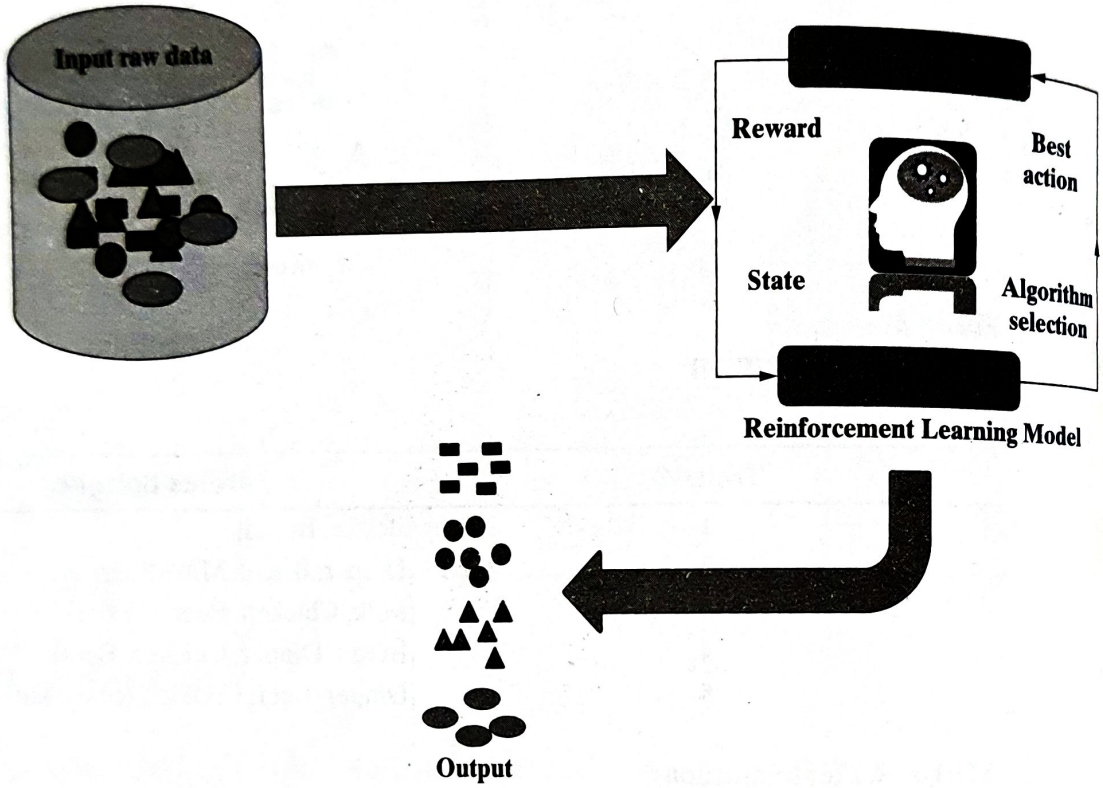

FIG. 1.10
Reinforcement learning

One contemporary example of reinforcement learning is self-driving cars. The critical information which it needs to take care of are speed and speed limit in different road segments, traffic conditions, road conditions, weather conditions, etc. The tasks that have to be taken care of are start/stop, accelerate/decelerate, turn to left / right, etc.

Further details on reinforcement learning have been kept out of the scope of this book.

Points to ponder:

- Reinforcement learning is getting more and more attention from both industry and academia. Annual publications count in the area of reinforcement learning in Google Scholar support this view.

- While Deep Blue used brute force to defeat the human chess champion, AlphaGo used RL to defeat the best human Go player.
- RL is an effective tool for personalized online marketing. It considers the demographic details and browsing history of the user real-time to show most relevant advertisements.

■

1.5.4 Comparison – supervised, unsupervised, and reinforcement learning

SUPERVISED	UNSUPERVISED	REINFORCEMENT
This type of learning is used when you know how to classify a given data, or in other words classes or labels are available.	This type of learning is used when there is no idea about the class or label of a particular data. The model has to find pattern in the data.	This type of learning is used when there is no idea about the class or label of a particular data. The model has to do the classification – it will get rewarded if the classification is correct, else get punished.
Labelled training data is needed. Model is built based on training data.	Any unknown and unlabelled data set is given to the model as input and records are grouped.	The model learns and updates itself through reward/ punishment.
The model performance can be evaluated based on how many misclassifications have been done based on a comparison between predicted and actual values.	Difficult to measure whether the model did something useful or interesting. Homogeneity of records grouped together is the only measure.	Model is evaluated by means of the reward function after it had some time to learn.
There are two types of supervised learning problems – classification and regression.	There are two types of unsupervised learning problems – clustering and association.	No such types.
Simplest one to understand.	More difficult to understand and implement than supervised learning.	Most complex to understand and apply.
Standard algorithms include • Naïve Bayes • *k*-nearest neighbour (kNN) • Decision tree • Linear regression • Logistic regression • Support Vector Machine SVM), etc.	Standard algorithms are • *k*-means • Principal Component Analysis (PCA) • Self-organizing map (SOM) • Apriori algorithm • DBSCAN etc.	Standard algorithms are • Q-learning • Sarsa
Practical applications include • Handwriting recognition • Stock market prediction • Disease prediction • Fraud detection, etc.	Practical applications include • Market basket analysis • Recommender systems • Customer segmentation, etc.	Practical applications include • Self-driving cars • Intelligent robots • AlphaGo Zero (the latest version of DeepMind's AI system playing Go)

1.6 PROBLEMS NOT TO BE SOLVED USING MACHINE LEARNING

Machine learning should not be applied to tasks in which humans are very effective or frequent human intervention is needed. For example, air traffic control is a very complex task needing intense human involvement. At the same time, for very simple tasks which can be implemented using traditional programming paradigms, there is no sense of using machine learning. For example, simple rule-driven or formula-based applications like price calculator engine, dispute tracking application, etc. do not need machine learning techniques.

Machine learning should be used only when the business process has some lapses. If the task is already optimized, incorporating machine learning will not serve to justify the return on investment.

For situations where training data is not sufficient, machine learning cannot be used effectively. This is because, with small training data sets, the impact of bad data is exponentially worse. For the quality of prediction or recommendation to be good, the training data should be sizeable.

1.7 APPLICATIONS OF MACHINE LEARNING

Wherever there is a substantial amount of past data, machine learning can be used to generate actionable insight from the data. Though machine learning is adopted in multiple forms in every business domain, we have covered below three major domains just to give some idea about what type of actions can be done using machine learning.

1.7.1 Banking and finance

In the banking industry, fraudulent transactions, especially the ones related to credit cards, are extremely prevalent. Since the volumes as well as velocity of the transactions are extremely high, high performance machine learning solutions are implemented by almost all leading banks across the globe. The models work on a real-time basis, i.e. the fraudulent transactions are spotted and prevented right at the time of occurrence. This helps in avoiding a lot of operational hassles in settling the disputes that customers will otherwise raise against those fraudulent transactions.

Customers of a bank are often offered lucrative proposals by other competitor banks. Proposals like higher bank interest, lower processing charge of loans, zero balance savings accounts, no overdraft penalty, etc. are offered to customers, with the intent that the customer switches over to the competitor bank. Also, sometimes customers get demotivated by the poor quality of services of the banks and shift to competitor banks. Machine learning helps in preventing or at least reducing the customer churn. Both descriptive and predictive learning can be applied for reducing customer churn. Using descriptive learning, the specific pockets of problem, i.e. a specific bank or a specific zone or a specific type of offering like car loan, may be spotted where maximum churn is happening. Quite obviously, these are troubled areas where further investigation needs to be done to find and fix the root cause. Using predictive learning, the set of vulnerable customers who may leave the bank very soon, can be identified. Proper action can be taken to make sure that the customers stay back.

1.7.2 Insurance

Insurance industry is extremely data intensive. For that reason, machine learning is extensively used in the insurance industry. Two major areas in the insurance industry where machine learning is used are risk prediction during new customer onboarding and claims management. During customer onboarding, based on the past information the risk profile of a new customer needs to be predicted. Based on the quantum of risk predicted, the quote is generated for the prospective customer. When a customer claim comes for settlement, past information related to historic claims along with the adjustor notes are considered to predict whether there is any possibility of the claim to be fraudulent. Other than the past information related to the specific customer, information related to similar customers, i.e. customer belonging to the same geographical location, age group, ethnic group, etc., are also considered to formulate the model.

1.7.3 Healthcare

Wearable device data form a rich source for applying machine learning and predict the health conditions of the person real time. In case there is some health issue which is predicted by the learning model, immediately the person is alerted to take preventive action. In case of some extreme problem, doctors or healthcare providers in the vicinity of the person can be alerted. Suppose an elderly person goes for a morning walk in a park close to his house. Suddenly, while walking, his blood pressure shoots up beyond a certain limit, which is tracked by the wearable. The wearable data is sent to a remote server and a machine learning algorithm is constantly analyzing the streaming data. It also has the history of the elderly person and persons of similar age group. The model predicts some fatality unless immediate action is taken. Alert can be sent to the person to immediately stop walking and take rest. Also, doctors and healthcare providers can be alerted to be on standby.

Machine learning along with computer vision also plays a crucial role in disease diagnosis from medical imaging.

1.8 STATE-OF-THE-ART LANGUAGES/TOOLS IN MACHINE LEARNING

The algorithms related to different machine learning tasks are known to all and can be implemented using any language/platform. It can be implemented using a Java platform or C / C++ language or in .NET. However, there are certain languages and tools which have been developed with a focus for implementing machine learning. Few of them, which are most widely used, are covered below.

1.8.1 Python

Python is one of the most popular, open source programming language widely adopted by machine learning community. It was designed by Guido van Rossum and was first released in 1991. The reference implementation of Python, i.e. CPython, is managed by Python Software Foundation, which is a non-profit organization.

Python has very strong libraries for advanced mathematical functionalities (NumPy), algorithms and mathematical tools (SciPy) and numerical plotting (matplotlib). Built on these libraries, there is a machine learning library named **scikit-learn**, which has various classification, regression, and clustering algorithms embedded in it.

1.8.2 R

R is a language for statistical computing and data analysis. It is an open source language, extremely popular in the academic community – especially among statisticians and data miners. R is considered as a variant of S, a GNU project which was developed at Bell Laboratories. Currently, it is supported by the R Foundation for statistical computing.

R is a very simple programming language with a huge set of libraries available for different stages of machine learning. Some of the libraries standing out in terms of popularity are plyr/dplyr (for data transformation), caret ('Classification and Regression Training' for classification), RJava (to facilitate integration with Java), tm (for text mining), ggplot2 (for data visualization). Other than the libraries, certain packages like Shiny and R Markdown have been developed around R to develop interactive web applications, documents and dashboards on R without much effort.

1.8.3 Matlab

MATLAB (matrix laboratory) is a licenced commercial software with a robust support for a wide range of numerical computing. MATLAB has a huge user base across industry and academia. MATLAB is developed by MathWorks, a company founded in 1984. Being proprietary software, MATLAB is developed much more professionally, tested rigorously, and has comprehensive documentation.

MATLAB also provides extensive support of statistical functions and has a huge number of machine learning algorithms in-built. It also has the ability to scale up for large datasets by parallel processing on clusters and cloud.

1.8.4 SAS

SAS (earlier known as 'Statistical Analysis System') is another licenced commercial software which provides strong support for machine learning functionalities. Developed in C by SAS Institute, SAS had its first release in the year 1976.

SAS is a software suite comprising different components. The basic data management functionalities are embedded in the Base SAS component whereas the other components like SAS/INSIGHT, Enterprise Miner, SAS/STAT, etc. help in specialized functions related to data mining and statistical analysis.

1.8.5 Other languages/tools

There are a host of other languages and tools that also support machine learning functionalities. Owned by IBM, **SPSS** (originally named as Statistical Package for

the Social Sciences) is a popular package supporting specialized data mining and statistical analysis. Originally popular for statistical analysis in social science (as the name reflects), SPSS is now popular in other fields as well.

Released in 2012, **Julia** is an open source, liberal licence programming language for numerical analysis and computational science. It has baked in all good things of MATLAB, Python, R, and other programming languages used for machine learning for which it is gaining steady attention from machine learning development community. Another big point in favour of Julia is its ability to implement high-performance machine learning algorithms.

1.9 ISSUES IN MACHINE LEARNING

Machine learning is a field which is relatively new and still evolving. Also, the level of research and kind of use of machine learning tools and technologies varies drastically from country to country. The laws and regulations, cultural background, emotional maturity of people differ drastically in different countries. All these factors make the use of machine learning and the issues originating out of machine learning usage are quite different.

The biggest fear and issue arising out of machine learning is related to privacy and the breach of it. The primary focus of learning is on analyzing data, both past and current, and coming up with insight from the data. This insight may be related to people and the facts revealed might be private enough to be kept confidential. Also, different people have a different preference when it comes to sharing of information. While some people may be open to sharing some level of information publicly, some other people may not want to share it even to all friends and keep it restricted just to family members. Classic examples are a birth date (not the day, but the date as a whole), photographs of a dinner date with family, educational background, etc. Some people share them with all in the social platforms like Facebook while others do not, or if they do, they may restrict it to friends only. When machine learning algorithms are implemented using those information, inadvertently people may get upset. For example, if there is a learning algorithm to do preference-based customer segmentation and the output of the analysis is used for sending targeted marketing campaigns, it will hurt the emotion of people and actually do more harm than good. In certain countries, such events may result in legal actions to be taken by the people affected.

Even if there is no breach of privacy, there may be situations where actions were taken based on machine learning may create an adverse reaction. Let's take the example of knowledge discovery exercise done before starting an election campaign. If a specific area reveals an ethnic majority or skewness of a certain demographic factor, and the campaign pitch carries a message keeping that in mind, it might actually upset the voters and cause an adverse result.

So a very critical consideration before applying machine learning is that proper human judgement should be exercised before using any outcome from machine learning. Only then the decision taken will be beneficial and also not result in any adverse impact.

1.10 SUMMARY

- Machine learning imbibes the philosophy of human learning, i.e. learning from expert guidance and from experience.
- The basic machine learning process can be divided into three parts.
  - ❖ Data Input: Past data or information is utilized as a basis for future decision-making.
  - ❖ Abstraction: The input data is represented in a summarized way
  - ❖ Generalization: The abstracted representation is generalized to form a framework for making decisions.
- Before starting to solve any problem using machine learning, it should be decided whether the problem is a right problem to be solved using machine learning.
- Machine learning can be classified into three broad categories:
  - ❖ Supervised learning: Also called predictive learning. The objective of this learning is to predict class/value of unknown objects based on prior information of similar objects. Examples: predicting whether a tumour is malignant or benign, price prediction in domains such as real estate, stocks, etc.
  - ❖ Unsupervised learning: Also called descriptive learning, helps in finding groups or patterns in unknown objects by grouping similar objects together. Examples: customer segmentation, recommender systems, etc.
  - ❖ Reinforcement learning: A machine learns to act on its own to achieve the given goals. Examples: self-driving cars, intelligent robots, etc.
- Machine learning has been adopted by various industry domains such as Banking and Financial Services, Insurance, Healthcare, Life Sciences, etc. to solve problems.
- Some of the most adopted platforms to implement machine learning include Python, R, MATLAB, SAS, SPPSS, etc.
- To avoid ethical issues, the critical consideration is required before applying machine learning and using any outcome from machine learning.

SAMPLE QUESTIONS

MULTIPLE-CHOICE QUESTIONS (1 MARK EACH):

1. Machine learning is ___ field.
(a) Inter-disciplinary
(b) Single
(c) Multi-disciplinary
(d) All of the above

2. A computer program is said to learn from _____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 E.
(a) Training
(b) Experience
(c) Database
(d) Algorithm

3. _____ has been used to train vehicles to steer correctly and autonomously on road.
   (a) Machine learning
   (b) Data mining
   (c) Neural networks
   (d) Robotics

4. Any hypothesis find an approximation of the target function over a sufficiently large set of training examples will also approximate the target function well over other unobserved examples. This is called _____.
   (a) Hypothesis
   (b) Inductive hypothesis
   (c) Learning
   (d) Concept learning

5. Factors which affect performance of a learner system does not include
   (a) Representation scheme used
   (b) Training scenario
   (c) Type of feedback
   (d) Good data structures

6. Different learning methods does not include
   (a) Memorization
   (b) Analogy
   (c) Deduction
   (d) Introduction

7. A model of language consists of the categories which does not include
   (a) Language units
   (b) Role structure of units
   (c) System constraints
   (d) Structural units

8. How many types are available in machine learning?
   (a) 1
   (b) 2
   (c) 3
   (d) 4

9. The *k*-means algorithm is a
   (a) Supervised learning algorithm
   (b) Unsupervised learning algorithm
   (c) Semi-supervised learning algorithm
   (d) Weakly supervised learning algorithm

10. The Q-learning algorithm is a
    (a) Supervised learning algorithm
    (b) Unsupervised learning algorithm
    (c) Semi-supervised learning algorithm
    (d) Reinforcement learning algorithm

11. This type of learning to be used when there is no idea about the class or label of a particular data
    (a) Supervised learning algorithm
    (b) Unsupervised learning algorithm
    (c) Semi-supervised learning algorithm
    (d) Reinforcement learning algorithm

12. The model learns and updates itself through reward/punishment in case of
    (a) Supervised learning algorithm
    (b) Unsupervised learning algorithm
    (c) Semi-supervised learning algorithm
    (d) Reinforcement learning algorithm