# MACHINE LEARNING

# CHAPTER 1

## Machine Learning Landscape

# What is Human Learning ?Types of Human Learning

## 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 keeps 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 launches. Thus, with more learning, tasks can be performed more efficiently.

## 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
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.

## 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 newcomers 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.

## **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 the 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

## 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 checklist on things that we should do, and things that we should not do, based on our experiences.

# 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.

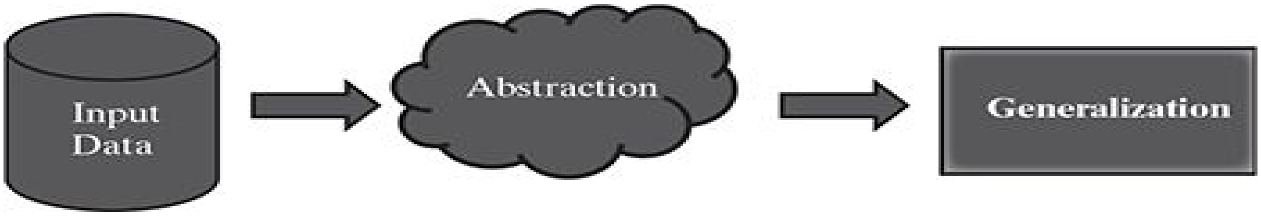
The first step in any project is defining your problem. Even if the most powerful algorithm is used, the results will be meaningless if the wrong problem is solved.

## 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

schematic representation of the machine learning process.



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 the typical process of learning from the 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
    1. Fishes: Always live in water and lay eggs
    2. Amphibians: Semi-aquatic i.e. may live in water or land; smooth skin; lay eggs
    3. Reptiles: Semi-aquatic like amphibians; scaly skin; lay eggs; cold-blooded
    4. Birds: Can fly; lay eggs; warm-blooded
    5. 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 the context of machine learning.

## 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 '$y = c1 + c2x$' (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 c1 and c2. 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.

## Generalization

The first part of the 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.

## 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 the problem is.

**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 is 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 the 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 lifetime 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.

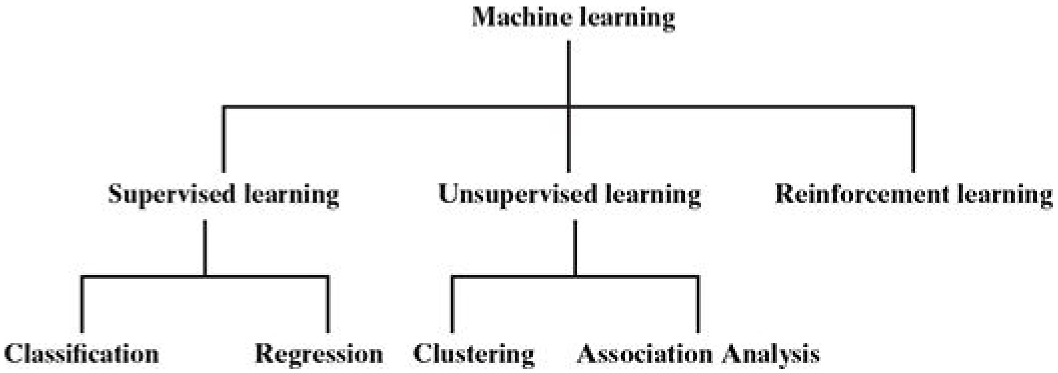
Summary

- **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.

## TYPES OF MACHINE LEARNING

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.



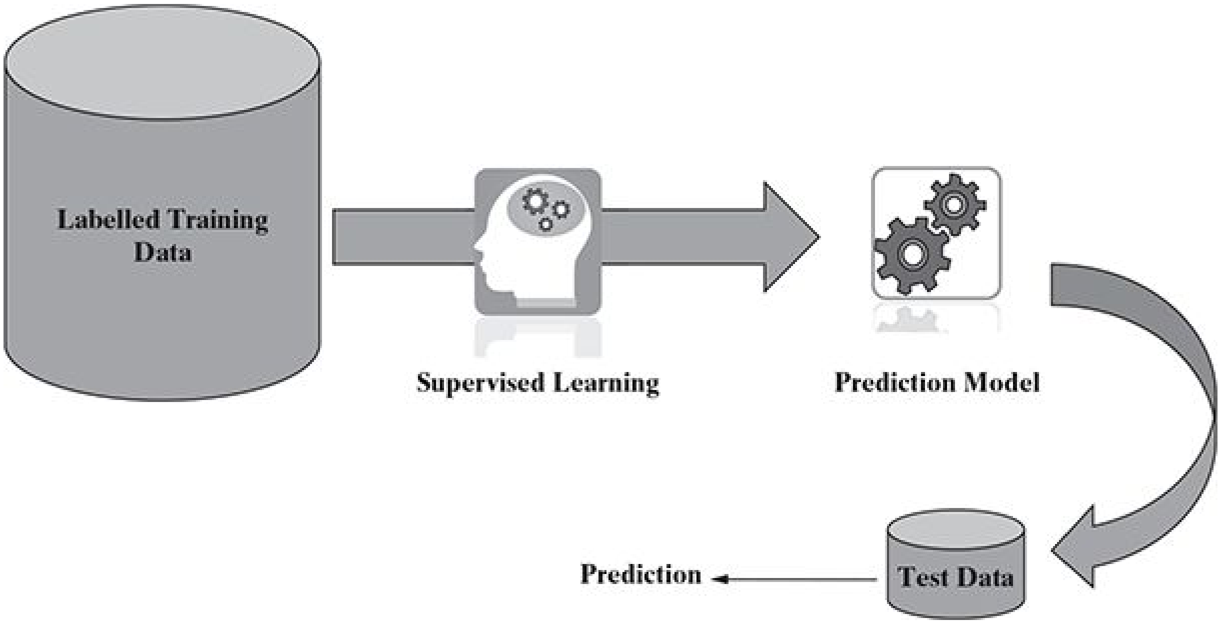
**Types of machine learning**

## 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 the 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.

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.



**Supervised learning**

Some examples of supervised learning are

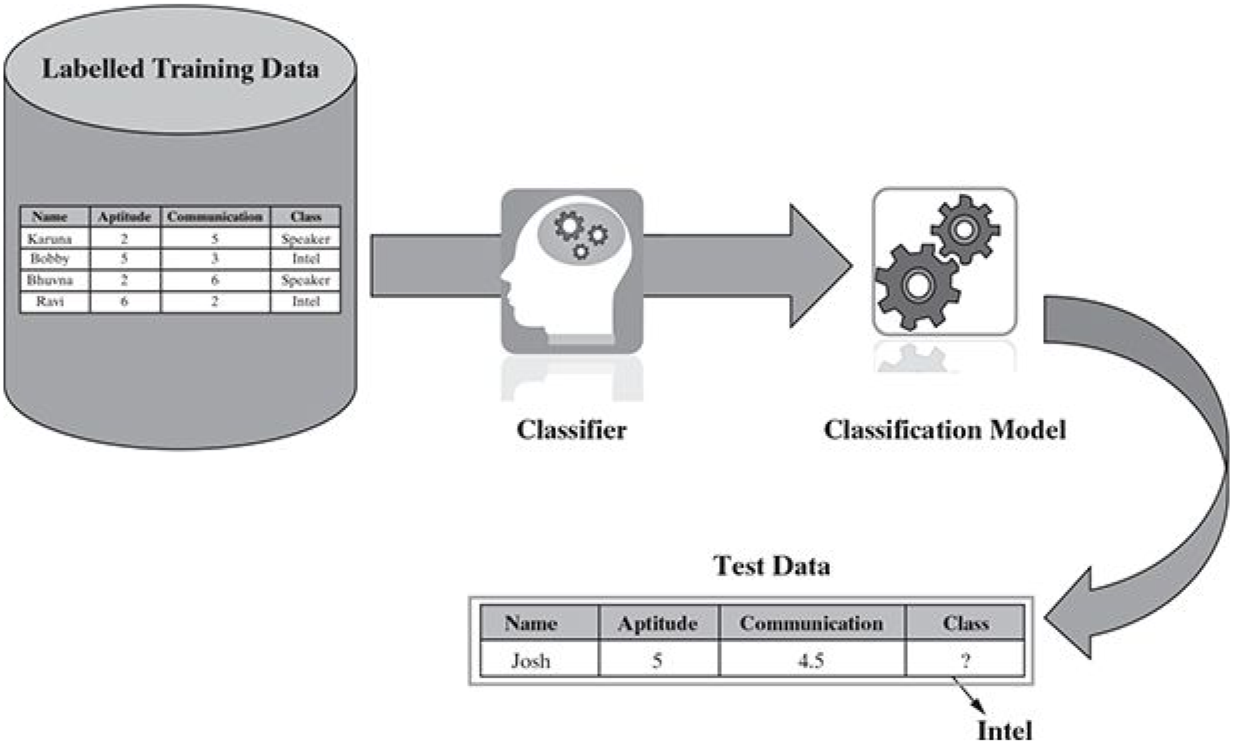
- Predicting the results of a game
- Predicting whether a tumour is malignant or benign
- Predicting the price of domains like real estate, stocks, etc.
- Classifying texts such as classifying a set of emails as spam or non-spam

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 the 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.

## 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 an 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 is a classification problem. Figure 1.5 depicts the typical process of classification.



| Name | Aptitude | Communication | Class |
|---|---|---|---|
| Karuna | 2 | 5 | Speaker |
| Bobby | 5 | 3 | Intel |
| Bhuvna | 2 | 6 | Speaker |
| Ravi | 6 | 2 | Intel |

| Name | Aptitude | Communication | Class |
|---|---|---|---|
| Josh | 5 | 4.5 | ? |

**Classification**

There are a 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 the 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

## 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 squares 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 predictions for the next year based on the sales figure of previous years vis-à-vis investment being put in. Obviously, the data related to the 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.

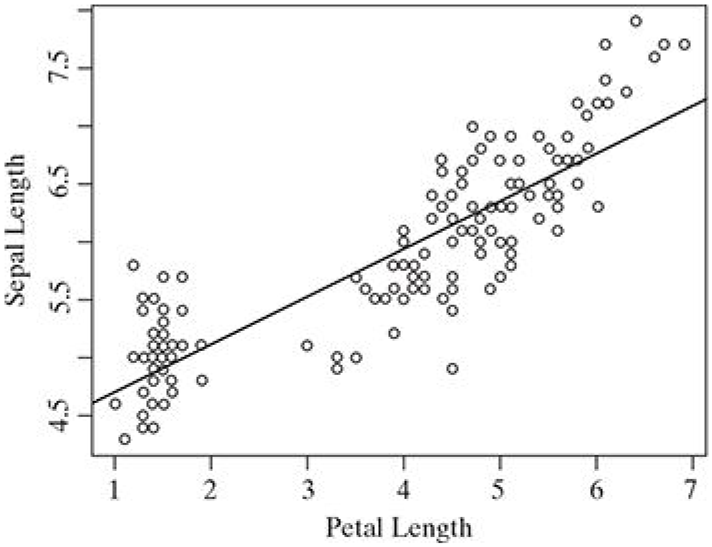
shows a typical simple regression model, where the regression line is fitted based on values of the target variable with respect to different values of the predictor variable. A typical linear regression model can be represented in the form –

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

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.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.



Regression

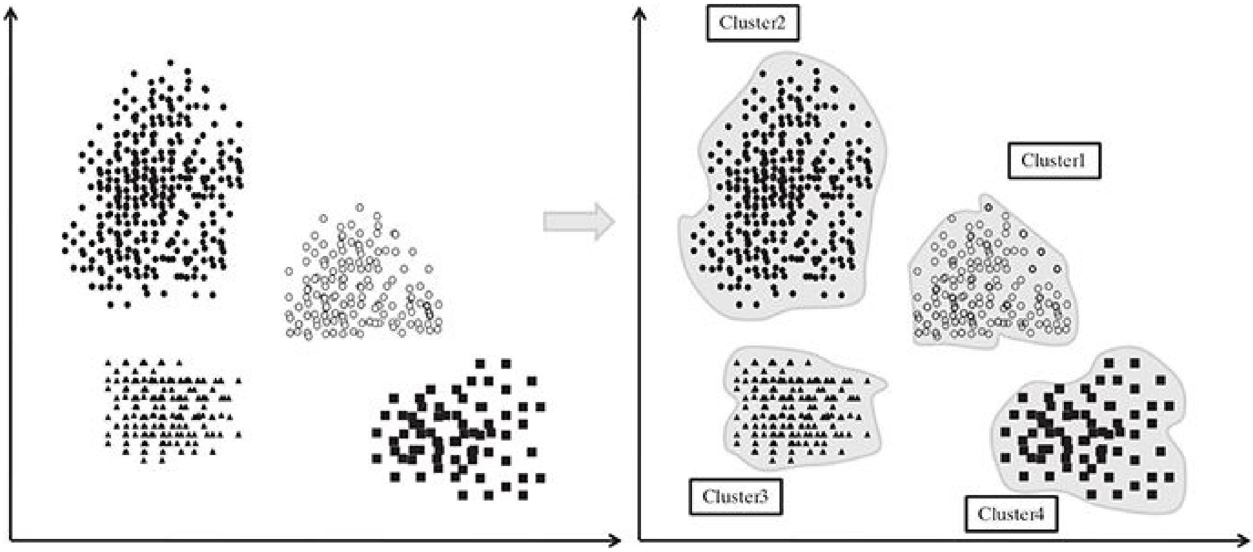
Typical applications of regression can be seen in

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

## Unsupervised learning

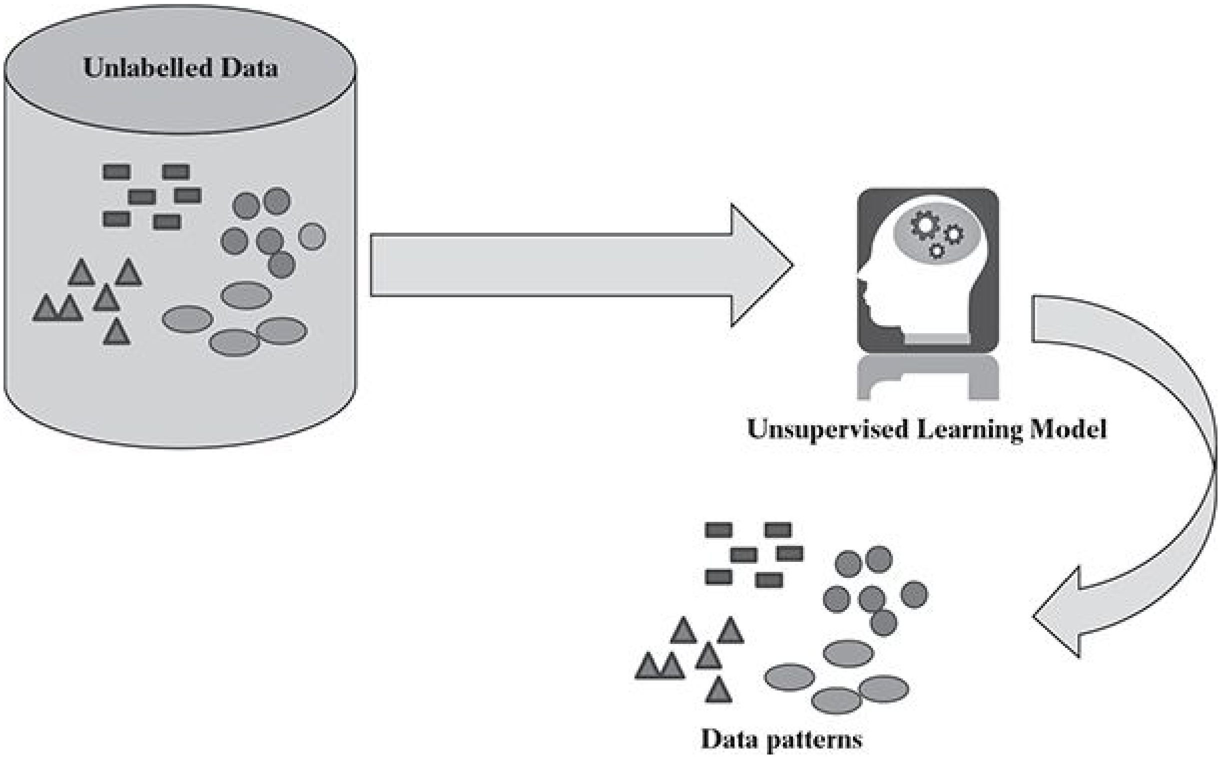
Unlike supervised learning, in unsupervised learning, there is no labelled training data to learn from and no prediction to be made. In unsupervised learning, the objective is to take a dataset as input and try to find natural groupings or **patterns** within the data elements or records. Therefore,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

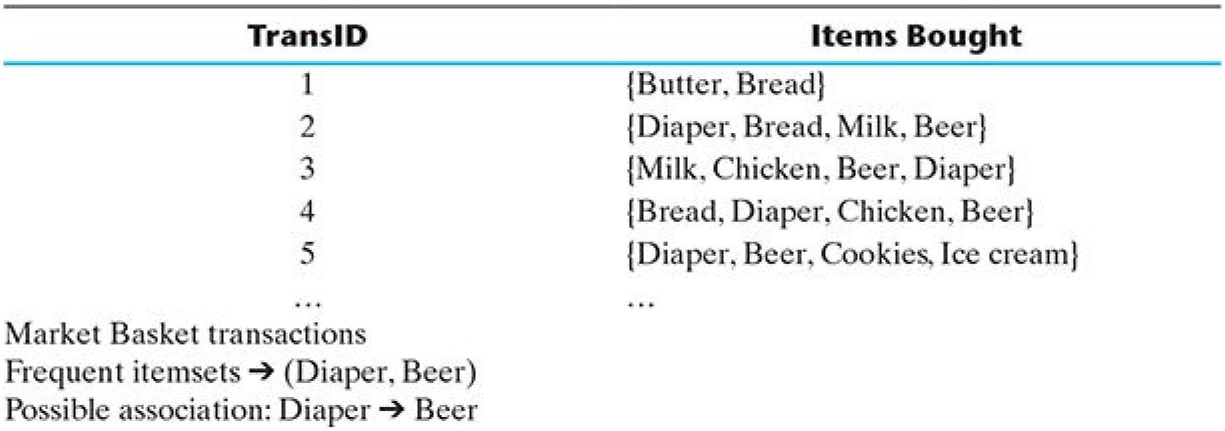


Distance-based clustering

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 . 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 the sales pipeline, hence a critical input for the sales group. Critical applications of association analysis include market basket analysis and recommender systems.



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

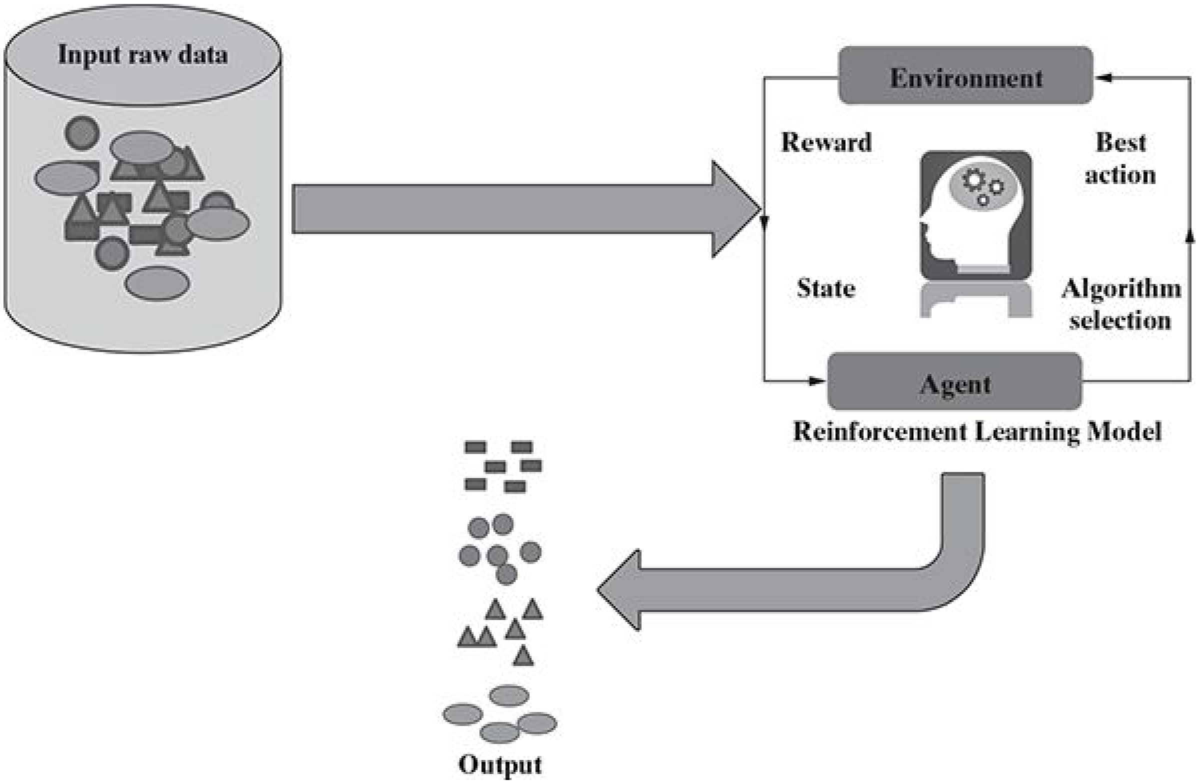
Market basket analysis

## 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 over the obstacle, their parents are elated and appreciate the baby with loud claps / or maybe 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.

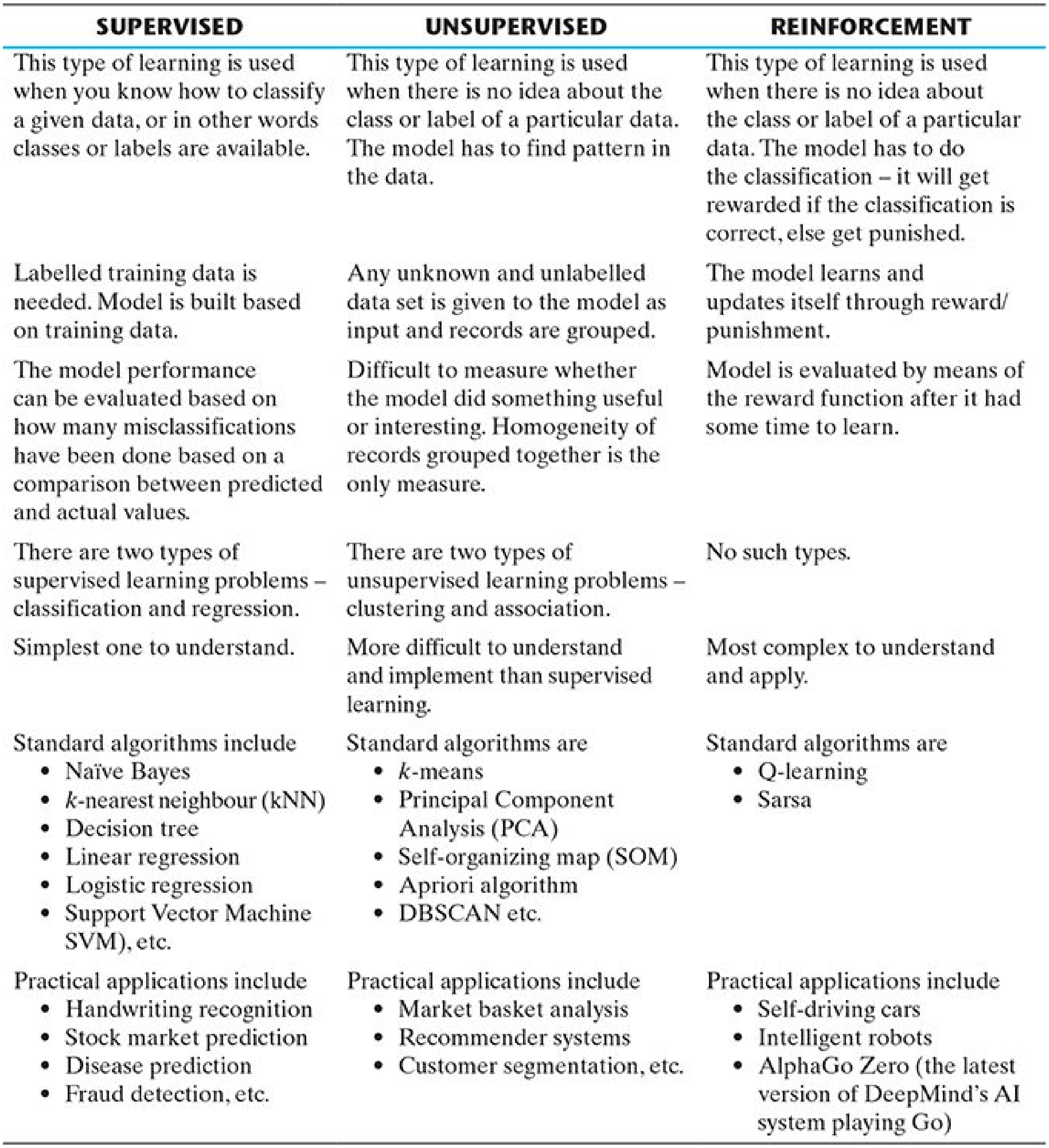
In the same way, machines often learn to do tasks autonomously. Let’s try to understand in context of the example of the child learning to walk. The action tried to be achieved is walking, the child is the agent and the place with hurdles on which the child is trying to walk resembles the environment. It tries to improve its performance by 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. captures the high-level process of reinforcement learning.



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.

## 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<br>• Naïve Bayes<br>• *k*-nearest neighbour (kNN)<br>• Decision tree<br>• Linear regression<br>• Logistic regression<br>• Support Vector Machine SVM), etc. | Standard algorithms are<br>• *k*-means<br>• Principal Component Analysis (PCA)<br>• Self-organizing map (SOM)<br>• Apriori algorithm<br>• DBSCAN etc. | Standard algorithms are<br>• Q-learning<br>• Sarsa |
| Practical applications include<br>• Handwriting recognition<br>• Stock market prediction<br>• Disease prediction<br>• Fraud detection, etc. | Practical applications include<br>• Market basket analysis<br>• Recommender systems<br>• Customer segmentation, etc. | Practical applications include<br>• Self-driving cars<br>• Intelligent robots<br>• AlphaGo Zero (the latest version of DeepMind's AI system playing Go) |

# MACHINE LEARNING ACTIVITIES

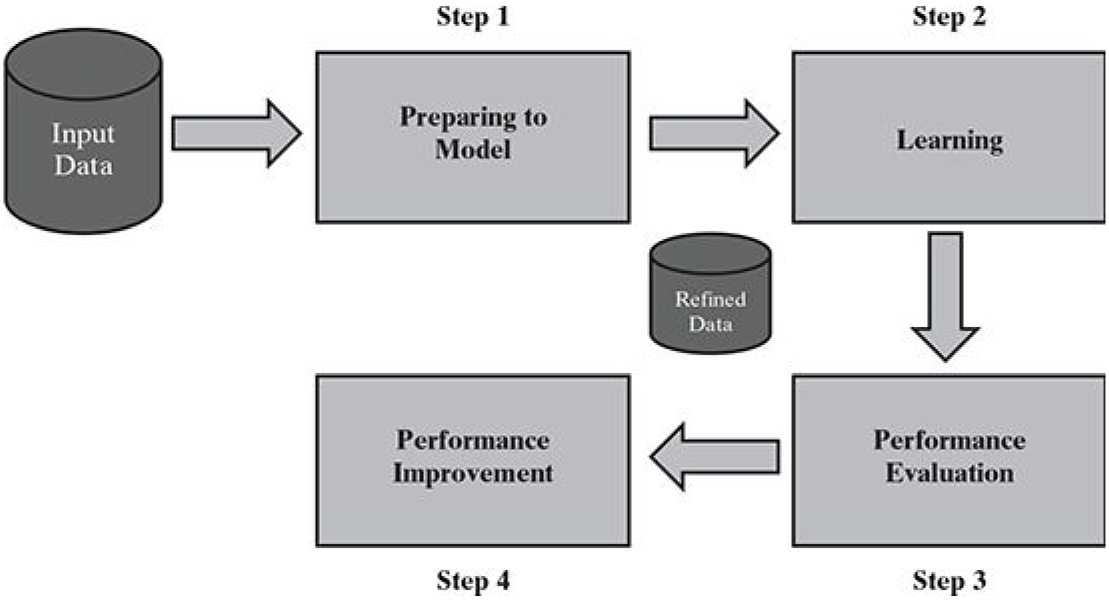
The first step in machine learning activity starts with data. In case of supervised learning, it is the labelled training data set followed by test data which is not labelled. In case of unsupervised learning, there is no question of labelled data but the task is to find patterns in the input data. A thorough review and exploration of the data is needed to understand the type of the data, the quality of the data and relationship between the different data elements. Based on that, multiple pre-processing activities may need to be done on the input data before we can go ahead with core machine learning activities. Following are the typical **preparation** activities done once the input data comes into the machine learning system:

- Understand the type of data in the given input data set.
- Explore the data to understand nature and quality.
- Explore the relationships amongst the data elements, e.g. inter-feature relationship.
- Find potential issues in data.
- Do the necessary remediation, e.g. impute missing data values, etc., if needed.
- Apply pre-processing steps, as necessary.
- Once the data is prepared for modelling, then the learning tasks start off. As a part of it, do the following activities:

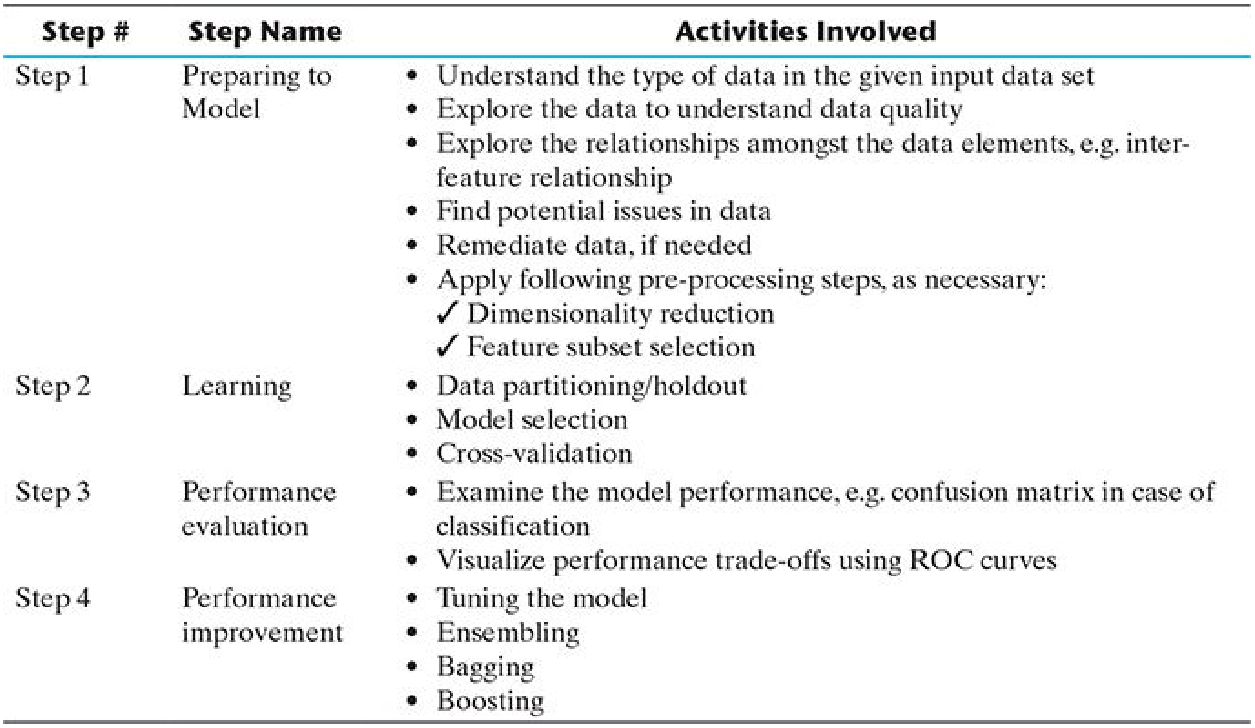
- ○ The input data is first divided into parts – the training data and the test data (called holdout). This step is applicable for supervised learning only.
- ○ Consider different models or learning algorithms for selection.
- ○ Train the model based on the training data for supervised learning problems and apply to unknown data. Directly apply the chosen unsupervised model on the input data for unsupervised learning problems.

After the model is selected, trained (for supervised learning), and applied on input data, the performance of the model is evaluated. Based on options available, specific actions can be taken to improve the performance of the model, if possible.

Depicts the four-step process of machine learning.



**Activities in Machine Learning**

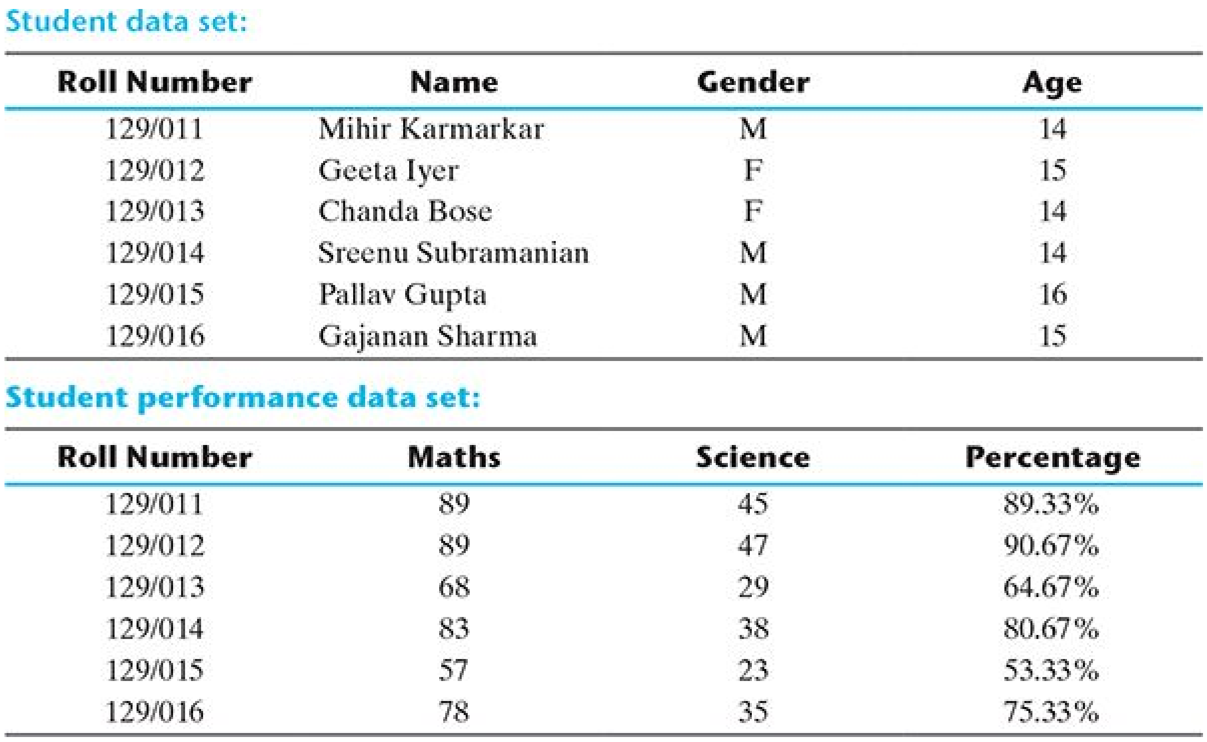


| Step # | Step Name | Activities Involved |
|---|---|---|
| Step 1 | Preparing to Model | • Understand the type of data in the given input data set<br>• Explore the data to understand data quality<br>• Explore the relationships amongst the data elements, e.g. inter-feature relationship<br>• Find potential issues in data<br>• Remediate data, if needed<br>• Apply following pre-processing steps, as necessary:<br>✓ Dimensionality reduction<br>✓ Feature subset selection |
| Step 2 | Learning | • Data partitioning/holdout<br>• Model selection<br>• Cross-validation |
| Step 3 | Performance evaluation | • Examine the model performance, e.g. confusion matrix in case of classification<br>• Visualize performance trade-offs using ROC curves |
| Step 4 | Performance improvement | • Tuning the model<br>• Ensembling<br>• Bagging<br>• Boosting |

## BASIC TYPES OF DATA IN MACHINE LEARNING

Before starting with types of data, let’s first understand what a data set is and what are the elements of a data set. A data set is a collection of related information or records. The information may be on some entity or some subject area. we may have a data set on students in which each record consists of information about a specific student. Again, we can have a data set on student performance which has records providing performance, i.e. marks on the individual subjects.

Each row of a data set is called a record. Each data set also has multiple attributes, each of which gives information on a specific characteristic. For example, in the data set on students, there are four attributes namely Roll Number, Name, Gender, and Age, each of which understandably is a specific characteristic about the student entity. Attributes can also be termed as feature, variable, dimension or field. Both the data sets, Student and Student Performance, are having four features or dimensions; hence they are told to have four-dimensional data space. A row or record represents a point in the four-dimensional data space as each row has specific values for each of the four attributes or features. Value of an attribute, quite understandably, may vary from record to record. For example, if we refer to the first two records in the Student data set, the value of attributes Name, Gender, and Age are different.



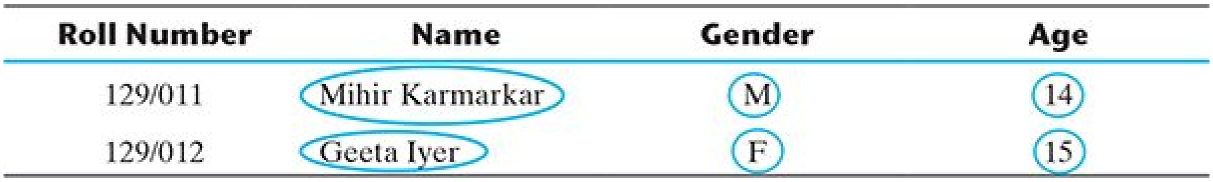
Student data set:

| Roll Number | Name | Gender | Age |
|---|---|---|---|
| 129/011 | Mihir Karmarkar | M | 14 |
| 129/012 | Geeta Iyer | F | 15 |
| 129/013 | Chanda Bose | F | 14 |
| 129/014 | Sreenu Subramanian | M | 14 |
| 129/015 | Pallav Gupta | M | 16 |
| 129/016 | Gajanan Sharma | M | 15 |

Student performance data set:

| Roll Number | Maths | Science | Percentage |
|---|---|---|---|
| 129/011 | 89 | 45 | 89.33% |
| 129/012 | 89 | 47 | 90.67% |
| 129/013 | 68 | 29 | 64.67% |
| 129/014 | 83 | 38 | 80.67% |
| 129/015 | 57 | 23 | 53.33% |
| 129/016 | 78 | 35 | 75.33% |

Examples of data set



| Roll Number | Name | Gender | Age |
|---|---|---|---|
| 129/011 | Mihir Karmarkar | M | 14 |
| 129/012 | Geeta Iyer | F | 15 |

Data set records and attributes

Now that a context of data sets is given, let’s try to understand the different types of data that we generally come across in machine learning problems. Data can broadly be divided into following two types:

1. Qualitative data
2. Quantitative data

**Qualitative data** provides information about the quality of an object or information which cannot be measured. For example, if we consider the quality of performance of students in terms of ‘Good’, ‘Average’, and ‘Poor’, it falls under the category of qualitative data. Also, the name or roll number of students are information that cannot be measured using some scale of measurement. So they would fall under qualitative data. Qualitative data is also called **categorical data**. Qualitative data can be further subdivided into two types as follows:

1. Nominal data
2. Ordinal data

**Nominal data** is one which has no numeric value, but a named value. It is used for assigning named values to attributes. Nominal values cannot be quantified. Examples of nominal data are

1. Blood group: A, B, O, AB, etc.
2. Nationality: Indian, American, British, etc.
3. Gender: Male, Female, Other

**Ordinal data**, in addition to possessing the properties of nominal data, can also be naturally ordered. This means ordinal data also assigns named values to attributes but unlike nominal data, they can be arranged in a sequence of increasing or decreasing value so that we can say whether a value is better than or greater than another value. Examples of ordinal data are

1. Customer satisfaction: ‘Very Happy’, ‘Happy’, ‘Unhappy’, etc.
2. Grades: A, B, C, etc.
3. Hardness of Metal: ‘Very Hard’, ‘Hard’, ‘Soft’, etc.

Like nominal data, basic counting is possible for ordinal data. Hence, the mode can be identified. Since ordering is possible in case of ordinal data, median, and quartiles can be identified in addition. Mean can still not be calculated.

**Quantitative data** relates to information about the quantity of an object – hence it can be measured. For example, if we consider the attribute ‘marks’, it can be measured using a scale of measurement. Quantitative data is also termed as numeric data. There are two types of quantitative data:

1. Interval data
2. Ratio data

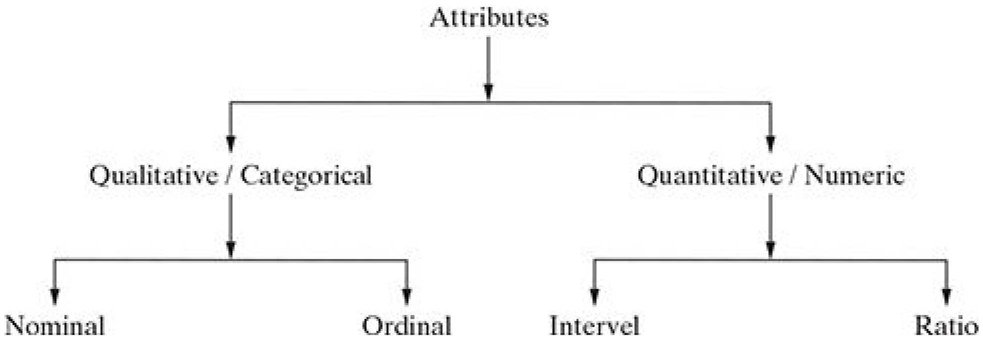
**Interval data** is numeric data for which not only the order is known, but the exact difference between values is also known. An ideal example of interval data is Celsius temperature. The difference between each value remains the same in Celsius temperature. For example, the difference between 12°C and 18°C degrees is measurable and is 6°C as in the case of difference between 15.5°C and 21.5°C. Other examples include date, time, etc.

For interval data, mathematical operations such as addition and subtraction are possible. For that reason, for interval data, the central tendency can be measured by mean, median, or mode. Standard deviation can also be calculated.

However, interval data do not have something called a 'true zero' value. For example, there is nothing called '0 temperature' or 'no temperature'. Hence, only addition and subtraction applies for interval data. The ratio cannot be applied. This means, we can say a temperature of 40°C is equal to the temperature of 20°C + temperature of 20°C. However, we cannot say the temperature of 40°C means it is twice as hot as a temperature of 20°C.

**Ratio data** represents numeric data for which exact value can be measured. Absolute zero is available for ratio data. Also, these variables can be added, subtracted, multiplied, or divided. The central tendency can be measured by mean, median, or mode and methods of dispersion such as standard deviation. Examples of ratio data include height, weight, age, salary, etc.

summarized view of different types of data that we may find in a typical machine learning problem.



Apart from the approach detailed above, attributes can also be categorized into types based on a number of values that can be assigned. The attributes can be either discrete or continuous based on this factor.

Discrete attributes can assume a finite or countably infinite number of values. Nominal attributes such as roll number, street number, pin code, etc. can have a finite number of values whereas numeric attributes such as count, rank of students, etc. can have countably infinite values. A special type of discrete attribute which can assume two values only is called binary attribute. Examples of binary attributes include male/ female, positive/negative, yes/no, etc.

Continuous attributes can assume any possible value which is a real number. Examples of continuous attribute include length, height, weight, price, etc.