## Foundations of Machine Learning 7Hrs

Introduction to Artificial Intelligence, What is machine learning? , Applications of Machine learning, Understand Data, Types of machine learning: Supervised, Unsupervised, Reinforcement Learning, Theory of learning: feasibility of learning, error and noise, training versus testing, theory of generalization, bias and variance, learning curve

Section 1

### Introduction to Artificial Intelligence

Artificial Intelligence refers to the technique involved with an objective to make computers intelligent. The Intelligence that is expressed with the behavior of computer while performing any task. Computer is machine which simply follows the instructions already fed by the programmer. The kind of job that a computer could do was mainly to assist humans popularly it did numerical computation. The first program we learn to write is add two numbers and so on. Later it moved to many other applications like database systems, web. Traditional programs are logic based or algorithm based which takes input and then applies the algorithm based on certain logic. But every problem may not have a definite answer. For example given two numbers let us say 8 and 5 what's the sum? Answer is 13 and it is always 13 irrespective of the situation, condition etc. But there are few kind of problems which has no definite answer and need some analysis, decision making , to anwer the question. For example, who will win in IPL? That style of computing is now changing with the introduction of artificial intelligence. Now, we want to use computers not just to assist humans but to replace humans. The machines which should have the characteristics like, Reasoning, Learning, Problem Solving, Perception, Linguistic Intelligence etc. AI is mainly used to understand and mimic the natural intelligence of human in taking decisions. The focus is now on the situations or the data available. Data becomes the main source for working of such intelligence.

### Difference between Human and Machine Intelligence

- Humans perceive by patterns whereas the machines perceive by set of rules and data.
- Humans store and recall information by patterns, machines do it by searching algorithms. For example, the number 40404040 is easy to remember, store, and recall as its pattern is simple.
- Humans can figure out the complete object even if some part of it is missing or distorted; whereas the machines cannot do it correctly.

Artificial intelligence is a science and technology based on disciplines such as Computer Science, Biology, Psychology, Linguistics, Mathematics, and Engineering.

Applications of AI: Gaming, NLP, Expert Systems, Vision Systems, Speech Recognition, Handwriting recognition, Face Recognition, Intelligent Robots

**Artificial Intelligence and machine Learning**

Machine learning is one of the approach to implement Artificial Intelligence.
Machine Learning is a current application of AI based around the idea that we should really just be able to give machines access to data and let them learn for themselves.

Machine leanring has become the very popular approach and may soon cover all the applications of AI
Two reasons behind the rise of Machine Learning, (Author: Bernard Marr)

1. One of these was the realization – credited to Arthur Samuel in 1959 – that rather than teaching computers everything they need to know about the world and how to carry out tasks, it might be possible to teach them to learn for themselves. (we have become lazy to teach)
2. The second, more recently, was the emergence of the internet, and the huge increase in the amount of digital information being generated, stored, and made available for analysis. (data available everywhere)

**Section 2**

**Introduction to Machine Leanring**

> Machine Learning is an approach for Artificial Intelligence where the goal is to use example data (considered as past experience) to solve a given problem.
>
> Machine learns continuously,

a machine learns whenever it changes its structure, program, or data (based on its inputs or in response to external information) in such a manner that its expected future performance improves.

Arthur Samuel (1959). Machine Learning: Field of study that gives computers the ability to learn without being explicitly programmed. •
Tom Mitchell (1998) Well-posed Learning Problem: A computer program is said to learn from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E.

**Section 3**
**Applications of Machine Learning**

1. Self driving cars
2. collaborative filtering. Internet book-stores such as Amazon, or video rental sites such as Netflix use this informa-tion extensively to entice users to purchase additional goods
3. automatic translation of documents.
4. Speech Recognition

5. Fraud detection

## Section 4
## Understand Data

In Machine Learning, "It's not who has the best algorithm that wins, it's who has the most data"

1. Select Data: text, Audio, Images, Video, Structured Data (Tables), Data generated in sensors
2. Preprocess Data: Formatting, Cleaning and Sampling
3. Transform Data: Scaling, Decomposition, Aggregation (depending on the types of algorithm)

It is useful to characterize learning problems according to the type of data they use. This is a great help when encountering new challenges, since quite often problems on similar data types can be solved with very similar techniques.

1. **Vectors** constitute the most basic entity we might encounter in our work. For instance, a life insurance company might be interesting in obtaining the vector of variables (blood pressure, heart rate, height, weight, cholesterol level, smoker, gender) to infer the life expectancy of a potential customer. A farmer might be interested in determining the ripeness of fruit based on (size, weight, spectral data).
2. **Lists:** In some cases the vectors we obtain may contain a variable number of features. For instance, a physician might not necessarily decide to perform a full battery of diagnostic tests if the patient appears to be healthy.
3. **Sets** may appear in learning problems whenever there is a large number of potential causes of an e_ect, which are not well determined. For instance, it is relatively easy to obtain data concerning the toxicity of mushrooms. It would be desirable to use such data to infer the toxicity of a new mushroom given information about its chemical compounds. However, mushrooms contain a cocktail of compounds out of which one or more may be toxic. Consequently we need to infer the properties of an object given a set of features, whose composition and number may vary considerably.
4. **Matrices** are a convenient means of representing pairwise relationships. For instance, in collaborative _ltering applications the rows of the matrix may represent users whereas the columns correspond to products. Only in some cases we will have knowledge about a given (user, product) combination, such as the rating of the product by a user.
5. **Images** could be thought of as two dimensional arrays of numbers, that is, matrices. This representation is very crude, though, since they exhibit spatial coherence (lines, shapes) and (natural images exhibit) a multiresolution structure. That is, downsampling an image leads to an object which has very similar statistics to the original image. Computer vision and psychooptics have created a raft of tools for describing these phenomena.
6. **Video** adds a temporal dimension to images. Again, we could represent them as a three dimensional array. Good algorithms, however, take the temporal coherence of the image sequence into account.

7. **Trees and Graphs** are often used to describe relations between collections of objects. Both examples above describe estimation problems where our observations are vertices of a tree or graph. However, graphs themselves may be the observations. For instance, the DOM-tree of a webpage, the call-graph of a computer program, or the protein-protein interaction networks may form the basis upon which we may want to perform inference.
8. **Strings** occur frequently, mainly in the area of bioinformatics and natural language processing. They may be the input to our estimation problems, e.g. when classifying an e-mail as spam, when attempting to locate all names of persons and organizations in a text, or when modeling the topic structure of a document. Equally well they may constitute the output of a system.
9. **Compound structures** are the most commonly occurring object. That is, in most situations we will have a structured mix of different data types. For instance, a webpage might contain images, text, tables, which in turn contain numbers, and lists, all of which might constitute nodes on a graph of webpages linked among each other.

**Define Machine Learning ( Mathematically)**

Algorithms learn from data automatically. Data is represented as vector, matrix etc. Let us say we have a input **X** = (x1, x2, … , xn) with n-dimension that is with n number of parameters and we have a output **Y.** The learning is to learn the relationship between X and Y that is an optimized function f which leads to Y=f(X). Y represents the expected output. Sometimes we may not have the expected output.

Input X → called as features

Output Y → called as target variable

The set of data is available.

## Section 5

## Types of Machine Learning

1. Supervised Learning: In **supervised learning**, each example is a pair consisting of an input object (typically a vector) and a desired output value (also called the supervisory signal).
    a. A **supervised learning** algorithm analyzes the training data and produces an inferred function, which can be used for mapping new examples.
        i. Classification
        ii. Regression

2. Un Supervised Learning: **Unsupervised learning** is a type of **machine learning** algorithm used to draw inferences from datasets consisting of input data without labeled responses.
    a. The most common **unsupervised learning** method is cluster analysis, which is used for exploratory data analysis to find hidden patterns or grouping in data.

3. **Reinforcement learning** is often used for robotics, gaming and navigation.
    a. With reinforcement learning, the algorithm discovers through trial and error which actions yield the greatest rewards.
    b. This type of learning has three primary components: the agent (the learner or decision maker), the environment (everything the agent interacts with) and actions (what the agent can do).
    c. The objective is for the agent to choose actions that maximize the expected reward over a given amount of time.
    d. The agent will reach the goal much faster by following a good policy.
    e. So the goal in reinforcement learning is to learn the best policy.

**Supervised Learning**

Let us consider an example developing an ML application for predicting the health condition of people as Healthy (H) or Not-Healthy (NH). It is an example of classification.
First we need to think of the parameters that can be given as input
To make it simple, let us consider the two parameters, Height and Weight for deciding the person is Healthy or Not-healthy
**So input X is of two dimension [x1 x2]** $\in \mathbb{R}$
**Y is one dimension and holds a discrete values** $\in$ **(1,2)**
Pair (X, Y) is the training example

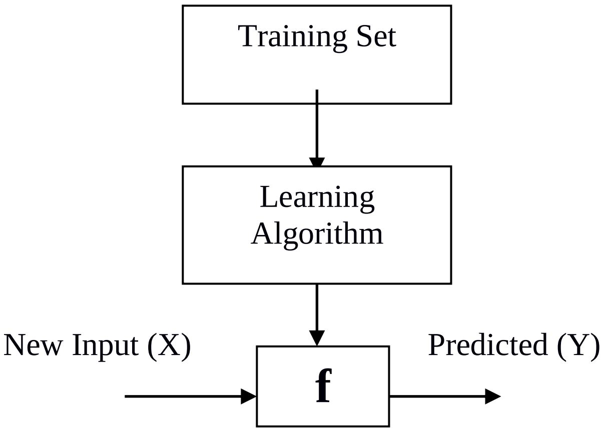
Let us say, we have population of data which contains the height and weight information along with health condition of students in a class.
We retrieve a sample from the population given as below,

| | Height x1 | Weight x2 | Health Condition | Target Variable |
|---|---|---|---|---|
| X1 | 162 | 45 | NH | 2 (Y1) |
| X2 | 154 | 50 | H | 1 (Y2) |
| X3 | 170 | 85 | H | 1 (Y3) |
| X4 | 150 | 60 | H | 1 (Y4) |
| X5 | 175 | 50 | NH | 2 (Y5) |
| X6 | 160 | 55 | NH | 2 (Y6) |
| X7 | 161 | 48 | ? | ? (Y7)<br>Actually known as (1) |
| X8 | 153 | 56 | ? | ?(Y8)<br>Actually known as (2) |
| X9 | 160 | 48 | ? | ?(Y9) |

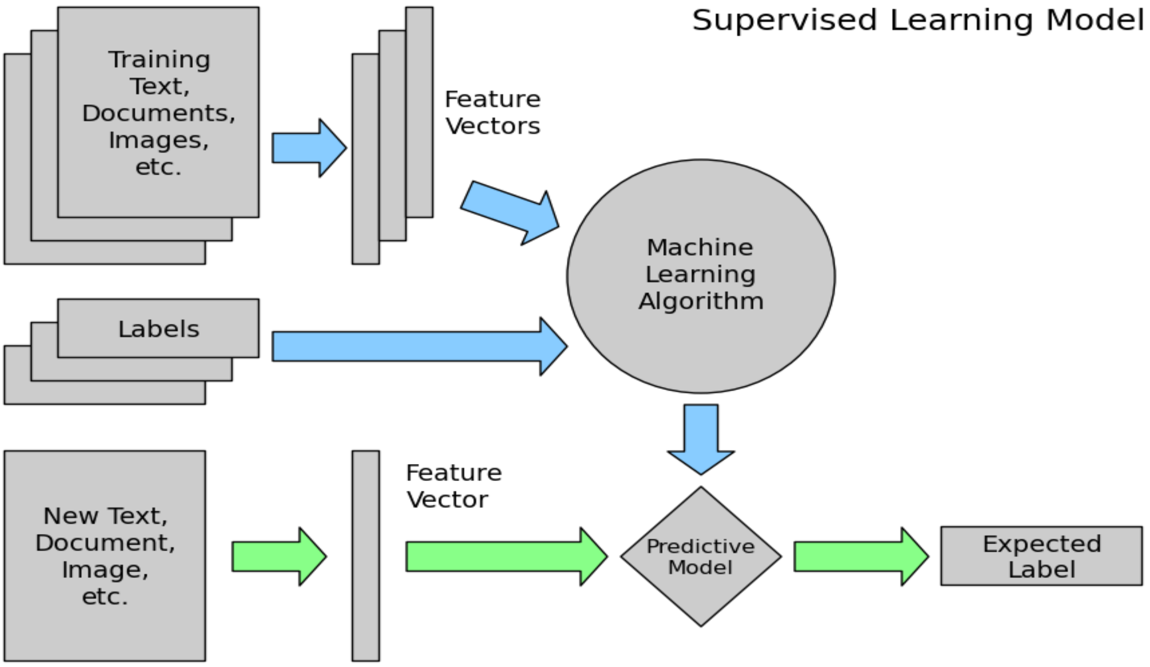
| | | | | |
|---|---|---|---|---|
| | | | | Actually known as (1) |
| X10 | 155 | 59 | ? | ?Y10<br>Actually known as (2) |

Dataset is (X1, Y1), (X2, Y2), … , (XN, YN)
Target function y-=f(x) where f is unknown
Set of all possible f is also referred as Hypothesis

For the example, we know both input and output.
Machine learning is done by understanding the relationship between the input and output.



Feature Extraction: It is a technique for transforming the raw data into a compact representation with numeric values which represents the original data


Supervised Learning Model
Training Text, Documents, Images, etc.
Feature Vectors
Machine Learning Algorithm
Labels
New Text, Document, Image, etc.
Feature Vector
Predictive Model
Expected Label

**Types of input**

| | |
|---|---|
| Image | X is n-dimensional [x1, x2, ….., xn] where each xi represents a pixel value |
| Audio | X is n-dimensional [x1, x2, ….., xn] where each xi represents a sample amplitude value |
| Text | X may be a binary vector representing presence/absence of a particular word<br>X may be n-dimensional [x1, x2, ….., xn] representing number for occurrence of a word |

The example that we discussed is of classification

If Y takes discrete values then we call it as **Classification**
If Y takes continuous values, then we call it as **Regression** where $Y \in \mathbb{R}$

**Example for regression (from coursera)**
Predicting the price of the house based on the parameters like, Living area size, No. of bedrooms
We already have examples of house that is sold,

| Living area | No. of Bedroom | Price |
|---|---|---|
| 2104 | 03 | 400 |
| 1600 | 03 | 330 |
| 2400 | 03 | 369 |
| 1416 | 02 | 232 |
| 3000 | 04 | 540 |
| 2540 | 03 | ? |

Examples (Supervised Learning) ,

- Problem 1: You have a large inventory of identical items. You want to predict how many of these items will sell over the next 3 months (?).
- Problem 2: You'd like software to examine individual customer accounts, and for each account decide if it has been hacked/compromised (?)

# Supervised Learning

- **Given:** Training examples $\langle \mathbf{x}, f(\mathbf{x}) \rangle$ for some unknown function $f$.
- **Find:** A good approximation to $f$.

## Example Applications

- **Credit risk assessment**

  $\mathbf{x}$: Properties of customer and proposed purchase.

  $f(\mathbf{x})$: Approve purchase or not.

- **Disease diagnosis**

  $\mathbf{x}$: Properties of patient (symptoms, lab tests)

  $f(\mathbf{x})$: Disease (or maybe, recommended therapy)

- **Face recognition**

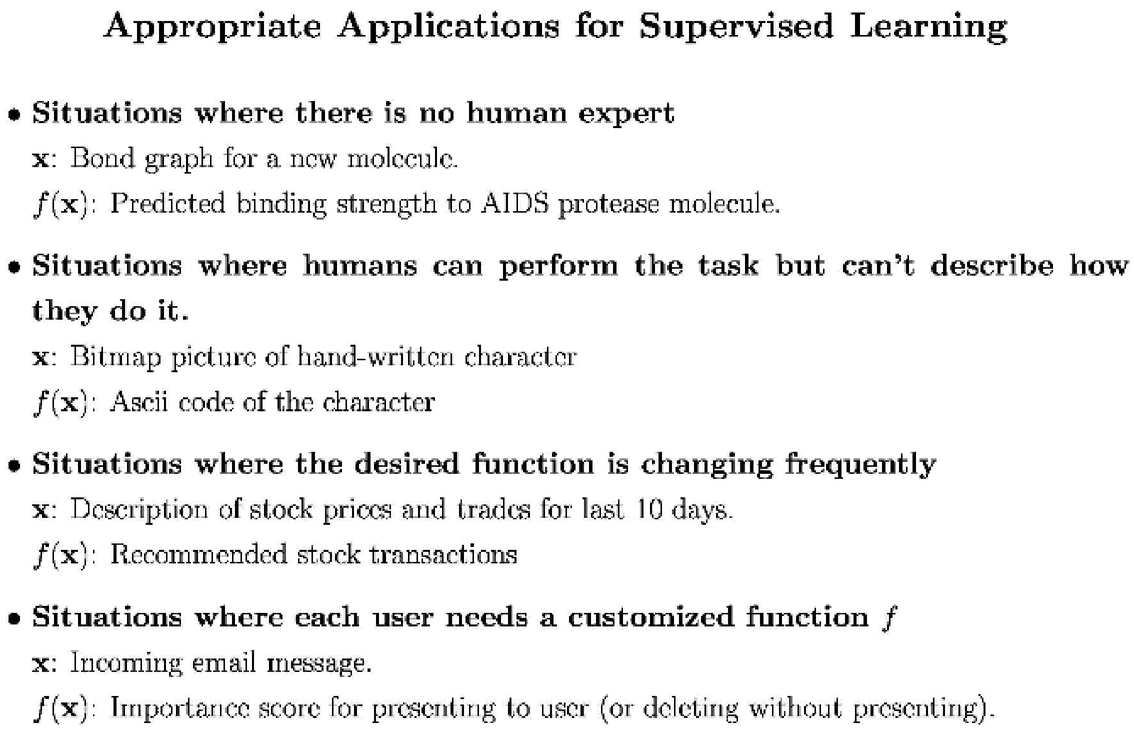
  $\mathbf{x}$: Bitmap picture of person's face

  $f(\mathbf{x})$: Name of the person.

- **Automatic Steering**

  $\mathbf{x}$: Bitmap picture of road surface in front of car.

  $f(\mathbf{x})$: Degrees to turn the steering wheel.



## Appropriate Applications for Supervised Learning

- **Situations where there is no human expert**

  $\mathbf{x}$: Bond graph for a new molecule.

  $f(\mathbf{x})$: Predicted binding strength to AIDS protease molecule.

- **Situations where humans can perform the task but can't describe how they do it.**

  $\mathbf{x}$: Bitmap picture of hand-written character

  $f(\mathbf{x})$: Ascii code of the character

- **Situations where the desired function is changing frequently**

  $\mathbf{x}$: Description of stock prices and trades for last 10 days.

  $f(\mathbf{x})$: Recommended stock transactions

- **Situations where each user needs a customized function $f$**

  $\mathbf{x}$: Incoming email message.

  $f(\mathbf{x})$: Importance score for presenting to user (or deleting without presenting).

**Unsupervised Learning**

Here we have the input X but there is no prediction for out Y.

Instead we are looking at similarities in the objects of X based on some parameters.

Example: group the Houses based on the similarity

| Living area | No. of Bedroom |
|---|---|
| 2104 | 03 |
| 1600 | 02 |
| 2400 | 03 |
| 1416 | 02 |
| 3000 | 04 |
| 2540 | 03 |

Clustering is one the technique which uses Unsupervised learning. There is another technique called as Principal Component Analysis.

When we a large collection of data without having any idea about the output response (unlabelled data), unsupervised learning helps in grouping the objects.

Example,

1. groups of shoppers characterized by their browsing and purchase histories,

2. Movies grouped by the ratings assigned by movie viewers.
3. Market customer segmentation
4. Social Network Analysis
5. Astronomical data analysis
6. CCTV footages inside ATM

**Unsupervised learning works based on the distance measure**

## Section 6
## Theory of Learning

In this section we learn about some theoretical terms and concepts related to machine learning.

### Feasibility of Learning

"The **term** "**population**" is used in statistics to represent all possible measurements or outcomes that are of interest to us in a particular study."
Population: Population need not refer to a living always. Statisticians also speak of a population of objects, or events, or procedures, or observations.

Size of the population is infinite (very large).

Sample: The **term** "**sample**" refers to a portion of the **population** that is representative of the **population** from which it was selected.

The data that we considered however large it may be, it is a sample and not population.

As the population includes all possible values, even the unseen data is a part of the population.

The Machine Learning model is developed by considering the examples in Sample.

The question is whether the developed application using sample can be used successfully for the entire population?

**Answer is yes**

### Training versus Testing

Machine learning is about learning some properties of a data set and applying them to new data.

This is why a common practice in machine learning to evaluate an algorithm is to split the data at hand in two sets, one that we call a training set on which we learn data properties, and one that we call a testing set, on which we test these properties.

The example dataset (sample) is divided into two sets, called as training set and testing set.
Training is the step where train set is used to build the Machine learning Model (function f)

During testing, the developed machine model is tested using the test set to generate the predicted target labels.

**The output that we get during testing is called as Predicted target label**

If the predicted target label is matched with that of the actual target label, then the output is declared to be correct otherwise error.

Number of such correct output is used to compute the accuracy (in %) for the machine learning model

For the below example, Sample size is 10.
It can be divided into train set of size 6 and test set of size 4.
During testing if 3 out 4 are correctly classified, then accuracy is 75%.

| | Height x1 | Weight x2 | Health Condition | Target Variable |
|---|---|---|---|---|
| X1 | 162 | 45 | NH | 2 (Y1) |
| X2 | 154 | 50 | H | 1 (Y2) |
| X3 | 170 | 85 | H | 1 (Y3) |
| X4 | 150 | 60 | H | 1 (Y4) |
| X5 | 175 | 50 | NH | 2 (Y5) |
| X6 | 160 | 55 | NH | 2 (Y6) |
| X7 | 161 | 48 | ? | ? (Y7)<br>Actually known as (1) |
| X8 | 153 | 56 | ? | ?(Y8)<br>Actually known as (2) |
| X9 | 160 | 48 | ? | ?(Y9)<br>Actually known as (1) |
| X10 | 155 | 59 | ? | ?Y10<br>Actually known as (2) |

## Error and Noise

The **error** measure is used to show the difference between the outputs that we get and the actual output from the training data is used to guide the learning process

The error term is usually defined as Mean squared error (MSE)

$$MSE = \frac{1}{N}\sum_{p=1}^{N}\left\|y_p - \hat{y}_p\right\|^2$$

Sum squared error

$$MSE = \sum_{p=1}^{N}\left\|y_p - \hat{y}_p\right\|^2$$

**Learning Curve**

(Refer: https://www.researchgate.net/publication/247934703_Learning_Curves_in_Machine_Learning)

A learning curve shows a measure of predictive performance on a given do-main as a function of some measure of varying amounts of learning effort. The most common form of learning curves in the general field of machine learning shows predictive accuracy on the test examples as a function of the number of training examples
as in Figure,

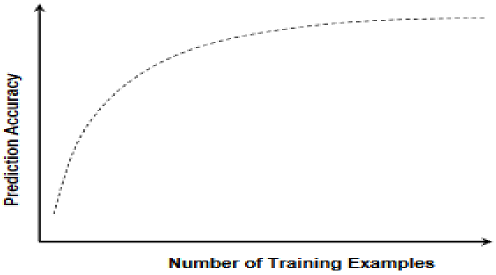


Figure 1: Stylized learning curve showing the model accuracy on test examples as function of the number of training examples.

**Noise** is any unwanted anomaly in the data and due to noise, the pattern may be difficult to learn and zero error may be infeasible with a simple hypothesis class. There are several interpretations of noise:

- There may be imprecision in recording the input attributes, which may shift the data points in the input space

- There may be errors in labeling the data points, which may relabel positive instances as negative and vice versa.
- There may be additional attributes, which we have not taken into account, that affect the label of an instance.

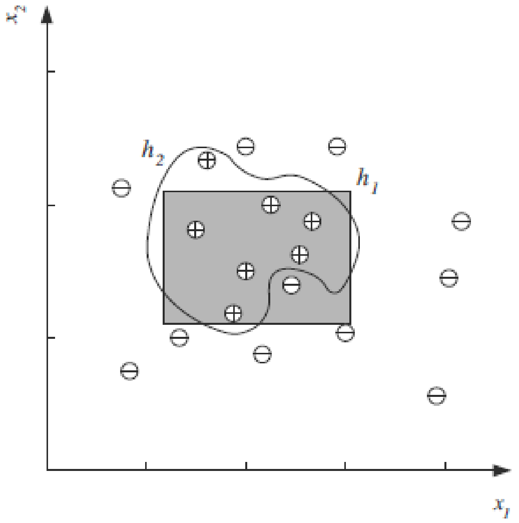


Figure 2.8 When there is noise, there is not a simple boundary between the positive and negative instances, and zero misclassification error may not be possible with a simple hypothesis. A rectangle is a simple hypothesis with four parameters defining the corners. An arbitrary closed form can be drawn by piecewise functions with a larger number of control points.

**Noise - difference between the data and the true function**

**Class noise** randomly alters the value of the function;
**Attribute noise** randomly alters the values of the components of the input vector.

**Theory of generalization**

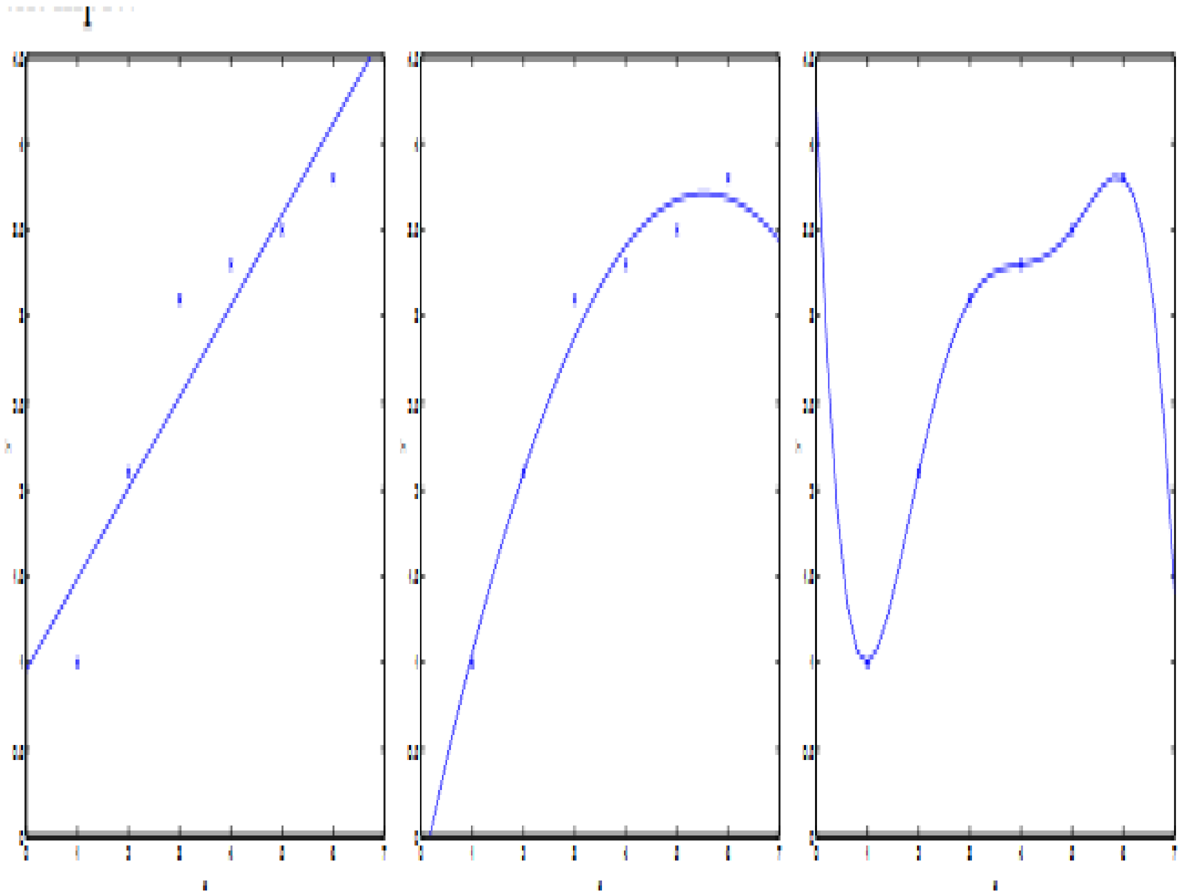
The model is developed using the examples of the training set. The model should be able to generate the right output for an input instance outside the training set.
How well a model trained on the training set, predicts the right output for new instances outside training set is called generalization.

**Overfitting (too much learning)** happens when a model learns the detail and noise in the training data to the extent that it negatively impacts the performance of the model on new data.
This means that the noise or random fluctuations in the training data is picked up and learned as concepts by the model.
The problem is that these concepts do not apply to new data and negatively impact the models ability to generalize.

**Underfitting (too lazy)** refers to a model that can neither model the training data nor generalize to new data.

An underfit machine learning model is not a suitable model and will be obvious as it will have poor performance on the training data.

## Bias and Variance



## Bias Error

Bias are the simplifying assumptions made by a model to make the target function easier to learn.

Here, the real world problem which is extremely complicated is approximated by a much simpler model.

Generally, parametric algorithms have a high bias making them fast to learn and easier to understand but generally less flexible. In turn, they have lower predictive performance on complex problems that fail to meet the simplifying assumptions of the algorithms bias.

The bias of a model to be the expected generalization error even if we were to fit it to a very (say, infinitely) large training set. Thus, for the problem above, the linear model suffers from large bias, and may underfit (i.e., fail to capture structure exhibited by) the data.

- **Low Bias**: Suggests less assumptions about the form of the target function.
- **High-Bias**: Suggests more assumptions about the form of the target function.

Let X be a sample from a population specified up to a parameter $\theta$, and
let $d = d($X$)$ be an estimator of $\theta$.
To evaluate the quality of this estimator, we can measure how much it is different from $\theta$, that is, $(d(X)-\theta)^2$.
But since it is a random variable (it depends on the sample), we need to
average mean square error this over possible X and bias of an estimator is defined as,

$$b_\theta(d) \; = \; E[d(X)] \; - \; \theta$$

$If\ b_\theta(d) = 0, for\ all\ \theta$ values, then we say that d is an unbiased estimator of $\theta$.

**Variance Error**

Variance is the amount that the estimate of the target function will change if different training data was used.

It refers to the amount by which $\hat{f}$ would change is we estimated it using a different training data set. (Actual function is $f$)

The target function is estimated from the training data by a machine learning algorithm, so we should expect the algorithm to have some variance. Ideally, it should not change too much from one training dataset to the next, meaning that the algorithm is good at picking out the hidden underlying mapping between the inputs and the output variables.

Machine learning algorithms that have a high variance are strongly influenced by the specifics of the training data. This means that the specifics of the training have influences the number and types of parameters used to characterize the mapping function.

- **Low Variance**: Suggests small changes to the estimate of the target function with changes to the training dataset.

- **High Variance**: Suggests large changes to the estimate of the target function with changes to the training dataset.

There is no escaping the relationship between bias and variance in machine learning.

- Increasing the bias will decrease the variance.
- Increasing the variance will decrease the bias.

There is a trade-off at play between these two concerns and the algorithms you choose and the way you choose to configure them are finding different balances in this trade-off for your problem

In reality, we cannot calculate the real bias and variance error terms because we do not know the actual underlying target function. Nevertheless, as a framework, bias and variance provide the tools to understand the behavior of machine learning algorithms in the pursuit of predictive performance.

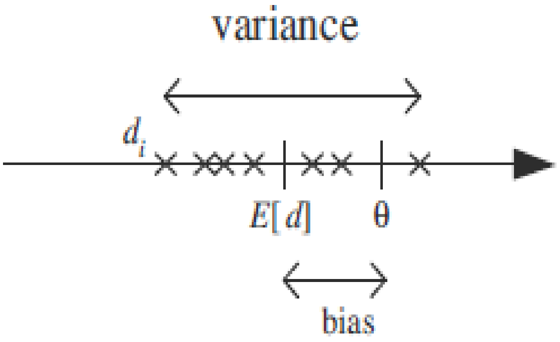


**Figure 4.1** $\theta$ is the parameter to be estimated. $d_i$ are several estimates (denoted by '×') over different samples $X_i$. Bias is the difference between the expected value of $d$ and $\theta$. Variance is how much $d_i$ are scattered around the expected value. We would like both to be small.

## Probability and Distributions (Review)

**A random experiment** is one whose outcome is not predictable with certainty in advance

The set of all possible outcomes is known as the ***sample space*** *S*. Sample space is like a Universal set.

A sample space is ***discrete*** if it consists of a finite (or countably infinite) set of outcomes; otherwise it **is *continuous***.

Any subset *E* of *S* is an ***event***.

**Example (discrete):**

Tossing a coin ---------------- (Random experiment)<br>
S= {Head, Tail} ---------------- (Sample space – discrete)<br>
E={ Head} ---------------- (Event)

Tossing a coin twice<br>
Tossing a die

**Example (Continuous)**

Experiment :: Observe the height in ft of a randomly chosen UF student.<br>
Sample Space ::S= [4,7] i.e. all real numbers between 4 to 7<br>
This is an example of a continuous or uncountable sample space.

Time elapsed between the arrival of two customers in a bank<br>
S= { x∈ ℝ | x>0}

## Probability (Frequency Approach)

When an experiment is continually repeated under the exact same conditions, for any event *E*, the proportion of time that the outcome is in *E* approaches some constant value. This constant limiting frequency is the probability of the event, and we denote it as *P(E).*

Classical Approach:

It is defined as if an event may occurs in ‘h’ different ways out of total number of ‘n’ different ways, then the probability of an event is h/n.

Frequency Approach:

If after ‘n’ repetitions of an experiment , an event is observed to occur in ‘h’ different ways, then the probability of an event is h/n.

Event A= { head}
P(A)=?

**Axioms of Probability**

1. 0 ≤ *P(E)* ≤ 1. If *E*1 is an event that cannot possibly occur, then *P(E*1*)* =0. If *E*2 is sure to occur, *P(E*2*)* = 1.
2. *S* is the sample space containing all possible outcomes, *P(S)* = 1.
3. If *Ei, i* = 1, . . . , *n* are mutually exclusive (i.e., if they cannot occur at the same time, as in *Ei* ∩*Ej* =∅*, j ≠ i*, where ∅ is the *null event* that does not contain any possible outcomes), we have

$$P(\bigcup_{i=1}^{n} E_i) = \sum_{i=1}^{n} P(E_i)$$

**Conditional Probability**

*P(E|F)* is the probability of the occurrence of event *E* given that *F* occurred and is given as

$$P(E|F) = \frac{P(E \ \cap \ F)}{P(F)}$$

$$P(E \cap F) = P(E|F)P(F) = P(F|E)P(E)$$

If E and F are Independent then $P(E|F) = P(E)$
In that case ,

$$P(E \cap F) = P(E)P(F)$$

**Random Variables**

A *random variable* is a function that assigns a number to each outcome in the sample space of a random experiment.

Sample S={head, tail} , we can write it as S={0,1}

We define a random variable X can take values (0,1)

X can be a discrete Random variable or a Continuous Random variable

We can have more than one random variables , X1, X2, ….., Xn in case of multiple random experiment.

**Tossing a coin twice**

**S={HH, HT, TH, TT}**

**X → Number of heads**
**X → 2, 1, 1, 0**

**P(X=0) = ?**
**P(X=1) = ?**
**P(X=2) = ?**

**Or we can define**

**X → Number of tails**
**X → Number of head –(minus) number of tails**

**In case of continuous RV , probability is always for a range of values**
**P(a <= X <=b) = ?**

**P(X=a)=0 (RV cannot take a single definite value)**

**Probability Distributions, Probability Density, Distribution Function**

**Probability Function / Probability Distribution**

Let X be a discrete random variable, and suppose that the possible values that it can assume are given by x1,x2,x3, . . . , arranged in some order. Suppose also that these values are assumed with probabilities given by

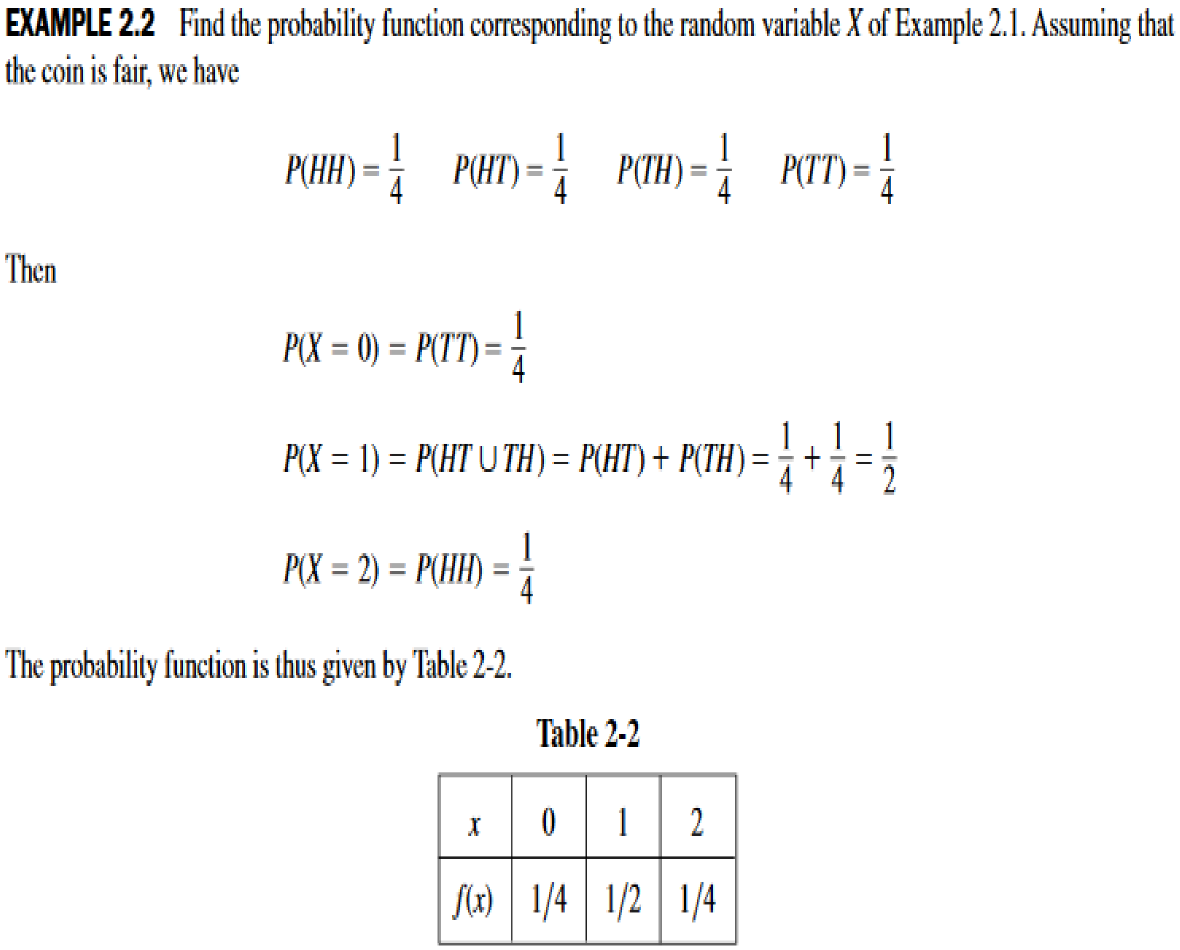
$P(X=x_k)=f(x_k)$, k=1, 2, . . .

(1)

It is convenient to introduce the probability function , also referred to as probability distribution , given by

P(X=x)=f(x)

1. f(x)>=0
2. $\sum_x f(x) = 1$



**EXAMPLE 2.2** Find the probability function corresponding to the random variable $X$ of Example 2.1. Assuming that the coin is fair, we have

$$P(HH) = \frac{1}{4} \qquad P(HT) = \frac{1}{4} \qquad P(TH) = \frac{1}{4} \qquad P(TT) = \frac{1}{4}$$

Then

$$P(X = 0) = P(TT) = \frac{1}{4}$$

$$P(X = 1) = P(HT \cup TH) = P(HT) + P(TH) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}$$

$$P(X = 2) = P(HH) = \frac{1}{4}$$

The probability function is thus given by Table 2-2.

**Table 2-2**

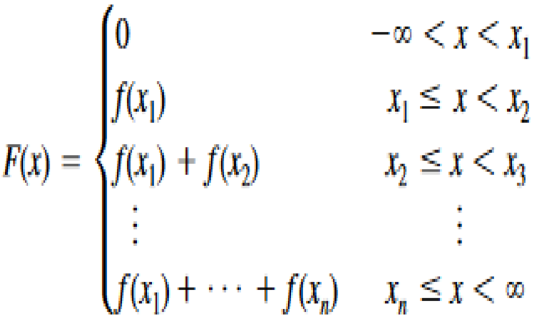
| $x$ | 0 | 1 | 2 |
|---|---|---|---|
| $f(x)$ | 1/4 | 1/2 | 1/4 |

**Distribution Function (Cumulative Distribution Function)**

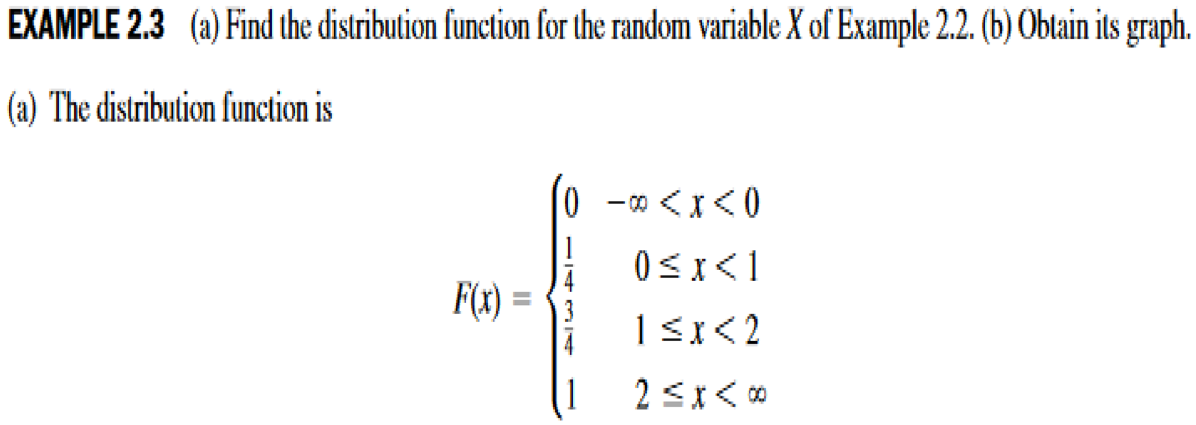
**Discrete**

The distribution function for a discrete random variable X can be obtained from its probability function by noting that, for all x ,

$$F(x) = P(X \leq x) = \sum_{u \leq x} f(u)$$



$$F(x) = \begin{cases} 0 & -\infty < x < x_1 \\ f(x_1) & x_1 \leq x < x_2 \\ f(x_1) + f(x_2) & x_2 \leq x < x_3 \\ \vdots & \vdots \\ f(x_1) + \cdots + f(x_n) & x_n \leq x < \infty \end{cases}$$



**EXAMPLE 2.3** (a) Find the distribution function for the random variable $X$ of Example 2.2. (b) Obtain its graph.

(a) The distribution function is

$$F(x) = \begin{cases} 0 & -\infty < x < 0 \\ \frac{1}{4} & 0 \leq x < 1 \\ \frac{3}{4} & 1 \leq x < 2 \\ 1 & 2 \leq x < \infty \end{cases}$$

**Continuous**

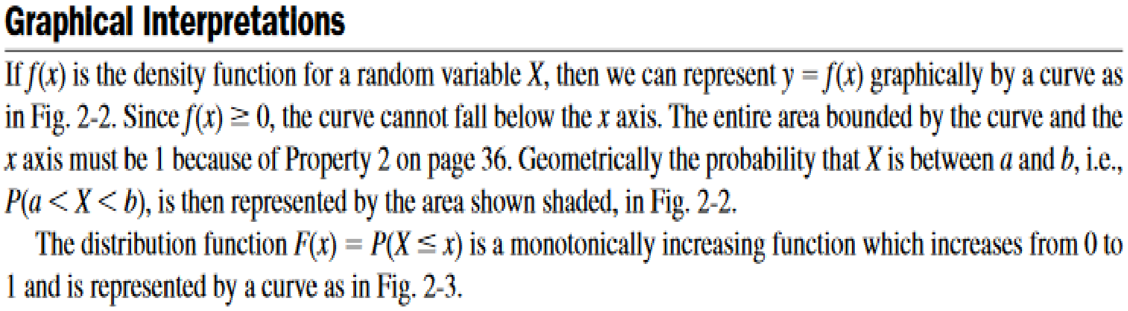
A nondiscrete random variable X is said to be absolutely continuous , or simply Continuous , if its distribution function may be represented as

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(u)\, du$$

In this case,

$$\int_{-\infty}^{\infty} f(x)dx = 1$$

$$P(a \leq X \leq b) = \int_{a}^{b} f(x)dx$$



## Graphical Interpretations

If $f(x)$ is the density function for a random variable $X$, then we can represent $y = f(x)$ graphically by a curve as in Fig. 2-2. Since $f(x) \geq 0$, the curve cannot fall below the $x$ axis. The entire area bounded by the curve and the $x$ axis must be 1 because of Property 2 on page 36. Geometrically the probability that $X$ is between $a$ and $b$, i.e., $P(a < X < b)$, is then represented by the area shown shaded, in Fig. 2-2.

The distribution function $F(x) = P(X \leq x)$ is a monotonically increasing function which increases from 0 to 1 and is represented by a curve as in Fig. 2-3.

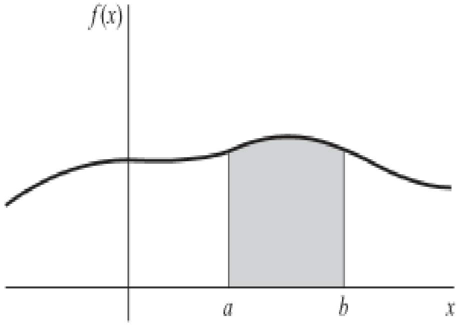


Fig. 2-2

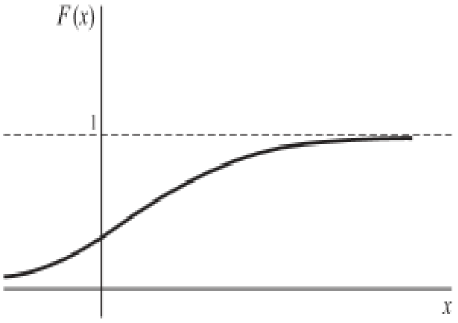


Fig. 2-3

**Expectations and Variance**

$$E[X] = \sum xf(x)$$

$$E[X] = \int_{-\infty}^{\infty} xf(x)dx$$

VAR[X]=E[$X^2$]-(E[X])$^2$