CHAPTER I

INTRODUCTION TO ARTIFICIAL NEURAL NETWORK

1.1 INTRODUCTION

In this Chapter, we discuss, in brief, the fundamentals of ANN models and neural computing. Much of the inspiration for these types of models comes from neuroscience. Therefore, first we discuss the biological neural network in Section 2. In Section 3, the concept of ANN is introduced and the necessary details are discussed in Section 4. The history of development of ANN is given in Section 5. Finally, the Chapter ends with some interesting applications of ANN which are reported in the literature during last five years.

1.2 BIOLOGICAL NEURAL NETWORK

Before studying ANN, it is helpful to first consider the biological origins of neurons. In this Section, therefore, we describe in brief, the biological neural network. The discussion is mainly based on Zurada(1992), Haykin(1994), and Schalkoff(1997). At this stage, we point out that though the original motivation for the development of ANN was to model the brain, the current research in neural network is totally different from that of modelling the working of brain.

1.2.1 *Biological Neuron*

Nerve Cells

The nervous system consists of two classes of cells, namely 'neurons' or 'nerve cells' and 'glia' or 'glial cells'. Neurons are the basic building blocks of biological information processing system. So, we concentrate only on neurons.

The neurons of the brain are classified according to their function : i) 'Sensory neurons' which provide input to the nervous system (for example, optic nerves), ii) 'Motor neurons' which transmit control signals to muscles and glands, and iii) 'Internuronal neurons' propagate signals from one side to another and constitute by far the largest class of cells in the nervous system.

Fig. 1.1 shows typical structure of a neuron. It has three major portions, each of which makes a specific contribution to the processing of signals:

1) The 'cell body' consists of the cell nucleus and perikaryon.

2) The 'axon' begins at the 'axon hillock'. The axon generates the cell action potential. The axon is the main transmission mechanism of the brain.

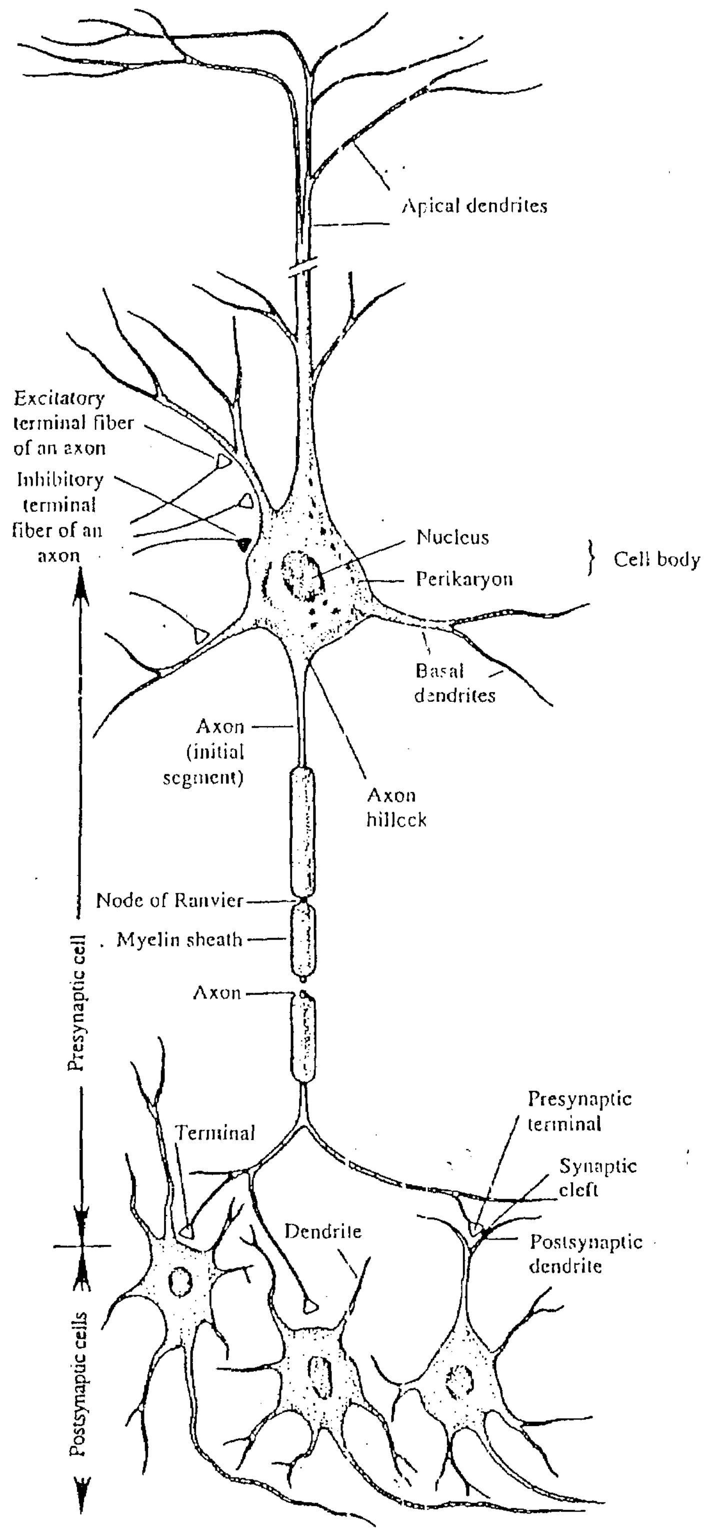


Figure 1.1 Biological Neuron (Reproduced from Schalkoff, 1997)

3) 'Dendrites' are similar to that of the tree branches. Most neurons have multiple dendrites. The dendrites of one neuron are connected to the axon of other neuron via 'synaptic connections' or 'synapses'. There are usually between 1000 and 10,000 synapses on each neuron and there are about 10^{11} neurons in a brain.

Synaptic Activity

Synaptic transmission involves complicated chemical and electrical process. Sensory or chemical stimuli create a change in synaptic potential. This is the basis by which one neuron influences the state of others. In the cell body, this activity integrates and determines axon potential. One of the most interesting aspects of cell behaviour is that the cell body forms weighted input action potential to an all-or-nothing output. If the overall cell stimulus is below a 'threshold', no signal is produced. If the total stimulus is above the threshold, some output is produced or the neuron begins to 'fire'. Some neuron transmitters are 'excitatory' i.e. they cause an increase in 'action potential' of the receiving neurons and some are 'inhibitory' which means that they do not allow to increase the receiving neurons action potential.

Information Processing In The Brain

It is well established fact that, a human brain has tremendous capacity to perform certain computations, for example, pattern recognition, perception and motor control and so on. Consider, for instance, human vision which is an information-processing task. It is the function of visual system to provide a representation of the environment around us and more important, to supply the information we need to interact with the environment. Now question is, how does a human brain do it?

The information processing takes place in the brain as follows:

The signals are received, which are referred to as 'inputs' (through eyes, ears etc.) from outside world by the brain and certain processing is done on these signals. It is biologically observed that the inputs are summed up by giving appropriate weightages to the inputs. Each neuron receives inputs or signals from large number of other neurons. After combining these inputs it sends output signals to a large number of other neurons and lastly final outputs or responses are given through sensory organs depending on the 'strength' of the summed inputs.

The above overall process is done in central nervous system and the basic central processing units are neurons. Further, in addition to this, the following also are special computational features associated with brain :

1. The brain integrates and stores the past experience, which could be previous classification of input data. In this sense, it self-organizes experience.

2. The brain considers new experiences in the context of stored experiences.

3. The brain is able to make accurate predictions about new situations on the basis of previously stored information or self-organized experience. This suggests a generalization capacity of a brain.

4. The brain does not require perfect information.

Consider the following examples which clearly help to understand the working of the 'process' that takes place in the brain:

Example 1 : To teach a child the alphabets, first step is to show him or her a letter. Then based on his or her response, provide feedback to the child. This process is repeated for each letter until the child recognizes the alphabet correctly. In this learning process, the brain stores the necessary information to recognize an alphabet. Such type of learning is generally called 'supervised learning'. Thus once the learning process is complete, then, in future, even when a distorted alphabet is presented to a child, he or she correctly identifies it.

Example 2 : When a child touches a hot heating coil, he or she soon learns, without any external teaching, not to touch it. Moreover, based on this experience, the child associates a bright red glow with hot, and learns to avoid touching objects with this feature. This implies self-organized experience. Such a type of learning is called 'unsupervised learning ' or 'learning without a teacher'.

The concept of 'learning' through experience forms the basis of Artificial Neural Networks model. In the Section that follows, we will introduce the concept of ANN and discuss it in detail.

1.3 ARTIFICIAL NEURAL NETWORK (ANN)

Artificial Neural Network is a 'machine' that is designed to model a particular task or function of interest. The network is usually implemented using electronic components (or simulated in a software on a digital computer), which performs useful computations through a process of learning. A neural network is a network of a massive interconnections of simple computing units called as 'neurons' or 'processing units'.

In the literature, there are many definitions of ANN and we will give below some important definitions :

Definition 1 (Haykin, 1994)

A neural network is a massively parallel distributed processor that has natural tendency for storing experimental "Knowledge" and making it available for use. It resembles the brain in two respects:

1. Knowledge is acquired by the network through a learning process.
2. Interneuron connection strengths known as 'weights' are used to store the knowledge.

Definition 2 (Schalkoff, 1997)

A neural network is a structure (network) composed of a number of interconnected units (artificial neurons). Each unit has an input/output (I/O) characteristics and it implements a local computation or function. The output of any unit is determined by its I/O characteristics, its interconnections to other units, and (possibly) external inputs.

Analogous to the neurons in the brain, the artificial neuron is a fundamental unit to the operation of neural network. It was designed to mimic the characteristics of the biological neurons.

In the following, we discuss the structure (or architecture) of different artificial neuron models.

1.3.1 *Artificial Neuron Model*

Every neuron model consists of a processing element with input connections and a single output. The signal flow of neurons input, say x_i (for $i = 1,2,3,...,n$), is considered to be unidirectional as indicated by arrows, as is a neuron's output signal flow. A general neuron symbol is shown diagrammatically in Fig. 1.2

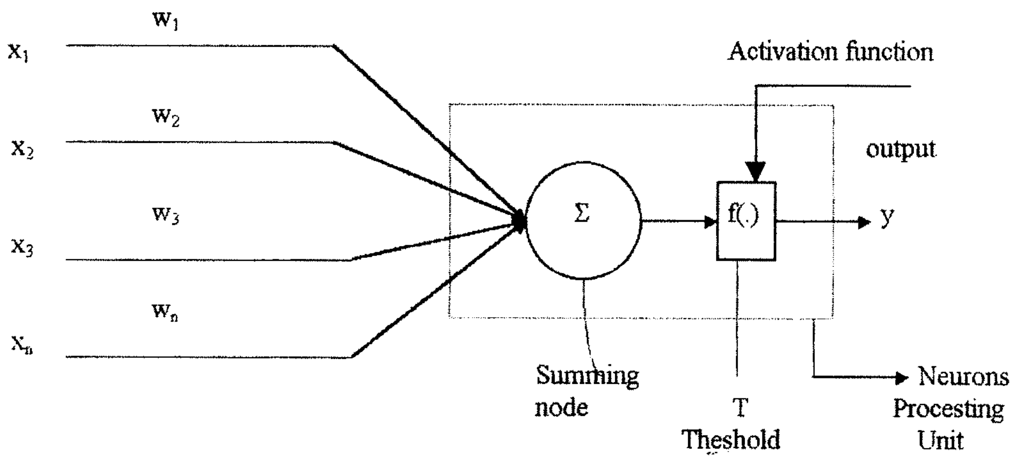


Figure. 1.2 Artificial Neuron Model

This network consists of three basic elements namely :

1. A set of 'synapses' each of which is characterized by a weight or strength of its own called as 'synaptic weight'. (specifically, a signal x_i at the input of synapse i connected to node and is multiplied by the synaptic weight w_i called weighted input).

2. All the weighted inputs are summed up, called as 'netinput' or 'net', to determine the 'activation level' of neuron.

3. An 'activation function' for limiting the amplitude (largeness) of the output of a neuron. The activation function is also referred to as a 'squashing function'. Because, it squashes (limits) the allowable amplitude range of the output signals to some finite value.

The model shown in Fig.1.2 also includes an externally applied threshold T that has effect of lowering the netinput of the activation function. Sometimes, for increasing the netinput of the activation function a 'bias' term rather than threshold is introduced. That is, bias is the negative of threshold.

The symbolic representation in Fig 1.2 shows a set of weights w_i's and neuron's processing unit or node. The neuron output signal, y, is given by

$$y = f(\text{net}), \tag{1.3.1}$$

where

$$\text{net} = \underline{w}' \underline{x} , \tag{1.3.2}$$

and

$$\underline{w} = [w_1, \ldots, w_n]' , \tag{1.3.3}$$

and $\underline{x}$ is the input vector,

$$\underline{x} = [x_1, x_2, \ldots, x_n]' \tag{1.3.4}$$

The function f(.) is often referred to as an activation function. Its domain is a set of activation value $\underline{w}'\underline{x}$, of the neuron model.

Note that temporarily the threshold value is not explicitly shown in (1.3.1) and (1.3.2), but this is only for notational convenience. We have assumed that the modeled neuron has n-1 actual synaptic connections that come from actual variable inputs $x_1, x_2, \ldots, x_{n-1}$. And also we assumed that $x_n = -1$ and $w_n = T$, where T is threshold as defined earlier. We will see later, the threshold as a separate neuron model parameter, because it plays an important role in some models.

1.3.2 Types of Activation Functions

The activation function, denoted by f(.) defines the output of a neuron in terms of the activity level at its input. In other words, activation function is defined as mapping the activation value into the artificial units output. This mapping may be as simple as using identity function, or as complex as using nonlinear mapping function. (The class of nonlinear squashing functions is however, more interesting and useful). Below we discuss variety of activation functions :

1. <u>Threshold Function</u> :

For this type of activation function, as shown in Fig. 1.3 (a), we have

$$f(net) = \begin{cases} 1, & \text{if } net \geq 0 \\ 0, & \text{if } net < 0 \end{cases} \qquad (1.3.5)$$

In this case, if the internal activity level of the neuron is nonnegative, the output of neuron takes value 1 and 0 otherwise. Historically, threshold function is used to compute Boolean functions, for example, 'AND', 'OR' and 'NOT'. (This part will be discussed in depth in Ch. II).

2. <u>Piecewise Linear Function</u> :

For the piecewise linear function, described in Fig. 1.3.(b), we have

$$f(net) = \begin{cases} 1, & \text{if } net \geq 1/2 \\ net, & \text{if } 1/2 > net > -1/2 \\ 0, & \text{if } net \leq -1/2 \end{cases} \qquad (1.3.6)$$

This form of activation function is non-linear, although it is composed of linear segments. (We note that, the piecewise linear function reduces to a threshold function if the amplification factor of the linear region is made infinitely large)

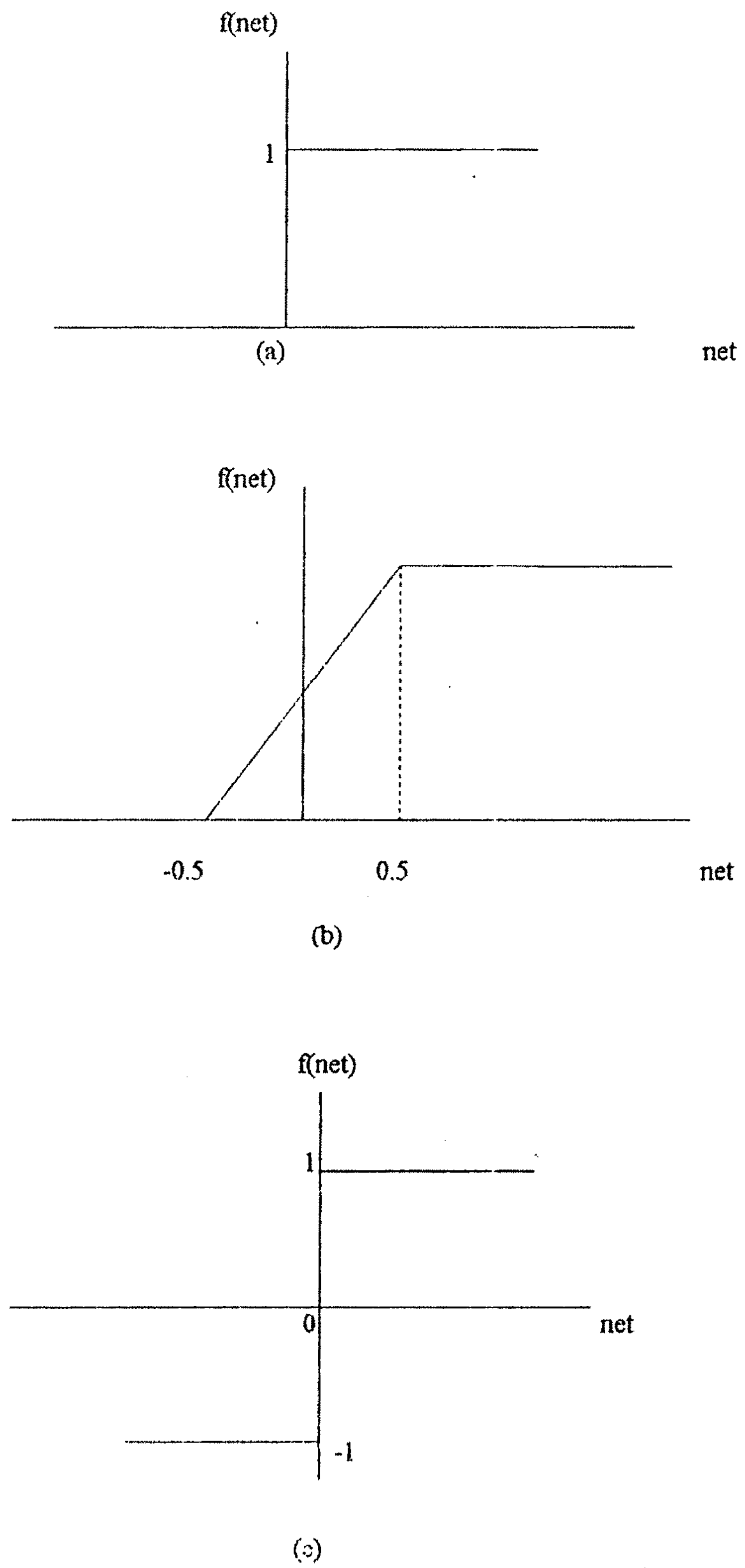
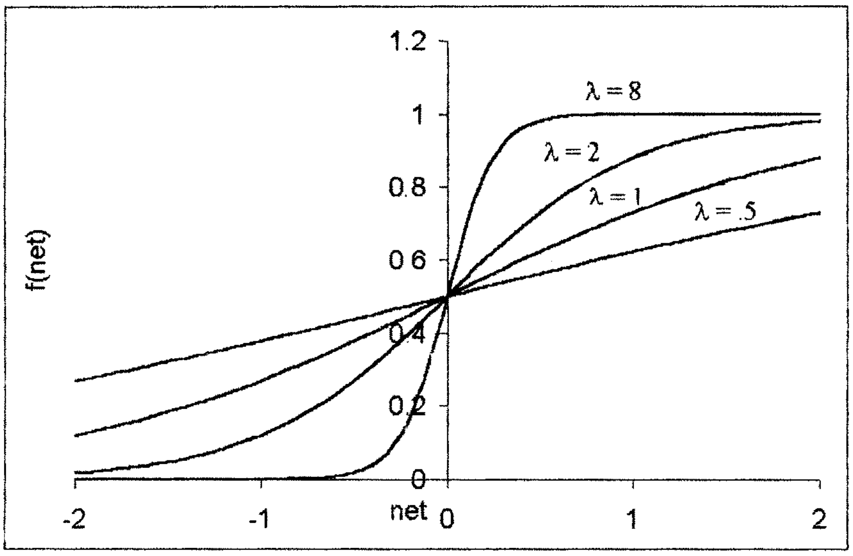
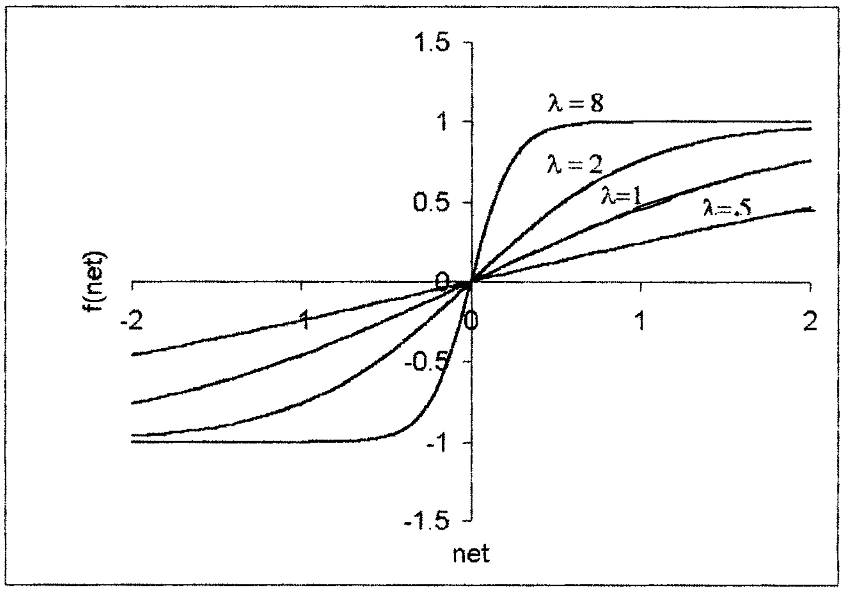


Figure 1.3 Common Activation Functions



(d)



(e)

Figure 1.3 (Cont.)

3. <u>The Sigmoid(logistic) Activation Function</u> :

The sigmoid function is the most common form of activation function used in the construction of artificial neural networks. It is defined as a strictly increasing function that displays smoothness and asymptotic properties. An example of the sigmoid is the 'logistic function', defined by

$$f(net) = \frac{1}{1 + \exp(-\lambda\ net)}, \qquad (1.3.7)$$

where $\lambda > 0$ is the 'slope parameter' of the sigmoid function. It is proportional to the neuron gain determining the steepness of continuous function f(net) near net=0. By varying the parameter λ, we obtain sigmoid function of different slope, as shown in Fig 1.3.(d). In fact, slope at the origin equals $\lambda/4$. As the slope parameter approaches infinity, the sigmoid function becomes simply a Threshold function(1.3.5). Whereas, the Threshold function takes value of 0 or 1, a sigmoid assumes a continuous range of values from 0 to 1.

There are several reasons for using the sigmoid function in the literature on neural network and these are :

1. It has a biological basis. The output or excitation function of biological neuron follows a sigmoidal characteristics.

2. It squashes(limits) the allowable range of the output signal to some finite value.

3. It is nondecreasing and differentiable everywhere, a desirable property in many applications.

4. It is expressible in closed form.

5. Modifications or extensions relate to other squashing functions.

4. <u>Other Activation Functions</u> :

The activation functions defined in Eqs (1.3.5), (1.3.6) and (1.3.7) ranges from 0 to 1. Furthermore, if an output range from -1 to 1 is desired, in which case activation function assumes an antisymmetric form with respect to the origin. Specifically, the Threshold function of Eq. (1.3.5) is redefined as

$$f(net) = \begin{cases} 1 , & \text{if } net \geq 0 \\ -1 , & \text{if } net < 0 \end{cases} , \qquad (1.3.8)$$

which is commonly referred to as 'signum function' as shown in Fig. 1.3.(c). Also, sigmoid function with output range (-1,1) is shown in Fig. 1.3(e) and defined as

$$f(net) = \frac{2}{1 + \exp(-\lambda \; net)} - 1, \qquad (1.3.9)$$

where $\lambda > 0$.

The 'hyperbolic tangent function' is given by

$$f(net) = \tanh\left(\frac{net}{2}\right) = \frac{1 - \exp(-net)}{1 + \exp(-net)}, \tag{1.3.10}$$

It is similar in shape to that of sigmoid function and is often used by biologists as a mathematical model of nerve cell activation.

NOTE :- 1. Activation functions (1.3.5) and (1.3.7) are called as 'unipolar binary' and 'unipolar continuous' activation functions respectively. The word "unipolar " is used to point out that only positive responses of neurons are produced for this definition of activation functions.

2. Activation functions (1.3.8) and (1.3.9) are called 'bipolar binary' and 'bipolar continuous' activation functions respectively. Because both positive and negative responses of neurons are produced.

1.4 LEARNING PROCESS OF ANN

In this Section, we discuss a process called 'learning' or 'training' of neural network. Learning is an important aspect of ANN. We define learning in the context of ANN as follows : consider the weight vector $\underline{w}$, where

$$\underline{w} = [w_1, w_2, \ldots, w_n]',$$

and the input vector $\underline{x}$

$$\underline{x} = [x_1, x_2, \ldots, x_n]',$$

with corresponding desired output $\underline{d}$

$$\underline{d} = [d_1, d_2, \ldots, d_k]'$$

Definition : Learning Process

Learning is defined as a process by which the weights or synaptic weights in neural network are adjusted to solve the problem presented to the network. This process is analogous to the process of learning an alphabet in the example given earlier.

The two types of learning associated with neural network are namely:

1. Supervised learning.
2. Unsupervised learning.

Below we discuss the two types:

Supervised Learning :

In supervised learning, we assume that at each instant of time when input is applied to the network, desired response $\underline{d}$ of the system provided by the 'teacher'(the term teacher: as having "knowledge" about the set of input-output examples presented to the network) is known. So supervised learning, also called learning with a teacher occurs when there is a known set of target values associated with each input in the training set. As an illustration of this, we refer to Example 1 given in Section 2.

Unsupervised Learning :

Unsupervised learning is more plausible method of training in the biological system. For this, we refer to Example 2 given in Section 2. In this process, the training set consists of only input vectors. The training algorithm modifies network weights to produce output vectors that are consistent.

In this dissertation, however, we will be mainly considering supervised learning rule.

1.5 BRIEF HISTORY OF ANN

We conclude this introductory Chapter with a brief history of ANN. The discussion is based on Zurada(1992), Haykin (1994) and Schalkoff (1997).

The modern era of neural network is said to have begun with the pioneering work of McCulloch and Pitts (1943). They outlined the first formal model of an elementary computing neuron and it laid the groundwork for future development of ANN.

In 1948, Wiener's famous bock 'Cybernetics' was published, describing some important concepts for control, communication and statistical signal processing.

The next major development in neural networks took place in 1949. Hebb (1949) first proposed a learning rule for updating neurons connections. He stated that the information can be stored in connections and postulated the learning technique that had a great impact on future development in this field.

During the 1950s, the first neurocomputers were built and tested. They modified connections automatically. During this stage neuron like element called a 'perceptron' was proposed by Rosenblatt in 1958. It was a machine capable of learning to classify certain patterns by modifying connections to the threshold elements.

In the early 1960s, a device called ADALINE (for ADAptive LINEar combiner) was introduced. And a new powerful learning rule called the 'LMS (Least Mean Square) learning rule' was developed by Widra Widraw and Hoff (1960,1962). The rule minimized the summed mean square error during training involving pattern classification.

During the period of perceptron in the 1960s, it seemed as if neural network could do anything. But then came the book by Minsky and Papert in 1969, who used mathematical tools to demonstrate that there are fundamental limits on what a (single layer) perceptron can compute. In brief, they stated that, many of the desirable mapping were unachievable with the perceptron.

The problem was not solved until mid - 1980s. Hence ANN research became less active during 1969 - 1980s.

In the 1980s, major contributions to the theory and design of neural network were made. During the period from 1982 until 1986, several publications for future development were published. The era of renaissance started with Hopfield (1982, 1984) introducing a 'recurrent' neural network structure for 'associative memory'. His paper formulated computational properties of a fully connected network. This particular class of networks attracted a great deal of attention in the 1980s and referred to as 'Hopfield network'.

During the period from 1982 until 1986, several important articles were published that significantly furthered the potential of neural networks.

In 1986, the development of the 'backpropagation algorithm' was reported by Rumelhart, Hinton and Williams, which became the most popular learning algorithm for training multilayer networks.

Perhaps more than any other publications, the 1982 paper by Hopfield and the 1986 two-volume book by Rumelhart and McClelland were the most influenced publications responsible for the resurgence of interest in neural networks in the 1980s.

Of late, ANNs are being used as alternative methods to statistical tools (Ripley, 1994) and active research in this field is being undertaken by many Statisticians.

1.6 NEURAL COMPUTATIONS : SOME APPLICATIONS

We now give some applications of neural network reported in the literature.

1. Pattern recognition :- Neural networks have been used for feature extraction, radar signal classification and analysis, speech recognition and understanding. These are also used for fingerprint identification, character (letter or number) recognition and handwriting analysis.

2. Image processing and computer vision :- It includes image matching, processing computer vision (e.g. circuit board inspection), image compression, stereo vision, and processing and understanding of time varying images.

3. Medicine :- Some of the properties of neural network can be used to assist the medicinal chemist in the design of drugs(Harget and Bodor, 1992). They have been used for diagnosis of various diseases and medical image processing, electrocardiographic signal analysis and understanding. They are also used in predicting heart problems in patients.

4. Planning control and search :- ANNs are used for parallel implementation of Constraint Satisfaction Problems (CSPs), solutions to Traveling Salesman like CSPs, and control and robotics.

5. Financial Systems :- Financial systems, including stock market analysis, real estate appraisal, credit card authorization and securities trading etc.

6. Military Systems :- It includes, radar clutter classification and technical speaker recognition.

7. Power Systems :- Power systems, including of system state estimation, transient detection and classification, fault detection and recovery, load forecasting and security assessment.