Density and Approximation by Using Feed Forward Artificial Neural Networks

R.S.Naoum , L.N.M.Tawfiq

College of Education, Ibn Al-Haitham, University of Baghdad

Abstract

In this paper ,we will consider the density questions associated with the single hidden layer feed forward model. We proved that a FFNN with one hidden layer can uniformly approximate any continuous function in C(k)(where k is a compact set in R^n) to any required accuracy.

However, if the set of basis function is dense then the ANN's can has almost one hidden layer. But if the set of basis function non-dense, then we need more hidden layers. Also, we have shown that there exist localized functions and that there is no theoretical lower bound on the degree of approximation common to all activation functions(contrary to the situation in the single hidden layer model).

Introduction

Artificial Neural networks(ANN)effectively is a mapping which creates a function of several variables from a number of sums and products of functions of one variable. It may contain several layers which can be used for calculations, corresponding to the layers of the ANN's and it may modify the parameters in the resulting form of the approximation by a learning or training procedure which passes through the ANN's repeatedly or which moves forward and back ward through the ANN's.

Let inputs vectors be in R^n and the output Y of the ANN be a vector in R^m, where usually $m << n$ (In many cases $m = 1$). The ANN thus computes a function $g : R^n \longrightarrow R^m$ which we regard as an approximation to some other function $f : R^n \longrightarrow R^m$. We wish to know if our set of possible functions g, corresponding to our particular class of ANN, is dense in some suitable function space which includes our target function f. In view of the difficulty of dealing with non differential and discontinuous functions, it is usual to use a smooth activation function, instead of a threshold, for the units.

Note that the activation function $\sigma : R \longrightarrow R$ required a various restrictions to be putted on σ to make it practical for the ANN, and we will introduce these as required. We will discuss how a multi layer FFNN with a single hidden layer of K neuron units can be used to approximate a function of several variables, where the weight at the j-th hidden neuron is denoted W_j.

Thus for a given input vector x, the input to this unit is $W_j^T x$. We assume that each of the hidden units has identical activation function σ, but that 'threshold like' shift of the argument by a real scalar c_j is permitted. So the output from the j-th hidden unit is $\sigma(W_j^T x + c_j)$.

Now we denote the weight connecting the j-th hidden unit to the output by v_j. The out put function g of the ANN is therefore (see Fig.(1), (1):

$$g(x) = \sum_{j=1}^{k} v_j \, \sigma(W_j^T x + c_j) \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots [1.1]$$

activation function σ used in practice have the property of being monotonic increasing, bounded and sigmoidal, which means that the limits at $+\infty$, $-\infty$ are 1 and 0 respectively. Except for the threshold function, they are also continuous and smooth. The most popular choice is the sigmoid function:

$\sigma(x) = 1/(1 + \exp(-\alpha x)) \ldots\ldots\ldots\ldots\ldots\ldots\ldots [\ 1.2]$

where α can be adjusted so that we can avoid the local minimum.

However, the density proofs do not use all these conditions. For the basic results only continuity or uniform continuity is required, plus the condition that σ be sigmoidal.

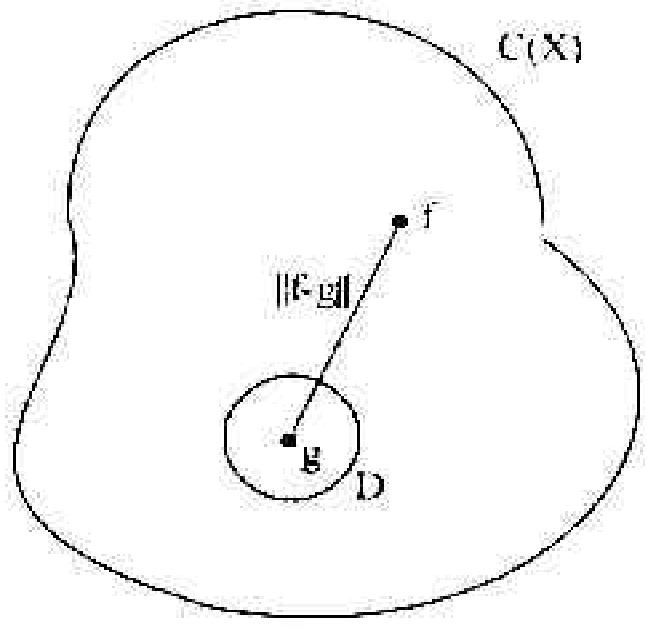
Density

In this section we will consider density questions associated with the single hidden layer feed forward neural model. That is, for an activation function, σ, and, for any $f \in C(R^n)$,k compact subset of R^n, and any $\varepsilon > 0$, there exists $g(x) = \sigma(w.x - \theta)$, where $\theta \in R$, $w \in R^n$, such that: $\max_{x \in K} |f(x) - g(x)| < \varepsilon$

Firstly, we introduce definition of density:

Definition (2.1)

A subset D in C(X) is ***dense*** if and only if:



$\forall$ $f \in C(X)$, $\forall$ compact set $K \subset X$ and $\forall$ $\varepsilon > 0$, $\exists$ $g \in D$, – such that: $|f g||_K < \varepsilon$

Remarks

Density is the theoretical means, the ability to approximate well.

1. Density dose not imply a good, efficient scheme for the approximation
2. Lack density means that it is impossible to approximate a large class of functions, and this effectively precludes any scheme based there on being in the least useful.
3. C(X) becomes a topological algebra. Since X is normed linear space, it is not a compact topological space and usual Stone-

Weienstrass theorem does not apply to it. However, there are a suitable substitute for it.

Now, we state Kolmogorov's theorem:

Theorem (2.2) Kolmogorov's mapping Neural Network Existence Theorem)

Given any continuous function $f : [0,1[^n \longrightarrow R^M$, $f(x) = y$, f can be implemented exactly by a three-layer feed forward neural network having n processing elements in the first (x-input) layer, (2n + 1) processing elements in the middle layer, and m processing elements in the top (y-output) layer.

Proof : The proof can be found in (2) ζ

As stated in the above theorem, the Kolmogorov mapping network consists of three layers of processing elements (input layer-hidden layer and output layer). The first layer (input layer) consists of n input units. The second layer consists of 2n + 1 semi linear units (i.e., the transfer function of these units is similar to a linear weighted sum). Finally, the third (output) layer has m processing elements with highly nonlinear transfer functions. The second layer implement the following transfer function $z_k = \sum_{j=1}^{n} \lambda^k \psi(x_j + k\varepsilon) + k$

where the real constant λ and the continuous real monotonically increasing function ψ are independent of f(although they do depend on n). The constant ε is a rational number $0 < \varepsilon \leq \delta$, where δ is an arbitrarily chosen positive constant. The m top layer processing elements (output units) have the following transfer functions: $y_i = \sum_{k=1}^{2n+1} g_i(z_k)$

where the functions g_i, $i = 1,2,\ldots m$ are real and continuous (and depend on f and ε). No specific example of a function ψ and constant ε are known (still an open problem). No example of a g function is known. The proof of the theorem is not constructive, so it does not tell us how to determine these quantities. It is strictly an existence

theorem. It tells us that such a three layer mapping network must exist, but it doesn't tell us how to find it.

However, the direct usefulness of this result is doubtful, because no constructive method for developing the g_i functions is known.

Direct Approaches to Density

In this section, we introduce several proofs of the density result, by considering the one-dimensional case. We start with the following definition.

Definition (3.1)

Let K be a compact set in R and $f \in C(K)$. The modulus of continuity of f is defined as:

$$w(f.\delta) = \sup_{\substack{x,y \in K \\ |x-y|<\delta}} |f(x) - f(y)|$$

If we choose K to be the interval [0, 1], $n \in N$ and we consider the step function $h_n(x)$ which takes the value f(v/n) in the interval $v/n \leq x \leq (v+1)/n$. Obviously this gives :

$\|f - h_n\|_\infty \leq w(f, 1/n)$. It is convenient to write:

$$h_n(x) = f(0) + \sum_{v=1}^{\mu} \{f(v/n) - (f(v-1)/n)\} \ldots\ldots\ldots\ldots\ldots\ldots [3.1]$$

where μ is the largest integer which does not exceed nx.

Definition (3.2)

If σ is a continuous function on R and sigmoidal, we define A_n to be the smallest positive integer, such that:

$$|\,\sigma(x) \le \frac{1}{n} \text{ for } x \le -A_n \text{ and } (1 - \frac{1}{n}) \le \sigma(x) \le (1 + \frac{1}{n}) \text{ for } x \ge A_n$$

Then they define the quasi interpolate g_n as:

$$g_n(x) = f(0) + \sum_{v=1}^{n} \{f(v/n) - f((v-1))/n\} \sigma(A_n(nx - v)) \text{......[3.2]}$$

for $x \in [0, 1]$. Now, we introduce the following theorem:

Theorem (3.3)

There exists a constant c such that $\forall\ f \in C[0,1]$: $\|f - g_n\|_\infty \le cw(f, 1/n)$.

(Note that: Here the uniform norm is taken on the interval [0, 1] and c is independent of f).

Proof : We have:

$$\|f - g_n\|_\infty = \|f - h_n + h_n - g_n\|_\infty \le \|f - h_n\|_\infty + \|h_n - g_n\|_\infty$$

We already know $\|f - h_n\|_\infty \le w(f, 1/n)$ so only the second term need be considered. Now for any $x \in [0. 1]$ with μ defined as in [3.1]we have:

$$h_n(x) - g_n(x) = f(0) + \sum_{v=1}^{\mu} \{f(v/n) - f((v-1)/n)\} - f(0) - \sum_{v=1}^{n} \{f(v/n) - f((v-1)/n)\} \sigma(A_n(nx - v))$$

$$= \sum_{v=1}^{\mu} \{f(v/n) - f((v-1)/n)\}(1 - \sigma(A_n(nx - v))) + \sum_{v=\mu+1}^{n} \{f(v/n) - f((v-1)/n)\} \sigma(A_n(nx - v))$$

Now from definition (3.2) we have:

If $v \leq \mu - 1$ implies $v + 1 \leq \mu$ but $\mu \leq nx$, thus $v + 1 \leq nx$, then $nx - v \geq 1$

So $|1 - \sigma(A_n(nx - v))| \leq \frac{1}{n}$, by the definition of A_n.

Similarly if $v \geq \mu + 2$, implies $|\sigma(A_n(nx - v))| \leq \frac{1}{n}$. Thus:

$$|h_n(x)-g_n(x)| \leq$$

$$\sum_{v=1}^{\mu} \left|\{f(v/n) - f((v-1)/n)\}\right| \left|1 - \sigma(A_n(nx - v))\right| + \sum_{v=\mu+1}^{n} \left|\{f(v/n) - f((v-1)/n)\}\right| \left|\sigma(A_n(nx - v))\right|$$

$$= \frac{4}{n} w(f,1/n)$$

That is, $\|f - g_n\|_\infty \leq w(f, 1/n) + (4/n)\, w(f, 1/n) = (1 + 4/n)w(f, 1/n)$

That is c may be chosen as (1+4/n).Which complete the proof $.\zeta$

Remarks

1. There are two well known methods of passing from one-dimensional to higher- dimensional approximations: the blending operator and the tensor product (3).
2. Suppose we have two sets of basis functions} $\varphi_1, \varphi_2, \ldots, \varphi_\mu\}$ and $\{\psi_1, \psi_2, \ldots, \psi_v\}$ where $\varphi_i, \psi_j : R \longrightarrow R$. The tensor product basis is the set of $\mu \times v$ functions:

   $\alpha_{i,j}(x, y) = \varphi_i(x)\, \psi_j(y)$.

Sometimes one can construct a two-dimensional approximation using the tensor product basis by applying a one-dimensional approximation operator in each dimension.

In practice the two sets are usually the same type of function (e.g. both polynomials or both trigonometric functions) although μ and ν may of course be different. Now. what happens if we apply this construction to ridge functions. For simplicity we assume that the same function σ is to be used for x and y. So typical one-dimensional ridge functions will be $\sigma(a_i x + c_i)$ and $\sigma(b_j y + d_j)$. The tensor product basis thus consists of functions of the form $\sigma(a_i x + c_i)$ $\sigma(b_j y + d_j)$. In

general this does not give a two-dimensional ridge function so we will not land up with a ANN approximation of the form (1.1).However, there is one particular choice of σ for which the construction does work, namely $\sigma(x) = \exp(x)$. Then we get:

$$\sigma(a_i x + c_i)\ \sigma(b_j y + d_j) = \exp(a_i x + c_i)\ \exp(b_j y + d_j)$$

$$= \exp\ (a_i x + b_j y + c_i + d_j)\ = \sigma(a_i x + b_j y + c_i + d_j)$$

The above observation has been used by several authors to produce an n-dimensional ridge function approximations. The basic idea is to prove that the density of the ridge functions for the special case of $\sigma(x) = \exp(x)$ and then to use a. One-dimensional result such as theorem 3.3 to approximate the exponential function by linear Combinations of the desired σ. Now, we introduce the following definition:

Definition (3.4)

A set of functions is said to be fundamental in a given space if a linear combinations of them are dense is that space.

Theorem (3.5)

Let k be a compact set in R^n. Then the set E of functions of the form $\mu(x) = \exp(a^T x)$, where $a \in R^n$, is fundamental in C(K).

Proof

By the Stone-Weierstrass theorem we need only show that the set forms an algebra and separates points. Suppose $x \in K$. First, we have:

$$\exp(a^T x)\ \exp(b^T x) = \exp(a^T x + b^T x) = \exp((a^T + b^T)x).$$

The set also contains the function'1'simply choose a= 0. This establishes that E is an algebra. It remains to show that E separates the points of K. So let $x, y \in K$ with $x \neq y$. Set $a=(x-y)$.Then $a^T(x - y) \neq 0$, so $a^T x \neq a^T y$.Thus $\exp(a^T x) \neq \exp(a^T y)$.The proof is complete.$\zeta$

Before considering more constructive versions of this result we complete the density proof.

Theorem (3.6)

Let K be a compact set in R^n. Then the set F of functions of the form g (x),defined by (1.1) with σ as a continuous sigmoidal function is dense in C(K).

Proof

Let $f \in C(K)$. For any $\varepsilon > 0$, there exists (by theorem 3.5) a finite number m of vectors a_i, such that: $\left\| f - \sum_{i=1}^{m} \exp\left(a_i^T x\right) \right\|_\infty < \frac{\varepsilon}{2}$

since there are only m scalars $a_i^T x$, we may find a finite interval including all of them. Thus there exists a number Γ such that exp $(a_i^T x)$ = $\exp(\Gamma y)$ where $y = (a_i^T x / \Gamma) \in [0,1]$. Then theorem (3.3) tells us that the function exp (Γy) can be approximated by linear combinations functions of the form $\sigma(W_j^T x + c_j)$ with a uniform error less than $\varepsilon/2m$, from which the desired result easily follows. ζ

Remarks

1. Theorem (3.5) tells us one hidden layer is sufficient to approximate any continuous function to any required accuracy.
2. Γ in the proof of theorem (3.6) can be chosen to be an integer, and the numbers A_n, n and v in (3.2) are also integers.
3. The question of rate of convergence of approximations is obviously of considerable importance. If f is smooth and we use smooth approximating functions such as [1.2] we might

hope to get better convergence than the simple $O(1/n)$ which implied by theorem (3.3).

Now, we introduce the following interesting result about density.

Let X a normed vector space over R. The (bounded) dual space of X, denoted by X′, is the space of all bounded linear functionals on X. (A linear functional is a linear mapping from X to R). It has a natural norm defined by: $\| \ell \| = \sup_{\substack{x \in X \\ \|x\|=1}} |\ell(x)|$

X′ is always a Banach space, even if X is not.

Now let V be a subspace of X. We wish to know whether V is dense in X. The relevance of the dual space is shown by the following theorem.

Theorem 3.7 (4)

V is dense in X if and only if the only linear functional $\ell \in X'$ for which $\ell(v) = 0$ for all $v \in V$ is the trivial one $\ell(x) = 0$.

Proof : Suppose first that v is dense in X. Suppose also that ℓ is a linear functional ℓ such that $\ell(v) = 0$ for $v \in V$. Let $x \in X$. For any $\varepsilon > 0$, we have $v \in V$, with $\|x - v\| < \varepsilon$.

Then $|\ell(x)| = |\ell(x) - \ell(v)| = |\ell(x - v)| \le \|\ell\| \, \|x - v\| < \|\ell\|\varepsilon$.

Since this is true for any $\varepsilon > 0$, we must have $\ell(x) = 0$.

This establishes the "only if" part of the theorem.

Now suppose that V is not dense in X. Then there is a $x \in X$ and a number $\delta > 0$ such that $\|x - v\| > \delta$, for all v in V. Let W be the space spanned by x and the space V, i.e., the set of all linear combinations $\alpha x + v$, where $\alpha \in R$ and $v \in V$. Note that $\alpha \in R$ is unique, for if $\alpha_1 x + v_1 = \alpha_2 x + v_1$, we have $(\alpha_1 - \alpha_2)x = v_1 - v_1$, whence we must have $\alpha_1 = \alpha_2$ since $x \notin V$. Thus we can define the following linear functional on W : $\ell(\alpha x + \nu) = \alpha$. Note that $\ell(x) = 1$ and $\ell(\nu)$=0,forall $v \in V$. Now if $w = \alpha x + \nu$ with $\alpha \neq 0$.

$\|w\| = \|\alpha x + \nu\| = |\alpha| \, \|x + \alpha^{-1} v\| \geq |\ell(w)|\delta = |\alpha|\delta \Rightarrow \|w\| \geq |\ell(w)|\delta$ .So $|\ell(w)| \leq \|w\| /\delta$.

On the other hand if $\alpha = 0$, that is $|\ell(w)| = 0$, so the inequality above holds trivially. This shows that ℓ is a non-trivial bounded linear functional on the whole of X. This completes the proof .ζ

Thus if we want to establish density of V in X we need only to show that any linear functional which annihilates V is, in fact, the zero functional.

Poggio, Girosi and Jane in (5) shows:

For approximating a d-variable functions, $f(x_1, x_2, \ldots, x_d)$ with S continuous derivatives, the best achievable approximation accuracy (rate of convergence) is $O(m^{-s/d})$, where $f_m(x, w) = \sum_{j=1}^{m} w_j \varphi_j$, that is they relate the smoothness of the function, the number of basis function, m, and the dimension of the domain, d. It is clear that if m is fixed then as the number of variable increased we get smaller error bound and this is one of the advantage of using ANN, to approximate $f(x_1, x_2, \ldots, x_d)$, instead of using other methods such FEM, where such upper bound get poorer as d increase. However for a given approximation error the number of parameter m exponentially increases with d (for a fixed measure of complexity S). It implies that the number of samples needed for accurate estimation of m (number of basis function or dimension of the space) parameters also grows exponentially with dimensionality d. This result constitutes the curse of dimensionality. If we view S/d as the complexity index of possible trade- off between the smoothness and dimensionality which is the rate of convergence and the number of samples needed for accurate estimation that increases exponentially with complexity index s/d. Thus fast rate of convergence for high-dimensional problems can be obtained, in principle, by imposing stronger smoothness constraints.

Theorem (3.8) (Maiorov and Pinkus, 1999) (6)

There exist $\sigma \in C^{\infty}(R)$ that are sigmoidal and strictly increasing, and have the property that for every $g(x) = \sum_{i=1}^{r} g_i(a^i.x)$, $a^i \in R^n$, $g_i \in C(R)$ and $\varepsilon > 0$ there exist c_i , $\theta_i \in R$ and $W^i \in R^n$, $i = 1, 2, \ldots.\ r - n + 1$ satisfying:

$$\left| g(x) - \sum_{i=1}^{r+n+1} c_i \sigma(W^i x - \theta_i) \right| < \varepsilon$$

for all $x \in B^n$, where B^n denote the unit ball in R^n, that is:

$$B^n = \{x : \|x\|_2 = (x_1^2 + x_2^2 + \ldots + x_n^2)^{1/2} \leq 1\}$$

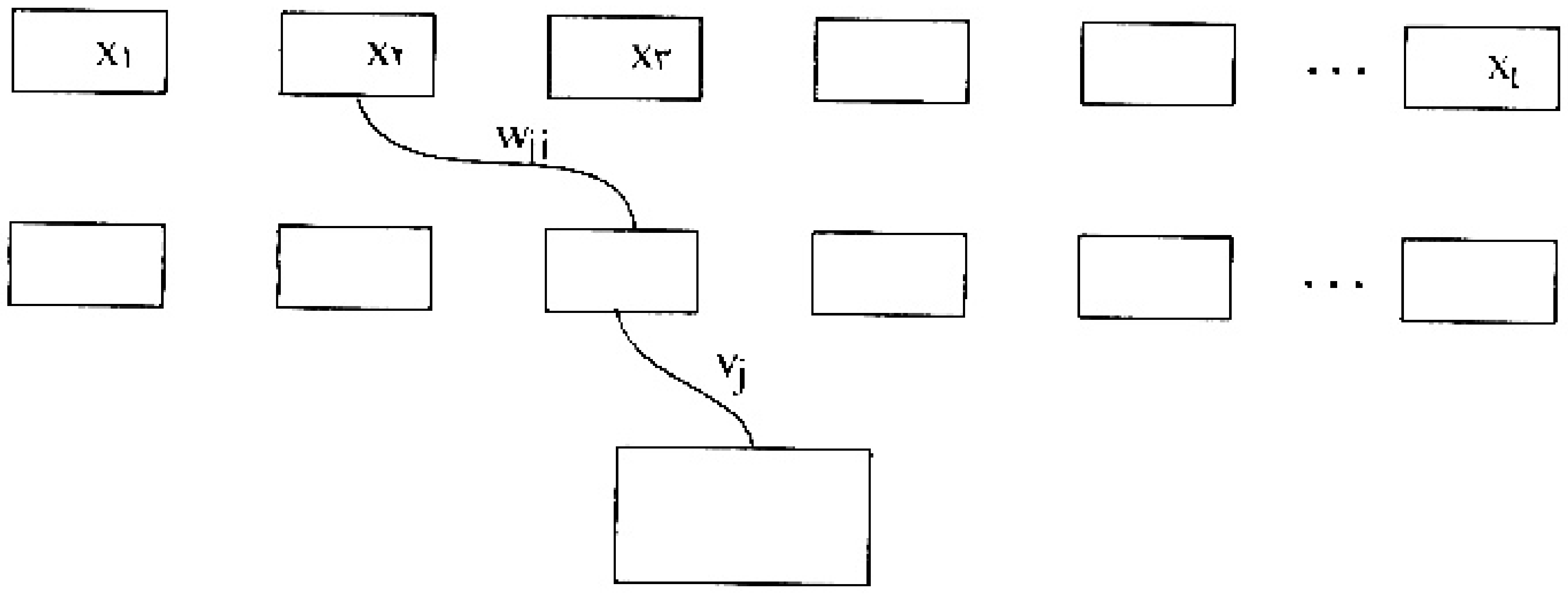
Remark

In practice any approximation process depends not only on the degree (order) of approximation, but also on the possibility, complexity and cost of finding good approximates. The above activation functions are very smooth and give the best degree of approximation. However, they are very difficult to be implemented.

References

1. Cheney ,E. W. and Light, W. (2000), Course in approximation theory, Pub. Books, Colepub. Company,.
2. Kolmogorov ,A . (1950). Mapping Networks: Multi - Layer Data Transformation Structures ,
3. Ellacott, S.W. (1994).A Spects of the Numerical Analysis of Neural Networks, Acta Numerica, P.145 – 202,.
4. Fukuda, T. and Shibata,T. (1992). IEEE Transactions on Industrial elect ronics, <u>39</u> (6),
5. Poggio ,T.; Girosi F . and Jone, M. (1995). J. of Neural Comp.<u>7</u>: 219 - 269,

6. Pinkus,A. (1999). IEEE Transactions on Neural Networks, P. 143-195,



$$\sum_{j=1}^{k} v_j \sigma\left(\sum_{i=1}^{t} w_{ji} x_i + c_j\right) = f(x_1, x_2, \ldots, x_t)$$

Fig.(1) Layers in ANN.

الكثافة والتقريب بأستخدام الشبكات العصبية الصناعية ذات التغذية التقدمية

رياض شاكر نعوم و لمى ناجي توفيق
كلية التربية ابن الهيثم ، جامعة بغداد

الخلاصة

يتضمن البحث دراسة العلاقة بين الكثافة ونموذج الشبكات العصبية ذات التغذية التقدمية حيث يتضمن البحث دراسة كيفية عن تقريب دالة $C(K)$ f حيث أن K مجموعة مرصوصة في R^n

هذا وأن النتائج العددية التي توصلنا اليها تنص على أن الشبكات العصبية الصناعية ذات التغذية التقدمية والتي تحتوي على طبقة خفية واحدة يمكن استخدامها لتقريب أي دالة مستمرة في $C(K)$ و لأي دقة مطلوبة. كذلك تم مناقشة درجة التقريب لبعض المسائل المفتوحة .

كما يتضمن البحث دراسة العلاقة بين الطبقات الخفية في الشبكات العصبية الصناعية ومجموعة الدوال الأساسية والعدد الشرطي للنظام الناتج .