ON FEED FORWARD NEURAL NETWORK WITH RIDGE BASIS FUNCTION

L.N.M.Tawfiq [*]
College of Education (Ibn Al-Haitham)
University of Baghdad

Q.H. *Eqhaar* [**]
College of Science
University of Al-Qadisiya

1. ABSTRACT

In this paper, we show the degree of approximation by a single hidden layer feed forward model with n units in the hidden layer is bounded below by the degree of approximation by a linear combination of n ridge functions. We prove that there exists an analytic, strictly monotone, sigmoidal activation function for which this lower bound is essentially attained.

Also we extend the Kolmogorov's existence theorem to be apply at any compact set, (i.e., closed and bounded set) also we prove that a FFNN with one hidden layer can uniformly approximate any continuous function of several variable, $f(x_1, x_2, \ldots, x_n)$, which is defined in compact set to any required accuracy.

2. INTRODUCTION

A ridge function is a multivariate function $h : R^n \to R$ of the simple form

$h(x_1, \ldots, x_n) = g(a_1x_1 + \ldots + a_nx_n) = g(a \cdot x)$,

where $g : R \to R$ and $a = (a_1, \ldots, a_n) \in R^n \setminus \{0\}$. In other words, it is a multivariate function constant on the parallel hyperplanes $a \cdot x = c$, $c \in R$.

The vector $a \in R^n \setminus \{0\}$ is generally called the direction. Ridge functions appear in various areas and under various guises. We find them in the area of partial differential equations (where they have been known for many, many years under the name of plane waves). We also find them used in computerized tomography , in statistics (where they appear in projection pursuit algorithms), in neural networks, and of course in approximation theory. More about ridge functions may be found in Pinkus, and references therein.

When dealing with ridge functions, one is generally interested in one of three possible sets of functions.[1],[2] The first is given by

$$R(a^1, \ldots, a^m) = \{ \sum_{i=1}^{m} g_i (a^i \cdot x) : g_i \in C(R);\ i = 1, \ldots, m) \}.$$

That is, we fix a finite number of directions and consider linear combinations of ridge functions with these directions. The functions g_i are the "variables". This is a linear space.

The second set is

$$R_m = \{ \sum_{i=1}^{m} g_i (a^i \cdot x) : a^i \in R^n \setminus \{0\},\ g_i \in C(R),\ i = 1, \ldots, m \}$$

Here, we fix m and choose both the functions g_i and the directions a^i. This is not a linear space.

The third set is motivated by a model in neural networks. It is a subset of the second. We fix $\sigma \in C(R)$, called the transfer function in neural network literature, and let

$$N_m = \{ \sum_{i=1}^{m} c_i\ \sigma (a^i \cdot x - b_i) : a^i \in R^n \setminus \{0\},\ c_i, b_i \in R,\ I = 1, \ldots, m \}.$$

Here we also fix m and choose both the directions a^i (called the weights), and the shifts b_i (called the thresholds). This is not a linear space.

3. 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 (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:

1. Density is the theoretical means the ability to approximate well.
2. Density dose not imply a good, efficient scheme for the approximation.
3. 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.

Now, we state Kolmogorov's theorem:

Theorem (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 [3].

As stated in the above theorem, the Kolmogorov mapping network consists of three layers of processing elements (input layer-hiddenlayer and output layer). The first layer (input layer) consists of n input units. The second layer consists of 2n + 1 semilinear 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

No specific example of a function ψ and constant ε are known (still an open problem). 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.

4. 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 theorem:

Theorem (3)

Let B^d denote the unit ball in R^d, i.e., $B^d = \{x: \|X\|_2 \le 1\}$ and S^{d-1} its boundary ,i.e., $S^{d-1} = \{ x : \|X\|_2 = 1 \}$.

There exists a function Φ which is C^∞, strictly increasing and sigmoidal satisfying the follawing . Given $f \in C[-1,1]$ and $\varepsilon > 0$, there exist real constants c_i ,integers r_i and vectors $W^i \in S^{d-1}$, i =1,…, d+1 ,such that $\left| f(x) - \sum_{i=1}^{d+1} c_i \phi(W^i.x - r_i) \right| < \varepsilon$ for all $x \in B^d$.

Proof

The space C[-1,1] is separable .That is ,it contains a countable dense subset .Let $\{u_k\}_{k=1}^{\infty}$ be a subset .Thus to each $f \in C[-1,1]$ and each $\varepsilon > 0$ there exists m (dependent upon f and ε) for which $|f(t) - u_m(t)| \prec \varepsilon$ for all t $\in$ [-1,1] .Assume each u_k is in C^∞ [-1,1] .(we can ,for example , choose the $\{u_k\}_{k=1}^{\infty}$ from among the set of all polynomials with rational cofficients).

We will now construct a sigmoidal function Φ , i.e., for which $\lim_{t\to-\infty} \Phi(t)=0$ and $\lim_{t\to\infty} \Phi(t)=1$, which is strictly increasing and in C^∞ and is such that for each $f \in C[-1,1]$ and each $\varepsilon > 0$ there exists an integer m and real coefficients a_1^m, a_2^m such that $|f(t) - (a_1^m \phi(t) + a_2^m \phi(t))| < \varepsilon$ for all t $\in$ [-1,1].

We do this by constructing ϕ so that $a_1^k \phi(t) + a_2^k \phi(t) = u_k(t)$, for each k .

Let h be any C^∞ ,strictly monotone (with h' (x)>0 for all x) ,sigmoidal function .

We define $\phi(t) = b_k + c_k t + d_k u_k(t)$ for t $\in$ [-1,1]. Where we choose the constants b_k, c_k, d_k so that

1) $\phi(k)=h(k)$
2) $0<\phi'(t) \le h'(t)$ on [k,k+2] .

This is easily done. We make one further assumption .On the intervals [-1,1]and [-2,0] we demand that ϕ again satisfy conditions (1), (2), as above , and be linear , and that $\phi(t)$ be linearly independent on [-1,1] .From the construction there exists , for each $k \ge 1$, real a_1^k ,a_2^k, for which $a_1^k \phi(t) + a_2^k \phi(t) = u_k(t)$ .for all t $\in$ [-1,1]

Thus for some $f \in C[-1,1]$ and $a^j \in S^{d-1}$, j =1,2,…, d+1.

From the above construction of ϕ there exist constants b_1^j, b_2^j and an integer r_j such that $\left| f(x) - (b_1^j \phi(a^j.x - r_j) + b_2^j \phi(a^j.x - r_j)) \right| < \varepsilon$. for all $x \in B^d$. Now each $\phi(a^j.x - r_j)$, j = 1, 2,…, d+1 ,is a linear function , i.e., a linear combination of 1, x_1, x_2 ,…, x_d . Thus $\left| f(x) - \sum_{i=1}^{d+1} c_i \phi(W^i.x - r_i) \right| < \varepsilon$ for all $x \in B^d$.

Theorem (4) [4]

There exists a constant c such that $\forall$ $f \in C[0,1]$.

$\|f - g_n\|_\infty \le c\, w(f, 1/n)$.

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

Remarks

1. There are two well known methods of passing from one- dimensional to higher- dimensional approximations: the blending operator and the tensor product [5]. We can not illustrate both the idea here.
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 v 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 :

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

Where v_j denote the weight connecting j-th hidden unit to the output and 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.

However, there is one particular choice of σ for which the construction does work, namely $\sigma(x) = \exp(x)$. Then we get:

$$\begin{aligned}\sigma(a_i x + c_i)\, \sigma(b_j y + d_j) &= \exp(a_i x + c_i)\, \exp(b_j y + d_j) \\ &= \exp(a_i x + b_j y + c_i + d_j) \\ &= \sigma(a_i x + b_j y + c_i + d_j)\end{aligned}$$

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 (4) to approximate the exponential function by linear Combinations of the desired σ.

Now, we introduce the following definition:

Definition (5)

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

Theorem (6)

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. □

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

Theorem (7)

Let K be a compact set in R^n. Then the set F of functions of the form g(x), defined by (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 6) 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 (6) tells us that the function $\exp(\Gamma y)$ 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. □

Remarks

1. Theorem (7) tells us one hidden layer is sufficient to approximate any continuous function to any required accuracy.
2. Γ in the proof of theorem (7) can be chosen to be an integer
3. The only open problem of the previous paragraph is to show that the vectors a in theorem (6) can be chosen with rational elements.
4. 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 (2) we might hope to get better convergence than the simple O(1/n) which implied by theorem (4).

5. INTERPOLATION

The ability to have a good approximation to a continuous function f is related to the ability to be interpolated by another simpler function (e.g., polynomial). If one can approximate well, then one expects to be able to interpolate (the inverse need not, in general hold).

Assume we are given $\sigma \in C(R)$. For k distinct points $\{x_i\}_{j=1}^{k} \subset R^n$, and associated data $\{d_i\}_{i=1}^{k} \subset R$, can we always find m, $\{w_j\}_{j=1}^{m} \subset R^n$ and $\{c_j\}_{j=1}^{m}$, $\{\theta_j\}_{j=1}^{m} \subset R$ for which

$$\sum_{j=1}^{m} c_j \sigma(W_j x_i - \theta_j) = d_i \quad \text{, for i = 1, 2, …,k.}$$

Furthermore, what is the relationship between k and m ?

If σ is sigmoidal, continuous and non-decreasing, one can always interpolate with m = k. But the open problem is extend this result to any bounded, continuous, non-linear σ which has a limit at infinity. In other word we can define the interpolation as the following:

Given a set of k ordered pairs (x_i, d_i), i = 1, 2, …,k with $x_i \in R^n$ and $d_i \in R$, the problem of interpolation is to find a function $F : R^n \longrightarrow R$ that satisfies the interpolation condition $F(x_i) = d_i$, i = 1, 2, …, k. For strict interpolation, the function F is constrained to pass through all the k data points. The definition can be easily extended to the case where the output is m-dimensional. The desired function is then $F : R^n \longrightarrow R^m$.

In practice, the function F is unknown and must be determined by using the given data (x_i, d_i), i = 1, 2, …, k. A typical neural network implementation of this problem is a two-step process: Training, where the neural network learns how to construct the function F from the given training data $\{x_i, d_i\}$, and generalization, where the neural network predicts the output for a lest input.

6.COMPARISON OF RBF NETWORKS AND MULTILAYER FEED FORWARD NEURAL NETWORK WITH RIDGE BASIS FUNCTION:-

Both RBFNN and FFNN with ridge basis function are non linear layered networks having universal approximation properties ,

The most important differences between them are :

1- An RBFNN has a single hidden layer , while an FFNN with ridge basis function can have several hidden layers.

2- The computational nodes in the FFNN with ridge basis function are similar in various layers , while in the RBFNN they are quite different in the output and hidden layers .

3- In the RBFNN , the output layer is linear , while it may be nonlinear in an FFNN with ridge basis function .This has two consequences :

(a) RBF output layer training is simple .

(b) Model selection becomes almost analytic for RBF's , where as Cross Validation with full network optimization is needed for FFNN with ridge basis functions .

4- In each hidden node ,the activation function of RBFNN computes an Euclidean distance ,while in FFNN with ridge basis functions an inner product between the input and the weight vector is computed.

5- FFNN with ridge basis function construct global approximations ,while RBFNN 's approximation locally nonlinear input-output mappings .

6- FFNN with ridge basis function may require less parameters than the RBFNN for

achieving the same accuracy .

7- RBFNN are usually faster to train .

7.References

[1] E.W,Cheney ,C.K.chui and L.L.Schumaker (eds.),"**Ridge Functions**", Sigmodal Functions and Neural Networks , Approximation Theory VII,pp.158-201.

[2] M.Buhmann and A.Pinkus, "**Identifying Linear Combinations of Ridge Functions Mathematic**", LS8,Universitat Dortmund, Germany, mdb@math.uni-dortmund.de

[3] A.Kolmogorov,"**Mapping Networks : Multi-Layer Data Transformation Structures**", 1950.

[4] L.Naji ,"**On Design And Training of Artificial Neural Networks For Solving Differential Equations**", phD.Thesis , College of Education Ibn Al - Haitham , Bahgdad University, 2004.

.

[5] S.W.Ellacott ,"**A Spaect of the Numerical Analysis of Neural Networks**", Acta Numerca, pp.145-202 ,1994.

حول الشبكات العصبية الصناعية ذات التغذية التقدمية ذو دوال الاساس الصلبة

أ.م.د.لمى ناجي محمد توفيق & قصي حاتم عگار الگفاري

المستخلص

في هذا البحث أثبتنا إن درجة التقريب بواسطة شبكات ذو التغذية التقدمية ذات طبقة خفية واحدة متضمنةn وحدة تكون مقيدة من الأسفل بدرجة تقريب التركيب الخطي لـ n من الدوال الصلبة ,الرتيبة, ذات استثارة نشطة.

كذلك وسعنا نظرية كولموكروف كي تطبق على أي مجموعة مرصوصة أيضا أثبتنا إن الشبكات العصبية ذات التغذية التقدمية والتي تحتوي على طبقة خفية واحدة يمكن استخدامها لتقريب أي دالة مستمرة متعددة المتغيرات معرفة على مجموعة مرصوصة ولأي دقة مطلوبة .