Boletín de Matemáticas
Nueva Serie Vol.I No.2 (1994) y Vol.II No.1 (1995) (35-58)

REDES NEURONALES Y APROXIMACIÓN DE FUNCIONES

LUZ GLORIA TORRES

Departamento de Matemáticas y Estadística
Universidad Nacional de Colombia

RESUMEN. En este artículo se dan a conocer resultados que ofrecen una sustentación teórica al poder computacional que las Redes Neuronales Artificiales han demostrado en muchas aplicaciones en campos diversos. Puesto que la relación entrada/salida de una red neuronal se puede describir en términos de una función, el éxito que estos sistemas han tenido en la solución de gran cantidad de problemas de difícil tratamiento con métodos convencionales, se puede explicar estudiando su capacidad para aproximar funciones. La teoría desarrollada se limita a una arquitectura específica que corresponde a las redes neuronales llamadas perceptrones multicapa.

1. INTRODUCCION

Las redes neuronales artificiales son sistemas de procesamiento masivamente paralelo, compuestas de una gran cantidad de unidades simples que son modelos simplificados de las neuronas biológicas de donde toman su nombre, altamente interconectadas cuyo comportamiento global, debido a interacciones locales, emula algunos procesos de la mente.

Una Red Neuronal Artificial (RNA) se define formalmente como una tripla

$$\mathcal{N} = < D, \{f_i\}, A >$$

donde D es un digrafo contable, localmente finito, con arcos etiquetados, cuyos vértices corresponden a las neuronas, los arcos a las conexiones sinápticas y las etiquetas de los arcos llamadas *pesos*, corresponden a las intensidades de las conexiones sinápticas en los sistemas neuronales naturales; w_{ij} indica el peso de la conexión de la neurona j a la neurona i.

A es un conjunto llamado de *Activación.* Contiene los elementos de "entrada" de las unidades. Posee una estructura de módulo sobre el anillo de pesos, W. En la mayor parte de las aplicaciones, A es el conjunto de números reales $\mathbb{R}$. Una entrada típica a la neurona i está dada por

$$\text{net}_i = \sum_j w_{ij} x_j$$

donde x_j es el valor de salida de la unidad j.

$\{f_i : A \longrightarrow A \, i \in V\}$ (V, conjunto de vértices del grafo D) es una colección de funciones llamadas de *activación* o *transferencia.* La dinámica local de la red neuronal está dada por

$$x_i := f_i(\text{net}_i - \theta_i)$$

donde θ_i es el *valor de umbral* que indica, en el caso más sencillo, que la neurona se activa si la entrada excede este valor y permanece inactiva en caso contrario.

En toda red neuronal se consideran básicamente tres tipos de unidades:

1. *Unidades de Entrada* que, como su nombre lo indica, reciben los valores de entrada del sistema y los pasan a otras unidades.
2. *Unidades de salida* que contienen los valores de "respuesta" de la red neuronal después de cada proceso computacional.
3. *Unidades ocultas* que no tienen comunicación con el "ambiente externo" y que el sistema utiliza para representaciones internas.

Las redes neuronales más comunes tanto en el tratamiento teórico como en las aplicaciones son las llamadas *perceptrones multicapa.* En estos sistemas

las neuronas se organizan en "capas", las unidades de una capa se conectan con unidades de las capas siguientes. El patrón de conexión más utilizado es aquel donde las unidades de una capa se conectan únicamente con las unidades de la capa siguiente (ver figura 1)

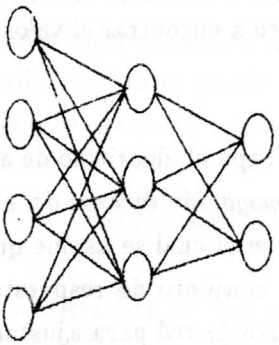

Figura 1

Las funciones de activación de mayor utilización son las que llamaremos *sigmoidales*. Una función sigmoidal es una aplicación

$$\sigma : \mathbb{R} \longrightarrow I \quad (I = [0,1])$$

tal que $\lim_{t\to+\infty} \sigma(t) = 1$ y $\lim_{t\to-\infty} \sigma(t) = 0$

Las siguientes funciones son sigmoidales:

La función de Heaviside
$$f(x) = \begin{cases} 1 & \text{si } x > 0 \\ 0 & \text{en otro caso} \end{cases} \tag{1}$$

La función sigmoide o logística
$$f(x) = \frac{1}{1+e^{-x}} \tag{2}$$

La función "rampa unidad"
$$f(x) = \begin{cases} 1 & \text{si } x \geq 1 \\ x & \text{si } 0 \leq x < 1 \\ 0 & \text{si } x < 0 \end{cases} \tag{3}$$

La función coseno sigmoidal
$$f(x) = \begin{cases} 0 & \text{si } -\infty < x \leq -\frac{\pi}{2} \\ \frac{\cos(x+\frac{3\pi}{2})+1}{2} & \text{si } -\frac{\pi}{2} \leq x \leq \frac{\pi}{2} \\ 1 & \text{si } \frac{\pi}{2} \leq x < \infty \end{cases} \tag{4}$$

Una de las características de las redes neuronales, que establece una diferencia fundamental con modelos convencionales de computación, es la capacidad de realizar tareas complejas mediante *aprendizaje* y *generalización* a partir de ejemplos. En estos sistemas neuronales la información se almacena en las conexiones sinápticas, lo que significa que el aprendizaje en un problema específico se reduce a encontrar el valor apropiado para los pesos de estas conexiones.

Para los perceptrones multicapa el algoritmo de aprendizaje más común es el llamado *algoritmo de propagación inversa del error*, basado en el método de descenso en gradiente y en el cual se asume que se conoce tanto el conjunto de entradas como el conjunto de respuestas correctas correspondientes, que serán utilizadas por la red para ajustar los pesos, de tal manera que, el error entre los valores de salida de la red y los valores correctos dados, sea mínimo.

Según este algoritmo los pesos de las conexiones se cambian de acuerdo con la fórmula

$$w_{ij}^{\text{nuevo}} = w_{ij}^{\text{anterior}} + \eta \Delta w_{ij}$$

donde $\Delta w_{ij} = -\frac{\partial E}{\partial w_{ij}}$, η es una constante positiva llamada *parámetro* o *rata de aprendizaje*; E es la función de error dada en términos de la suma del cuadrado de las diferencias entre los valores correctos y los valores de salida de la red: Si para la entrada $\mathbf{x} = (x_1, x_2, \ldots, x_n)$, $\mathbf{d} = (d_1, d_2, \ldots, d_m)$ corresponde a los valores correctos y $\mathbf{y} = (y_1, y_2, \ldots, y_m)$ a los valores de salida de la red, entonces E está dada por

$$E = \frac{1}{2} \sum_j (d_j - y_j)^2$$

En un perceptrón multicapa tanto la entrada como la salida se pueden considerar como vectores con componentes reales y cuya dimensión está determinada por el número de unidades en la capa respectiva. La relación

entrada/salida permite expresar la operación básica de la red en términos de una función

$$f : \mathbb{R}^n \longrightarrow \mathbb{R}^m$$

lo cual conduce a un estudio de las redes neuronales a la luz de la teoría de aproximación de funciones pues es natural preguntarse sobre el tipo de funciones que un perceptrón multicapa puede representar exactamente, o aproximar, en el sentido que se precisará más adelante.

2. Preliminares

El éxito en tareas de aproximación de funciones en muchas aplicaciones de las redes neuronales, llevó a los investigadores a buscar en la matemática, resultados que puedan fundamentar teóricamente estas capacidades.

Robert Hecht Nielsen [15] (1987) llamó la atención sobre un teorema de Kolmogorov [26,35,36] quien, hacia 1957, junto con el también ruso Arnold, trabajaron en la solución del problema 13 de Hilbert donde se conjetura que las raíces de la ecuación

$$x^7 + ax^3 + bx^2 + cx + 1 = 0$$

como funciones de los coeficientes a, b, y c, no son representables como sumas y superposiciones de funciones ni siquiera de dos variables.

Kolmogorov demostró que una función continua de valor real definida en el cubo n-dimensional I^n ($I = [0, 1]$) puede representarse como sumas y composiciones de funciones de una sola variable:

Teorema 2.1 (Kolmogorov 1957). *Para cada entero $n \geq 2$ existen $n \times (2n + 1)$ funciones monótonas crecientes $\psi_{pq}, p = 1, 2, \ldots, n$ y $q = 1, 2, \ldots, 2n + 1$ con la siguiente propiedad:*

Para cada función continua de valor real, $f : I^n \longrightarrow \mathbb{R}$ existen funciones

continuas $\phi_q, q = 1, 2, \ldots, 2n+1$, *tales que*

$$f(x_1, x_2, \ldots, x_n) = \sum_{q=1}^{2n+1} \phi_q \Big[\sum_{p=1}^{n} \psi_{pq}(x_p) \Big]$$

Las funciones ψ_{pq} *son universales, es decir, no dependen de* f. *Las funciones* ϕ_q *son continuas, de una sola variable y dependen de* f.

Sprecher [35] (1965) mejoró la representación de Kolmogorov demostrando que las funciones ψ_{pq} pueden ser reemplazadas por funciones de la forma $\lambda^p \psi$ donde λ es una constante y ψ es una función monótona creciente que satisface una condición de Lipschitz ($|f(x) - f(y)| \leq c|x - y|^\alpha$, $\alpha = \frac{\ln 2}{\ln(2n+2)}$, c constante)

Hecht-Nielsen aplicó la versión de Sprecher a la neurocomputación; de esta manera demuestra la existencia de redes neuronales que representan o implementan funciones continuas.

Teorema 2.2 (Hecht-Nielsen 1987). *Dada una función continua* $f : I^n \longrightarrow \mathbb{R}^m$, f *puede ser implementada por una red neuronal de tres capas, con* n *unidades en la capa de entrada,* $2n+1$ *unidades en la capa oculta,* m *nodos en la capa de salida.*
El valor de salida de los nodos de la capa oculta está dado por

$$z_k = \sum_{j=1}^{n} \lambda^k \psi(x_j + \epsilon k) + k$$

donde λ *es constante,* ψ *es una función real monótona creciente,* λ *y* ψ *son independientes de* f *(aunque dependen de* n*),* ϵ *es un número racional,* $0 < \epsilon \leq \delta$, δ *es una constante positiva arbitraria y* ψ *satisface una condición de Lipschitz,* $|\psi(x) - \psi(y)| \leq c|x - y|^\alpha$, $0 < \alpha \leq 1$
Los m *elementos de la capa de salida tienen la siguiente función de activación:*

$$y_i = \sum_{k=1}^{2n+1} \psi_i(z_k)$$

donde las funciones ψ_i, $i = 1, \ldots, m$ *son reales y continuas y dependen de* f *y* ϵ

Recientemente Sprecher [36] (1993) dio una versión más fuerte de su teorema donde demuestra que la función ψ no depende de n. Este resultado es válido también en el teorema de existencia de redes neuronales que representan funciones continuas.

Uno de los aspectos interesantes en las aplicaciones de redes neuronales multicapa es que, hasta el momento, se carece de un método general en la determinación del número óptimo de unidades en la capa oculta para la solución de un determinado problema. Curiosamente el teorema de Hecht-Nielsen establece un número exacto de unidades en la capa interna pero a costa de precisión en las funciones de activación de estas unidades.

El problema de aproximación de funciones se formula en términos de densidad de conjuntos de funciones en espacios topológicos definidos naturalmente por métricas, siendo las más comunes, la métrica derivada de la norma del supremo y las derivadas de las normas en espacios L_p, $p \geq 1$:

Denotemos por $\mathcal{C}[X]$ al espacio de funciones continuas de valor real definidas en un conjunto $X \subset \mathbb{R}^n$. Si $f \in \mathcal{C}[X]$, se define la norma del supremo de la siguiente manera:

$$\|f\| = \sup_{\mathbf{x} \in X} \{|f(\mathbf{x})|\}$$

Esta norma induce la métrica

$$d(f, g) = \|f - g\| = \sup\{|f(\mathbf{x}) - g(\mathbf{x})|\}$$

Si $(\Omega, \mathcal{F}, \mu)$ es un espacio de medida, p un número real, $p \geq 1$, el espacio L_p está formado por las funciones medibles f tales que

$$\int_{\Omega} |f|^p d\mu < \infty$$

Se define la norma de f como $\|f\|_p = \left(\int_\Omega |f|^p d\mu\right)^{\frac{1}{p}}$ la cual induce la seudométrica $d_p(f,g) = \|f-g\|_p$. Si dos funciones que difieren solo en un conjunto de medida cero se consideran equivalentes, entonces, d_p es una métrica.

L_∞ es el espacio de las funciones *esencialmente acotadas*, es decir, las funciones acotadas por fuera de un conjunto de medida cero. En este caso, $\|f\|_\infty = esssup|f|$ donde

$$esssup\ g = \inf\{c : \mu\{w : g(w) > c\} = 0\}$$

que corresponde al número c más pequeño tal que $g \leq c$, casi en todas partes.

El interés de los investigadores se ha centrado en estudiar las capacidades de aproximación de los perceptrones multicapa con una o más capas ocultas, cuyas unidades poseen la misma función de activación y las unidades de salida tienen una función de activación lineal. Es suficiente considerar perceptrones con una sola unidad de salida; en este caso la red implementa una aplicación

$$f : \mathbb{R}^n \longrightarrow \mathbb{R}$$

La generalización a redes neuronales con varias unidades de salida es inmediata: debido a la universalidad de las funciones ψ_{pq} en el teorema de Kolmogorov, la representación de una función $g : \mathbb{R}^n \longrightarrow \mathbb{R}^m$ se puede lograr a partir de m perceptrones multicapa con una sola unidad de salida, donde cada uno de los cuales lleva a cabo la implementación de una función $f_i : \mathbb{R}^n \longrightarrow \mathbb{R}$, $i = 1, \ldots, m$; en este caso $g(\mathbf{x}) = (f_1(\mathbf{x}), \ldots, f_m(\mathbf{x}))$.

Puesto que el teorema de Hecht-Nielsen no precisa el tipo de funciones de activación de la capa oculta, la caracterización de tales funciones ha sido tema de investigación desde 1987 en el estudio de las capacidades de aproximación de las redes neuronales, pero más que considerar las funciones de activación adecuadas para una representación exacta, se ha hecho énfasis

en el tipo de funciones de activación mediante las cuales la red neuronal aproxima convenientemente ciertas funciones. Se reseñan a continuación algunos resultados en este sentido:

Gallant y White [10] (1988) demuestran que una red neuronal de tres capas es capaz de efectuar una aproximación de una función por series de Fourier. Utiliza como función de activación para las unidades de la capa intermedia una función coseno sigmoidal definida previamente en (4).

Funahashi [9] (1989) prueba que una función continua definida en un conjunto compacto de $\mathbb{R}^n$ puede ser aproximada por una red neuronal de tres capas. Las funciones de activación son no constantes, acotadas, monótonas crecientes y continuas. Sus demostraciones se basan en el análisis de Fourier y los teoremas de Paley-Wiener

Cybenko [7] (1989) obtiene los mismos resultados de Funahashi pero utilizando el teorema de Hahan-Banach y el teorema de representación de Riesz (ver Teorema 3.2).

Hornik, Stinchcombe y White [18] (1989) demuestran que una red neuronal de tres capas cuya capa oculta tiene funciones de activación sigmoidales puede aproximar funciones continuas y medibles. Los mismos autores en [19] (1990) prueban que redes neuronales con la misma arquitectura pueden aproximar funciones y sus derivadas en espacios de Sobolev. En [20] (1991) las condiciones de las funciones de activación son debilitadas: basta que sean funciones acotadas y no constantes.

Blum y Li [3] (1991) muestran cómo una red neuronal de dos capas internas y con funciones de Heaviside como funciones de activación aproxima funciones continuas y medibles (ver Teorema 4.1).

Hornik [21] (1993) amplía el conjunto de funciones de activación. Percep-

trones de tres capas con funciones de activación localmente integrables (en el sentido de Riemman) y no polinomiales pueden aproximar una función continua y con funciones de activación esencialmente acotadas y no polinomiales son capaces de aproximar funciones medibles (ver Teoremas 3.6 y 3.7).

Es conveniente anotar que en los resultados obtenidos la cantidad de unidades de las capas internas no es acotada. Se admite siempre la posibilidad de tener un "número suficiente" de neuronas para lograr una "buena aproximación".

3. Aproximacion con funciones sigmoidales

Los perceptrones multicapa que calculan funciones de la forma

$$f(\mathbf{x}) = \sum_{i=1}^{k} a_i \sigma(\mathbf{b}_i \cdot \mathbf{x} + c_i) \tag{5}$$

donde σ es una función sigmoidal, $a_i, c_i \in \mathbb{R}$, $\mathbf{b}_i, \mathbf{x} \in \mathbb{R}^n$, $k \in \mathbb{N}$, son ampliamente usados en las aplicaciones de redes neuronales.
Denotaremos $S(\sigma)$ al conjunto de funciones de tipo (5) con $f : \mathbb{R} \longrightarrow I$, $I = [0, 1]$. Kùrková [25] (1992) utiliza la propiedad de las funciones en $S(\sigma)$ de aproximar una función continua en un intervalo cerrado ($S(\sigma)$ es denso en $C([a, b])$ con la topología de la convergencia uniforme), para establecer un teorema de aproximación derivado del teorema de representación de Kolmogorov.

Teorema 3.1 (Kùrková 1992). *Sean $n \in \mathbb{N}$, $n \geq 2$, $\sigma : \mathbb{R} \longrightarrow I$ una función sigmoidal continua, $f \in C(I^n)$ y ϵ un número real positivo, entonces existen $k \in \mathbb{N}$ y funciones ϕ_i, $\psi_{pi} \in S(\sigma)$ tales que*

$$\left| f(x_1, \ldots, x_n) - \sum_{i=1}^{k} \phi_i \left(\sum_{p=1}^{n} \psi_{pi}(x_p) \right) \right| < \epsilon$$

para todo $(x_1, \ldots, x_n) \in I^n$

Demostración. Según el teorema de Kolmogorov existen funciones continuas monótonas crecientes ψ_{pq} y funciones continuas ϕ_p tales que

$$f(x_1, \ldots, x_n) = \sum_{q=1}^{2n+1} \phi_q \left(\sum_{p=1}^{n} \psi_{pq}(x_p) \right)$$

Sea $[a, b] \subset \mathbb{R}$ tal que, para $p = 1, \ldots, n,\ q = 1, \ldots, 2n+1\ \ \psi_{pq}(I) \subset [a, b]$

Puesto que el conjunto de funciones $f : [a, b] \longrightarrow \mathbb{R}$, con $f(x) = \sum_{i=1}^{k} w_i \sigma(v_i x + u_i)$ donde $\sigma : \mathbb{R} \longrightarrow I$ es una función sigmoidal continua y $w_i,\ v_i,\ u_i\ i = 1, \ldots, k$ son números reales, es denso en $C[a, b]$, para cada $q = 1, \ldots, 2n+1$, existe $g_p \in S(\sigma)$ tal que

$$|g_q(x) - \phi_q(x)| < \frac{\epsilon}{(2n)(2n+1)} \text{ para todo } x \in [a, b] \tag{6}$$

ya que las funciones g_q son uniformemente continuas, existe $\delta > 0$ tal que

$$\text{si } |x - y| < \delta \text{ entonces } |g_q(x) - g_q(y)| < \frac{\epsilon}{(2n)(2n+1)} \tag{7}$$

Además, para $p = 1, \ldots, n,\ q = 1, \ldots, 2n+1$ existen $h_{pq} \in S(\sigma)$ tales que

$$|h_{pq}(x) - \psi_{pq}(x)| < \delta \text{ para todo } x \in I \tag{8}$$

De (8) y (7) se concluye

$$\left| \sum_{p=1}^{n} h_{pq}(x) - \sum_{p=1}^{n} \psi_{pq}(x) \right| < n\delta$$

y

$$\left| \sum_{q=1}^{2n+1} g_q \left(\sum_{p=1}^{n} h_{pq}(x_p) \right) - \sum_{q=1}^{2n+1} g_q \left(\sum_{p=1}^{n} \psi_{pq}(x_p) \right) \right| < \frac{\epsilon}{2}$$

La condición (6) conduce a

$$\left|\sum_{q=1}^{2n+1}\phi_q\left(\sum_{p=1}^{n}\psi_{pq}(x_p)\right)-\sum_{q=1}^{2n+1}g_q\left(\sum_{p=1}^{n}\psi_{pq}(x_p)\right)\right|<\frac{\epsilon}{2}$$

Por lo tanto,

$$\left|f(x_1,x_2,\ldots,x_n)-\sum_{q=1}^{2n+1}g_q\left(\sum_{p=1}^{n}h_{pq}(x_p)\right)\right|$$

$$=\left|\sum_{q=1}^{2n+1}\phi_q\left(\sum_{q=1}^{n}\psi_{pq}(x_p)\right)-\sum_{q=1}^{2n+1}g_q\left(\sum_{p=1}^{n}h_{pq}(x_p)\right)\right|$$

$$\leq\left|\sum_{q=1}^{2n+1}\phi_q\left(\sum_{p=1}^{n}\psi_{pq}(x_p)\right)-\sum_{q=1}^{2n+1}g_q\left(\sum_{p=1}^{n}\psi_{pq}(x_p)\right)\right|$$

$$+\left|\sum_{q=1}^{2n+1}g_q\left(\sum_{p=1}^{n}\psi_{pq}(x_p)\right)-\sum_{q=1}^{2n+1}g_q\left(\sum_{p=1}^{n}h_{pq}(x_p)\right)\right|$$

$$<\frac{\epsilon}{2}+\frac{\epsilon}{2}=\epsilon$$

En la expresión (5) se sugiere que la red neuronal que calcula la función f tiene a σ como función de activación en las unidades de la capa oculta. Ya que en las aplicaciones muy difícilmente aparecen funciones que no sean continuas o medibles, para estudiar las capacidades de aproximación de este tipo de redes, basta investigar las condiciones que debe satisfacer σ para que el conjunto de las funciones (5) sea denso en $\mathcal{C}(K)$, $K\subset\mathbb{R}^n$, K compacto y las condiciones bajo las cuales ese conjunto es denso en $L_p(K)$, $K\subset\mathbb{R}^n$, K medible.

$\mathcal{M}(I^n)$ denota el espacio de medidas regulares de Borel, signadas, definidas en I^n

Recordamos que una medida μ definida sobre un espacio topológico Ω es *regular* si

$$\mu(E) = \inf\{\mu(V) : V \supset E,\ V \text{ abierto}\}$$
$$= \sup\{\mu(C) : C \subset E,\ C \text{ cerrado}\}$$

para todo conjunto de Borel $E \subset \Omega$

Una medida *signada* o *con signo* se obtiene a partir de una medida, si admite valores positivos y negativos.

Definición. Una función $\sigma : \mathbb{R} \longrightarrow \mathbb{R}$ es *discriminatoria* si para una medida $\mu \in \mathcal{M}(I^n)$

$$\int_{I^n} \sigma(\mathbf{y} \cdot \mathbf{x} + \theta) d\mu(\mathbf{x}) = 0 \quad \text{para todo } \mathbf{y} \in \mathbb{R}^n,\ \theta \in \mathbb{R}$$

implica $\mu = 0$

Teorema 3.2 (Cybenko 1989). *Sea σ una función continua discriminatoria. Entonces el conjunto de funciones $f : I^n \longrightarrow \mathbb{R}$*

$$S = \left\{ f : f(\mathbf{x}) = \sum_{j=1}^{k} \alpha_j \sigma(\mathbf{y}_j \cdot \mathbf{x} + \theta_j),\ k \in \mathbb{N},\ \alpha_j \in \mathbb{R},\ \mathbf{y}_j \in \mathbb{R}^n, \theta_j \in \mathbb{R} \right\}$$

es denso en $C(I^n)$

Demostración. Es claro que $S \subset C(I^n)$. Afirmamos que $\overline{S} = C(I^n)$. Si $\overline{S} \subsetneq C(I^n)$, por el teorema de Hahn Banach ([1] pag.141), existe un funcional lineal acotado $F \neq 0$ sobre $C(I^n)$ tal que

$$F(\overline{S}) = F(S) = 0$$

Por el teorema de representación de Riesz ([32] pag.121) este funcional lineal es de la forma

$$F(h) = \int_{I^n} h(\mathbf{x}) d\mu(\mathbf{x})$$

para alguna medida diferente de cero $\mu \in \mathcal{M}(I^n)$ y para todo $h \in C(I^n)$

Ya que $\sigma(\mathbf{y} \cdot \mathbf{x} + \theta) \in \mathbb{R}$ para todos $\mathbf{y} \in \mathbb{R}^n$ y $\theta \in \mathbb{R}$ entonces

$$\int_{I^n} \sigma(\mathbf{y} \cdot \mathbf{x} + \theta) d\mu(\mathbf{x}) = 0$$

para todos $\mathbf{y} \in I^n$, $\theta \in \mathbb{R}$

Puesto que σ es discriminatoria, se tiene $\mu = 0$, lo cual es imposible. Por lo tanto, S es denso en $C(I^n)$

Se demuestra ahora que una función sigmoidal acotada y medible es discriminatoria. Con esto queda probado, como caso especial, que un perceptrón multicapa con una capa oculta y con funciones de activación sigmoidales continuas, aproxima cualquier función continua de valor real definida en I^n

Lema 3.1. *Una función sigmoidal medible y acotada $\sigma : \mathbb{R} \longrightarrow \mathbb{R}$ es discriminatoria. En particular, las funciones sigmoidales continuas son discriminatorias.*

Demostración. Sea $\sigma_\lambda(\mathbf{x}) = \sigma(\lambda(\mathbf{y} \cdot \mathbf{x} + \theta) + \phi)$ entonces

$$\lim_{\lambda \to +\infty} \sigma_\lambda(\mathbf{x}) = \begin{cases} 1 \text{ si} & \mathbf{y} \cdot \mathbf{x} + \theta > 0 \\ 0 \text{ si} & \mathbf{y} \cdot \mathbf{x} + \theta < 0 \end{cases}$$

y $\sigma_\lambda(\mathbf{x}) = \sigma(\phi)$si $\mathbf{y} \cdot x + \theta = 0$

Es decir, σ_λ converge puntualmente a la función

$$\alpha(\mathbf{x}) = \begin{cases} 1 & \text{si } \mathbf{y} \cdot \mathbf{x} + \theta > 0 \\ 0 & \text{si } \mathbf{y} \cdot \mathbf{x} + \theta < 0 \\ \sigma(\phi) & \text{si } \mathbf{y} \cdot \mathbf{x} + \theta = 0 \end{cases}$$

Suponemos que $\int_{I^n} \sigma(\lambda(\mathbf{y} \cdot \mathbf{x} + \theta) + \phi) d\mu(\mathbf{x}) = 0$ para una medida $\mu \in \mathcal{M}(I^n)$ y demostraremos que $\mu = 0$

Sea $\mu \in \mathcal{M}(I^n)$, por el teorema de convergencia de Lebesgue, se tiene

$$\int_{I^n} \sigma(\lambda(\mathbf{y} \cdot \mathbf{x} + \theta) + \phi) d\mu(\mathbf{x}) = \int_{I^n} \sigma_\lambda(\mathbf{x}) d\mu(\mathbf{x}) = \int_{I^n} \alpha(\mathbf{x}) d\mu(\mathbf{x})$$

$$= \int_{P_{\mathbf{y},\theta}} d\mu(\mathbf{x}) + \int_{H_{\mathbf{y},\theta}} \sigma(\phi) d\mu(\mathbf{x}) = \mu(H_{\mathbf{y},\theta}) + \sigma(\phi)\mu(P_{\mathbf{y},\theta}) \qquad (9)$$

para todos ϕ, θ, donde $H_{\mathbf{y},\theta} = \{\mathbf{x} : \mathbf{y}\cdot\mathbf{x}+\theta = 0\}$, $P_{\mathbf{y},\theta} = \{\mathbf{x} : \mathbf{y}\cdot\mathbf{x}+\theta > 0\}$ Si

$$\mu(H) + \mu(P) = 0 \qquad (10)$$

para todo hiperplano H y "semiplano" P, entonces $\mu = 0$. En efecto,
Sea $\mathbf{y}$ fijo. Para una función acotada medible h, se define un funcional lineal F de la manera siguiente:

$$F(h) = \int_{I^n} h(\mathbf{y}\cdot\mathbf{x}) d\mu(\mathbf{x})$$

Si h es la función indicadora sobre $[0, +\infty)$, entonces $F(h) = \mu(H_1) + \mu(P_1)$ donde $P_1 = \{\mathbf{x} : \mathbf{y}\cdot\mathbf{x} - \theta > 0\}$ y $H_1 = \{\mathbf{x} : \mathbf{y}\cdot\mathbf{x} - \theta = 0\}$. Por (10) se tiene que para esta función h, $F(h) = 0$. Igual resultado se obtiene si h es la función indicadora sobre $(0, \infty)$ o sobre cualquier intervalo, y ya que F es lineal, $F(h) = 0$ si h es una función simple (suma de funciones indicadoras sobre intervalos).

Puesto que $F \in L_\infty(\mathbb{R})$ (μ es finita) y el conjunto de funciones simples es denso en $L_\infty(\mathbb{R})$, entonces $F = 0$

En particular para las funciones medibles acotadas $s(u) = \text{sen}(\mathbf{m}\cdot\mathbf{u})$ y $c(u) = \cos(\mathbf{m}\cdot\mathbf{u})$ se tiene:

$$F(s + ic) = \int_{I^n} (\cos(\mathbf{m}\cdot\mathbf{x}) + i\text{sen}(\mathbf{m}\cdot\mathbf{x}))\, d\mu(\mathbf{x})$$

$$= \int_{I^n} e^{i\mathbf{m}\cdot\mathbf{x}} d\mu(\mathbf{x}) = 0 \quad \text{para todo } \mathbf{m}$$

luego la transformada de Fourier de μ es 0 y por lo tanto $\mu = 0$ ([34] pag. 187) y se concluye que σ es discriminatoria.

Aunque las funciones sigmoidales continuas son las más utilizadas por cuanto permiten que la red neuronal aprenda con el algoritmo de propagación inversa , las funciones sigmoidales discontinuas como la función de

Heaviside son de interés porque están relacionadas con el perceptrón simple clásico (que consta de una capa de entrada y una capa de salida cuyas unidades poseen funciones de activación no continuas). El teorema 3.3 nos permite concluir que un perceptrón de tres capas cuya capa interna posee funciones de activación sigmoidales *continuas o discontinuas* es capaz de aproximar funciones más generales.

Si la definición de función discriminatoria se modifica de la manera siguiente:

Para $h \in L_\infty(I^n)$, si $\int_{I^n} \sigma(\mathbf{y} \cdot \mathbf{x} + \theta)h(\mathbf{x})d\mathbf{x} = 0$ para todo y y todo θ, entonces $h(\mathbf{x}) = 0$ casi en todas partes, se sigue de inmediato que las funciones sigmoidales en general son discriminatorias (las medidas $h(\mathbf{x})d\mathbf{x} \in \mathcal{M}(I^n)$) y se tiene el siguiente resultado:

Teorema 3.3 (Cybenko 1989). *Sea σ una función sigmoidal acotada y medible. El conjunto de funciones*

$$\left\{ f : I^n \longrightarrow \mathbb{R} | f(\mathbf{x}) = \sum_{j=1}^{k} \alpha_j \sigma(\mathbf{y}_j \cdot \mathbf{x} + \theta_j),\ \alpha_j \in \mathbb{R}, \theta_j \in \mathbb{R},\ \mathbf{y}_j \in \mathbb{R}^n,\ k \in \mathbb{N} \right\}$$

es denso en $L_1(I^n)$

4. Aproximacion con otras funciones de activacion

Hornik [20] (1991) extendió los resultados anteriores al caso de redes neuronales cuyas funciones de activación en la capa interna son acotadas y no constantes. Estas redes son capaces de aproximar funciones continuas definidas en conjuntos compactos y funciones definidas en conjuntos medibles.

Teorema 3.4 (Hornik 1991). *Si $\sigma : \mathbb{R} \longrightarrow \mathbb{R}$ es continua, acotada y no constante entonces el conjunto de funciones $f : \mathbb{R}^n \longrightarrow \mathbb{R}$*

$$S = \left\{ f : f(\mathbf{x}) = \sum_{j=1}^{k} a_j \sigma(\mathbf{y}_j \cdot \mathbf{x} + \theta_j),\ k \in \mathbb{N},\ \mathbf{y}_j \in \mathbb{R}^n,\ a_j \in \mathbb{R},\ \theta_j \in \mathbb{R} \right\}$$

es denso en $C(K)$ para todo subconjunto compacto $K \subset \mathbb{R}^n$

Teorema 3.5 (Hornik 1991). *Si $\sigma : \mathbb{R} \longrightarrow \mathbb{R}$ es acotada y no constante entonces el conjunto de funciones S es denso en $L_p(\mu)$ para todas las medidas finitas μ sobre $\mathbb{R}^n$*

La demostración de los teoremas 3.4 y 3.5 sigue los mismos pasos de la demostración del teorema 3.2 una vez que se tenga el siguiente resultado:

Lema 3.2. *Si $\sigma : \mathbb{R} \longrightarrow \mathbb{R}$ es acotada y no constante, entonces σ es discriminatoria.*

Recientemente Hornik [21] (1993) demostró que las capacidades de aproximación de una red neuronal de tres capas se conservan si las unidades internas poseen funciones de activación integrables (en el sentido de Riemann) y no polinomiales.

Sean $A \subset \mathbb{R}^n$, $\Theta \subset \mathbb{R}$. Denotamos

$$S_A^\Theta = \left\{ f : f(\mathbf{x}) = \sum_{j=1}^{k} a_j \sigma(\mathbf{y}_j \cdot \mathbf{x} + \theta_j),\ a_j \in \mathbb{R},\ \mathbf{y}_j \in A,\ \theta_j \in \Theta,\ k \in \mathbb{N} \right\}$$

donde $f : \mathbb{R}^n \longrightarrow \mathbb{R}$

Teorema 3.6 (Hornik 1993). *Sea $\sigma : \mathbb{R} \longrightarrow \mathbb{R}$ una función Riemann-integrable y no polinomial sobre un conjunto compacto Θ. Supongamos que A contiene una vecindad del origen, entonces, para todo compacto $K \subset \mathbb{R}^n$, S_A^Θ contiene un subconjunto denso en $C(K)$*

La demostración de este teorema requiere de dos lemas que se enuncian a continuación (Ver [21])

Denotamos

$$J_\epsilon(\sigma(t)) = \int_{\{|u| \leq 1\}} w(u)\sigma(t - \epsilon u)du$$

donde

$$w(t) = \begin{cases} ce^{-\frac{1}{1-t^2}} & \text{si } |t| < 1 \\ 0 & \text{si } |t| \geq 1 \end{cases}$$

c es una constante que satisface $\int_{\mathbb{R}} w(t)dt = 1$

Lema 3.3. *Sea $\epsilon > 0$ y sea T un intervalo compacto de longitud positiva. Supongamos que $\sigma : \mathbb{R} \longrightarrow \mathbb{R}$ es Riemann-integrable sobre $N_\epsilon(T) = \{s : |s-t| \leq \epsilon$ para algún $t \in T\}$ (σ es acotada y continua casi en todas partes en $N_\epsilon(T)$), entonces $J_\epsilon(\sigma)$ puede ser aproximada (uniformemente) sobre T por una combinación lineal de las funciones*

$$\phi(t) = \sigma(t-s), \ |s| \leq \epsilon$$

Lema 3.4. *Sea $\epsilon_0 > 0$ y T un intervalo compacto de longitud positiva. Sea $\sigma \in L_\infty(N_{\epsilon_0}(T))$ y no polinomial (casi en todas partes) sobre T. Para $0 < \epsilon < \epsilon_o$ sea*

$$m_\epsilon(T) = \inf\{m \geq 0 : D^m J_\epsilon(\sigma(t)) = 0, \ \ t \in \ T\}$$

D^m indica la derivada de orden m, entonces $\limsup_{\epsilon \to 0+} m_\epsilon(T) = \infty$

Demostración del Teorema 3.6. Sean $K \subset \mathbb{R}^n$ compacto, $M = \max\{|\mathbf{x}| : \mathbf{x} \in K\}$. Se seleccionan η_o, $\epsilon_o > 0$ y un intervalo compacto Θ_o de tal forma que

$$A_o = \{\mathbf{a} : |\mathbf{a}| < \eta_o\} \subset A, \quad N_{\eta_o M + \epsilon_o}(\Theta_o) \subset \Theta$$

Por hipótesis σ es no polinomial en Θ_o. Por el lema 3.3, la función

$$\Phi(\mathbf{x}) = J_\epsilon \sigma(\mathbf{a} \cdot \mathbf{x} + \theta), \ \mathbf{a} \in A_o, \ \theta \in \Theta_o, \ 0 < \epsilon \leq \epsilon_o$$

puede ser aproximada por una función en $S_{A_o}^{N_{\epsilon_o}(\Theta_o)}$. Afirmamos que $S_{A_o}^{N_{\epsilon_o}(\Theta_o)}$ es denso en $C(K)$.

Supongamos que no lo es, entonces existe una medida μ, finita signada, sobre K, diferente de cero, tal que

$$H(\mathbf{a},\theta,\epsilon) = \int_K J_\epsilon \sigma(\mathbf{a}\cdot\mathbf{x}+\theta)d\mu(\mathbf{x}) = 0 \qquad (11)$$

para todos $\mathbf{a} \in A_o$, $\theta \in \Theta_o$, $0 < \epsilon < \epsilon_o$

Para cada θ y cada ϵ, H es $\mathcal{C}^\infty$ en $\mathbf{a}$. Sea $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_n)$ un multi-índice. Tomando derivadas parciales de orden α, se tiene

$$\int_K D^\alpha J_\epsilon \sigma(\mathbf{a}\cdot\mathbf{x}+\theta)d\mu(\mathbf{x}) = 0,\ D^\alpha = \left(\left(\frac{\partial}{\partial x_1}\right)^{\alpha_1}, \ldots, \left(\frac{\partial}{\partial x_n}\right)^{\alpha_n}\right)$$

Si $\mathbf{a} = 0$

$$\int_K D^\alpha J_\epsilon \sigma(\mathbf{a}\cdot\mathbf{x}+\theta)d\mu(\mathbf{x}) = D^\alpha J_\epsilon \sigma(\theta)\int_K \mathbf{x}^\alpha d\mu(\mathbf{x}) = 0\ (\mathbf{x}^\alpha = x_1^{\alpha_1}x_2^{\alpha_2}\ldots x_n^{\alpha_n})$$

Por el lema 3.4, existen $\theta \in \Theta_o$ y $0 < \epsilon < \epsilon_o$ tales que $D^m J_\epsilon \sigma(\theta) \neq 0$ para todo $m \geq 0$ luego $\int_K \mathbf{x}^\alpha d\mu(\mathbf{x}) = 0$ para todos los multi-índices α. Pero $\mathbf{x}^\alpha$ aproxima una función continua con soporte compacto (Ejerc.6.1 [35]), entonces

$$\int_K \Phi d\mu(\mathbf{x}) = 0$$

para toda función Φ continua con soporte compacto sobre K. Esto implica que $\mu = 0$ ([35] pag. 188), lo cual contradice la hipótesis. Se concluye entonces que $S_{A_o}^{N_{\epsilon_o}(\Theta_o)}$ es denso en $\mathcal{C}(K)$.

Teorema 3.7 (Hornik 1993). *Sea σ una función esencialmente acotada y no polinomial en un intervalo compacto Θ. Supongamos que A contiene una vecindad del origen. Entonces para toda medida finita, μ, sobre $\mathbb{R}^n$ con soporte compacto, $S_A^\Theta(\sigma)$ contiene un subconjunto denso en $L_p(\mu)$*

Demostración. Sea K el soporte de μ. Se seleccionan η_o, ϵ_oy Θ_o como en el teorema 8. Supongamos que $S_{A_o}^{N_{\epsilon_o}(\Theta_o)}$ no es denso en $L_p(\mu)$, entonces existe

una medida finita signada no nula μ sobre K tal que $\int_K \sigma(\mathbf{a}\cdot\mathbf{x}+\theta)d\mu(\mathbf{x}) = 0$, $\mathbf{a} \in A_o$, $\theta \in N_{\epsilon_o}(\Theta_o)$

Reemplazando θ por $\theta - \epsilon u$, se tiene

$$\int_K J_\epsilon \sigma(\mathbf{a}\cdot\mathbf{x}+\theta)d\mu(\mathbf{x})$$

$$= \int_K \left(\int_{\mathbb{R}} \sigma(\mathbf{a}\cdot\mathbf{x}+\theta-\epsilon u)w(u)du \right) d\mu(\mathbf{x})$$

$$= \int_{\mathbb{R}} \left(\int_K \sigma(\mathbf{a}\cdot\mathbf{x}+\theta-\epsilon u)d\mu(\mathbf{x}) \right) w(u)du$$

$\mathbf{a} \in A_o$, $\theta \in \Theta_o$, $0 < \epsilon < \epsilon_o$

los pasos siguientes son los mismos de la demostración del teorema 3.6.

4. Aproximacion de funciones con funciones de Heaviside

En los resultados presentados hasta ahora, se ha probado la existencia de redes neuronales que aproximan ciertas funciones pero las demostraciones no han incluido la construcción explícita de la red. En la demostración del teorema 4.1 se construye un perceptrón con dos capas internas y cuyas unidades están provistas de las funciones de Heaviside, que aproxima funciones continuas.

Llamaremos *simple con rango finito* a una función $g : K \subset \mathbb{R}^n \longrightarrow \mathbb{R}$ si $K = \bigcup_{i=1}^m D_i$. $D_i \bigcap D_j = \Phi$, $i \neq j, m \in \mathbb{N}$ tal que g es constante sobre cada D_i $i = 1, \ldots, m$

Son las funciones más sencillas que se usan para aproximar funciones. Sabemos que

$$S(K) = \{g : K \longrightarrow \mathbb{R},\ g \text{ simple } \}$$

es denso en $C(K)$, K compacto y puesto que $C(K)$ es denso en $L_2(K)$, K medible, entonces S(K) es denso en $L_2(K)$.

Recordamos que las funciones constantes a trozos sobre particiones finitas son simples y que si $f : \mathbb{R}^n \longrightarrow R$ es arbitraria y las funciones $g_1, g_2, \ldots, g_n$ son simples , entonces la compuesta $f(g_1, g_2, \ldots, g_n)$ es simple.

Teorema 4.1 (Blum y Li 1991). *Sea $f \in C(K)$, $K \subset \mathbb{R}^n$ compacto. Para cualquier $\epsilon > 0$ existe una red neuronal con dos capas internas, cuyas unidades tienen la función de Heaviside como función de activación, con una sola unidad de salida con función de activación lineal, que aproxima a f con error menor que ϵ*

Demostración. Sabemos que f puede ser aproximada (con error menor que ϵ) por una función constante a trozos, g, sobre una partición rectangular sobre K. En este caso g es una función simple y las funciones simples son densas en $C(K)$.

Si se demuestra que g puede ser representada exactamente por una red neuronal con las especificaciones dadas, se prueba de esta manera que esta red efectúa una aproximación de f.

Sea $\mathbf{x} = (x_1, x_2, \ldots, x_n) \in K$. Construimos una red neuronal con n unidades en la capa de entrada, $2n$ unidades en una primera capa oculta, 1 unidad en una segunda capa oculta y una unidad de salida. Cada $x_i, i = 1, \ldots, n$ es el valor del nodo i de entrada de la red. Consideremos una partición rectangular que recubra a K de la forma

$$a_{i_1} < a_{i_2} < \cdots < a_{i_n}, \quad i = 1, 2, \ldots, n$$

Cada nodo de entrada i, se conecta a dos unidades de la primera capa interna. Si $a_{i_j} < x_i \leq a_{i_{j+1}}$, la salida de estas unidades está dada por

$$y_{i1} = H(x_i - a_{i_j}), \qquad y_{i2} = H_1(a_{i_{j+1}} - x_i)$$

donde H es la función de Heaviside y

$$H_1(t) = \begin{cases} 0 & \text{si } t > 0 \\ 1 & \text{si } t \leq 0 \end{cases}$$

Entonces

$$(y_{i1}, y_{i2}) = \begin{cases} (0,1) & \text{si } x_{i_j} \leq a_{i_j} \\ (1,1) & \text{si } a_{i_j} < x_i \leq a_{i_{j+1}} \\ (1,0) & \text{si } x_i > a_{i_{j+1}} \end{cases}$$

Los $2n$ nodos con salida y_{i1}, y_{i2}, $i = 1, \ldots, n$ se conectan a la unidad en la segunda capa oculta. El valor de las conexiones se fija en 1 y el valor de umbral se hace igual a $2n - 1$. La salida de esta unidad está dada por

$$z(\mathbf{x}) = H\left(\sum_{i=1}^{n}(y_{i1} + y_{i2}) - (2n-1))\right)$$

Observamos que $z(\mathbf{x}) = 1$ si el punto $\mathbf{x}$ está en la "caja" de la partición determinada por $a_{i_j} < x_i \leq a_{i_{j+1}}$, y $z(\mathbf{x}) = 0$ en caso contrario.

Esta construcción se hace por cada rectángulo n-dimensional de la partición. Luego, cada salida z se conecta a la única unidad de salida con pesos w_z que coinciden con el valor de g en la "caja" correspondiente. La salida de la red neuronal es entonces

$$g(\mathbf{x}) = \sum w_z z(\mathbf{x})$$

El número de unidades internas que la red neuronal requiere es bastante grande. Por ejemplo, Si $f : [0,1]\times[0,1] \longrightarrow \mathbb{R}$ y si se considera una partición de m^2 "cuadrados", según la construcción del teorema 10, se necesitan $4m^2$ unidades en la primera capa oculta y m^2 en la segunda. Sin embargo, en la primera capa oculta, hay muchos nodos con el mismo valor (para puntos correspondientes a valores extremos de los intervalos de la partición), lo que permite reducir el número de unidades a $4(m+1)$ en esta capa. Puede verse que si $f : [0,1]^n \longrightarrow \mathbb{R}$ y si hay m^n "cajas" en la partición, entonces se necesitan a lo más $m^n + 2n(m+1)$ unidades en las dos capas internas.

Bibliografía

1. R. B. Ash, *Real Analysis and Probability*, Academic Press, New York, 1972.
2. P. Billingsley, *Probability and Measure*, Wiley, New York, 1979.

3. E. K. Blum y L. K. Li, *Approximation Theory and Feedforward Networks*, Neural Networks **4** (1991), 511-515.
4. P. Cardialaguet, G. Euvarard, *Approximation of a Function and its Derivative with a Neural Network*, Neural Networks **5** (1992), 207-220.
5. N. E. Cotter, *The Stone Wierstrass Theorem and its Application to Neural Networks*, IEEE Transactions on Neural Networks **1** (1990), 290-295.
6. Neil E. Cotter and Thierry J. Guillerm, *The CMAC and a Theorem of Kolmogorov*, Neural Networks **5** (1992), 221-228.
7. G. Cybenko, *Approximation by Superposition of a Sigmoidal Function*, Math. Contr., Signals Syst. **2** (1989), 303-314.
8. J. Dugundji, *Topology*, Allyn and Bacon, Boston, 1966.
9. K.I. Funahashi, *On the Approximate Realization of Continuous Mappings by Neural Networks*, Neural Networks **2** (1989), 183-192.
10. A.R. Gallant and H. White, *There exists a Neural Network that does not make Avoidable Mistakes*, IEEE Second International Joint Conference on Neural Networks **I** (1987), 593-608.
11. A.R. Gallant and H. White, *On Learning the Derivatives of an Unknown Mapping with Multilayer Feedforward Networks*, Neural Networks **5** (1992), 129-138.
12. M. Garzón, *Analysis of Cellular Automata and Neural Networks*, Department of Mathematical Sciences, Memphis State University, 1990.
13. B. Geva and J. Sitte, *A Constructive Method for Multivariate Functions Approximation by Multilayer Perceptrons*, IEEE Transactions on Neural Networks **3** (1992), 621-624.
14. R. Goldberg, *Methods of Real Analysis*, Xerox College Publishing, Masachussets, 1964.
15. P.R. Halmos, *Measure Theory*, Springer-Verlag, Berlin, 1974.
16. R. Hecht-Nielsen, *Kolmogorov's Mapping Neural Networks Existence Theorem*, IEEE First International Joint Conference on Neural Networks **I** (1987), 593-608.
17. R. Hecht-Nielsen, *Neurocomputing*, Addison Wesley, 1990.
18. J. Hertz, A. Krogh and R. Palmer, *Introduction to the Theory of Neural Computation*, Addison Wesley, 1991.
19. K. Hornik, M. Stinchcombe and H. White, *Multilayer Feedforward Networks are Universal Approximators*, Neural Networks **2** (1989), 359-366.
20. K.Hornik, M.Stinchcombe and H.White, *Universal Approximation of an Unknown Mapping and its Derivatives Using Multilayer Feedforward Networks*, Neural Networks **3** (1990), 551-560.
21. K. Hornik, *Approximation Capabilities of Multilayer Feedforward Networks*, Neural Networks **4** (1991), 251-257.
22. K. Hornik, *Some New Results on Neural Network Approximation*, Neural Networks **6** (1993), 1069-1072.
23. Y. Ito, *Representation of Functions by Superposition of a Step or Sigmoid Function and their Applications to Neural Network Theory*, Neural Networks **4** (1991), 385-394.
24. Y. Ito, *Approximation of Functions on a Compact Set by Finite Sums of a Sigmoid Functions without Scaling*, Neural Networks **4** (1991), 817-826.
25. H. Katsuura and D. Sprecher, *Computational Aspects of Kolmogorov Superposition Theorem*, Neural Networks **7** (1994), 451-461.

26. P. Koiran, *On the Complexity of Approximating Mappings using Feedforward Networks*, Neural Networks **6** (1993), 649-653.
27. V. Kürková, *Kolmogorov's Theorem and Multilayer Neural Networks*, Neural Networks **5** (1992), 501-506.
28. M. Leshno et al, *Multilayer Feedforward Networks with Nonpolinomial Activation Function Can Approximate Any Function*, Neural Networks **6** (1993), 861-867.
29. R.P. Lippman, *An Introduction to Computing with Neural Nets*, IEEE ASSP Magazine **4** (1987), 4-22.
30. B. Müller and J.Reinhardt, *Neural Networks An Introduction*, Springer-Verlag, Berlin, 1990.
31. T.Poggio and F. Girosi, *Networks for Approximation and Learning*, Proceedings of the IEEE **78** (1990), 1481-1496.
32. C.S. Rees, S.M. Shah and C.V. Stanojvic, *Theory and applications of Fourier Analysis*, Marcel Dekker, INC, New York, 1981.
33. H.L. Royden, *Real Analysis*, The Macmillan Company Collier Macmillan Limited, London, 1970.
34. W. Rudin, *Principles of Mathematical Analysis*, McGraw-Hill, New York, 1976.
35. W. Rudin, *Functional Analysis*, McGraw-Hill, New York, 1991.
36. D.A. Sprecher, *On the Structure of Continuous Functions of Several Variables*, Transactions of the American Mathematical Society **115** (1965), 340-355.
37. D. Sprecher, *A Universal Mapping for Kolmogorov Superposition Theorem*, Neural Networks **6** (1993), 1089-1094.
38. L. G. Torres , G. Hernández y L. F. Niño, *Redes Neuronales*, X Coloq. Distrital de Matemáticas y Estadística, Santafé de Bogotá, Diciembre 1993.