UAB
Universitat Autònoma
de Barcelona

Treball Final de
Grau en Matemàtiques

Machine learning: mathematical foundations

Alícia Chimeno Sarabia

Supervisor
Roberto Rubio

Any
2022/23

Convocatòria
Juny

Abstract

In today's world, many people employ machine learning models, yet only a few understand the underlying mathematics that support them. How can we find a predictive function from a given dataset and ascertain the existence of such a function? This research seeks to address these concerns by exploring the mathematical foundations of function approximation in machine learning. Especially focus on function approximation using neural networks. Our research presents a significant finding, demonstrating that a multilayer feedforward network equipped with a non-polynomial activation function can effectively approximate any continuous function. Through this study, we aim to bridge the gap between the practical application of machine learning and the mathematical principles that underpin its success.

Resum

blslalblallb en català

Preface

Inpirat per **?**.

blablalbla amb glosary Universitat Autònoma de Barcelona (UAB) i ara curt UAB.

Contents

Chapter 1

Introduction

> Computers are like a bicycle for our minds.
>
> — Steve Jobs, *Michael Lawrence Films*

Our brain is constantly classifying and recognizing. For instance, when we spot a dog on the street, one easy classification we can make is {dog, not dog}, which is probably too easy for our brain—it's almost instantaneous. However, things get a bit more complex when we read the teacher's whiteboard. What happens when we encounter a symbol that confuses us because it resembles another? We can interpret the mathematics behind this reasoning ºas the brain seeking/creating a function that provides us with the certainty of recognizing that particular letter. Eventually, we reach a point where we feel confident enough to write it down in our notes.

Artificial intelligence aims to replicate the remarkable capabilities of our brains. It seeks to develop computational models and algorithms that can perform tasks such as classification, recognition, and decision-making with a level of accuracy and efficiency comparable to human intelligence. When AI first emerged, one of the initial challenges was hand-written digit recognition, exemplified by the MNIST digits dataset. This dataset comprises 60,000 examples of handwritten digits from 0 to 9. To enable a machine learning model to recognize these digits, it must effectively map each image to its corresponding number. This problem naturally aligns with a mathematician's perspective of function learning, where the goal is to approximate a function based on a given dataset consisting of points in space.

Neural Networks are a key approach used in artificial intelligence to tackle such problems. The theory of function approximation through neural networks has a long history dating back to the work by McCulloch and Pitts

This Bachelor's thesis aims to dig into the mathematical foundations of machine learning, Our main ... is to demonstrate that the "real-world" functions we seek to approximate can be effectively approximated by a specific type of functions.

Chapter 2

Machine Learning

2.1 Machine Learning Basics

Machine Learning focuses on the development of algorithms and models that enable computers to learn from data with the aim of making predictions without being explicitly programmed.

We can think about learning as the way we understand it as a human. We can classify a learning problem based on the degree of feedback. Machine learning models fall into three primary categories:

- Supervised learning, where we have immidiate feedback.
- Reinforcement learning, where we have indirect feedback. For example whenwe are playing the game of chess.
- Unsupervised learning, where we have non feedback signal. For example,deducing which dog belongs to each owner.

Machine learning models simplify reality for the purposes of understanding or prediction. This prediction can be either a numerical prediction or a classifications prediction. A number of machine learning algorithms are commonly used.

2.1.1 Motivation

Consider the problem of assessing the eligibility of a consumer for a credit. We are provided with the following set of data:

Costumer application:	
Age	23 years
Gender	Male
Annual Salary	$30,000
Years in Residence	1 year
Years in Job	1 year
Current Debt	$15,000
...	...

- Input: $x_c = (x_{c_1}, ..., x_{c_d})$ "attributes of the costumer that we want to classify".
- Output:
$$y = \begin{cases} approve \\ deny \end{cases}$$
- Target funtion: f "ideal credit approval formula"
- Data: The set $\{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$ corresponds of a historical records of credit customers where x_i is the attributes of the costumer and y_i classification awarded.

We are looking for the function f such that $f(x_c) = y$.

A fundamental problem of machine learning is the following. Given data of the form $\{(x_i, y_i)\}_{i=1}^m \subset \mathbb{R}^n \times \mathbb{R}$, drawn randomly from a probability distribution μ, find a model P such that $P(x_i) = y_i$. An important aspect of machine learning is that *many supervised learning tasks are about function learning.*.

Example 1. An example of a supervised learning task is digit recognition. The objective is to identify handwritten digits (0-9) based on input images. In this task, we aim to learn a probability distribution function denoted as f, which maps a set of pixel values ranging from 0 (black) to 255 (white), representing a 28x28 image, to a probability distribution over the digits 0 to 9.

$$f : \{0, ..., 255\}^{28\times 28} \longrightarrow \text{probability distribution on } \{0, 1, ..., 9\}$$

Example 2. Example of a classification problem. We want to classify if an image is a dog or not a dog. We would like to produce a value which is correlated with the probability of this image being a dog or not a dog. We can approach the problem in the following way. We want to find a function that takes very high values when dog-image and very low val ues when non dog images and takes the value 0 when its uncertain.

$$d : \mathbb{R}^{\#\text{pixels in image}} \to \mathbb{R}$$

such that $\mathbb{P}(d(\text{image})) =$ probability that the image is a dog.

That is what we mean by many problems can be recast as function learning. Note that there is not a god-given reason why this function should exist. We know that certain points in space, and they have certain values associated to them, but we dont know that there is some big function.

Important principle II: Sometimes function learning can be recast as a classification problem.

Binary classification problem. Rather learning $\mu : \mathbb{R}^{\#\text{bits}} \to \mathbb{R}$ where big values correspond to likely and small values to unlikely. It is better to learn $\mu : \mathbb{R}^{\#\text{bits}} \to$ probability distribution on $\{-1, 0, 1\}$. In number theory the function $\mu(n)$ it is called Möebius function

$$\mu(n) = \begin{cases} 0 & \text{if } n \text{ has a repeated square factor,} \\ -1 & \text{if } n \text{ has an odd number of distinct prime factors,} \\ 1 & \text{if } n \text{ has an even number of distinct prime factors.} \end{cases}$$

2.1.2 Linear Regression

A linear regression algorithm is used to predict numerical values, based on a linear relationship between different values. A simple linear model has an outcome, denoted by y, also known as a response variable, and a predictor, x. It is defined by the following equation.

$$y_i = \beta_0 + \beta_1 x_i + \epsilon$$

where $i = 1, ..., n$ indexes the observations from 1 to n in the dataset.

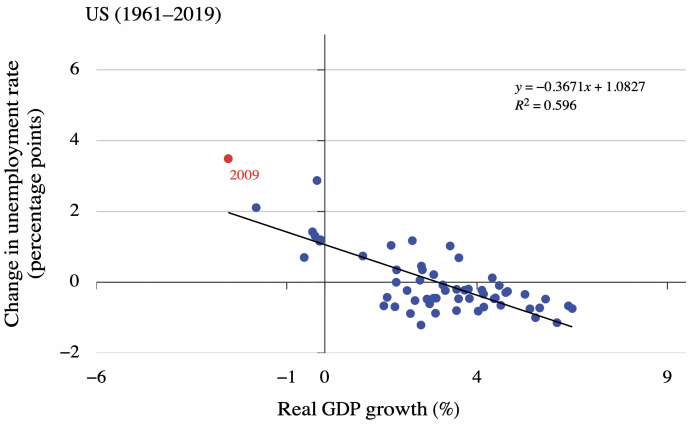


Figure 2.1: y - response variable: unemployment rate , x predictor: GDP growth ,

We can add additional p predictors to a simple linear model, transforming it into a multivariate linear model, which we define as follows:

$$y_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{pi} + \epsilon_i$$

2.1.3 Logistic Regression

Logistic regression is a model for predicting the probability that a binary response is 1. It is suitable for classification tasks, as well as for prediction of probabilities. From a statistical perspective, it is defined by assuming that the distribution of the binary response variable, y, given the features, x, follows a Bernoulli distribution with success probability p.

$$P(y = 1|X = x) = p$$

We need to define the concept of sigmoid function that will be important along the work. A *sigmoid function* is a mathematical function that maps input values to a range between 0 and 1. We have the following sigmoid function, the logit inverse function:

$$\text{logit} : (0,1) \to \mathbb{R} \quad \text{and is expressed as:} \quad \text{logit}(x) = \log\left(\frac{x}{x-1}\right)$$

$$\text{logit}^{-1} : \mathbb{R} \to (0,1) \quad \text{and is expressed as:} \quad \text{logit}^{-1}(x) = \frac{e^x}{1+e^x}$$

The linear predictor, $w^T x + \theta$, fluctuates between $(-\infty, \infty)$ where x represents all predictors in the model. To address this difference in scale, the outcome variable is transformed using the logit function. The transformed result, $\text{logit}(x)$, is expressed in logarithms of probabilities. The logistic regression model assumes a linear (affine) relationship between the feature vector x_i and the log odds of p. Namely,

$$\text{logit}(p) = w^T x + \theta$$

The logistic model can be alternatively expressed using the inverse logit function:

$$P(y = 1|X = x) = \text{logit}^{-1}(w^T x + \theta)$$

2.2 Multilayer Feedforward Networks

Artificial Neural Networks (ANN) are the quintessential deep learning models, especially *multilayer feedforward networks*. They are widely used for nonlinear function approximation. The goal of a artificial neural network is to approximate some function f^*. For example, for a classifier, $y = f^*(x)$ maps an input x to a category y.

The term *neural* refers to the fact that this model was originally inspired by how biological neurons process information. These artificial neurons mimic the processing of information in biological neurons.

The term *feedforward* indicates the direction of information flow within the network, moving only forward in contraposition to backwards. Each layer processes the input data and passes its output to the next layer, creating a sequence of transformations until the final output is produced. $f(f_1(f_2(f_3)))$

The term *network* refers to the interconnected structure of artificial neurons. An multilayer network consists of multiple layers, including an input layer, one or more hidden layers, and an output layer.

The architecture of the network entails determining its *depth*, *width*, and *activation functions* used. Depth is the number of hidden layers. Width is the number of units (nodes) on each hidden layer.

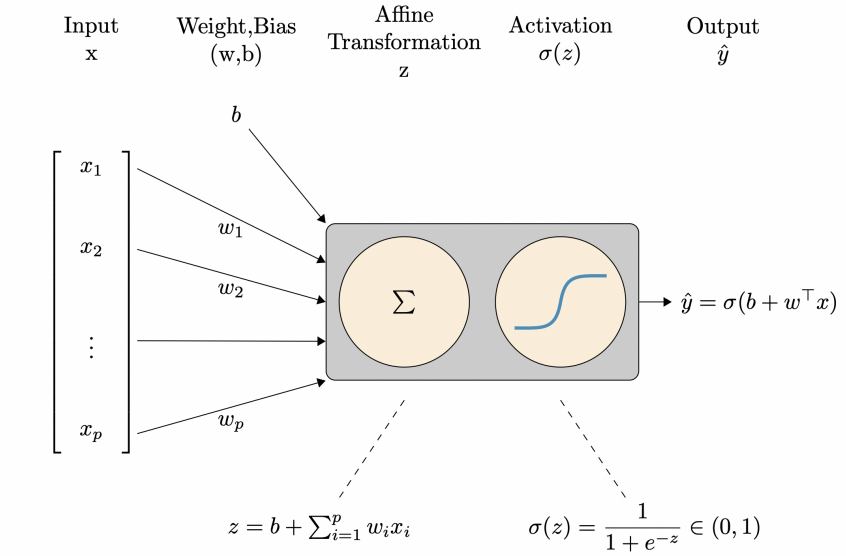
2.3 Architecture of a Multilayer Feedforward Network

2.3.1 Artificial neuron

The equation

$$y = \sigma(w^T x + \theta) \tag{2}$$

represents what we may call a *single layer of a deep learning model*, also called an *artificial neuron*. Observe that the artificial neuron is composed of an affine transformation $z = w^T x + \theta$ followed by a (generally) non-linear transformation $\sigma(z)$. In more detail, $x \in \mathbb{R}^n$ is the input vector and represents a set of n features or predictors, $w \in \mathbb{R}^n$ is the weights vector. Each element of the weights vector w_i corresponds to the weight or importance assigned to the corresponding input feature x_i. θ is the bias and σ is the activation function. The result variable is an scalar output $y \in \mathbb{R}$.



2.3.2 Activation Function

The introduction of the activation function in ANN was inspired by biological neural networks whose purpose is to decide whether a particular neuron fires or not. The

simple addition of such a nonlinear function can tremendously help the network to exploit more, thereby learning faster. There are various activation functions proposed in the literature, and it is difficult to find the optimal activation function that can tackle any problem.

Note that a logistic regression is an artificial neuron where the activation function σ is logit^{-1}. A few very popular activation function are the Hyperbolic Tangent and the ReLu:

$$\tanh : \mathbb{R} \to (-1, 1)$$

and

$$\text{ReLU} : \mathbb{R} \to (0, \infty) \quad \text{and is expressed as:} \quad \text{ReLU}(x) = \max(0, x)$$

We now get into more details on the precise definition of a deep neural network, which is after all a purely mathematical object.

Definition 1. A *multilayer feedforward network* is the function

$$f(x) = \sum_{j=1}^{k} \beta_j \cdot \sigma(w_j \cdot x - \theta_j)$$

where $x \in \mathbb{R}^n$ is the input vector, $k \in \mathbb{N}$ is the number of processing units in the hidden layer, $w_j \in \mathbb{R}^n$ is the weight vector that connects the input to processing unit j in the hidden layer, $\sigma : \mathbb{R} \to \mathbb{R}$ is an activation function, $\theta_j \in \mathbb{R}$ is the threshold (or bias) associated with processing unit j in the hidden layer, and $\beta_j \in \mathbb{R}$ is the weight that connects processing unit j in the hidden layer to the output of the network.

Let N_w be the family of all functions implied by the network's architecture. If we can show that N_w is dense in $C(\mathbb{R}^n)$, we can conclude that for every continuous function $g \in C(\mathbb{R}^n)$ and each compact set $K \subset \mathbb{R}^n$, there is a function $f \in N_w$ such that f is a good approximation to g on K.

Under which necessary and sufficient conditions on σ will the family of networks N_w be capable of approximating to any desired accuracy any given continuous function?

Chapter 3

Function Approximation

Creating a machine learning model to predict/classify from a given data is a similar process than when we calculate a function from a given points in the space. This is called function approximation and among the most famous techniques of function approximation, we find interpolation: such as Taylor polynomial, Chebyshev polynomial, the method of least squares, or spline approximation.

In this chapter we present some mathematical definitions and results of function approximation. If we want to approximate functions, we need to define the following notions: metric spaces, distance between functions, density,

3.1 Definitions and some results

Definition 2. We dentoe by $\mathcal{C}(\mathbb{R}^n)$ the set of continuous functions defined on $\mathbb{R}^n$.

Definition 3. The support of a function u is denoted by

$$supp(u) = \overline{\{x|u(x) \neq 0\}}$$

Definition 4. We denote by $\mathcal{C}_0^\infty$ the set of infinitely differentiable functions $\mathcal{C}^\infty$, also called smooth functions, with compact support.

3.1.1 Metric spaces

Definition 5. A *metric* (or *distance*) on a set X is a function $d : X \times X \to \mathbb{R}$ such that for all $s, t, u \in X$ the following properties are satisfied:

1. $d(s,t) \geq 0$ and $d(s,t) = 0$ if and only if $s = t$.
2. $d(s,t) = d(t,s)$.
3. $d(s,t) \leq d(s,u) + d(u,t)$ (*triangle inequality*).

Definition 6. A *metric space* is a pair (X, d), where X is a set and d is a distance in X.

If we take X to be a set of functions, the metric $d(f, g)$ will enable us to measure the distance between functions $f, g \in X$.

Proposition 7. Let (X, d) be a metric space and $\mathcal{B}$ be a basis of open sets for X. Then, for all open set U in X, we have:

$$U = \bigcup \{B_x : x \in U, B_x \subseteq U, B_x \in \mathcal{B}\}$$

This proposition states that any open set U in a metric space X can be expressed as the union of basis elements B_x.

Definition 8. Let (X, d) be a metric space. The closure of a set A is defined as follows:

$$closure(A) = \overline{A} = \{t \,|\, \forall \epsilon > 0, \exists a \in A,\, d(a, t) < \epsilon\}.$$

Proposition 9. Let (X, d) be a metric space and $A \subseteq X$. Then, $a \in \overline{A}$ if and only if there exists a sequence (a_n) in A such that for every $\epsilon > 0$, there exists $N \in \mathbb{N}$ such that $d(a_n, a) < \epsilon$ for all $n \geq N$.

Proposition 10. Let ρ be a metric defined on the set $\mathcal{C}_0^\infty[a, b]$ as follows:

$$\rho(\varphi_1, \varphi_2) = \sum_{n=0}^{\infty} 2^{-n} \frac{\|\varphi_1 - \varphi_2\|_n}{1 + \|\varphi_1 - \varphi_2\|_n}$$

where

$$\|\varphi\|_n = \sum_{j=0}^{n} \sup_{x \in [a,b]} |\varphi^{(j)}(x)|.$$

Then the metric space $(\mathcal{C}_0^\infty[a, b], \rho)$ is complete, also known as a Fréchet space.

Proof in 196 [1967]

3.1.2 Baire's Theorem

Definition 11. Let (X, d) be a metric space and $A \subseteq X$ subset. A is said to be *nowhere dense* if for every (nonempty) open subset $U \subseteq X$, the intersection $U \cap \overline{A}$ is not dense in U, meaning that U contains a point that is not in the closure of A.

Definition 12. A set it is said to be category I if it can be written as a countable union of nowhere-dense sets. Otherwise it is said to be of *category II*

Theorem 13. *(Baire's Category Theorem) Any complete metric space is of category II.*

3.1.3 Lebesgue space

Definition 14. Let Σ be a σ-algebra over a set Ω. A *measure* over Ω is any function

$$\mu : \Sigma \longrightarrow [0, \infty]$$

satisfying the following properties:

1. $\mu(\varnothing) = 0$.
2. *σ-additivity*: If $(A_n) \in \Sigma$ are pairwise disjoint, then:

$$\mu\left(\bigsqcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \mu(A_n)$$

Definition 15. A box in $\mathbb{R}^d$ is a set of the form

$$Q = [a_1, b_1] \times ... \times [a_d, b_d] = \prod_{i=1}^{d} [a_i, b_i]$$

and the volume of the box is

$$vol(Q) = (b_1, a_1)...(b_d - a_d) = \prod_{i=1}^{d} (b_i - a_i).$$

Definition 16. The *exterior measure* (or outer measure) of a set $E \subseteq \mathbb{R}^d$ is

$$|E|^* = \inf\{\sum_k vol(Q_k)\}$$

where the infimum is taken over all finite or countable collection of boxes $\{Q_k\}$ such that $E \subseteq \cup_k Q_k$

Definition 17. A set $E \subseteq \mathbb{R}^n$ is *Lebesgue mesurable* (or mesurable) if $\forall \epsilon > 0$, there exist U open set such that $E \subseteq U$ and $|U \setminus E|^* < \epsilon$

Definition 18. A set $N \subset \mathbb{R}^n$ is called a *null set* if $|N|^* = 0$

Definition 19. We say that a property holds almost everywhere (a.e.) if the set of points that doesn't hold it is null.

Definition 20. A function u that is measurable on $\Omega \in \mathbb{R}^n$ is said to be *essentially bounded* on Ω if there is a constant λ such that $|u(x)| \leq \lambda$ a.e on Ω. The greatest lower bound of such constants λ is called the essential supremum of $|u|$ on Ω and is denoted by $\operatorname{ess\,sup}_{x\in\Omega} |u(x)|$. We denote by $L^\infty(\Omega)$ the space of all functions u that are essentially bounded on Ω. We denote the norm on $L^\infty(\Omega)$ by $\|\|_{L^\infty(\Omega)}$ defined by

$$\|u\|_{L^\infty(\Omega)} = \operatorname*{ess\,sup}_{x\in\Omega} |u(x)|.$$

Definition 21. A function u defined almost everywhere on a domain Ω (a domain is an open set in $\mathbb{R}^n$) is said to be *locally essentially bounded* on Ω if for every compact set $K \subset \Omega$, $u \in L^\infty(K)$. We denote $u \in L^\infty_{loc}(K)$.

Definition 22. Let $\mathcal{M}$ denote the set of functions which are in $L^\infty_{loc}(\mathbb{R})$ and have the following property. The closure of the set of points of discontinuity of any function in $\mathcal{M}$ is of zero Lebesgue measure.

This implies that for any $\sigma \in \mathcal{M}$, interval $[a, b]$. and $\delta > 0$, there exists a finite number of open intervals, the union of which we denote by U, of measure δ, such that σ is uniformly continuous on $[a, b]/U$.

Definition 23. We say that a set of functions $F \subset L^{\infty}_{loc}(\mathbb{R})$ is *dense* in $C(\mathbb{R}^n)$ if for every function $g \in C(\mathbb{R}^n)$ and for every compact $K \subset \mathbb{R}^n$, there exist a sequence of functions $f_j \in F$ such that

$$\lim_{j\to\infty} \|g - f_j\|_{L^{\infty}(K)} = 0.$$

Definition 24. $\varphi : I \to \mathbb{R}$ is uniformly continuous on I if $\forall \epsilon > 0 \exists\ \delta > 0$ such that $|\varphi(x) - \varphi(y)| < \epsilon$ whenever $|x - y| < \delta$

3.1.4 Convolution

Definition 25. Let f, g be real-valued functions with compact support. We define the *convolution* of f with g as

$$(f * g)(x) = \int f(x - t)g(t)\, dt$$

Proposition 26. If f is a smooth function that is compactly supported and g is a distribution, then $f * g$ is a smooth function defined by

$$\int_{\mathbb{R}^d} f(y)g(x - y)\, dy = (f * g)(x) \in C^{\infty}(\mathbb{R}^d).$$

See Adams et al. [2003]

Proposition 27.

$$\frac{\partial}{\partial x_i}(f * g) = \frac{\partial f}{\partial x_i} * g = f * \frac{\partial g}{\partial x_i}.$$

3.1.5 Annihilator

Definition 28. (Annihilator of a Subset of a Vector Space). Let V be a vector space over a field F, and $X \subseteq V$ subset. The *annihilator of X* is defined as the set of all linear functionals $V \to F$ that evaluate to zero on every element of X:

$$X^0 = \{\varphi \in V^* : \quad \text{for all}\, x \in X \quad \varphi(x) = 0\}$$

Definition 29. (Annihilator of a subset of V^*). Let V be a finite-dimensional vector space over a field F, and let $F \subset V^*$. The annihilator of F is defined as the set of all vectors in V that are annihilated by every functional in F:

$$F^0 = \{x \in V : \quad \text{for all}\, \varphi \in F \quad \varphi(x) = 0\}$$

Chapter 4

Theorem and proof

The work revolves around the following theorem and its proof.

4.1 Theorem

Theorem 30. *Let $\sigma \in \mathcal{M}$. Set*

$$\Sigma_n = span\{\sigma(w \cdot x + \theta) : w \in \mathbb{R}^n, \theta \in \mathbb{R}\}.$$

Then Σ_n is dense in $\mathcal{C}(\mathbb{R}^n)$ if and only if σ is not a polynomial.

4.1.1 Previous results

There has been significant research on the approximation capabilities of feedforward networks prior to the proof of this theorem. Previous studies have demonstrated that if the activation functions of the network satisfy certain explicit assumptions, then the network can be proven to be a universal approximator. For instance, Hornik [1991] have established two results, which are as follows:

Theorem 31. *(Hornik Theorem 1). Multilayer feedforward networks with a bounded and nonconstant activation function can approximate any function in $L^p(\mu)$ arbitrary well, given a sufficiently large number of hidden units.*

Theorem 32. *(Hornik Theorem 2) Multilayer feedforward networks with a continuous, bounded and nonconstant activation function can approximate any continuous function on X arbitrarily well (with respect to the uniform distance) given a sufficiently large number of hidden units.*

The theorem generalizes Hornik's Theorem 2 by establishing necessary and sufficient conditions for universal approximation. Note that the theorem merely requires "nonpolynomiality" in the activation function. Unlike Hornik's result, the activation functions do not need to be continuous or smooth. This has an important biological interpretation because the activation functions of real neurons may well be discontinuous or even non-elementary.

4.2 Proof

4.2.1 If σ is not a polynomial then Σ_n is dense in $\mathcal{C}(\mathbb{R}^n)$

Consider that σ is not an algebraic polynomial and we aim to show that Σ_n is dense in $\mathcal{C}(\mathbb{R}^n)$. In order to show that, we need the following Lemmas.

Lemma 1. If we have that $\sigma * \varphi$ is a polynomial for all $\varphi \in \mathcal{C}_0^\infty$, then the degree of the polynomial $\sigma * \varphi$ is finite, i.e. there exists an $m \in \mathbb{N}$ such that $deg(\sigma * \varphi) \leq m$ for all $\varphi \in \mathcal{C}_0^\infty$.

Proof. We first prove the claim in the case of $\varphi \in \mathcal{C}_0^\infty[a, b]$, for some $a < b$.

By Proposition 10 we have that $(\mathcal{C}_0^\infty[a, b], \rho)$ is a complete metric space.

Consider the following set,

$$V_k = \{\varphi \in \mathcal{C}_0^\infty[a, b] \,|\, deg(\sigma * \varphi) \leq k\}.$$

It is clear that this set $V_k \subseteq \mathcal{C}_0^\infty[a, b]$. We want to show that $\mathcal{C}_0^\infty[a, b] = V_k$.

The set V_k fulfills the following properties: $V_k \subset V_{k+1}$, V_k is a closed subspace, $\cup_{k=0}^\infty V_k = \mathcal{C}_0^\infty[a, b]$ and V_k is a vector space. We can easily see that $\mathcal{C}_0^\infty[a, b]$ is also a vector space.

As $\mathcal{C}_0^\infty[a, b]$ is a complete metric space, by Baire's Category Theorem 13, this set is of category II, i.e. $\mathcal{C}_0^\infty[a, b]$ cannot be written as a countable union of nowhere-dense sets. Recall that $\mathcal{C}_0^\infty[a, b]$ can be written as a countable union of V_k, therefore some V_m is not a nowhere-dense set, that is, there exists an open set $U \subseteq C_0^\infty[a, b]$ that is contained in the closure of V_m, but, as V_m is closed, for that we have that $U \subseteq V_m$. For topology results, any open set of a vector space contains a basis of the vector space, in our case U contains a baisis of $\mathcal{C}_0^\infty[a, b]$, and $U \subseteq V_m$, therefore V_m contains a baisis of $\mathcal{C}_0^\infty[a, b]$. Now we can conclude that $\mathcal{C}_0^\infty[a, b] \subseteq V_k$. And consequently $\mathcal{C}_0^\infty[a, b] = V_k$. This means that any $\varphi \in \mathcal{C}_0^\infty[a, b]$ also satisfies $\varphi \in V_k$, that means that the convolution $\sigma * \varphi$ has degree finite.

For the general case where $\varphi \in \mathcal{C}_0^\infty$, we note that the number m does not depend on the interval $[a, b]$. This can be seen as follows. By translation m depends at most of the length of the interval. Let $[A, B]$ be any interval. For $\varphi \in \mathcal{C}_0^\infty[A, B]$ we can find $\varphi_i \in \mathcal{C}_0^\infty[a_i, b_i]$ for $i = 1, ..., k$ such that $[A, B] \subset \cup_{i=1}^k [a_i, b_i]$ where $b_i - a_i = b - a$ and $\varphi = \sum_{i=1}^k \varphi_i$ Thus

$$\sigma * \varphi = \sum_{i=1}^{k} \sigma * \varphi_i$$

and for every $i = 1, ..., k$ we have that $\sigma * \varphi$ is a polynomial of degree less than or equal to m. Therefore $deg(\sigma * \varphi) \leq m$. □

Lemma 2. If $\sigma * \varphi$ is a polynomial such that $deg(\sigma * \varphi) \leq m$ for all $\varphi \in \mathcal{C}_0^\infty$, then σ is a polynomial of degree at most m.

Proof. If $\sigma * \varphi$ is a polynomial of degree m. For all $\varphi \in \mathcal{C}_0^\infty$, using (27) we have that

$$(\sigma * \varphi)^{(m+1)}(x) = \int \sigma(x-y)\varphi^{(m+1)}(y)\,dy = 0$$

From standard results in Distribution Theory [pp 57 Friedman [1963]], σ is itself a polynomial of degree at most m (a.e.). □

Proposition 1. If σ is not a polinomial then $\sigma * \varphi$ is not a polynomial for some $\varphi \in \mathcal{C}_0^\infty$.

Proof. We will show the contrapositive. Suppose that the convolution $\sigma * \varphi$ is a polynomial for all $\varphi \in \mathcal{C}_0^\infty$, by Lemma 1 the degree of the convolution is finite. Now we have that $\sigma * \varphi$ is a polynomial of finite degree, by Lemma 2 we have that σ is a polynomial. □

Lemma 3. For each $\varphi \in \mathcal{C}_0^\infty$, $\sigma * \varphi \in \overline{\Sigma_1}$.

Proof. We recall the definition of the set

$$\Sigma_1 = \text{span}\{\sigma(w \cdot x + \theta) : w \in \mathbb{R}, \theta \in \mathbb{R}\}. \tag{1}$$

Without loss of generality, assume that supp $\varphi \subseteq [-\alpha, \alpha]$. To show that $\sigma * \varphi \in \overline{\Sigma_1}$ we will use the characterization for elements in the closure. We will to prove that there exists a sequence in Σ_1 such that uniformly converges to $\sigma * \varphi$ on $[-\alpha, \alpha]$. Note that we usually denote for un$\rightrightarrows$)

We shall consider the following sequence:

$$h_m = \sum_{i=1}^{m} \varphi(y_i)\Delta y_i \sigma(x - y_i).$$

Which satisfies $h_j \in \Sigma_1$ for $j = 1, ..., m$. Note that we have $w_i = 1, \theta_i = -y_i$ and $\beta_i = \varphi(y_i)\Delta y_i$.

We will define a partition of the interval $[-\alpha, \alpha]$ to be the following, where

$$y_i = -\alpha + \frac{2i\alpha}{m} \quad i = 1, ..., m$$

and $\Delta y_i = \frac{2\alpha}{m}$.

Given $\epsilon > 0$, we choose $\delta > 0$ such that

$$10\delta \|\sigma\|_{L^\infty\{-2\alpha, 2\alpha\}} \|\varphi\|_{L^\infty} \leq \frac{\epsilon}{3}.$$

We know that $\sigma \in M$. Hence, for the previous choosen $\delta > 0$ and $[-\alpha, \alpha]$ interval, there exists $r(\delta)$ finite number of intervals the measure of whose union U is δ such that σ is uniformly continuous on $[-2\alpha, 2\alpha]/U$. We now choose m_i sufficiently large so that

1. $m_1\delta > \alpha r(\delta)$. We can do this by Archimedes' principle.
2. From the uniform continuity of φ we know that
   if $|s-t| \leq \delta_2 = \frac{2\alpha}{m_2}$ then
   $$|\varphi(s) - \varphi(t)| \leq \epsilon_2 = \frac{\epsilon}{2\alpha\|\sigma\|_{L^\infty[-2\alpha,2\alpha]}}$$
3. From the previous, σ is uniformly continuous on $[-2\alpha, 2\alpha]/U$. We chose m such that, if $s,t \in [-2\alpha,2\alpha]/U$ and $|s-t| \leq \delta_3 = \frac{2\alpha}{m_3}$ then
   $$|\sigma(s) - \sigma(t)| \leq \epsilon_3 = \frac{\epsilon}{\|\varphi\|_L}$$

We choose m such that $m = max\{m_1, m_2, m_3\}$.

Now, fix $x \in [-\alpha, \alpha]$. Set $\Delta_i = [y_{i-1}, y_i]$ where $y_0 = \alpha$.

First, recall that,

$$\int \sigma(x-y)\varphi(y)dy = \sum_{i=1}^{m} \int_{\Delta_i} \sigma(x-y)\varphi(y)dy.$$

Consider the following difference

$$\begin{aligned}
&\left| \int \sigma(x-y)\varphi(y)dy - \sum_{i=1}^{m} \int_{\Delta_i} \sigma(x-y_i)\varphi(y)dy \right| \\
&\qquad = \left| \sum_{i=1}^{m} \int_{\Delta_i} \sigma(x-y)\varphi(y)dy - \sum_{i=1}^{m} \int_{\Delta_i} \sigma(x-y_i)\varphi(y)dy \right| \\
&\qquad = \left| \sum_{i=1}^{m} \int_{\Delta_i} \varphi(y)\Big(\sigma(x-y) - \sigma(x-y_i)\Big)dy \right| \\
&\qquad \leq \sum_{i=1}^{m} \int_{\Delta_i} |\varphi(y)|\,|\sigma(x-y) - \sigma(x-y_i)|\,dy.
\end{aligned}$$

If $x - \Delta_i \cap U = \varnothing$. Since $x - y \notin U$, $x - y_i \notin U$ and $x - y_i \in [-2\alpha, 2\alpha]$. For choice of m in property 2, we have

$$\begin{aligned}
\sum_{i=1}^{m} \int_{\Delta_i} |\varphi(y)|\,|\sigma(x-y) - \sigma(x-y_i)|\,dy &\leq \frac{\epsilon}{\|\varphi\|_{L_1}} \sum_{i=1}^{m} \int_{\Delta_i} |\varphi(y)| \\
&= \frac{\epsilon}{3\|\varphi\|_{L_1}} \int |\varphi(y)|\,dy \\
&= \frac{\epsilon}{3\|\varphi\|_{L_1}} |\varphi(y)|_{L_1} \\
&= \frac{\epsilon}{3}.
\end{aligned}$$

If $x - \Delta_i \cap U \neq \varnothing$ then we will denote such intervals by $\widetilde{\Delta_i}$.

$\sum_i |\widetilde{\Delta_i}| = \sum_i |(x - \Delta_i \cap U)| \leq |U| + 2|\Delta_i| r(\delta) \leq \delta + 2 \cdot \frac{2\alpha}{m} r(\delta) \leq \delta + 4\delta = 5\delta$

We used the property $m\delta > \alpha r(\delta)$ indeed $\delta > \frac{\alpha \cdot r(\delta)}{m}$.

$$\begin{aligned}
&\sum_{i=1}^{m} \int_{\widetilde{\Delta_i}} |\varphi(y)| \, |\sigma(x-y) - \sigma(x-y_i)| \, dy \\
&\qquad \leq \sum_{i=1}^{m} \int_{\widetilde{\Delta_i}} \|\varphi\|_{L^\infty} \, 2\|\sigma\|_{L^\infty[-2\alpha,2\alpha]} \\
&\qquad = \|\varphi\|_{L^\infty} \, 2\|\sigma\|_{L^\infty[-2\alpha,2\alpha]} \sum_i |\widetilde{\Delta_i}| \\
&\qquad \leq \|\varphi\|_{L^\infty} \, 2\|\sigma\|_{L^\infty[-2\alpha,2\alpha]} \, 5\delta \leq \epsilon/3
\end{aligned}$$

$$\begin{aligned}
&\left| \sum_{i=1}^{m} \int_{\Delta_i} \sigma(x-y_i)\varphi(y)dy - \sum_{i=1}^{m} \sigma(x-y_i)\varphi(y_i)\Delta y_i \right| \\
&\qquad = \left| \sum_{i=1}^{m} \int_{\Delta_i} \sigma(x-y_i)[\varphi(y) - \varphi(y_i)]dy \right| \\
&\qquad \leq \sum_{i=1}^{m} \int_{\Delta_i} |\sigma(x-y_i)| \, |\varphi(y) - \varphi(y_i)| \, dy \\
&\qquad \leq \sum_{i=1}^{m} \int_{\Delta_i} |\sigma(x-y_i)| \, dy \left[\frac{\epsilon/3}{2\alpha\|\sigma\|_{L^\infty[-2\alpha,2\alpha]}} \right] \leq \frac{\epsilon}{3}
\end{aligned}$$

Finally, we have the result $h_m \rightrightarrows \sigma * \varphi$ because

$$\left| \int \sigma(x-y)\varphi(y)dy - \sum_{i=1}^{m} \sigma(x-y_i)\varphi(y_i)\Delta y_i \right| \leq \epsilon$$

for all $x \in [-\alpha, \alpha]$. □

Lemma 4. If $\sigma \in \mathcal{C}^\infty$, then Σ_1 is dense in $\mathcal{C}(\mathbb{R})$.

Proof. We suppose that $\sigma \in \mathcal{C}^\infty$ and recall by the theorem hypothesis σ is not a polynomial. We can write any function f of the set Σ_1 as

$$f = \sum_i \beta_i \sigma_i(w_i x + \theta_i) = \beta_1 \sigma_1(w_1 x + \theta_1) + ...$$

We can see that the function

$$\frac{\sigma([w+h]x + \theta) - \sigma(wx + \theta)}{h} \in \Sigma_1$$

because is a linear combination, where $\beta_1 = \frac{1}{h}, \beta_2 = \frac{-1}{h}$.

By hypothesis, $\sigma \in \mathcal{C}^\infty$. By definition of derivative we have

$$\frac{d}{dw}\sigma(wx+\theta) = \lim_{h\to 0} \frac{\sigma([w+h]x+\theta) - \sigma(wx+\theta)}{h} \in \overline{\Sigma_1}$$

because the limit of a set belongs to the closure of the set.

By the same argument,

$$\frac{d^k}{dw^k}\sigma(wx+\theta) \in \overline{\Sigma_1}$$

for all $k \in \mathbb{N}, w, \theta \in \mathbb{R}$.

If we differentiate this expression k times, we obtain

$$\frac{d^k}{dw^k}\sigma(wx+\theta) = \sigma^{(k)}(wx+\theta)\cdot x^k$$

We are going to see that if σ is not a polynomial (by hypothesis) then there exists a $\theta_k \in \mathbb{R}$ such that $\sigma^{(k)}(\theta_k) \neq 0$. To show that, let us assume that does not exist any $\theta_k \in \mathbb{R}$ such that $\sigma^{(k)}(\theta_k) \neq 0$. This means that the k-th derivative at every point is 0, i.e.

$$\sigma^{(k)}(\theta) = 0 \quad \forall \theta \in \mathbb{R}$$

If we integrate this expression, we will have $\int \sigma^{(k)} = \int 0$. This implies that

$$\sigma^{(k-1)}(x) = C_1$$

for some constant C_1, as integrating zero results in a constant. If we integrate again, we have:

$$\sigma^{(k-2)}(x) = C_1 x + C_2$$

for some constants C_1 and C_2.
Continuing this process, we arrive at

$$\sigma(x) = C_1 x^{k-1} + C_2 x^{k-2} + \ldots + C_{k-1}x + C_k$$

for constants $C_1, C_2, \ldots, C_k$. Hence, σ is a polynomial of degree $k-1$, which contradicts our assumption that σ is not a polynomial. Therefore, there always exists a point where the derivative does not vanish.

Thus, we evaluate at the point where the derivative does not vanish, we call it θ_k.

$$\sigma^{(k)}(\theta_k)\cdot x^k = \frac{d^k}{dw^k}\sigma(wx+\theta)\Big|_{w=0,\theta=\theta_k} \in \overline{\Sigma_1}$$

This implies that $\overline{\Sigma_1}$ contains all polynomials, because the expression $\sigma^{(k)}(\theta_k)x^k$ generates all polynomials. By the Weierstrass theorem, we know that the polynomials are dense in $\mathcal{C}(\mathbb{R})$. This concludes that the set $\overline{\Sigma_1}$ contains a set which is dense in $\mathcal{C}(\mathbb{R})$, therefore Σ_1 is dense in $\mathcal{C}(\mathbb{R})$.

□

Lemma 5. If for some $\varphi \in \mathcal{C}_0^\infty$ we have that $\sigma * \varphi$ is not a polynomial, then Σ_1 is dense in $\mathcal{C}(\mathbb{R})$.

Proof. From Lemma 3, $\sigma * \varphi \in \overline{\Sigma_1}$ for each $\varphi \in \mathcal{C}_0^\infty$. It thus follows that, $\sigma * \varphi(wx + \theta) \in \overline{\Sigma_1}$, for each $w, \theta \in \mathbb{R}$. Now for σ and $\varphi \in \mathcal{C}_0^\infty$, we have that $\sigma * \varphi \in \mathcal{C}^\infty$, see results in distributions 26. From Lemma 4 applied in $\sigma = \sigma * \varphi$, if $\sigma * \varphi \in \mathcal{C}^\infty$, then Σ_1 dense in $\mathcal{C}(\mathbb{R})$. □

We will prove that approximating a $\mathcal{C}(\mathbb{R})$ function with one from the set Σ_1 implies approximating a function $\mathcal{C}(\mathbb{R}^n)$ from the set Σ_n. Therefore, it is only necessary to approximate a continuous function. We can see this from the density characterization:

Lemma 6. If Σ_1 is dense in $\mathcal{C}(\mathbb{R})$, then Σ_n is dense in $\mathcal{C}(\mathbb{R}^n)$.

Proof. Consider the set

$$V := span\{f(ax) : a \in \mathbb{R}^n, f \in \mathcal{C}(\mathbb{R})\}.$$

First, we shall see that V is dense in $\mathcal{C}(\mathbb{R}^n)$. If we show that V contains the polynomials, which are dense in $C(\mathbb{R}^n)$ for Weierstrass Theorem, that will be enough.

In fact, the set

$$H := span\{p(ax) : a \in \mathbb{R}^n, p \in R[x]\} \subseteq V.$$

We only need to show that $H = R[x]$, in other words, that the polynomials of degree k can be generated by $(a \cdot x)^k$.

Let $g \in \mathcal{C}(\mathbb{R})$, for any compact subset $K \subset \mathbb{R}^n$, V dense in $\mathcal{C}(K)$. That is, given $\epsilon > 0$, there exist $f_i \in \mathcal{C}(\mathbb{R})$ and $a_i \in \mathbb{R}^n$ $_{i\,=\,1,\ldots,k}$ such that

$$\left|g(x) - \sum_{i=1}^{k} f_i(a^i \cdot x)\right| < \frac{\epsilon}{2}$$

for all $x \in K$. We now consider the set of all the points in the compact K multiplied by the vector a^i. That is $\{a^i \cdot x | x \in K\} \subseteq [\alpha_i, \beta_i]$ for some finite interval $[\alpha_i, \beta_i]$, $i = 1, ..., k$. By hipothesis Σ_1 dense in $\mathcal{C}(\mathbb{R})$, specifically Σ_1 is dense in $[\alpha_i, \beta_i]$ $i = 1, ..., k$. Hence there exist constants c_{ij}, w_{ij} and θ_{ij}, $j = 1, ..., m_i$, $i = 1, ..., k$ such that

$$\left|\ f_i(y) - \sum_{j=1}^{m} c_{ij}\sigma(w_{ij}y + \theta_{ij})\right| < \frac{\epsilon}{2k}$$

for all $x \in K$.

Therefore,

$$\left|g(x) - \sum_{i=1}^{k}\sum_{j=1}^{m} c_{ij}\sigma(w_{ij}(a^i \cdot x) + \theta_{ij})\right| < \epsilon$$

□

We showed that to approximate a $\mathcal{C}(\mathbb{R}^n)$ function we only need to approximate a $\mathcal{C}(\mathbb{R})$ function with the set Σ_1.

4.3 Proof of Theorem 28

Proof.

$\Rightarrow$ To prove this implication statement, we aim to show that if Σ_n is dense in $\mathcal{C}(\mathbb{R}^n)$, then σ is not a polynomial. We will proceed to prove the contrapositive statement, assuming that σ is indeed a polynomial, and demonstrate that in this case, Σ_n cannot be dense in $\mathcal{C}(\mathbb{R}^n)$.

Let σ be a polynomial of degree k, then $\sigma(wx + \theta)$ is a polynomial of degree k for every w, θ. Recall that

$$\Sigma_n = span\{\sigma(w \cdot x + \theta) : w \in \mathbb{R}^n, \theta \in \mathbb{R}\}$$

that is the set of algebraic polynomials of degree at most k. To show that Σ_n is not dens in $\mathcal{C}(\mathbb{R}^n)$, for the definition of density, we need to find a function $f(x) \in \mathcal{C}(\mathbb{R}^n)$, $\epsilon > 0$ and K such that $\|p - f\| > \epsilon$ for all p polynomial of degree k. For example, let $f(x) = cos(x)$, and let $p(x) = \sigma(wx + \theta)$ that has degree at most k. This implies has maximum k roots. We can find a interval where $cos(x)$ has k+1 roots. Therefore, Σ_n is not dense in $\mathcal{C}(\mathbb{R}^n)$.

$\Leftarrow$ In order to prove this implication, we need to show that if σ is not an algebraic polynomial, then Σ_n is dense in $\mathcal{C}(\mathbb{R}^n)$.

By hypothesis, σ is not a polynomial, by Proposition 1 this implies that $\sigma * \varphi$ is not a polynomial. For Lemma 3 we have that $\sigma * \varphi \in \overline{\Sigma_1}$. We showed in Lemma 5 that $\sigma * \varphi \in \mathcal{C}^\infty$ and for Lemma 4 and Lemma 5 we have that Σ_1 is dense in $\mathcal{C}(\mathbb{R})$. Finally, for Lemma 6 Σ_n is dense in $\mathcal{C}(\mathbb{R}^n)$.

□

4.4 About the theorem

4.4.1 Why does it not contradict the Weierstrass approximation theorem?

Theorem 33. *(Weierstrass approximation theorem). Let $f : [a, b] \to \mathbb{R}$ be a continuous function. Then, there exists polynomials $p_n \in \mathcal{R}[x]$ such that the sequence (p_n) converge uniformly to f on $[a, b]$.*

Corollary 34. The set of polynomial functions $\mathcal{R}^n[x]$ is dense in the space of continuous functions on a compact set $K \subset \mathbb{R}^n$, $\mathcal{C}(K)$. So any continuous function on a compact set can be approximated arbitrarly well by a polynomial.

The theorem states that: if Σ_n is dense in $\mathcal{C}(\mathbb{R}^n)$ then σ is not an algebraic polynomial. But why this statment does not contradict the Weierstrass approximation theorem ? This does not work because σ has degree fixed k, then any element in the set Σ_n has degree at most k. Hence, the set Σ_n is a finite vector space and can not be dense

in $\mathcal{C}(\mathbb{R}^n)$. Not all contiunous functions can be approximated with a polynimial of degree fixed, for example: (comment per afegir : per exemple una funcio que sigui continua que no es pugui approximar per un polinomi de com a molt grau k , una k tingui grau mes gran que k polinomi de k+1 ??)

4.4.2 Conclusions

The theorem only requires for the activation function to be nonpolynomial, we do nott need continuity on sigma. For example, let σ be continuous with a jump discontinuity at 0 such that:

$$\lim_{x \to 0^-} \sigma(x) = 0 \qquad \lim_{x \to 0^+} \sigma(x) = 1$$

Given $f \in \mathcal{C}(\mathbb{R})$ and $K \subset \mathbb{R}$ compact, it is possible to approximate f from Σ_1 on K. Let $h \in \overline{\Sigma_1}$ defined as $\sigma(wx)$ letting $w \to 0$, the function:

$$h(x) = \begin{cases} 0 & \text{for } x < 0 \\ 1 & \text{for } x > 0 \end{cases}$$

Linear combinations of h and its translates can unifformly approximate any continuous function on any finite interval (and thus an), compact subset of R).

Our finding that activation function need not be continuous or smooth also has an important biological interpretation, because the activation functions of real neurons may well be discontinuous,or even nonelementary.¡

4.4.3 Results

Definition 35. The set $L^p(\mu)$ contains all mesurable functions f such that:

$$\|f\|_{L^p}(\mu) = \Big(\int_{R^n} |f(x)|^p d\mu(x) \Big)^{1/p} < \infty$$

Proposition 36. Let μ be a non-negative finite measure on $\mathbb{R}$ with compact support, absolutely continous with respect to Lebesgue measure. Then Σ_n is dense in $L_p(\mu)$, $1 \leq p < \infty$, if and only of, σ is not a polynomial.

Proposition 37. If $\sigma \in M$ is not a polynomial (a.e) then,

$$\Sigma_n(\mathcal{A}) = span\{\sigma(\lambda w \cdot x + \theta) : \lambda, \theta \in \mathbb{R}, w \in \mathcal{A} \}$$

is dense in $\mathcal{C}(\mathbb{R}^n)$ for some $\mathcal{A} \subset \mathbb{R}^n$ if and only if there does not exist a nontrivival polynomial vanishing on $\mathcal{A}$.

Leshno et al. [1993]

Chapter 5

References

Fréchet spaces. volume 25 of *Pure and Applied Mathematics*, pages 85–94. Elsevier, 1967.

R. A. Adams, J. J. F. Fournier, and J. J. F. Fournier. *Sobolev Spaces*. Elsevier, 2003.

A. Friedman. *Generalized Functions and Partial Differential Equations*. Englewood Cliffs, 1963.

K. Hornik. Approximation capabilities of multilayer feedforward networks. *Neural Networks*, 4(2):251–257, 1991. ISSN 0893-6080.

M. Leshno, V. Y. Lin, A. Pinkus, and S. Schocken. Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. *Neural Networks*, 6(6):861–867, 1993.