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 is the science of programming computers so they can learn from data. The key is that it allows solving a problem without being explicitly programmed. Mitchell (1997) provides a definition for a machine learning algorithm: "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E."

2.1.1 Tasks

A machine learning task refers to the problem that the machine learning model aims to solve or accomplish. Machine learning tasks can vary widely depending on the nature of the data and the desired outcome. Some common machine learning tasks include classification, regression, clustering, and anomaly detection.

2.1.2 Performance Measure

2.1.3 Experience

2.2 Types of Learning

But what do we mean by learning? 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. The three main types are:

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

2.2.1 Supervised learning

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.

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

2.3 Function Learning

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. 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

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

2.4 Artificial neural networks

Artificial neural networks (ANNs) are widely used for nonlinear function approximation. (nonlinear classifier.) They were initially inspired by the way biological neurons process information.

They are composed of interconnected processing units called artificial neurons, which are organized in layers and are capable of learning and generalizing from data.

Chapter 3

Function Approximation

Among the most famous techniques for function approximation, we find interpolation: such as Taylor polynomial, Chebyshev polynomial, the method of least squares, or spline approximation. In this chapter we are going to talk about ...

3.1 ????

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

Definition 1. 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*).

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 $f, g \in X$.

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

Definition 3. We denote by $\mathcal{C}_0^\infty$ functions $\mathcal{C}^\infty$ with compact support. Recall that the support of a function u is denoted by $supp(u) = \{x|u(x) \neq 0\}$

Proposition 4. Let ρ be a metric on the set $\mathcal{C}_0^\infty[a,b]$ defined by

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

where $\|\varphi\|_n = \sum_{j=0}^{n} \sup_{x \in [a,b]} |\varphi^{(j)}(x)|$. We can show that $(\mathcal{C}_0^\infty[a,b], \rho)$ is a complete metric space (Fréchet space).

Lebesgue measure

Definition 5. A box in $\mathbb{R}^d$ is a set of the form

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

The volume of the box is

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

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

Definition 8. A function u defined almost everywhere on a measurable set $\Omega \in \mathbb{R}^n$ is said to be *essentially bounded* on Ω if $|u(x)|$ is bounded almost everywhere on Ω. We denote $u \in L^\infty(\Omega)$ with the norm

$$\|u\|_{L^\infty(\Omega)} = inf(\lambda|\{x : |u(x)| \geq \lambda\} = 0) = ess\sup_{x \in \Omega} |u(x)|$$

We have that $L^\infty(\mathbb{R})$ is the space of essentially bounded functions.

Examples and counterexamples of functions essentially bounded.

- $f : \Omega \to$

Definition 9. 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 10. 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

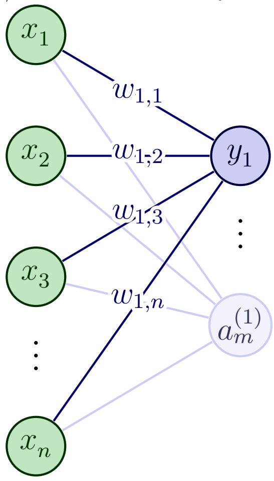
$$\lim_{j \to \infty} \|g - f_j\|_{L^\infty(K)} = 0.$$

Definition 11. Let f, g be real-valued functions with compact support. We define the *convolution* of f with g as

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

3.2 Multilayer Feedforward Network

Multilayer feedforward networks are a type of artificial neural network that consist of several layers of interconnected nodes, with each node taking input from the previous layer and producing output for the next layer. The general architecture of a multilayer feedforward network, MFN, consist of: input layer: n-input units, one/more hidden layers : intermediate processing units, output layer: m output-units.



$$\begin{pmatrix} a_1^{(1)} \\ a_2^{(1)} \\ \vdots \\ a_m^{(1)} \end{pmatrix} = \beta_j \cdot \sigma \left[\begin{pmatrix} w_{1,0} & w_{1,1} & \dots & w_{1,n} \\ w_{2,0} & w_{2,1} & \dots & w_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ w_{m,0} & w_{m,1} & \dots & w_{m,n} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} + \begin{pmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_m \end{pmatrix} \right]$$

$$a^{(1)} = \beta_j \cdot \sigma \left(\mathbf{W}^{(0)} a^{(0)} - \theta_{\mathbf{j}} \right)$$

Definition 12. (Multilayer feedforward networks) The function that a MFN compute is:

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

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 applied element-wise to the vector $w_j^T x - \theta_j$, where $\theta_j \in \mathbb{R}$ 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 4

Theorem and proof

The work revolves around the following theorem and its proof.

4.1 Theorem

Theorem 13. *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 an algebraic polynomial.*

4.2 Pre-pre

Definition 14. 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 $[a, b]/U$.

4.3 Proof

4.3.1 Left $\Leftarrow$??

Consider that σ is not an algebraic polynomial and we aim to show that Σ_n is dense in . We will divide the proof into

σ is not a polynomial $\Rightarrow$ $\sigma * \varphi$ is not a polynomial.

We will proove that if σ is not a polynomial (true by hypothesis) then $\sigma * \varphi$ is not a polynomial. We will show the contrapositive: if the convolution is a polynomial, then σ is a polynomial.

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 $\varphi \in \mathcal{C}_0^\infty[a,b]$, where $\mathcal{C}_0^\infty[a,b]$ is the set of functions $\mathcal{C}_0^\infty$ with support in $[a,b]$ for any $a < b$.

For () we have that $(\mathcal{C}_0^\infty[a,b], \rho)$ is a complete metric space. By assumption, we have that $\sigma * \varphi$ is a polynomial (for any $\varphi \in \mathcal{C}_0^\infty[a,b]$).

Consider the following set, which has the property that we want to show.

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

Clearly, if $\varphi \in V_k$, then $deg(\sigma * \varphi) \leq k$. We want to show that $\mathcal{C}_0^\infty[a,b] \subseteq V_k$. This set fulfills the following properties, $V_k \subset V_{k+1}$, V_k is a closed subspace and $\cup_{k=0}^\infty V_k = \mathcal{C}_0^\infty[a,b]$. As $\mathcal{C}_0^\infty[a,b]$ is a complete metric space, for Blaire's Category Theorem then there exists an integer m such that $V_m = \mathcal{C}_0^\infty[a,b]$.

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 transition 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 $\varphi \in \mathcal{C}_0^\infty$, for (31) we have that

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

From standard results in Distribution Theory [e.g., Friedman(1963, 57-59)], σ is itself a polynomial of degree at most m (a.e.). □

Conclusion: Consider σ is not a polynomial implies that $\sigma * \varphi$ is not a polynomial.

$\sigma * \varphi \in \overline{\Sigma_1}$

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

Proof. We recall that set

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

Consider

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

The sequence (h_m) satisfies $h_j \in \Sigma_1$ for $j = 1, ..., m$. $(w_i = 1, \theta_i = -y_i, \beta_i = \varphi(y_i)\Delta y_i)$.

Where $y_i = -\alpha + \frac{2i\alpha}{m}$, $\Delta y_i = \frac{2\alpha}{m}$ for $i = 1, ..., m$. Partition of the interval $[-\alpha, \alpha]$

We want to show that $h_m \rightrightarrows \sigma * \varphi$ in $[-\alpha, \alpha]$.

Given $\epsilon > 0$, we choose $\delta > 0$ such that $10\delta\|\sigma\|_{L^\infty\{-2\alpha,2\alpha\}}\|\varphi\|_{L^\infty} \le \frac{\epsilon}{3}$. Note that ...

We know that $\sigma \in M$. Hence, for this given $\delta > 0$ and $[-\alpha, \alpha]$ interval, there exists $r(\delta)$ finite number of intervals the measure of whose union $\mathfrak{U}$ is δ such that σ is uniformly continuous on $[-2\alpha, 2\alpha]$. 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.

3. From the previous, σ is uniformly continuous on $[-2\alpha, 2\alpha]$.

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 $y_0 = \alpha$ .. dibuix.

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

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

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

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

$$= \frac{\epsilon}{3\|\varphi\|_{L_1}} \int |\varphi(y)| \, dy$$

$$= \frac{\epsilon}{3\|\varphi\|_{L_1}} |\varphi(y)|_{L_1}$$

$$= \frac{\epsilon}{3}$$

If $x - \Delta_i \cap U \neq \varnothing$

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

True by our choice of m, satisfies $m\delta > \alpha r(\delta) \iff \delta > \frac{\alpha \cdot r(\delta)}{m}$

$$\sum_{i=1}^{m} \int_{\widetilde{\Delta_i}} |\varphi(y)| \, |\sigma(x-y) - \sigma(x-y_i)| \, dy \leq$$

$$\leq \sum_{i=1}^{m} \int_{\widetilde{\Delta_i}} \|\varphi\|_{L^\infty} \, 2\|\sigma\|_{L^\infty[-2\alpha,2\alpha]}$$

$$= \|\varphi\|_{L^\infty} \, 2\|\sigma\|_{L^\infty[-2\alpha,2\alpha]} \sum_i |\widetilde{\Delta_i}|$$

$$\leq \|\varphi\|_{L^\infty} \, 2\|\sigma\|_{L^\infty[-2\alpha,2\alpha]} \, 5\delta \leq \epsilon/3$$

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

$$= \left| \sum_{i=1}^{m} \int_{\Delta_i} \sigma(x-y_i)[\varphi(y) - \varphi(y_i)]dy \right|$$

$$\leq \sum_{i=1}^{m} \int_{\Delta_i} |\sigma(x-y_i)| \, |\varphi(y) - \varphi(y_i)| \, dy$$

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

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 $x \in [-\alpha, \alpha]$ $\square$

4.3.2 Σ_1 dense in $\mathcal{C}(\mathbb{R})$

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

Proof. Recall the set Σ_1 with (1). We can write any function $h \in \Sigma_1$ as

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

We can see that

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

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 will see by reduction to absurdity that if σ is not a polynomial (by hypothesis) then there exists a $\theta_k \in \mathbb{R}$ such that $\sigma^{(k)}(\theta_k) \neq 0$.
If σ is not a polynomial and $\sigma \in \mathcal{C}^\infty$, lets assume that $\nexists \theta_k \in \mathbb{R}$ such that $\sigma^{(k)}(\theta_k) \neq 0$. This means that the k-th derivative at every point is 0,

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

If we integrate k times this expression,

$$\int \sigma^{(k)} = \int 0 \Rightarrow \sigma^{(k-1)} = C$$

,

$$\int \sigma^{(k-1)} = \int C \Rightarrow \sigma^{(k-2)} = Cw$$

, then we end up σ is a polynomial. Contradiction. 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.

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

*$\overline{\Sigma_1}$ denotes the clausure of the set Σ_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, $\sigma * \varphi \in \overline{\Sigma_1}$. Clearly, $\sigma * \varphi(wx + \theta) \in \overline{\Sigma_1}$, for each $\theta \in \mathbb{R}$. For σ and $\varphi \in \mathcal{C}_0^\infty$ we have that $\sigma * \varphi \in \mathcal{C}^\infty$. This follows by results in distribution (30). From Lemma 4 applied in $\sigma = \sigma * \varphi$, if $\sigma * \varphi \in \mathcal{C}^\infty$, then Σ_1 dense in $\mathcal{C}(\mathbb{R})$. □

4.3.3 Σ_1 is dense in $\mathcal{C}(\mathbb{R})$, then Σ_n is dense in $\mathcal{C}(\mathbb{R}^n)$

We will proof 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. Let

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

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 would be enough.

!!mirar Let $L(a)$ denote the span of the n rows of a for each $a \in \mathbb{R}^n$. Set $L(R^n) = \cup L(a)$. Let

$$H_k^n = \{\sum c_m s^m\}$$

denote the set of homogeneous polynomials of n variables of total degree k, and

$$H^n = \cup_{k=0}^\infty H_k^n$$

the set of all homogeneous polynomials of n variables.

Assume that for a given $k \in \mathbb{N}$ no non-trivial $p \in H_k^n \subseteq V$ for all $k \in \mathbb{Z}$, then V contains all polynomials. For that we have V dense in $\mathcal{C}(\mathbb{R}^n)$. Now, we only need to show that $H_k^n \subseteq V$. SOS

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 $a_i \in \mathbb{R}^n$ $_{i=1,\dots,k}$ 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.

4.4 Proof

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 (only) 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, for Lemma 1 and Lemma 2 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 $\mathcal{C}(\mathbb{R}^n)$.

□

Leshno et al. [1993]

Chapter 5

About theorem

5.0.1 Why does it not contradict the Weierstrass approximation theorem?

Theorem 15. *(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 16. 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 apparoximated 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 ??)

5.0.2 Previous results

The activation functions that were reported thus far in the literature.

Theorem 17. *(Hornik Theorem 1). Standard 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 18. *(Hornik Theorem 2) Standard 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.

5.0.3 Results

Definition 19. 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 20. 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 21. 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 $\mathcal{A}$.

Remark 1. The theorem only requires for the activation function to be nonpolynomial, we dont 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, letting $w \to 0$ in $\sigma(wx)$ the function

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

$h \in \overline{\Sigma_1}$.

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

Chapter 6

Results

$$t = x + y \tag{6.1}$$

Chapter 7

Conclusions

It is a mistake to confound strangeness with mystery.

— Sherlock Holmes, *A Study in Scarlet*

7.1 Summary

7.2 Outlook and Future Work

Hem trobat:

- Aaaaa
- Bbbbbb

Chapter 8

References

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.

Appendix A

Theory used

Definition 22. Riemann integral reminder. The Riemann integral is a method for calculating the volume under a curve of a continuous function on a closed, bounded domain in R^n. The method involves dividing the domain into smaller subregions and approximating the volume of each subregion with a rectangular solid whose height is the function value at a specific point in the subregion. The Riemann sum is the sum of the volumes of all the rectangular solids, and as the size of the subregions approaches zero, the Riemann sum converges to the Riemann integral.

Definition 23. 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 24. The closure of a set A of a metric space (X, d) is defined as follows:

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

Proposition 25. Let (X, τ) be a topological space and $A \subseteq X$ be a subset. Then, A is dense in (X, τ) if and only if $\overline{A} = X$.

Definition 26. A metric space (X, d) is said to be *complete* if every Cauchy sequence in X converges to a point in X.

Definition 27. We say that a property holds almost everywhere (a.e.) if the set of points that doesn't hold it is null.

Definition 28. $\varphi : I \to \mathbb{R}$ is uniformly continuous on I if $\forall \epsilon > 0 \exists \; \delta > 0$ such that $|\varphi(x) - \varphi(y)| < \epsilon$ whenever $|x - y| < \delta$

A.1 Blaire's category theorem

Definition 29. Let A be a subset of the metric space (X, d). 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.

Proposition 30. If f is a smooth function that is compactly supported and g is a distribution, then $f * g$ is a smooth function defined by

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

Proposition 31.

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

Definition 32. 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 33. *(Blaire's Category Theorem) Any complete metric space is of category II.*

Therefore, if we have $\mathcal{C}_0^\infty[a, b]$ complete metric space, we know that is of category II, i.e. $\mathcal{C}_0^\infty[a, b]$ cannot be written as a countable union of nowhere-dense sets. We have $\cup_{k=0}^\infty V_k = \mathcal{C}_0^\infty[a, b]$. Therefore, some V_m contains a nonempty open set. V_m is a vector space thus $V_m = \mathcal{C}_0^\infty[a, b]$. no entenc el final