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

To my colleagues,

...

Abstract

Exemple d'abstract no definitiu, x ficar algo.

The Leshno-Lin-Pinkus-Schocken (LLPS) theorem is a seminal result in the field of neural network theory. It states that multilayer feedforward neural networks with a nonpolynomial activation function can approximate any function to any desired level of accuracy, provided that the network has a sufficient number of hidden units. In this thesis, we present a detailed proof of the LLPS theorem, along with an explanation of its significance and implications for the design and training of neural networks

Resum

blslalblallb en català

Preface

Inpirat per **?**.

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

Contents

Chapter 1

Introduction

ficar una quote que quedi bé

— John S. Bell, *Against Measurement*

Early computers were used to perform exact computations with high accuracy and efficiency. Back in 1945, one of the first electronic computer was invented for ballistic calculations during World War II. Computers seemed to be limited to these exact computation tasks. However, over time, researchers started pushing the boundaries of what computers can do, eventually leading to the development of what we call now Artificial Intelligence. AI seeks to make computers do the sorts of things that minds can do.

Some of these (e.g. reasoning) are normally described as "intelligent." Others (e.g. vision) aren't. But all involve psychological skills—such as perception, association, prediction, planning, motor control—that enable humans and animals to attain their goals.

3. on les matematiques prenen lloc ? pq son importants . Can we suggest conjectures, relationships , theorems between fields ??? using ml as a tool to see unexpected relationships.

ML might become a bycicle for the mind !!

MATHS USING MCH LEARNING <–> ML USING MATHS

Chapter 2

Artificial Intelligence

2.1 What is Artificial Intelligence

2.2 What is Machine Learning

Machine Learning is the science of programming computers so they can learn from data.

(with the aim to solve a problem without being explicitly programmed.)

For example,

2.3 Types of Learning

We can think about learning as the way we understand it as a human. We can learn . That is how we can classify the Machine Learning problem, based on the degree of feedback.

2.3.1 Supervised Learning

2.3.2 Reinforcement Learning

2.3.3 Unsupervised Learning

2.4 Output

2.4.1 Classification Problem

2.4.2 Regression Problem

Chapter 3

Artificial Neural Networks

Definition 1. *(Essentially bounded). A function u defined almost everywhere with respect to Lebesgue measure v 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 2. *(Locally essentially bounded). A function u defined almost everywhere with respect to Lebesgue measure 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 3. *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 $lim_{j\to\infty}\|g - f_j\|_{L^\infty(K)}$*

Definition 4. *Let 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 M is of zero Lebesgue measure.*

Proposition 1. *This implies that for any $\sigma \in M$, 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 5. *suport*

Definition 6. *(Multilayer feedforward networks) The general architecture of a multilayer feedforward network, MFN, consist of:*

- *input layer: n-input units, x*
- *one/more hidden layers : intermediate processing units*
- *output layer: m output-units $f(x)$*

function that a MFN compute is:

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

- $x = (x_1, ..., x_n)$ *input-vector*
- k: *# of processing-units in the hidden layer*
- $w = (w_1, ..., w_n)$: *weights vector*
- $\sigma : \mathbb{R} \to \mathbb{R}$ *activation function*
- θ *treshold value: ???*
- β

We take $C(\mathbb{R}^n)$ to be the family of real world functions that one may wish to approximate with feedforward network architectures

Definition 7. $\mathcal{C}_0^\infty$ *functions* $\mathcal{C}^\infty$ *with compact support.*

Definition 8. *Convergència uniforme)*

Definition 9. $\varphi : I \to \mathbb{R}$ *is uniformly continuous on* I *if* $\forall \epsilon > 0 \exists\ \delta > 0$ *such that* $|\varphi(x) - \varphi(y)| < \epsilon$ *whenever* $|x - y| < \delta$

Definition 10. *Let* f, g *be real-valued functions with compact support. We define the* convolution *of* f *with* g *as*

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

Theorem 1. *Let* $\sigma \in M$. *Set*

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

Then $\sum_n$ *is dense in* $\mathcal{C}(\mathbb{R}^n)$ *if and only if* σ *is not an algebraic polynomial.*

Chapter 4

Lemmas and proof

This chapter/section presents the lemmas that are necessary to prove the main result. In order to prove the theorem, I will first demonstrate the inverse implication. Specifically, I will show that " ".

4.1 Part 1

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 $a < b$.

Let ρ be a metric on $\mathcal{C}_0^\infty[a,b]$ 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 $\|\varphi\|_n = \sum_{j=0}^{n} \sup_{x\in[a,b]} |\varphi^{(j)}(x)|$. $\mathcal{C}_0^\infty[a,b]$ 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 (has the property that we want to show).
$V_k = \{\varphi \in \mathcal{C}_0^\infty[a,b] \,|\, deg(\sigma * \varphi) \leq k\}$. 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, then tere exists an integer m such that $V_m = \mathcal{C}_0^\infty[a,b]$. Proof ??

For the general case where $\varphi \in \mathcal{C}_0^\infty$, we note that the number m does not deppend on the interval $[a,b]$. Por tanto hemos visto que el conjunto con la propiedad que queriamos es = que el que tenemos. we

□

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. For all $\varphi \in \mathcal{C}_0^\infty$, we have that

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

□

Conclusion: If we have that $\sigma * \varphi$ is a polynomial then σ is a polynomial. This contradicts the hypothesis. Therefore, $\sigma * \varphi$ will not be a polynomial.

4.2 Part 2

Lemma 3. *For each* $\varphi \in \mathcal{C}_0^\infty$, $\sigma * \varphi \in \overline{\sum_1}$.

Proof. Consider

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

The sequence (h_m) satisfies $h_j \in \sum_1$ 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 $10\delta\|\sigma\|_{L^\infty\{-2\alpha,2\alpha\}}\|\varphi\|_{L^\infty} \leq \frac{\epsilon}{3}$. 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 $\mathcal{U}$ 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 $\Delta_i = [y_{i-1},]$ where .. dibuix.

First, recall that (fer la integral es igual que sumar per intervals les integrals)

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

Consider the following difference

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

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

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

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

If $x - \Delta_i \cap U = \varnothing$. Since $x - y \notin U$, $x - y_i \notin U$ and $x - y_i \in [-2\alpha, 2\alpha]$, bc (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 $m\delta > \alpha r(\delta) \iff \delta > \frac{\alpha \cdot r(\delta)}{m}$

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

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

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

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

□

Lemma 4. *If $\sigma \in \mathcal{C}^\infty$, then $\sum_1$ is dense in $\mathcal{C}(\mathbb{R})$.*

Proof. We recall that set $\sum_1 = span\{\sigma(w \cdot x + \theta) : w \in \mathbb{R}, \theta \in \mathbb{R}\}$. We can write any function $h \in \sum_1$ as $h = \sum_i \beta_i \sigma_i(w_i x + \theta_i) = \beta_1 \sigma_1(w_1 x + \theta_1) + ...$

$\frac{\sigma([w+h]x+\theta)-\sigma(wx+\theta)}{h} \in \sum_1$ because is a linear combination, where $\beta_1 = \frac{1}{h}, \beta_2 = \frac{-1}{h}$....
By hypothesis, we have $\sigma \in \mathcal{C}^\infty$. By definition of derivative we have

$$\frac{d}{dw}\sigma(wx+\theta) = \lim_{h\to 0} \frac{\sigma([w+h]x+\theta) - \sigma(wx+\theta)}{h} \in \overline{\sum_1}^*$$

Because the limit of a set belongs to the closure of the set.
By the same argument, $\frac{d^k}{dw^k}\sigma(wx+\theta) \in \overline{\sum_1}$ for all $k \in \mathbb{N}, w, \theta \in \mathbb{R}$.
We observe that $\frac{d}{dw}\sigma(wx+\theta) = \sigma'(wx+\theta)\cdot x$. If we differentiate this expression k times, we obtain

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

Since σ is not a polynomial (theorem hypothesis) then there exists a $\theta_k \in \mathbb{R}$ such that $\sigma^{(k)}(\theta_k) \neq 0$
Lets see.**** 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, i.e, $\sigma^{(k)}(\theta) = 0\ \forall \theta \in \mathbb{R}$. If we integrate k times, $\int \sigma^{(k)} = \int 0 \iff \sigma^{(k-1)} = C$, $\int \sigma^{(k-1)} = \int C \iff \sigma^{(k-2)} = Cw$, 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 this point θ_k where the derivative does not vanish.

$$\sigma^{(k)}(\theta_k)\cdot x^k = \frac{d^k}{dw^k}\sigma(wx+\theta)\Big|_{w=0,\theta=\theta_k} \in \overline{\sum_1}$$

That implies that $\overline{\sum_1}$ contains all polynomials, because the expression $\sigma^{(k)}(\theta_k)x^k$ generates all polynomials. By the Weierstrass theorem, it follows that $\sum_1$ contains... falta mirar. □

Lemma 5. *If for some $\varphi \in \mathcal{C}_0^\infty$ we have that $\sigma * \varphi$ is not a polynomial, then $\sum_1$ is dense in $\mathcal{C}(\mathbb{R})$*

Proof. From Lemma 3, $\sigma * \varphi \in$ □

Lemma 6. *If $\sum_1$ is dense in $\mathcal{C}(\mathbb{R})$, then $\sum_n$ is dense in $\mathcal{C}(\mathbb{R}^n)$.*

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

Proof.

⇒ To prove the implication, we will use proof by contrapositive. We will see the following. If σ is a polynomial then $\sum_n$ is not 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. We have $\sum_n = span\{\sigma(w\cdot x+\theta) : w \in \mathbb{R}^n, \theta \in \mathbb{R}\}$ that is the set of algebraic polynomials of degree at most k.

*$\overline{\sum_1}$ denotes the clausure of the set $\sum_1$

$\sum_n$ is not dens in $\mathcal{C}(\mathbb{R}^n)$ if for a function $f(x) \in \mathcal{C}(\mathbb{R}^n)$ we can find $\epsilon > 0$ and K such that $\|p - f\| > \epsilon$ for all p polynomial of degree k. For example, let $f(x) = cos(x)$, and $p(x) = \sigma(wx + \theta)$ that has degree k. This implies has maximum k roots. We can find a interval where there is k+1 roots.

$\Leftarrow$

□

Leshno et al. [1993]

Chapter 5

Results

$$t = x + y \tag{5.1}$$

Chapter 6

Conclusions

> It is a mistake to confound strangeness with mystery.
>
> — Sherlock Holmes, *A Study in Scarlet*

6.1 Summary

6.2 Outlook and Future Work

Hem trobat:

- Aaaaa
- Bbbbbb

Chapter 7

References

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

Appendix A

Derivation of an integration

Definition 11. *(Complete metric space).*