Approximation by Superpositions of a Sigmoidal Function

M. Ruby, C. Salinas

Department of Mathematics
Purdue University

Purdue University Machine Learning Seminar, 2019

Introduction

In many Machine Learning applications the goal is to approximate a continuous function $f : \mathbf{R}^n \to \mathbf{R}$ by an expression of the form

$$\sum_{j=1}^{N} \alpha_j \sigma(y_j^{\mathsf{T}} x + \theta_j), \tag{1}$$

where $y_j \in \mathbf{R}^n$ and $\alpha_j, \theta_j \in \mathbf{R}$ are fixed.
In Machine Learning terminology, the y represent weights and the θ biases. Equation (1) models a fully connected network with one hidden layer.

The Main Result

The author demonstrates that, theoretically, such an architecture is sufficient to approximate all continuous functions on the unit cube. That is,

Theorem 1 (Theorem 2 in C89)

For σ a continuous sigmoidal function, finite sums of the form

$$\sum_{j=1}^{N} \alpha_j \sigma(y_j^{\mathsf{T}} x + \theta_j),$$

are dense in $C(I_n)$.

The proof will rely on the following more general result

Theorem 2 (Theorem 1 in C89)

Let σ be continuous and discriminatory. Then the set of finite sums of the form

$$\sum_{j=1}^{N} \alpha_j \sigma(y_j^{\mathsf{T}} x + \theta_j),$$

is dense in $C(I_n)$.

With this theorem at our disposal we need only show that sigmoidal functions are discriminatory; that is, we will show that

Lemma 3 (Lemma 1 in C89)

Any continuous sigmoidal function is discriminatory.

To prove Theorem 2 we will need a little bit of Functional Analysis. In particular, we will need the following corollary of the Hahn–Banach Theorem

Corollary 4

If V is a normed vector space, W a subspace of V, and $v_0 \in V \setminus W$ with $\operatorname{dist}(v_0, W)$ nonzero, then there exists a bounded linear functional L such that $L(v_0) = 1$, but $L(v) = 0$ for all $v \in W$.

Proof.

The proof follows from the Hahn–Banach Theorem applied to the quotient V/W. For details, see Corollary 6.8 in Conway, p. 79. □

We will also be needing use of the Riesz Representation Theorem

Theorem 5 (Riesz Representation Theorem)

If X is locally compact and μ a measure on X, the map

$$\mu \longmapsto \int_X f \, d\mu$$

is an isometric isomorphism from the space of measures on X, $M(X)$, to the dual of $C_0(X)$.

Proof.

See Conway, Theorem 5.7. $\square$

What the Riesz Representation Theorem is saying is that every bounded linear functional L on $C_0(X)$ is of the form $L(f) = \int_X f \, d\mu$ for some measure μ on X.

Definitions

Let's define a couple of these terms.

Definition 6 (Discriminatory functions)

We say σ is *discriminatory* if given a measure μ on I_n

$$\int_{I_n} \sigma(y^{\mathsf{T}}x + \theta)\, d\mu(x) = 0$$

for all $y \in \mathbf{R}^n$ implies $\mu = 0$.

And of course.

Definition 7 (Sigmoidal functions)

We say that σ is sigmoidal if

$$\begin{cases} \sigma(t) \to 1 & \text{as } t \to +\infty, \\ \sigma(t) \to 0 & \text{as } t \to -\infty. \end{cases}$$

Let $S = \left\{ \sum_{j=1}^{N} \alpha_j \sigma(y_j^\mathsf{T} x + \theta_j) \right\} \subset C(I_n)$. Clearly S is a subspace of $C(I_n)$.

The author shows that the closure, R, of S is $C(I_n)$.

By the aforementioned Corollary 4, there exists a functional L which is zero on S but nonzero on $C(I_n) \setminus S$.

By the Riesz Representation Theorem, L is of the form

$$L(h) = \int_{I_n} h(x)\, d\mu(x),$$

for some μ measure on I_n.

Proof of Theorem 1, Part 2

Since $\sigma(y^{\mathsf{T}} x + \theta)$ is in R for all y, θ,

$$\int_{I_n} \sigma(y^{\mathsf{T}} x + \theta)\, d\mu(x) = 0,$$

for all y, θ.

However, since σ was assumed to be discriminatory, $\mu = 0$.
This contradicts assumption that L is nonzero.

Therefore, $R = C(I_n)$.

Next the author shows that the sigmoidal functions that we care about are in fact discriminatory. More generally, he shows that

Lemma 8

Any bounded, measurable sigmoidal function is discriminatory. In particular, any continuous sigmoidal function is discriminatory.

First we note that

$$\sigma(\lambda(y^{\mathsf{T}}x+\theta)+\phi)\begin{cases}\to 1 & \text{for } y^{\mathsf{T}}x+\theta>0 \text{ as } \lambda\to+\infty,\\ \to 0 & \text{for } y^{\mathsf{T}}x+\theta<0 \text{ as } \lambda\to+\infty,\\ =\sigma(\phi) & \text{for } y^{\mathsf{T}}x+\theta=0 \text{ for all } \lambda.\end{cases}$$

So $\sigma_\lambda(x) := \sigma(\lambda(y^{\mathsf{T}}x+\theta)+\phi)$ converges pointwise and boundedly to the function

$$\gamma(x)=\begin{cases}1 & \text{for } y^{\mathsf{T}}x+\theta>0,\\ 0 & \text{for } y^{\mathsf{T}}x+\theta<0,\\ \sigma(\phi) & \text{for } y^{\mathsf{T}}x+\theta=0,\end{cases}$$

as $\lambda\to\infty$.

Now, let $\Pi_{y,\theta} = \{x \colon y^{\mathsf{T}} x + \theta = 0\}$ and let $H_{y,\theta} = \{x \colon y^{\mathsf{T}} x + \theta > 0\}$.

By the Lebesgue Bounded Convergence Theorem,

$$\begin{aligned} 0 &= \int_{I_n} \sigma_\lambda(x)\, d\mu(x) \\ &= \int_{I_n} \gamma(x)\, d\mu(x) \\ &= \sigma(\phi)\mu(\Pi_{y,\theta}) + \mu(H_{y,\theta}) \end{aligned}$$

for all ϕ, θ, y.

The last thing we need to show is that if $\mu(H_{y,\theta}) = 0$ for all y, θ implies $\mu = 0$.

To that end, fix y and define the linear functional F as follows

$$F(h) = \int_{I_n} h(y^T x)\, d\mu(x).$$

Note that F is a bounded functional on $L^\infty(\mathbf{R})$ since μ is a finite signed measure.

Let $h = \chi_{[\theta,\infty)}$. Then

$$F(h) = \int_{I_n} \chi_{[\theta,\infty)}(y^\mathsf{T} x) = \mu(\Pi_{y,-\theta}) + \mu(H_{y,-\theta}) = 0.$$

Since the simple functions are dense in $L^\infty(\mathbf{R})$, $F = 0$.

Proof of Lemma 3

By linearity $L(h) = 0$ for $h = \chi_I$ for I any interval in $\mathbf{R}$.

Since simple functions are dense in $L^\infty(\mathbf{R})$, $L = 0$.

In particular,

By the previous results the main theorem now follows as a corollary to theorem 1 and lemma 2. That is, we have proven

Theorem 9 (Theorem 2 in C89)

For σ a continuous sigmoidal function, finite sums of the form in Equation (1), *i.e.,*

$$\sum_{j=1}^{N} \alpha_j \sigma(y_j^{\mathsf{T}} x + \theta_j),$$

are dense in $C(I_n)$.