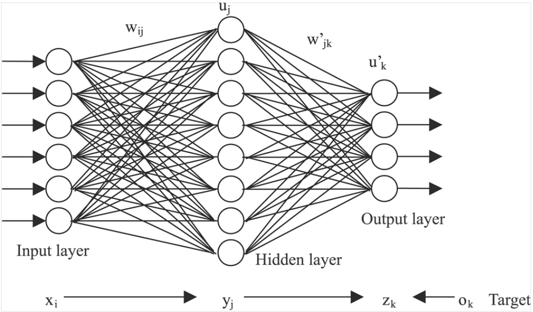
The Universal Approximation Theorem

Jacob Way

November 30, 2022

Background

- A neural network is a function approximation composed of simple "neuron" functions



- Passing different parameters to these simple functions affect the composition's properties
- In the field of machine learning, the parameters of a neural network are updated to "learn" a task using examples
- very large neural networks can carry out difficult tasks which would be impossible to design a simple function for, like facial recognition, object detection, and recommendation

Universal Approximation Theorem (Cybenko, '89)

Theorem

Define $I_n \in \mathbb{R}^n$ to be the n-dimensional unit box.
Let σ be any continuous sigmoidal function. Then finite sums of the form

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

are dense in $C(I_n)$.

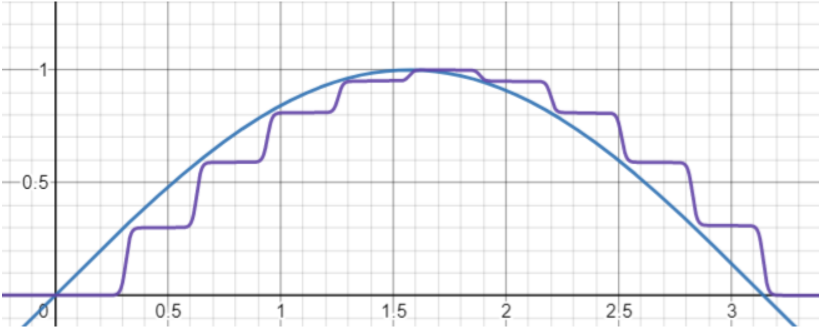
- In other words, for any $f(x)$ and $\epsilon > 0$, there exists such a function G such that $|f(x) - G(x)| < \epsilon$ for all $x \in I_n$

Example

Let σ be the logistic function $\sigma(x) = \frac{1}{1+e^{-100x}}$ and $f(x) = sin(x/pi)$.

Then we can make an arbitrarily close approximation of $f(x)$ using a finite sum.

$$G(x) = \sum_{j=1}^{n} \left(sin(\frac{j\pi}{n}) - sin(\frac{(j-1)\pi}{n}) \right) \sigma(x - \frac{j\pi}{n}) \qquad (2)$$



Increasing n will make this approximation more accurate.

Properties of Functionals

- A functional is a mapping from a vector space to the reals.

Let L be a functional.

- L is linear if $L(ax + by) = aL(x) + bL(y)$ for all scalars a and b and all vectors x and y.
- L is sub-linear if $L(ax) = aL(x)$ and $L(x + y) \leq L(x) + L(y)$ for all scalars $a \geq 0$ and all vectors x and y.

If L is a functional whose domain is a function space

- L is positive if $f(x) \geq 0$ for all x implies that $L(f) \geq 0$

Hahn-Banach Theorem

Theorem (Hahn-Banach)

Let X be a vector space and $p : X \to \mathbb{R}$ be a sublinear functional, $X_0 \subseteq X$ a linear subspace. If $\varphi_0 : X_0 \to R$ is a linear functional and is dominated by p on X_0, then there exists a linear extension $\varphi : X \to \mathbb{R}$ which is dominated by p on all of X.

Corollary

Let $X_0 \subseteq X$ be a subspace of a normed linear space X. If $y \in X \backslash X_0$ is a nonzero unit vector, then there exists a linear functional $\varphi : X \to \mathbb{R}$ such that $\varphi|_{X_0} = 0$ and $\varphi(y) = 1$.

Riesz-Markov Representation Theorem

Theorem (Riesz-Markov)

Let X be a compact metric space. If $\ell : C(X) \to \mathbb{R}$ is a positive linear functional on $C(X)$, then there exists a unique (positive) regular Borel measure on μ on X such that*

$$\ell(f) = \int_X f(x) d\mu(x).$$

Discriminatory Functions

Theorem

Let σ be a continuous sigmoidal function. Then σ is discriminatory, meaning that for a signed regular Borel measure μ on I_n:

$$\int_{I_n} \sigma(y^T x + \theta) d\mu(x) = 0 \tag{3}$$

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

- Basically this means that the integral of σ must be nonzero over some interval whenever μ is nonzero.

Proof

Let $S \subset C(I_n)$ be the set of functions of the form

$$G(x) = \sum_{j=1}^{N} \alpha_j \sigma(y_j^T x + \theta_j) \tag{4}$$

S is a linear subspace of $C(I_n)$.
Assume for the sake of contradiction that S is not dense in $C(I_n)$.
Then the closure of S, R, is a closed proper subspace of $C(I_n)$.

Proof

By the Hahn-Banach Theorem, there is a bounded linear functional L on $C(I_n)$ such that $L \neq 0$ but $L(R) = L(S) = 0$.
By the Riesz-Markov Representation Theorem, L is of the form:

$$L(h) = \int_{I_n} h(x) d\mu(x) \tag{5}$$

for a positive measure μ, and for all $h \in C(I_n)$.
Since $\sigma(y^T x + \theta) \in R$ for all y and θ, we must have that

$$\int_{I_m} \sigma(y^T x + \theta) d\mu(x) = 0 \tag{6}$$

for all y and θ.
However, since we assumed σ was discriminatory, this implies that $\mu = 0$ contradicting our assumption that μ is positive. Hence S must be dense in $C(I_n)$.

Further Issues

- Although this proof shows that a one-layer neural network can approximate any function, it does not say anything about the number of neurons needed to do so.
- In practice, the same number of neurons organized into several layers can approximate a function more accurately and can learn a function in fewer steps than a one-layer network.