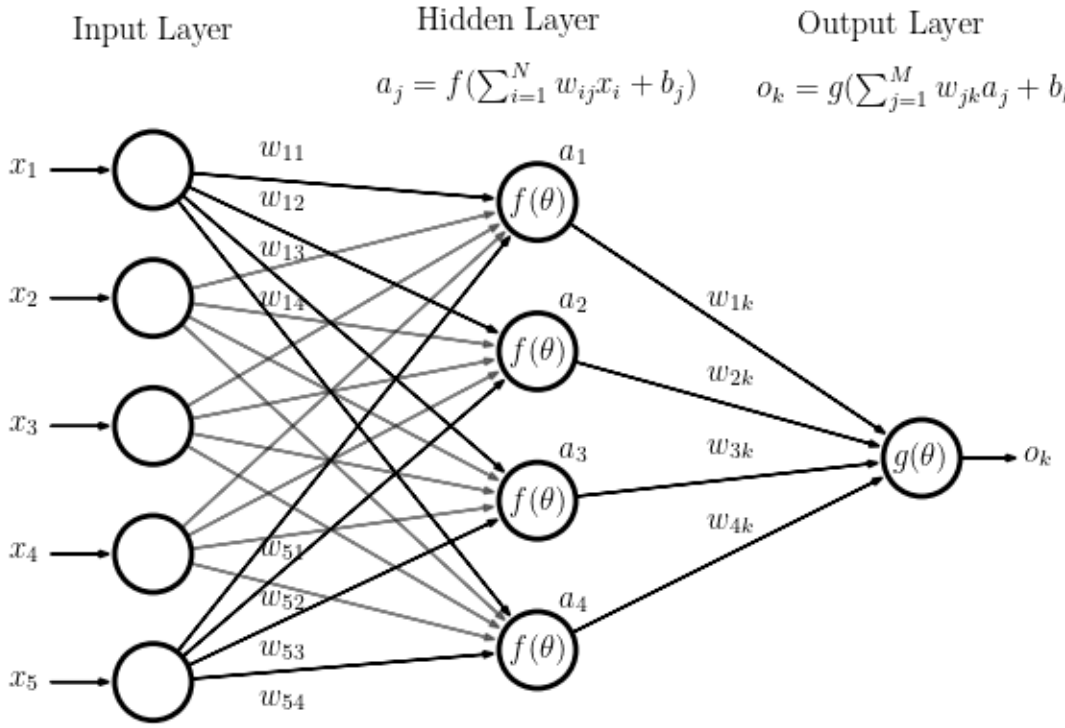
Neural Networks and Universal Approximation of Continuous Functions

William Rasmussen

What is an artifical neural network?

An artifical neural network is a computing system that mimics the way the brain works by using interconnected nodes that act like neurons in the brain, and using algorithms, the neural network "learns" over time. They were first theorized in 1943 by Warren McCulloch and Walter Pitt, but the study of neural networks and their application to artificial intelligence didn't fully take off until Kunihiko Fukushima developed the first multilayered neural network in 1975. Now they are used in many different applications such as facial and speech recognition, computer vision, fraud detection, financial modeling, and much more. Neural networks fall under the umbrella of artificial intelligence, along with machine learning. To briefly explain the differences of between the two, the structure of machine learning is different in that a machine learing model is made to be fed data as an input and using algorithims, it learns over time, and with more data and more time, its output becomes more precise. In the early stages of a machine learning model, it requires some human intervention, whereas a neural network requires less human intervention in the beginning. Where a machine learning model has an input and an output, a neural network is designed to have an input layer with many inputs, followed by at least one hidden layer, and then an output layer, making the structure quite different.



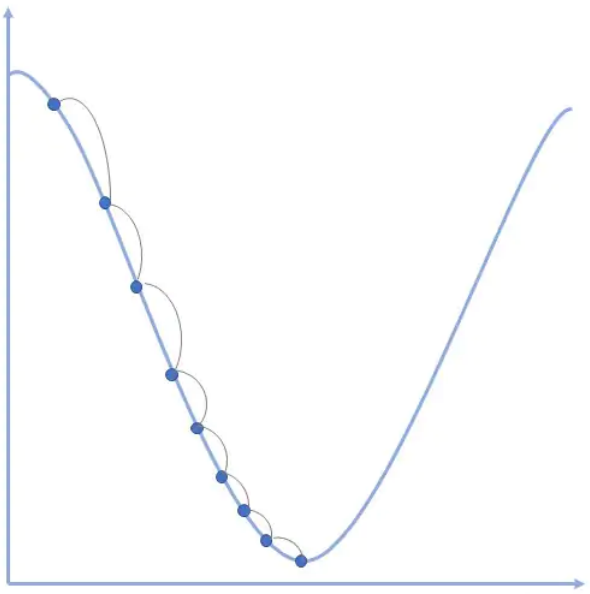
This diagram shows the interconnected nodes, or neurons, that are in the different layers. Data is fed into the input layer, where there is a neuron for each compontent in the input data, and it is then communicated to one or more hidden layers. In the hidden layers, a computation is done with weights and biases (w, b in the diagram respectively), and a weighted sum is calculated. That value is fed into an activation function, and the value of the activation function decides if the neuron should be activated or not, similar to how a brain neuron either fires or doesn't fire based off of an outside stimulus. In this paper, I will be using the sigmoid function as the activation function. The sigmoid function is defined as

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

The sigmoid function is useful because the output is between 0 and 1, so it gives a value that can be interpreted as either fire or don't fire. Sigmoid functions were originally the main activation function used in neural networks and machine learning models, but nowadays there are better activation functions, such as linear, softmax, and ReLU, with ReLU being one of the most widely used. These different activation functions generally have a constant gradient, whereas the sigmoid function's becomes increasingly small in, so more of a gradient allows for quicker learning. After the values are passed through the activation function, the output of the layer is used as the input in the next hidden layer and repeats until it reaches the output layer. In the calculation of the weighted sum, there are weights and biases applied to each. In the beginning stages of a neural network, these values are assigned randomly, but through back propogation, the weights and biases are updated in order to minimize error from expected values. In order to precisely assign weight and bias values, the gradient descent algorithm is used. This is defined as

$$p_{n+1} = p_n - \eta \nabla f(p_n)$$

Where η is called the learning rate and it scales the gradient and controls the step size. This allows for the error to be minimized.



Approximating functions using neural networks

Now we claim that for any continuous function f on a compact set K, there exists a feedforward neural network with a signle hidden layer, that uniformly approximates f to within $\epsilon > 0$ on K.

Definitions: Let I_n be an n^{th} dimensional unit cube. This can be represented by the Cartesian Product $[0,1]^n$. Let $C(I_n)$ be the space of continuous funcctions with codomain of $\mathbb{C}$, on I_n. Let $M(I_n)$ be the space of finite, signed Borel measures on I_n. A measure μ is regular if the following holds

1. $\mu(K) < \infty$ for all compact sets K
2. $\mu(E) = \inf\{\mu(U) : E \subseteq U, \text{where } U \text{ is open}\}$
3. $\mu(E) = \sup\{\mu(K) : K \subseteq E, \text{where } K \text{ is compact}\}$

A single measure gives us a notion of size on a set in that it tells us how much space a set takes up in relation to the larger set of which it is in. All integration we will use is done with respect to regular measures on $C(I_n)$. This becomes useful because the measure of a compact set has finite measure if the measure being applied is regular. In metric space, the Heine-Borel Theorem holds so a set is compact if and only if it is closed and bounded. Using regular measures is also beneficial because it allows for the use of the Riesz Representation Theorem. Note that the uniform norm of a function $f : A \to B$ is

$$||f|| = \sup\{|f(x)| : x \in A\}$$

So we can say that the uniform norm gives an upper bound on all values of f. For two functions f, g, $||f - g||$ gives a uniform bound on the amount that f and g differ from each other. To prove the claim, we want to show that a set of feedforward neural networks is dense in $C(I_n)$ with respect to the uniform norm. A set A is dense in K if $\overline{A} = K$, meaning the closure of A is the entire space K. Since we are working with a metric space, there is a notion of distance between functions, so A is dense in K if for every $x \in K$ there exists a sequence $(a_n)_{n\in\mathbb{N}} \in A$ such that $\lim_{n\to\infty} a_n = x$. Now we can restate the claim that for any continuous function f on a compact set K, there exists a feedforward neural network that uniformly approximates f to within $\epsilon > 0$ on K into three different statements.

1. The set of all feedfoward neural networks $\mathcal{N}$ is dense in $C(I_n)$
2. For every continuous function $f \in C(I_n)$, there exists a sequence of neural networks $(n_j) \in \mathcal{N}$ converging to f, as in $\lim_{j\to\infty} n_j = f$
3. For every continuous function $f \in C(I_n)$ and $\epsilon > 0$, there exists a neural network $g \in \mathcal{N}$ such that $||g - f|| < \epsilon$

Definition: A feedforward neural network having N neurons arranged in a single hidden layer is a function $y : \mathbb{R}^m \to \mathbb{R}$ defined by

$$y(\mathbf{x}) = \sum_{i=1}^{N} \alpha_i \sigma(\mathbf{w}_i^T \mathbf{x} + b_i),\ \mathbf{w}_i, \mathbf{x} \in \mathbb{R}^m,\ \alpha_i, b_i \in \mathbb{R}$$

Here, the $\mathbf{w}_i$ are the weights of the individual neurons and they are applied to the input $\mathbf{x}$. The α_i are the network weights and they are applied to the output of each neuron in the hidden layer. The b_i is the bias of neuron i. Recall that σ is the activation function $\sigma : \mathbb{R} \to \mathbb{R}$ defined as

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

and note that

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

Definition: σ is discriminatory if for $\mu \in M(I_n)$ and

$$\int_{I_n} \sigma(\mathbf{w}_i^T \mathbf{x} + b_i) d\mu(\mathbf{x}) = 0$$

for all $\mathbf{w}_i \in \mathbb{R}^m$, $b_i \in \mathbb{R}$ then $\mu = 0$. This means that for a nonzero μ, there exists $\mathbf{w}_i, b_i$ such that $\int_{I_n} \sigma(\mathbf{w}_i^T \mathbf{x} + b_i) d\mu(\mathbf{x}) \neq 0$, so this ensures that $\mathbf{w}_i \mathbf{x} + b_i$ can't be sent to a set of measure zero.

Theorems

Hahn-Banach Theorem: Let X be a real vector space, p a sublinear functional on X, M a subspace of X, and f a linear functional on M such that $f(x) \leq p(x)$ for all $x \in M$. Then there exists a linear functional F on X such that $F(x) \leq p(x)$ for all $x \in X$, and $F|M = f$. A linear functional is a linear map from a vector space X to A where $A = \mathbb{R}$ or $\mathbb{C}$. The Hahn-Banach Theorem gives a way to extend a bounded linear functional defined on $\mathcal{N}$ to one defined on all of $C(I_n)$. Also, $\int_{I_n}$ acts as a linear functional on $C(I_n)$.

Riesz Representation Theorem: If I is a positive linear functional on $C_c(X)$, there is a unique Radon measure μ on X such that $I(f) = \int f d\mu$ for all $f \in C_c(X)$. Moreover, μ satisfies

1. $\mu(U) = \sup \{I(f) : f \in C_c(X), 0 \leq f \leq 1, \text{supp}(f) \subset U\}$ for all open $U \subset X$ and
2. $\mu(K) = \inf \{I(f) : f \in C_c(X), 0 \leq f \leq 1, r \geq \chi_k\}$ for all compact $K \subset X$

Where $C_c(X)$ is the functions of compact support defined on X, and χ_K is the indicator function of the compact set K. In our case, $C(I_n)$ is σ compact so we don't have to deal with Radon measures, regular Borel measures work.

Lemma: Any bounded, measurable sigmoidal function σ is discriminatory. In particular, any continous sigmoidal funtion is discriminatory.

Theorem: If the σ in the neural network is a continuous, discriminatory function, then the set of all neural networks is dense in $C(I_n)$.

Proof: Let $\mathcal{N} \subset C(I_n)$ be the set of neural networks. $\mathcal{N}$ is a linear subspace of $C(I_n)$. Suppose $\overline{\mathcal{N}} \neq C(I_n)$, so $\overline{\mathcal{N}}$ is a closed, proper subspace of $C(I_n)$. By the Hahn-Banach Theorem, there exists a bounded linear functional on $C(I_n), L$ such that $L(\mathcal{N}) = L(\overline{\mathcal{N}}) = 0$, but $L \neq 0$. Here, $\int_{I_n}$ acts as a linear functional on $C(I_n)$. By the Riesz Representation Theorem, we can write the functional L as

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

for some $\mu \in M(I_n)$, for all $h \in C(I_n)$. By definition, any neural network is an element of $\mathcal{N}$, and L is identically zero on $\mathcal{N}$, so we have that

$$\int_{I_n} h(x) d\mu(x) = 0$$

σ is discriminatory, so μ must be equal to 0, which contradicts $L \neq 0$ because $\mu = 0$ implies that

$$\int_{I_n} h(x) d\mu(x) = 0 \text{ for all } h \in C(I_n)$$

Therefore $\mathcal{N}$ is dense in $C(I_n)$.

□