"Multilayer feedforward networks are universal approximators"

Kur Hornik, Maxwell Stinchcombe and Halber White
(1989)

Presenter: Sonia Todorova

Theoretical properties of multilayer feedforward networks

- universal approximators: standard multilayer feedforward networks are capable of approximating any measurable function to any desired degree of accuracy
- there are no theoretical constraints for the success of feedforward networks
- lack of success is due to inadequate learning, insufficient number of hidden units or the lack of a deterministic relationship between input and target

* rate of convergence as the number of hidden units grows
* rate of increase of the number of hidden units as the input dimension increases for a fixed accuracy

Main Theorems: Cybenko, 1989

Sums of the form

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

where $y_j \in R^n, \alpha_j, \theta_j \in R$, are dense in the space of continuous functions on the unit cube if σ is any continuous sigmoidal function.

Main Theorems: Hornik, 1989

English

Single hidden layer $\Sigma\Pi$ feedforward networks can approximate any measurable function arbitrarily well regardless of the activation function Ψ, the dimension of the input space r, and the input space environment μ.

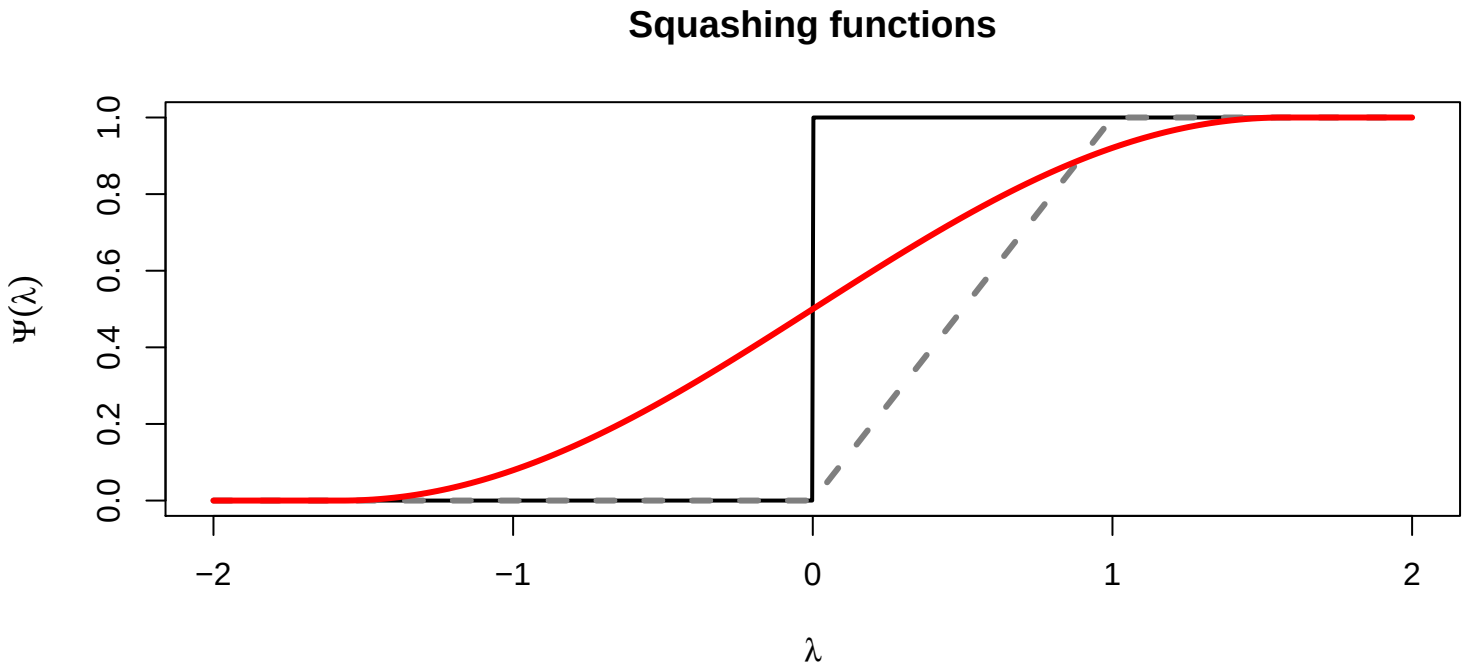
Math

For every squashing function Ψ, every r and every probability measure μ on (R^r, B^r), both $\Sigma\Pi^r(\Psi)$ and $\Sigma^r(\Psi)$ are uniformly dense on compacta in C^r and ρ_μ-dense in M^r

$$\sum_{j=1}^{q} \beta_j G(A_j(x)), x \in R^r, \beta_j \in R, A_j \in A^r, q \in N$$

$$\sum_{j=1}^{q} \beta_j \Pi_{k=1}^{l_j} G(A_{jk}(x)), x \in R^r, \beta_j \in R, A_{jk} \in A^r, l_j \in N$$

Def: A function $\Psi : R \longrightarrow [0, 1]$ is a **squashing function** if it is non-decreasing, $\lim_{\lambda \to \infty} \Psi(\lambda) = 1$ and $\lim_{\lambda \to -\infty} \Psi(\lambda) = 0$



indicator function,
ramp function,
cosine squasher (Gallant and White, 1988)

Main Theorems

English

There is a single hidden layer feedforward network that approximates any measurable function to any desired degree of accuracy on some compact set K.

Math

For every function g in M^r there is a compact subset K of R^r and an $f \in \sum^r(\Psi)$ such that for any $\epsilon > 0$ we have $\mu(K) < 1 - \epsilon$ and for every $X \in K$ we have $|f(x) - g(x)| < \epsilon$, regardless of Ψ, r, or μ.

Main Theorems

English
Functions with finite support can be approximated exactly with a single hidden layer.
Math
Let $\{x_1, ..., x_n\}$ be a set of distinct points in R^r and let $g : R^r \longrightarrow R$ be an arbitrary function. If Ψ achieves 0 and 1, then there is a function $f \in \Sigma^r(\Psi)$ with n hidden units such that $f(x_i) = g(x_i)$ for all i.

Main Theorems

Most results stated for a single-dimensional output but can be extended to multi-output networks.

Proofs

Stone-Weierstrass Theorem

Let A be an algebra of real continuous functions on a compact set K. If A separates points in K and if A vanishes at no point of K, then the uniform closure B of A consists of all real continuous functions on K (i.e. A is ρ_K-dense in the space of real continuous functions on K)

Namely, pick a set $\{f_1, f_2, ...\}$ of real continuous functions defined on a compact set K. Add all $f_j + f_k$, $f_j \cdot f_k$, and cf_j to form A. If for any $x \neq y$ there exists an $f \in A$ such that $f(x) \neq f(y)$, and for every x there exists an $f \in A$ such that $f(x) \neq 0$, then for any given real continuous function on K, A contains a function which approximates it arbitrarily well. The distance is measured as $\rho_K(f, g) = \sup_{x \in K} |f(x) - g(x)|$

See: *An Elementary Proof of the Stone-Weierstrass Theorem"*, Brosowski and Deutsch (1981)

Proof of Theorem 2.1

Let $K \subset R^r$ be any compact set.

For any G, $\Sigma\Pi^r(G)$ is an algebra on K, because any sum and product of two elements is in the same form, as are scalar multiples.

To show that $\Sigma\Pi^r(G)$ separates points on K, for any $x \neq y$, pick A such that $A(x) = a$, $A(y) = b$, for two constants $a \neq b$ such that $G(a) \neq G(b)$. Then, $G(A(x)) \neq G(A(y))$. This is why it is important that G be non-constant.

To show that $\Sigma\Pi^r(G)$ vanishes nowhere on K, pick $b \in R$ such that $G(b) \neq 0$ and let $A(x) = 0 * x + b$. For all $x \in K$, $G(A(x)) = G(b)$.

By the Stone-Weierstrass Theorem, $\Sigma\Pi^r(G)$ is ρ_K-dense in the space of real continuous functions on K

Lemma A1:
For any finite measure μ C^r is ρ_μ-dense in M^r

Recall: $\rho_\mu(f,g) = inf\{\epsilon > 0 : \mu\{x : |f(x) - f(y)| > \epsilon\} < \epsilon\}$

Proof Pick an arbitrary $f \in M^r$ and $\epsilon > 0$. Need to show that there is a $g \in C^r$ such that $\rho_\mu(f,g) < \epsilon$. For a sufficiently large number D, $\int min\{|f \cdot 1_{\{|f|<D\}} - f|, 1\}d\mu < \epsilon/2$

There exists a continuous g such that

$\int |f \cdot 1_{\{|f|<D\}} - g|d\mu < \epsilon/2$. Thus, $\int min\{|f-g|, 1\}d\mu < \epsilon$.

Theorem 2.5

Exact approximation of functions on a finite set in R^1
Let $\{x_1, ..., x_n\}$ be a set of distinct points in R^r and let $g : R^r \longrightarrow R$ be an arbitrary function. If Ψ achieves 0 and 1, then there is a function $f \in \Sigma^r(\Psi)$ with n hidden units such that $f(x_i) = g(x_i)$ for all i.

Proof

Let $\{x_1, x_2, ..., x_n\} \subset R^1$ and relabeling so that $x_1 < x_2 < ... < x_n$.
Pick $M > 0$ such that $\Psi(-M) = 1 - \Psi(M) = 0$.
Define A_1 as the constant affine functions $A_1 = M$ and set $\beta_1 = g(x_1)$.
Set $f^1(x) = \beta_1 \cdot \Psi(A_1(x))$.
Inductively define A_k by $A_k(x_{k-1}) = -M$ and $A_k(x_k) = M$.
Define $\beta_k = g(x_k) - g(x_{k-1})$.
Set $f^k(x) = \sum_{h=1}^{k} \beta_j \Psi(A_j(x))$. For $i \leq k f^k(x_i) = g_k(x_i)$.
The desired function is f^n.

Universal approximation bounds for superpositions of sigmoidal functions (Barron, 1993)

"For an artificial neural network with one layer of n sigmoidal nodes, the integrated squared error of approximation, integrating on a bounded subset of d variables, is bounded by c_f/n, where c_f depends on a norm of the Fourier transform of the function being approximated. This rate of approximation is achieved under growth constraints on the magnitudes of the parameters of the network. The optimization of a network to achieve these bounds may proceed one node at a time. Because of the economy of number of parameters, order nd instead of n^d, these approximation rates permit satisfactory estimators of functions using artificial neural networks even in moderately high-dimensional problems."