Nonparametric Estimation Using Neural Networks

Gábor Lugosi[1] Kenneth Zeger[2]

[1]Dept. of Mathematics, Faculty of Elect. Eng., Technical University of Budapest, 1521 Stoczek u. 2, Budapest, Hungary.
[2]Coordinated Science Laboratory, Dept. of Elect. and Comp. Engineering, University of Illinois, Urbana, IL 61801

Abstract — **We show that properly trained neural networks provide universally consistent nonparametric estimators. The results apply to regression estimation, conditional median estimation, curve fitting, pattern recognition and learning concepts. The estimators minimize the empirical L_p-error.**

Let the random variables X and Y take their values from $\mathbb{R}^d$ and $\mathbb{R}$, respectively. Denote the error of the L_p-optimal predictor by

$$J^* = \inf_m (\mathbf{E}|m(X) - Y|^p)^{1/p} = (\mathbf{E}|m^*(X) - Y|^p)^{1/p}.$$

Assume that we do not know anything about the distribution of the pair (X, Y), but a collection of independent, identically distributed copies $D_n = ((X_1, Y_1), \ldots, (X_n, Y_n))$ of (X, Y) is available, where D_n is independent of (X, Y). Our aim is to estimate good predictors from the data, that is, to construct a function $m_n(x) = m_n(x, D_n)$ such that its L_p-error $J(m_n) = \left(\mathbf{E}\left(|m_n(X) - Y|^p \middle| D_n\right)\right)^{1/p}$ is close to the optimum.

Definition 1 *A sequence of estimators $\{m_n\}$ is* consistent *for a given distribution of (X, Y), if $J(m_n) - J^* \to 0$ almost surely as $n \to \infty$. $\{m_n\}$ is* universally consistent *if it is consistent for* any *distribution of (X, Y) satisfying $\mathbf{E}|Y|^p < \infty$.*

The main results of the paper point out that properly trained neural networks provide estimators that are universally consistent, extending results by White [4], Haussler [2], and Faragó and Lugosi [1]. These estimators are based on empirical risk minimization.

Our method of constructing an estimator m_n is to choose it as a function from a class of functions $\mathcal{F}$ that minimizes the *empirical error* $J_n(f) = \left(\frac{1}{n}\sum_{j=1}^n |f(X_j) - Y_j|^p\right)^{1/p}$.

Formally, let $\{\mathcal{F}_n\}$ be a sequence of classes of functions, and define m_n as a function in $\mathcal{F}_n$ that minimizes the empirical error: $J_n(m_n) \le J_n(f)$ for $f \in \mathcal{F}_n$. For analyzing how close the error of the estimator $J(m_n)$ is to the optimum J^*, we will use the following decomposition:

$$J(m_n) - J^* = \left(J(m_n) - \inf_{f\in\mathcal{F}_n} J(f)\right) + \left(\inf_{f\in\mathcal{F}_n} J(f) - J^*\right).$$

The first term on the right hand side tells us about the "learnability" of $\mathcal{F}_n$, that is, how well the empirical minimization performs over this class. We will refer to this term as the *estimation error*. The second term, which we call the *approximation error* describes how rich the class $\mathcal{F}_n$ is, that is, how well the best function in the class performs. Here the main problem is to balance the trade-off between the approximation potential and the estimability of the class, that is, to determine, how fast the class should grow to get universally consistent estimators, if it is possible at all.

[1]Supported in part by NSF Grant No. NCR-92-96231.

A *neural network of one hidden layer with k hidden neurons* is a real function on $\mathbb{R}^d$ of the form

$$f_{\theta_k}(x) = \sum_{i=1}^k c_i\sigma(a_i^T x + b_i) + c_0,$$

where the *sigmoid* $\sigma : \mathbb{R} \to [0, 1]$ is a monotone nondecreasing function converging to 0 as $x \to -\infty$ and 1 as $x \to \infty$. $\theta_k = \{a_1, \ldots, a_k \in \mathbb{R}^d; b_1, \ldots, b_k, c_0, \ldots, c_k \in \mathbb{R}\}$. is the set of parameters that specify the network.

To handle the approximation error, we use a denseness theorem for feedforward neural networks, proved by Hornik [3]. To prove the convergence of the estimation error, we apply techniques based on the rich theory of empirical processes. Some difficulties arise in handling nonbounded random variables. We have the following consistency result.

Theorem 1 *Define a series of classes of neural networks $\mathcal{F}_1, \mathcal{F}_2, \ldots$ as*

$$\mathcal{F}_n = \left\{\sum_{i=1}^{k_n} c_i\sigma(a_i^T x + b_i) + c_0;\ a_i \in \mathbb{R}^d,\ b_i \in \mathbb{R},\ \sum_{i=1}^{k_n} |c_i|^p \le \beta_n^p\right\}$$

and let m_n minimize the empirical L_p-error over $\mathcal{F}_n$, i.e.

$$\frac{1}{n}\sum_{i=1}^n |m_n(X_i) - Y_i|^p \le \frac{1}{n}\sum_{i=1}^n |f(X_i) - Y_i|^p \quad \text{if } f \in \mathcal{F}_n.$$

Then if k_n and β_n satisfy $k_n \to \infty$, $\beta_n \to \infty$, and $\frac{1}{n}k_n^{2p-1}\beta_n^{2p}\log(k_n\beta_n) \to 0$, then $J(m_n) - J^* \to 0$ in probability, for all distributions of (X, Y). If there exists a $\delta > 0$ such that $k_n^{2p-2}\beta_n^{2p}/n^{1-\delta} \to 0$, then $J(m_n) - J^* \to 0$ a.s., that is, the estimate m_n is universally consistent.

References

[1] A. Faragó and G. Lugosi. Strong universal consistency of neural network classifiers. *IEEE Transactions on Information Theory*, vol.39, no.4, July, 1993.

[2] D. Haussler. Decision theoretic generalizations of the PAC model for neural net and other learning applications. *Information and Computation*, 100:78–150, 1992.

[3] K. Hornik. Approximation capabilities of multilayer feedforward networks. *Neural Networks*, 4:251–257, 1991.

[4] H. White. Connectionist nonparametric regression: multilayer feedforward networks can learn arbitrary mappings. *Neural Networks*, 3:535–549, 1990.

0 - 7803-2015-8/94/$4.00 ©1994 IEEE