Comment

Andrew R. Barron

Relationships between topics in statistics and artificial neural networks are clarified by Cheng and Titterington. There are fruitful concepts in artificial neural networks that are worthwhile for the statistical community to absorb. These networks provide a rich collection of statistical models, some of which are ripe for both mathematical analysis and practical applications. Many aspects of artificial neural networks are in need of further investigation. Here, I comment on approximation and computation issues and their impact on statistical estimation of functions.

APPROXIMATION

Attention is focussed on the most commonly studied feedforward networks (perceptrons) which have one or two "hidden" layers defined by composition of units of the form $\phi(wx + w_0)$, where ϕ is a hardlimiter or sigmoidal activation function and w_0, w denote the parameters (internal weights) that adjust the orientation, location and scale of the unit functions (Rosenblatt, 1962; Rumelhart, Hinton and Williams, 1986a). In the one hidden layer case, a linear combination of such units is taken with the internal weights adjusted so that the result approximates a target function. These networks may be regarded as an adjustable basis function expansion of ridge form similar to projection pursuit (Friedman and Stuetzle, 1981) and similar to sparse trigonometric series with adjustable frequency vectors. Linear combination of such adjustable basis functions can provide an accurate approximation with far fewer units than by linear combination of any fixed basis functions for certain classes of target functions when the number of input variables is greater than or equal to three (Barron, 1993). A consequence is that more accurate statistical function estimation is possible for such target functions (Barron, 1994).

These conclusions for one hidden layer networks are based, in part, on the following result developed in Jones (1992) and Barron (1993). Suppose a function $f(x)$ is such that $f(x)/V$ is in the closure of the convex hull of the set of units $\{\pm\phi(wx + w_0) : (w_0, w) \in R^{d+1}\}$, where ϕ is bounded by 1 for some positive number V. The closure is in the $L_2(\mu)$ norm, where μ is any given probability measure μ with bounded support on R^d. Then there are M such units with choices of weights depending on f and μ, such that their linear combination $f_M(x)$ (a single hidden layer network) achieves approximation error

$$||f - f_M|| \leq \frac{V}{\sqrt{M}},$$

where the norm is taken in $L_2(\mu)$. The surprising aspect is that the approximation rate as a function of M is independent of the dimension d. A subclass of functions that satisfy the condition are those that possess a bound on the first moment of the Fourier magnitude distribution. (This class includes all smooth positive definite functions and convex combinations of translates of such functions.) In contrast, approximation using any fixed M basis functions cannot achieve approximation error uniformly better than order $1/M^{1/d}$ for the same class of functions f, taking μ to be the uniform distribution on a d-cube (Barron, 1993).

It is of interest to characterize what classes of functions can be more parsimoniously approximated using two rather than one hidden layer in the network. Some functions such as the indicator of a cube or a ball are not accurately approximated by the ridge expansions represented by one-layer networks without resorting to a number of units exponentially large in the dimension. In these cases the network capabilities may be improved by inclusion of a second layer of threshold nonlinearities. Units on the second layer can provide indicators of the level sets of linear combinations of the first layer units. These level sets can be arranged to take arbitrary polygon shapes (Lippman, 1987). The linear combination of the outputs of the second layer then give piecewise constant approximations of a rather general form. One conclusion of the same flavor as above is that if a function f is such that $f(x)/V$ is in the closure of the convex hull of the set of signed indicators of K-sided polygons for some positive V, then there is a two hidden layer network function $f_{K,M}(x)$ with KM units on the first layer and M units on the second

Andrew R. Barron is Professor, Department of Statistics, Yale University, Box 2179, Yale Station, New Haven, Connecticut 06520.