NOTE Communicated by Eric Baum

Sigmoids Distinguish More Efficiently Than Heavisides

Eduardo D. Sontag
SYCON—Rutgers Center for Systems and Control, Department of Mathematics, Rutgers University, New Brunswick, NJ 08903 USA

Every dichotomy on a $2k$-point set in $\mathbb{R}^N$ can be implemented by a neural net with a single hidden layer containing k sigmoidal neurons. If the neurons were of a hardlimiter (Heaviside) type, $2k-1$ would be in general needed.

1 Introduction and Definitions

The main point of this note is to draw attention to the fact mentioned in the title, that sigmoids have different recognition capabilities than hard-limiting nonlinearities. One way to exhibit this difference is through a worst-case analysis in the context of binary classification, and this is done here. Results can also be obtained in terms of VC dimension, and work is in progress in that regard. For technical details and proofs, the reader is referred to Sontag (1989).

Let N be a positive integer. A *dichotomy* (S_-, S_+) on a set $S \subseteq \mathbb{R}^N$ is a partition $S = S_- \bigcup S_+$ of S into two disjoint subsets. A function $f : \mathbb{R}^N \to \mathbb{R}$ will be said to *implement* this dichotomy if it holds that

$$f(u) > 0 \text{ for } u \in S_+ \text{ and } f(u) < 0 \text{ for } u \in S_-$$

Let $\theta : \mathbb{R} \to \mathbb{R}$ be any function. We shall say that f is a *single hidden layer neural net with k hidden neurons of type θ* [or just that f is a "(k, θ)-net"] if there are real numbers $w_0, w_1, \ldots, w_k, \tau_1, \ldots, \tau_k$, and vectors $v_1, \ldots, v_k \in \mathbb{R}^N$ such that, for all $u \in \mathbb{R}^N$,

$$f(u) = w_0 + \sum_{i=1}^{k} w_i \theta(v_i.u - \tau_i) \tag{1.1}$$

where the dot indicates inner product.

For fixed θ, and under mild assumptions on θ, such neural nets can be used to approximate uniformly arbitrary continuous functions on compacts. See, for instance, Cybenko (1989) and Hornik *et al.* (1989). In particular, they can be used to implement arbitrary dichotomies.

In neural net practice, one often takes θ to be the *sigmoid*

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

or equivalently, up to translations and change of coordinates, the hyperbolic tangent $\theta(x) = \tanh(x)$. Another usual choice is the hardlimiter or *Heaviside* function

$$\theta(x) = \mathcal{H}(x) = \begin{cases} -1 & \text{if } x \leq 0 \\ 1 & \text{if } x > 0 \end{cases}$$

which can be approximated well by $\tanh(\gamma x)$ when the "gain" γ is large. Most analysis has been done for $\mathcal{H}$, but backpropagation techniques typically use the sigmoid (or equivalently tanh).

It is easy to see that arbitrary dichotomies on an l-element set can be implemented by $(l-1, \mathcal{H})$-nets, but that some dichotomies on sets of l elements cannot be implemented by nets with less than $l-1$ Heaviside hidden neurons.

We consider functions $\theta : \mathbf{R} \to \mathbf{R}$ that satisfy the following two properties:

(S1) $t_+ := \lim_{x \to +\infty} \theta(x)$ and $t_- := \lim_{x \to -\infty} \theta(x)$ exist, and $t_+ \neq t_-$.

(S2) There is some point c such that θ is differentiable at c and $\theta'(c) = \mu \neq 0$.

Note that the function $\mathcal{H}$ does not satisfy (S2), but the sigmoid of course does. The main result will be stated for these.

2 Main Result and Remarks

Theorem 1. *Let θ satisfy (S1) and (S2), and let S be any set of cardinality $l = 2k$. Then, any dichotomy on S can be implemented by some (k, θ)-net.*

Thus, using sigmoids we can reduce the number of neurons from $2k-1$ to k, a factor of 2 improvement. Of course this result should not really be surprising, since for Heaviside functions there are fewer free degrees of freedom [because $\mathcal{H}(\gamma x) = \mathcal{H}(x)$ for any $\gamma > 0$], and in fact its proof is very simple. The idea is to first classify using a net with $k-1$ Heaviside hidden neurons plus a direct connection from the inputs to the output, and then replacing these direct connections by just one nonlinear hidden neuron. The differentiablity assumption allows this replacement, since it means that at low gains any linear map can be approximated.

To conclude this note, we wish to remark that there are "universal" functions θ satisfying (S1)–(S2) and as differentiable as wanted, even real-analytic, such that, for each N and each dichotomy on any finite set $S \subseteq \mathbf{R}^N$, this dichotomy can be implemented by a $(1, \theta)$-net. Of course, the function θ is so complicated as to be purely of theoretical interest, but it serves to indicate that, unless further restrictions are made on (S1)–(S2), much better bounds can be obtained.

Acknowledgments

Supported in part by U.S. Air Force Grant AFOSR-88-0235.

References

Cybenko, G. 1989. Approximation by superpositions of a sigmoidal function. *Math. Control, Signals, Syst.* **2**, 303–314.

Hornik, K. M., Stinchcombe, M., and White, H. 1989. Multilayer feedforward networks are universal approximators. *Neural Networks* **2(5)**: 359–366.

Sontag, E. D. 1989. Sigmoids distinguish more efficiently than Heavisides. Report 89-12, *SYCON—Rutgers Center for Systems and Control,* August 1989. (Electronic versions available from sycon@fermat.rutgers.edu.)

Received 28 July 1989; accepted 15 September 1989.