The VC-Dimension and Pseudodimension of Two-Layer Neural Networks with Discrete Inputs

Peter L. Bartlett

Robert C. Williamson

Department of Systems Engineering

Research School of Information Sciences and Engineering

Australian National University

Canberra 0200

AUSTRALIA

March 24, 1994

Revised: September 16, 1994, July 28, 1995

Abstract

We give upper bounds on the Vapnik-Chervonenkis dimension and pseudodimension of two-layer neural networks that use the standard sigmoid function or radial basis function and have inputs from $\{-D, \ldots, D\}^n$. In Valiant's probably approximately correct (pac) learning framework for pattern classification, and in Haussler's generalization of this framework to nonlinear regression, the results imply that the number of training examples necessary for satisfactory learning performance grows no more rapidly than $W \log(WD)$, where W is the number of weights. The previous best bound for these networks was $O(W^4)$.

In using neural networks for pattern classification and regression tasks, it is important to be able to predict how much training data will be sufficient for satisfactory performance. Valiant's probably approximately correct (pac) framework [Val84] provides a formal definition of 'satisfactory learning performance' for $\{0,1\}$-valued functions. In [BEHW89], Blumer *et al.* present upper and lower bounds on the number of examples necessary and

sufficient for learning under this definition. These bounds depend linearly on the *Vapnik-Chervonenkis dimension* of the function class used for learning. In [Hau92], Haussler presents a generalization of the pac framework that applies to the problem of learning real-valued functions. In this case, the *pseudodimension* of the function class gives an upper bound on the number of examples necessary for learning.

References [BH89, Maa92, Sak93, Bar93, GJ93] present upper and lower bounds on the VC-dimension of threshold networks and networks with piecewise polynomial output functions. Most of these results can easily be extended to give bounds on the pseudodimension. However, these networks are not commonly used in applications because the most popular learning algorithm, the backpropagation algorithm, relies on differentiability of the units' output functions. In practice, sigmoid networks and radial basis function (RBF) networks are most widely used. Recently, Macintyre and Sontag [MS93] showed that the VC-dimension and pseudodimension of these networks are finite, and Karpinski and Macintyre [KM95] proved a bound of $O(W^2N^2)$, where W is the number of weights in the network and N is the number of processing units. In this note, we show that the VC-dimension and pseudodimension of two-layer sigmoid networks and radial basis function networks with discrete inputs is $O(W \log(WD))$, where the input domain is $\{-D, \ldots, D\}^n$. In many pattern classification and nonlinear regression tasks for which neural networks have been used, the set of inputs is a small finite set of this form. For the special case of sigmoid networks with binary inputs, our bound is within a log factor of the best known lower bounds [Bar93]. The result follows from the observation that a network with discrete inputs can be represented as a polynomially parametrized function class; hence VC-dimension bounds for such classes can be applied.

Suppose X is a set, and F is a class of real-valued functions defined on X. For $x = (x_1, \ldots, x_m) \in X^m$ and $r = (r_1, \ldots, r_m) \in \mathbb{R}^m$, we say that F shatters $(x_1, r_1, \ldots, x_m, r_m)$ if for all sign sequences $s = (s_1, \ldots, s_m) \in \{-1, 1\}^m$, there is a function f in F such that $s_i(f(x_i) - r_i) > 0$ for $i = 1, \ldots, m$. The pseudodimension of F is the length of the largest shattered sequence.

For pattern classification problems, we typically consider a class of $\{0, 1\}$-valued functions obtained by thresholding a class of real-valued functions. Define the threshold function, $\mathcal{H} : \mathbb{R} \to \{0, 1\}$, as $\mathcal{H}(\alpha) = 1$ if and only if $\alpha \geq 0$. If F is a class of real-valued functions, let $\mathcal{H}(F)$ denote the set $\{\mathcal{H}(f) : f \in F\}$.

The Vapnik-Chervonenkis dimension of a class F of real-valued functions defined on X is the size of the largest sequence of points that can be classified arbitrarily by $\mathcal{H}(F)$,

$$\mathrm{VCdim}(F) = \max\{m : \exists x \in X^m,\ \mathcal{H}(F) \text{ shatters } (x_1, 1/2, \ldots, x_m, 1/2)\}.$$

Clearly, $\mathrm{VCdim}(F) \leq \dim_P(F)$.

The function classes considered in this note can be indexed using a real vector θ of parameters. Let Θ and X be the parameter and input spaces respectively, and let $f : \Theta \times X \to \mathbb{R}$. The function f defines a parametrized class of real-valued functions defined on X, $\{f(\theta, \cdot) : \theta \in \Theta\}$. We also denote this function class by f.

Definition 1 *A two layer sigmoid network with n inputs, W weights, and a single real-valued output is described by the function $f_S : \mathbb{R}^W \times X \to \mathbb{R}$, where $X \subseteq \mathbb{R}^n$,*

$$f_S(\theta, x) = a_0 + \sum_{i=1}^{k} a_i/(1 + e^{-(b_i \cdot x + b_{i0})}),$$

with $a_i \in \mathbb{R}$, $b_i = (b_{i1}, \ldots, b_{in}) \in \mathbb{R}^n$, and $\theta = (a_0, \ldots, a_k, b_{10}, \ldots, b_{kn}) \in \mathbb{R}^W$. (For $x, y \in \mathbb{R}^n$, $x \cdot y = \sum_{i=1}^n x_i y_i$.) In this case, $W = kn + 2k + 1$.

A radial basis function (RBF) network is described by the function

$$f_{RBF}(\theta, x) = a_0 + \sum_{i=1}^{k} a_i e^{-\|x - c_i\|^2},$$

with $a_i \in \mathbb{R}$, $c_i = (c_{i1}, \ldots, c_{in}) \in \mathbb{R}^n$, and $\theta = (a_0, \ldots, a_k, c_{11}, \ldots, c_{kn}) \in \mathbb{R}^W$. (If $x \in \mathbb{R}^n$, $\|x\|^2 = x \cdot x$.) Here, $W = kn + k + 1$.

Theorem 2 *Let $X = \{-D, \ldots, D\}^n$ for some positive integer D. For the sigmoid and RBF networks $f_S, f_{RBF} : \mathbb{R}^W \times X \to \mathbb{R}$, we have*

$$\begin{aligned} \mathrm{VCdim}(f_S) &\leq \dim_P(f_S) < 2W \log_2(24eWD), \\ \mathrm{VCdim}(f_{RBF}) &\leq \dim_P(f_{RBF}) < 4W \log_2(24eWD). \end{aligned}$$

The proof of Theorem 2 follows from the simple observation that the function classes f_S and f_{RBF} can be expressed as a polynomial in some transformed set of parameters when the inputs are integers. We can then use an upper bound from [GJ93] on the VC-dimension of such a function class.

Proof Consider the function f_S defined in Definition 1. For any $\theta \in \Theta$, $x \in X$ and $r \in \mathbb{R}$, let

$$\begin{aligned} f_{S'}(\theta,(x,r)) &= (f_S(\theta,x)-r)\left(\prod_{i=1}^{k}\prod_{j=1}^{n} e^{-b_{ij}D}\right)\left(\prod_{i=1}^{k}\left(1+e^{-(b_i\cdot x+b_{i0})}\right)\right) \\ &= (a_0-r)\prod_{i=1}^{k}\left(\prod_{j=1}^{n} e^{-b_{ij}D}+e^{-b_{i0}}\prod_{j=1}^{n} e^{-b_{ij}(x_j+D)}\right)+ \\ &\quad \sum_{i=1}^{k} a_i\left(\prod_{j=1}^{n} e^{-b_{ij}D}\right)\prod_{\substack{h=1\\h\neq i}}^{k}\left(\prod_{j=1}^{n} e^{-b_{hj}D}+e^{-b_{h0}}\prod_{j=1}^{n} e^{-b_{hj}(x_j+D)}\right). \end{aligned}$$

Clearly, $f_{S'}(\theta,(x,r))$ always has the same sign as $f_S(\theta,x)-r$, since the denominators in $f_S(\theta,x)$ are always positive, so $\dim_P(f_S) \leq \mathrm{VCdim}(f_{S'})$. But $f_{S'}(\theta,(x,r))$ is polynomial in $\theta' = (a_0,\ldots,a_k,e^{-b_{10}},\ldots,e^{-b_{kn}})$, with degree no more than $2Dnk+k+Dn+1 < 3DW$. Theorem 2.2 in [GJ93] implies that $\mathrm{VCdim}(f_{S'}) < 2W\log_2(24eWD)$.

Similarly, $f_{RBF}(\theta,x)-r$ has the same sign as $f_{RBF'}(\theta,(x,r))$, where

$$\begin{aligned} f_{RBF'}(\theta,(x,r)) &= (f_{RBF}(\theta,x)-r)\prod_{i=1}^{k}\prod_{j=1}^{n} e^{2c_{ij}D} \\ &= (a_0-r)\prod_{i=1}^{k}\prod_{j=1}^{n} e^{2c_{ij}D}+\sum_{i=1}^{k} a_i\left(\prod_{\substack{h=1\\h\neq i}}^{k}\prod_{j=1}^{n} e^{2c_{hj}D}\right)\left(\prod_{j=1}^{n} e^{-x_j^2}e^{2c_{ij}(x_j+D)}e^{-c_{ij}^2}\right). \end{aligned}$$

Again, $f_{RBF'}(\theta,(x,r))$ is polynomial in

$$\theta' = (a_0,\ldots,a_k,e^{2c_{11}},\ldots,e^{2c_{kn}},e^{-c_{11}^2},\ldots,e^{-c_{kn}^2}),$$

with degree no more than $nD(k+2)+2 < 3DW$. As above, $\dim_P(f_S) < 4W\log_2(24eDW)$. □

The same argument can be used to show that two-layer sigmoid or RBF networks with W weights, $\{-D,\ldots,D\}$-valued inputs, and arbitrary connectivity have pseudodimension (and hence VC-dimension) $O(W\log(WD))$. From the results in [KM95], it is clear that the dependence on D, the size of the input set, is not necessary in these upper bounds.

Acknowledgements

This research was supported by the Australian Telecommunications and Electronics Research Board and by the Australian Research Council. Thanks to Sridevan Parameswaran for help in obtaining references.

References

[Bar93] P. L. Bartlett. Vapnik-Chervonenkis dimension bounds for two- and three-layer networks. *Neural Computation*, 5(3):353–355, 1993.

[BEHW89] A. Blumer, A. Ehrenfeucht, D. Haussler, and M. K. Warmuth. Learnability and the Vapnik-Chervonenkis dimension. *Journal of the Association for Computing Machinery*, 36(4):929–965, 1989.

[BH89] E. B. Baum and D. Haussler. What size net gives valid generalization? *Neural Computation*, 1:151–160, 1989.

[GJ93] P. Goldberg and M. Jerrum. Bounding the Vapnik-Chervonenkis dimension of concept classes parametrized by real numbers. In *Proceedings of the Sixth ACM Workshop on Computational Learning Theory*, pages 361–369, 1993.

[Hau92] D. Haussler. Decision theoretic generalizations of the PAC model for neural net and other learning applications. *Information and Computation*, 100:78–150, 1992.

[KM95] M. Karpinski and A. Macintyre. Quadratic VC-dimension bounds for sigmoidal networks. In *Proceedings of the 27th Annual Symposium on the Theory of Computing*, 1995.

[Maa92] W. Maass. Bounds for the computational power and learning complexity of analog neural nets. Technical report, Graz University of Technology, 1992.

[MS93] A. Macintyre and E. D. Sontag. Finiteness results for sigmoidal 'neural' networks. In *Proceedings of the 25th Annual Symposium on the Theory of Computing*, 1993.

[Sak93] A. Sakurai. Tighter bounds on the VC-dimension of three-layer networks. In *World Congress on Neural Networks*, 1993.

[Val84] L. G. Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1143, 1984.