On the Vapnik - Chervonenkis dimension of the Ising perceptron

1996 J. Phys. A: Math. Gen. 29 L199

(http://iopscience.iop.org/0305-4470/29/8/004)

J. Phys. A: Math. Gen. **29** (1996) L199–L204. Printed in the UK

LETTER TO THE EDITOR

On the Vapnik–Chervonenkis dimension of the Ising perceptron

S Mertens†

Institut für Theoretische Physik, Otto-von-Guericke Universität, Postfach 4120, D-39016 Magdeburg, Germany

Received 26 February 1996

Abstract. The Vapnik–Chervonenkis (VC) dimension of the Ising perceptron with binary patterns is calculated by numerical enumerations for system sizes $N \leqslant 31$. It is significantly larger than $\frac{1}{2}N$. The data suggest that there is probably no well-defined asymptotic behaviour for $N \to \infty$.

The Vapnik–Chervonenkis (VC) dimension is one of the central quantities used in both mathematical statistics and computer science to characterize the performance of classifier systems [1, 2]. In the case of feed-forward neural networks it establishes connections between the storage and generalization abilities of these systems [3–5]. In this letter we will discuss the VC dimension of the Ising perceptron with binary patterns.

The VC dimension d_{VC} is defined via the growth function $\Delta(p)$. Consider a set of instances x and a system C of binary classifiers c: $x \mapsto \{-1, 1\}$ that group all $x \in X$ into two classes labelled by 1 and -1, respectively. For any set $\{x^\mu\}$ of p different instances $x^1, \ldots, x^p$ we determine the number $\Delta(x^1, \ldots, x^p)$ of different classifications $c(x^1), \ldots, c(x^p)$ that can be induced by running through all classifiers $c \in C$. A set of instances is called *shattered* by the system C if $\Delta(x^1, \ldots, x^p)$ equals 2^p, the maximum possible number of different binary classifications of p instances. Large values of $\Delta(x^1, \ldots, x^p)$ roughly correspond to a large diversity of mappings contained in C. The growth function $\Delta(p)$ is now defined by

$$\Delta(p) = \max_{x^\mu} \Delta(x^1, \ldots, x^p)\,. \tag{1}$$

It is obvious that $\Delta(p)$ cannot decrease with p. Moreover, for small p one expects that there will be at least one shattered set of size p and hence $\Delta(p) = 2^p$. On the other hand, this exponential increase in the growth function is unlikely to continue for all p. The value of p where it starts to slow down gives a hint as to the complexity of the system C. In fact the Sauer lemma [1, 6] states that for all systems C of binary classifiers there exists a natural number d_{VC} (which may be infinite) such that

$$\Delta(p) \begin{cases} = 2^p & \text{if } p \leqslant d_{\mathrm{VC}} \\ \leqslant \displaystyle\sum_{i=0}^{d_{\mathrm{VC}}} \binom{p}{i} & \text{if } p \geqslant d_{\mathrm{VC}}\,. \end{cases} \tag{2}$$

† E-mail address: stephan.mertens@physik.uni-magdeburg.de

Here d_{VC} is called the VC dimension of the system C. Note that it will in general depend on the set X of instances to be classified.

A concrete example for a system of classifiers is given by the well known perceptrons defined by

$$\sigma = \mathrm{sign}\left(\sum_{i=1}^{N} J_i \xi_i\right) \tag{3}$$

where the weights $\boldsymbol{J} \in \mathbb{R}^N$ parameterize the perceptron and $\boldsymbol{\xi} \in \mathbb{R}^N$ is an instance or pattern to be classified. The multiplication of $\boldsymbol{J}$ by a constant factor does not affect the output σ, so the weights are usually restricted by $\boldsymbol{J}^2 = N$. For this *spherical perceptron* the exact result $d_{\mathrm{VC}} = N$ has been obtained analytically [7].

The *Ising perceptron* is a spherical perceptron with the additional constraint $J_i = \pm 1$ on the weigths. For real valued patterns $\boldsymbol{\xi} \in \mathbb{R}^N$ this constraint does not affect the VC dimension, i.e. $d_{\mathrm{VC}} = N$ still holds [8].

Since much of the interest in neural networks with discrete weights is due to their easy technical implementation it is important to consider not only binary weigths but also binary patterns $\xi_i = \pm 1$. To avoid problems with the sign function if $\boldsymbol{J} \cdot \boldsymbol{\xi}$ happens to be 0, one introduces a threshold $\Theta = \pm 1$ for N even: $\sigma = \mathrm{sign}(\boldsymbol{J} \cdot \boldsymbol{\xi} + \Theta)$. Since the VC dimension for the Ising perceptron with $N = 2n$ and $\Theta = \pm 1$ is the same as for $N = 2n + 1$ without threshold, we will consider only odd values of N throughout this letter.

The determination of the VC dimension of the Ising perceptron with binary patterns is a difficult problem. Analytical calculations based on the replica method [9] are not very helpful, since this method is suited to calculating *typical* or *average* quantities, whereas the VC dimension is an extremal concept due to the max in (1). For the spherical perceptron this difference does not really matter, but for networks with discrete weights it is crucial [8].

To get at least a lower bound for d_{VC} it suffices to find a large shattered set by a smart guess. Consider the set (N odd):

$$\begin{aligned}
\boldsymbol{\xi}^{(0)} &= (-1, -1, \ldots, -1, -1) \\
\boldsymbol{\xi}^{(1)} &= (-1, -1, \ldots, -1, +1) \\
\boldsymbol{\xi}^{(2)} &= (-1, -1, \ldots, +1, -1) \\
&\;\;\vdots \\
\boldsymbol{\xi}^{\frac{1}{2}(N+1)} &= (-1, \ldots, -1, +1, -1, \ldots, -1)\,.
\end{aligned} \tag{4}$$

Let $\boldsymbol{\sigma} = (\sigma_0, \ldots, \sigma_{\frac{1}{2}(N+1)})$ be an arbitrary output vector. To see how $\boldsymbol{\sigma}$ can be realized by the binary perceptron, we have to distinguish two cases:

First case: $\boldsymbol{\sigma} = (\sigma_0, \sigma, \ldots, \sigma)$ i.e. the output values for all patterns except $\boldsymbol{\xi}^{(0)}$ are the same. This output can be realized by the weights

$$\boldsymbol{J} = (-\sigma, \underbrace{\sigma_0, \ldots, \sigma_0}_{\frac{1}{2}(N-3)}, \underbrace{-\sigma_0, \ldots, -\sigma_0}_{\frac{1}{2}(N+1)})\,. \tag{5}$$

Second case: For all output vectors different from the first case, we can assert that

$$\left|\sum_{i=1}^{\frac{1}{2}(N+1)} \sigma_i\right| \leqslant \frac{1}{2}(N-3) \tag{6}$$

since at least one σ_i in the sum differs from the rest. As weights we choose

$$\boldsymbol{J} = (-\sigma_0, k_1, \ldots, k_{\frac{1}{2}(N-3)}, \sigma_{\frac{1}{2}(N+1)}, \ldots, \sigma_1) \tag{7}$$

where $\boldsymbol{k}$ can be any ± 1 vector with

$$\sum_{i=1}^{\frac{1}{2}(N-3)} k_i = -\sum_{i=1}^{\frac{1}{2}(N+1)} \sigma_i .$$

According to equation (6), such a vector can always be found. Again we have $\mathrm{sign}(\boldsymbol{J} \cdot \boldsymbol{\xi}^{(\mu)}) = \sigma_\mu$ for $\mu = 0, \ldots, \frac{1}{2}(N+1)$.

This proves that the set (4) is shattered and hence

$$d_{\mathrm{VC}} \geqslant \tfrac{1}{2}(N+3) \tag{8}$$

for the Ising perceptron with binary patterns. This value of d_{VC} agrees very well with numerical results obtained by a statistical enumeration method [10, 8]. For this method, one *randomly* draws p binary patterns and calculates $\Delta(\boldsymbol{\xi}^{(1)}, \ldots, \boldsymbol{\xi}^{(p)})$ by enumeration of all perceptrons $\boldsymbol{J} \in \{\pm 1\}^N$. If a single pattern set with $\Delta(\ldots) = 2^p$ is found, we know that $d_{\mathrm{VC}} \geqslant p$. Like the replica method, this method is not suited to calculating the VC dimension in cases where the maximum shattered sets are rare.

There is, however, a method that guarantees exact evaluation of the VC dimension: *exhaustive enumeration* of all shattered sets. The shattered pattern sets can be arranged as the nodes of a tree. The root of the tree is the empty pattern set (conveniently defined to be shattered). The children of a P pattern node are formed by all those shattered $(P+1)$ pattern sets that can be obtained from the parent by adding a new pattern. The recursive application of this definition gives the complete tree of all shattered sets. The VC dimension is the height of the tree. It can be measured by a traversal of the complete tree using standard algorithms.

The branching factor of the tree is $\mathrm{O}(2^N)$, its height is $\mathrm{O}(N)$, giving an overall complexity of $\mathrm{O}(2^{N^2})$. This exponential complexity limits the reachable size N very soon and calls for some tricks to reduce the number of nodes.

Before we can think of reducing the number of nodes, we must ensure that every node, i.e. every shattered set, is considered only once. A ± 1 pattern can be read as an N-bit integer (identifying -1 with 0), hence we have an *order relation* between the patterns. If we add only patterns to a set which are larger than the current elements of the sets, uniqueness of the nodes is guaranteed.

The first trick to reduce the number of nodes exploits the symmetry of the problem: a shattered set remains shattered if we multiply one of its elements, or the ith entry of all elements by -1. Therefore we may restrict ourselves to pattern sets where all elements start with -1: $\boldsymbol{\xi} = (-1, \ldots)$ and we can fix the set containing only the pattern $(-1, -1, \ldots, -1)$ as the root of the tree.

The second trick is of the *branch and bound* variety and exploits the fact that we are not interested in the complete tree but only in its height. Let us assume that we have an easy-to-calculate upper bound for the maximum height that can be reached from a given node. If this upper bound turns out to be lower than the maximum height already found during our traversal, we can safely discontinue exploration of the subtree rooted in this node!

The binary outputs of a set of P patterns can be interpreted as P-bit number c. Iterating over all 2^N binary weight vectors of our network, we get 2^N such output numbers c. If $P < N$, some of the c values must appear more than once. Let f_c denote the frequency of the output value c. The number of different classifications of this pattern set is given by the number of $f_c > 0$:

$$\Delta(\xi^1, \ldots, \xi^P) = \sum_{c=0}^{2^P-1} \Theta(f_c). \tag{9}$$

The f_c's have to be calculated at each node to test whether the pattern set is shattered or not. If

$$f_{\min} = \min_c \{f_c\} \tag{10}$$

is 0, the pattern set is not shattered (at least one classification c has not been realized). If $f_{\min} > 0$ the pattern set is shattered and we can try to enhance it. Each new pattern can split an existing classification into two (appending a -1 to c for some weight vectors and a $+1$ for others), i.e. from each classification c we get two new classifications c_1 and c_2 with $f_c = f_{c_1} + f_{c_2}$. One of the new frequencies is always $\leqslant \frac{1}{2} f_c$. Therefore we have $\log_2 f_{\min}$ as an upper bound for the number of patterns that can be added to a shattered set before we definitely get a non-shattered set.

This strategy allows us to prune many subtrees. For $N = 5$, branch and bound reduces the number of nodes from 77 to 4, for $N = 7$ from 8389 to 4625.

Even with these tricks, the complexity $\mathrm{O}(2^{N^2})$ is overwhelming. On an UltraSparc I 170, exhaustive enumeration for $N = 7$ takes less than a second. For $N = 9$, the running time is 6.5 hours! Nevertheless, the results obtained for $N \leqslant 9$ are already quite remarkable. For $N = 7$, the set

$$\begin{aligned}
\boldsymbol{\xi}^{(1)} &= (-1, -1-1, +1, +1, +1, +1)\\
\boldsymbol{\xi}^{(2)} &= (-1, +1+1, -1, -1, +1, +1)\\
\boldsymbol{\xi}^{(3)} &= (-1, +1+1, +1, +1, -1, -1)\\
\boldsymbol{\xi}^{(4)} &= (+1, -1+1, -1, +1, -1, +1)\\
\boldsymbol{\xi}^{(5)} &= (+1, -1+1, +1, -1, +1, -1)\\
\boldsymbol{\xi}^{(6)} &= (+1, +1-1, -1, +1, +1, -1)\\
\boldsymbol{\xi}^{(7)} &= (+1, +1-1, +1, -1, -1, +1)
\end{aligned} \tag{11}$$

is shattered, hence $d_{\mathrm{VC}} = 7$—the maximum possible value! Together with $d_{\mathrm{VC}} = 4$ for $N = 5$ and $d_{\mathrm{VC}} = 7$ for $N = 9$, these results do not allow a decent conjecture for the general expression. However, partial enumerations for larger values of N indicate, that d_{VC} is substantially larger than the value $\frac{1}{2}(N+3)$ provided by (4).

The largest shattered sets found by exhaustive and partial enumerations share a common feature: They can be transformed into quasi-orthogonal sets, i.e. into sets, where the patterns have minimum pairwise overlap†,

$$\boldsymbol{\xi}^{(\mu)} \cdot \boldsymbol{\xi}^{(\nu)} = \begin{cases} \pm 1 & \mu \neq \nu \\ N & \mu = \nu\,. \end{cases} \tag{12}$$

This observation leads to the idea of restricting the enumeration to quasi-orthogonal pattern sets.

† Exact orthogonality cannot be achieved for N odd.

To find such pattern sets, the notion of *Hadamard matrices* is useful (see, e.g., [11] or any textbook on combinatorics or coding theory). A Hadamard matrix is an $m \times m$ matrix H with ± 1 entries such that

$$HH^{\mathrm{T}} = mI \tag{13}$$

where I is the $m \times m$ identity matrix. The rows (or columns) of a Hadamard matrix form a set of m orthogonal binary patterns. This implies that m must be even, but the whole truth is more restrictive: If H is an $m \times m$ Hadamard matrix, then $m = 1$, $m = 2$ or $m \equiv 0 \bmod 4$. The reversal is a famous open question: Is there a Hadamard matrix of order $m = 4n$ for every positive n? The first open case is $m = 428$.

For special values of m there are rules for constructing Hadamard matrices [12], e.g.:

- $m = 2^n$ (Sylvester type);
- $m = q + 1$ where q is a prime power and $q \equiv 3 \bmod 4$ (Paley type);
- $m = 2(q + 1)$ where q is a prime power and $q \equiv 1 \bmod 4$ (Paley type).

These rules provide us with Hadamard matrices of sufficient size†. To get from a $4n \times 4n$ Hadamard matrix to quasi-orthogonal binary patterns we either cut out one column ($N = 4n - 1$) or add an arbitrary column ($N = 4n + 1$) and take the rows of the resulting matrix as patterns. The pattern set (11) is a result of this procedure applied to the 8×8 Hadamard matrix H_8 of Sylvester type:

$$H_8 = H_2 \otimes H_2 \otimes H_2 \tag{14}$$

with

$$H_2 = \begin{pmatrix} -1 & -1 \\ -1 & +1 \end{pmatrix}. \tag{15}$$

In (14) $\otimes$ denotes the usual Kronecker product.

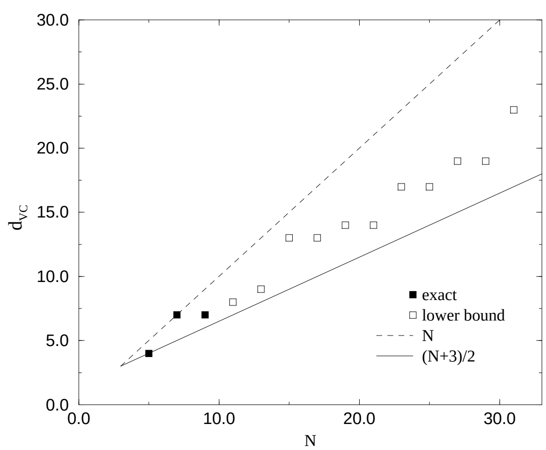

Figure 1. VC dimension of the Ising perceptron with binary patterns plotted against N. $d_{\mathrm{VC}} = N$ is an upper bound, $d_{\mathrm{VC}} = \frac{1}{2}(N + 3)$ is a lower bound provided by the set (4).

The restriction to quasi-orthogonal pattern sets allows us to consider larger values of N, but now the enumeration gives only lower bounds for d_{VC}. Results for $N \leqslant 31$ are

† The first value of $m = 4n$ where none of them applies is $m = 92$.

displayed in figure 1. The lower bound $d_{\mathrm{VC}} = \frac{1}{2}(N+3)$ achieved by the set (4) is exceeded for all $N > 5$, but the theoretic upper bound $d_{\mathrm{VC}} = N$ is attained only for $N = 7$. The data are not suited for a decent conjecture about a general expression for $d_{\mathrm{VC}}(N)$. Even the mere existence of a well defined asymptotic behaviour for $N \to \infty$ looks questionable. The VC dimension seems to be sensitive not only to the size but also to the number-theoretic properties of N: We observe a jump in $d_{\mathrm{VC}}(N)$ at $N = 2^n - 1$, i.e. at values of N where the corresponding Hadamard matrix is of Sylvester type.

The lower bounds in figure 1 do not rule out the possibility of a much more regular behaviour of the true $d_{\mathrm{VC}}(N)$, including well defined asymptotics. However, if the limit $\lim_{N\to\infty} d_{\mathrm{VC}}/N$ exists, it will probably be larger than 0.5.

The author appreciates fruitful discussions with A Engel. Thanks are also due to C Bessenrodt for her reference to Hadamard matrices.

References

[1] Vapnik V N and Chervonenkis A Y 1971 On the uniform convergence of relative frequencies of events to their probabilities *Th. Prob. Appl.* **16** 264
[2] Vapnik V N 1982 *Estimation of Dependences Based on Empirical Data* (Berlin: Springer)
[3] Haussler D, Kearns M and Schapire R 1991 Bounds on the sample complexity of Bayesian learning using information theory and the VC dimension *Proc. COLT'91* (San Mateo, CA: Morgan Kaufmann)
[4] Parrondo J M R and van den Broeck C 1993 Vapnik-Chervonenkis bounds for generalization *J. Phys. A: Math. Gen.* **26** 2211
[5] Engel A 1994 Uniform convergence bounds for learning from examples *Mod. Phys. Lett.* **8B** 1683
[6] Sauer N 1972 On the density of families of sets *J. Combinat. Theor.* A **13** 145
[7] Cover T M 1965 Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition *IEEE Trans. Electron. Comput.* **EC-14** 326
[8] Mertens S and Engel A 1996 On the VC-dimension of neural networks with binary weights, to be published
[9] Engel A and Weigt M 1996 Multifractal analysis of the coupling space of feed-forward neural networks *Phys. Rev.* E to appear
[10] Stambke J 1992 *Diploma Thesis* University of Giessen
[11] Krisement O 1990 A Hopfield model with Hadamard prototypes *Z. Phys.* B **80** 415
[12] Beth T, Jungnickel D and Lenz H 1985 *Design Theory* (Mannheim: Bibliograpisches Institut) ch I.9