On linear separability of random subsets of hypercube vertices

1991 J. Phys. A: Math. Gen. 24 L211

(http://iopscience.iop.org/0305-4470/24/4/010)

J. Phys. A: Math. Gen. 24 (1991) L211–L213. Printed in the UK

LETTER TO THE EDITOR

On linear separability of random subsets of hypercube vertices

Marco Budinich

Dipartimento di Fisica dell'Università di Trieste and Istituto Nazionale Fisica Nucleare Trieste, Via Valerio 2, I-34127 Trieste, Italy

Received 19 November 1990

Abstract. The classical Cover results on linear separability of points in $\mathbb{R}^d$ are a milestone in neural network theory. Nevertheless they are not valid for digital input networks because in this case the points are not, in general position, vertices of a d-dimensional hypercube. I show here that for large d all Cover findings can be extended to this case. It is also shown that for $n < O((d+1)^{3/2})$ the number of linear separations of n random hypercube vertices tends to that of n points in general position.

Feed-forward neural networks have frequently solicited studies on geometrical properties of their input space.

The values of the d input neurons can be thought of as coordinates of d-dimensional space $\mathbb{R}^d$ and then the set of all possible inputs is a subset of $\mathbb{R}^d$ (the pattern space). In the frequent case of digital inputs (0, 1 or ± 1) the pattern space shrinks to the set of the vertices of the d-dimensional hypercube $\mathbb{Q}^d \subset \mathbb{R}^d$.

The seminal Cover paper† [1] showed many interesting properties for sets of n points in general position in $\mathbb{R}^d$. The points are in general position if any k-tuple ($k \leq d+1$) of them is linearly independent.

Cover showed that the probability $P(n, d)$ that n random points in general position in $\mathbb{R}^d$ are linearly separable is‡

$$P(n,d) = \frac{\text{number of linear separations}}{\text{total number of separations}} = \frac{2\sum_{k=0}^{d}\binom{n-1}{k}}{2^n}. \tag{1}$$

From this formula Cover derives all of his interesting results directly applicable to one-layer feed-forward neural networks (perceptrons). The more important are (all for $d \to \infty$):

(a) the probability of linear separability of n random points falls to 0 when $n > 2(d+1)$

$$P(n,d) \to \Phi(-x) \qquad \text{for } d \to \infty \quad \text{and} \quad n = 2(d+1) + x\sqrt{2(d+1)}$$

where $\Phi(-x)$ is the cumulative normal distribution;

(b) the perceptron 'capacity' is $2(d+1)$, i.e. two random patterns per weight;

(c) the probability of 'non-ambiguous generalization' $\to 0$ if $n < 2(d+1)$ where n is the number of patterns already 'learned' by the network.

† For some more recent works with a similar approach see e.g. [2] and [3].

‡ This is the probability that exists an hyperplane separating a random partition of the n points in two sets. The n points are supposed to be in general position in $\mathbb{R}^d$. For more precise definitions see [1].

0305-4470/91/040211+03$03.50

If the pattern space is the set of vertices of $\mathbb{Q}^d$ (a very common situation in neural networks) equation (1) and all subsequent results are no longer valid. This happens because the points are usually not in general position.†

In what follows it is shown that for the identically defined probability $H(n, d)$ that n random vertices of $\mathbb{Q}^d$ are linearly separable,

$$H(n, d) \to P(n, d) \qquad \text{when } d \to \infty \tag{2}$$

holds, which extends (1) and related results to subsets of vertices of $\mathbb{Q}^d$ when $d \to \infty$ (the demonstration is similar to that used by Füredi in [4]).

Let $C_{\rm gp}(n, d)$ and $C(n, d)$ be the number of linear separations of a set Π_n of n points in $\mathbb{R}^d$ respectively with and without the hypothesis of general position. Füredi [4] obtains the following bounds from the geometrical theorem of Winder [5]

$$C_{\rm gp}(n, d) - \sum_{k=2}^{d+1} a_k(\Pi_n, d) \leq C(n, d) \leq C_{\rm gp}(n, d) \tag{3}$$

where $a_k(\Pi_n, d)$ is the number of linearly dependent k-tuples of points of the set Π_n.

To pass from (3) to the probabilities of (1) and (2) we have to average the quantities $C(n, d)$ and $a_k(\Pi_n, d)$ over all the possible Π_n and then to divide by the number of possible partitions, i.e. 2^n. With the hypothesis that the n points are vertices of $\mathbb{Q}^d$ we have $\binom{2^d}{n}$ possible choices for the set Π_n so (3) gives

$$P(n, d) - 2\sum_{\Pi_n}\sum_{k=2}^{d+1} a_k(\Pi_n, d)\left[2^n\binom{2^d}{n}\right]^{-1} \leq H(n, d) \leq P(n, d). \tag{4}$$

The quantity

$$\sum_{\Pi_n} a_k(\Pi_n, d)\left[\binom{2^d}{n}\binom{n}{k}\right]^{-1}$$

is, by definition, the probability that k points out of the n are not in general position. Since the points are vertices of $\mathbb{Q}^d$ this probability is bounded by the probability that a $(d+1)\times(d+1)$ random ± 1 matrix is singular and this probability is known [6] to tend to $O(1/\sqrt{d+1})$ when $d \to \infty$ so we have

$$\sum_{\Pi_n} a_k(\Pi_n, d)\binom{2^d}{n}^{-1} \leq \binom{n}{k} O\left(\frac{1}{\sqrt{d+1}}\right) \qquad \text{when } d \to \infty. \tag{5}$$

With this relation, observing that all quantities are positive, (4) gives

$$P(n, d) - O\left(\frac{1}{\sqrt{d+1}}\right)\sum_{k=2}^{d+1}\binom{n}{k}(2^{n-1})^{-1} \leq H(n, d) \leq P(n, d)$$

and being the fraction limited between 0 and 1 for every n this proves (2).

A similar argument can be used to study the number of linear separations of vertices of a hypercube‡. Intuitively it is known that for n random hypercube vertices two different cases exist. If $n \ll d$ hypercube symmetries are irrelevant and the number of linear separations will equal that of n points in general position while if $n \approx 2^d$ symmetries play a crucial role in diminishing the number of linear separations. In what

† The d dimensional hypercube is a highly symmetric figure where for example no $2d$ points in general position exist or where all points with a given number of 1's lay on just one hyperplane.

‡ In the past a lot of effort has been dedicated to this problem, i.e. to count the number of thresholding functions (see e.g. [5]).

follows the condition is proven that n has to satisfy (in the large d limit) to remain in the case where hypercube symmetries are marginal.

Starting from (3) we obtain

$$1-\sum_{\Pi_n}\sum_{k=2}^{d+1} a_k(\Pi_n, d)\left[C_{gp}(n,d)\binom{2^d}{n}\right]^{-1} \leqslant \frac{\langle C(n,d)\rangle}{C_{gp}(n,d)} \leqslant 1$$

where $\langle C(n,d)\rangle$ is the average value of $C(n,d)$. Using the definition of $C_{gp}(n,d)$, (1) and (5)

$$1-O\left(\frac{1}{\sqrt{d+1}}\right)\sum_{k=2}^{d+1}\binom{n}{k}\left[\sum_{k=0}^{d}\binom{n-1}{k}\right]^{-1} \leqslant \frac{\langle C(n,d)\rangle}{C_{gp}(n,d)} \leqslant 1$$

and from the asymptotic properties of this fraction for $n>2(d+1)$ and $d\to\infty$ we get

$$1-O\left(\frac{n}{(d+1)^{3/2}}\right) \leqslant \frac{\langle C(n,d)\rangle}{C_{gp}(n,d)} \leqslant 1$$

which proves that if $n<O((d+1)^{3/2})$ the average number of separating hyperplanes of n vertices of $\mathbb{Q}^d$ tends to $C_{gp}(n,d)$.

All this shows that as long as $n<O((d+1)^{3/2})$ while $d\to\infty$, hypercube symmetries are not important for the average number of separating hyperplanes. From this it follows that the probability of linear separability around $n=2(d+1)$ is not altered by hypercube symmetries. Both these properties derive from the result that the probability of a $d\times d$ binary matrix being singular is $O(1/\sqrt{d})$ when $d\to\infty$.

A final word of caution about the hypothesis of randomness in the choice of the n points that underlies all these results. In real life cases the patterns are highly correlated among themselves and these results do not apply directly.

References

[1] Cover T M 1965 Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition *IEEE Trans. Electron. Comput.* **EC-14** 326–34

[2] Baum E B 1988 On the capabilities of multilayer perceptrons *J Complexity* **4** 193–215

[3] Mitchison G J and Durbin R M 1989 Bounds on the learning capacity of some multi-layer networks *Biol. Cybern.* **60** 345–56

[4] Füredi Z 1986 Random polytopes in the *d*-dimensional cube *Discrete Comput. Geometry* **1** 315–9

[5] Winder R O 1966 Partitions of N-space by hyperplanes *SIAM J. Appl. Math.* **14** 811–8

[6] Komlós J 1967 On the determinant of (0, 1)-matrices *Studia Scientiarum Mathematicarum Hungarica* **2** 7–21

Bollobás B 1985 *Random Graphs* (New York: Academic) pp 347–50