TAGGING RARE EVENTS WITH SPARSELY ENCODED DATA

Jean-Pierre Nadal

Laboratoire de Physique Statistique* de l'Ecole Normale Supérieure
24, rue Lhomond -75231 Paris Cedex 05
France
*Laboratoire associé au C.N.R.S. (U.R.A. 1306)
et aux Universités Paris VI et Paris VII.

To appear in the proceedings of the workshop on
Neural networks: from biology to high energy physics
Isola d'Elba, Italy, June 5-14, 1991

Abstract

I give a review of recent results obtained on the associative properties of a simple perceptron in the sparse coding limit : simple Hebbian rules allow to reach optimal or near optimal performances, and these performances may be obtained in the regime where the network makes errors.

1. INTRODUCTION

In this conference, about one third of the talks beared on the application of artificial neural networks to high energy physics. I will introduce my paper in this particular context, hoping to make it readable by the high energy physicists. A typical potential application of neural networks to particle physics is the online identification of the presence of a b quark in the output of a collision. This interesting event is rare compared to the abundant number of other events, and its signature is not always easily recognizable. A standard approach, presented in several talks, is to choose a set of N variables which characterizes the observed events, and to train a feedforward neural network on a learning sample, that is on a set of events for which the class (b quark/ non b quark) they belong to is known. The network has N input neurons, may be one or two hidden layers and one output unit. After this learning phase, the network is tested on new events, the hope being that one has extracted the rule which allow to separate the signal from the background: the activity of the output unit is expected to be large for input patterns coding for a b quark, and close to 0 for all other event.
In this paper I will be concerned only with the learning stage, and in the particular case where the parameters which characterize the events are encoded as patterns of N bits, where the number M of one's is large but nevertheless very small as compared to the total number N of bits. It appears that for a simple perceptron (no hidden layer, figure 1), in the formal case where the input patterns are assumed to be randomly distributed, the optimal performance of simple, Hebbian, learning rules in this sparse coded limit can be computed exactly in the large N limit. The main outcome of such study is to emphasis the interest of using information theory criteria for evaluating the performance of a network, as also pointed out by G. Palm in this conference [1], and to show that, if the fraction of interesting events is very small and the data are sparsely encoded, Hebbian rules - which are "one-shot" learning rules- may lead to

optimal or near optimal performances. Before going into the details I will first introduce the criterion which I will use for characterizing the properties of the network.

2. INFORMATION THEORY CRITERION

Let us consider a simple perceptron with N inputs, whose activities will be denoted by V_j, j=1,...,N, and one output binary unit V =0,1 (figure 1).

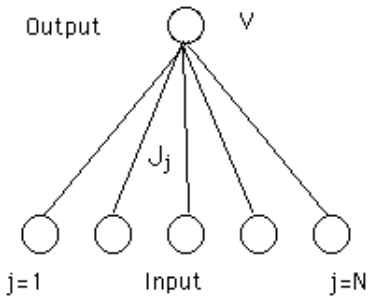

Figure 1: Perceptron type network with N input neurons and one output neuron.

The learning set consists of p patterns $\mathbf{V}^k = \{V_j, j=1,...,N\}$, k= 1,...,p. Among these p patterns, $p_+ = f' p$ are coding for an interesting event (the desired output is thus $V^k=1$) and $p_- = p - p_+ = (1-f')p$ are background events (for these patterns the desired output is $V^k=0$). The field h received by the output unit is

$$h = \sum_{j=1}^{N} J_j V_j \qquad (1)$$

where the J_j are the couplings. After some learning stage, the histograms $D_+(h)$ and $D_-(h)$ of the fields for signal and background events respectively may look as on figure 2.

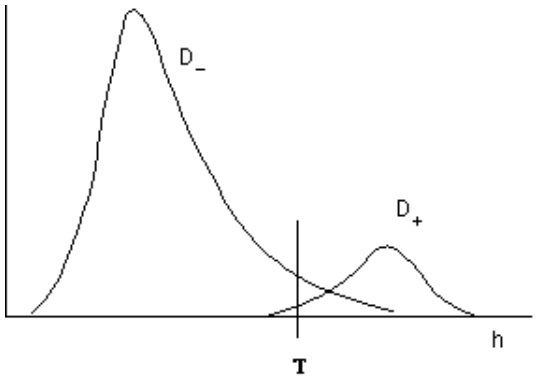

Figure 2: typical histograms of the fields for "bad" and "good" events.

Then one chooses a *cut*, that is a threshold **T**, so that the final output will be 1 if the field is larger than **T**, and 0 otherwise:

$$V = Y (h - \mathbf{T}) \quad (2)$$

(Y (x) being 1 for x > 0 and 0 otherwise). Note that, at this stage, it is equivallent to define the cut on the field h as above, or on an output given by a non linear, but monotonous, function of h. In that operating mode, if the D_+ and D_- histograms are overlapping the network is making errors. Let us denote by p_1 and p_2 the number of patterns which produce a positive output (V=1), among the patterns which belongs to the signal and those which belongs to the background respectively. The quantities usually considered in high energy physics are the *efficiency* E and the *purity* P defined by

$$E = p_1 / p_+ \quad (3)$$

$$P = p_1 / (p_1 + p_2) \quad (4)$$

One would like to reject as many bad events as possible (P large), keeping as many good events as possible (E large). What we suggest here is to use a global criterion, which characterizes the information content of the network. Its definition is [1-3, 13]:

$$I = (\ln C_p^{p_1+p_2} - \ln C_{p+}^{p_1} - \ln C_{p-}^{p_2}) / \ln 2 \quad (5)$$

where

$$C_N^M = N! / M! (N-M)!$$

If no error are made, $p_1 = p_+$ and $p_2=0$, I is the total information stored in the couplings. When errors are made, there is a loss of information due to the errors on the signal (second term in (5)) and a loss due to the errors on the background (last term in (5)). I will consider the information content i in bits *per coupling* (per synapse):

$$i = I / N \quad (6)$$

For large p it reads:

$$i = (p/N) \{ s((p_1 + p_2)/p) - f' s(p_1/p_+) - (1-f') s(p_2/p_-) \} \quad (7)$$

where s(.) is the mixing entropy in bits:

$$s(x) \ln 2 = - x \ln x - (1-x) \ln (1-x) \quad (8)$$

In terms of E and P, it can be written:

$$i = (p/N)\{ s(f' E / P) - f' s(E) - (1-f') s(E (1-P) f'/ P(1-f')) \} \quad (9)$$

For E close to one, there is no error on the signal and the second term vanishes; for P close to 1 there is no error on the background and the last term vanishes.
The strategy I will consider below is to choose the threshold **T** which maximizes the information stored.

In practical applications one may have additional constraints, such as a limitation on the fraction of events which can be stored due to the speed of data acquisition. This implies an upper bound on $(p_1+p_2)/p = f' E / P$, and one has to maximize i under this constraint.

3. ASSOCIATIVE PROPERTIES IN THE SPARSE CODING LIMIT

3.1 Continuous Couplings

In the evaluation of the performance of neural networks in associative tasks, many aspects of the sparse coding limit have been studied, for both attractor neural networks and simple perceptron type networks[1-15]. The regime of interest is when, in each pattern to be learned, the number M of active neurons (chosen at random with probability f) is negligible with respect to the total number of neurons N (in the large N limit), the coding rate f=M/N being of order lnN/N. For simplicity in all what follows I will consider the particular case of equal input and output coding rates, f=f', although the results can be extended to f' not equal to f.

Although the maximal number of such random patterns that can be stored with real valued synapses diverges with N like $(N/\ln N)^2$, the total amount of information stored per synapse tends to a finite limit, $i_m = 1/(2 \ln 2) = .721$ (bits per synapse)[16]. The general picture is that, when the coding rate goes from 1/2 to 0, the maximal information capacity, in bits per synapse, decreases from 2 to 1/ (2 ln2). Such results have been obtained with the "replica method" of statistical physics.

However, in the sparse coding limit simple learning rules, of the Hebb type, give very good performance, and in some cases the maximal theoretical capacity i_m can even be reached. Such exceptional performance is obtained with the Hebb rule:

$$J_j = \sum_{k=1}^{p} V^k V_j^k \qquad (10)$$

and with a well chosen threshold [13]. Another interesting property of this rule is its performance in the full error regime, that is when the number of stored patterns become very large. As one would expect, the system makes many errors and the information content decreases; however, the information content *per synapse* goes to a finite value, i_{as}[13]:

$$i_{as} = \frac{1}{\pi \ln 2} \qquad (11)$$

In fact, this is a common property of a family of Hebb rules defined for any value of the coding rate f (f between 0. and .5):

$$J_j = \sum_{k=1}^{p} (V^k - f)(V_j^k - f) \qquad (12)$$

For f=1/2 one recovers the Hopfield model. However, it is only for very small values of f that the information content goes through a maximum before going to its asymptotic value i_{as} (as the number of stored patterns is progressively increased). And it is only in the limit f -> 0 that this maximum becomes equal to the theoretical upper bound (as computed with the replica techniques).

3.2 Binary Couplings

3.2.1 A second look at the Willshaw model

One of the most interesting rules has been introduced long ago by Willshaw, Buneman and Longuet-Higgins [2]. It can be seen as a clipped version of the preceeding rule. The couplings are 0 or 1, with:

J_j = 1 if and only if for at least one pattern k $V^k = V_j^k = 1$ (13)

The threshold is given the value **T** = M, so that in retrieval there is no error on the signal: $p_1 = p_+$. In the sparse coding limit it allows to store up to

i_W = ln 2 = .693 bits per synapse, (14)

which is very close to i_m = .721 ! With G. Toulouse [13] we have reinvestigated the properties of this model puting the emphasis on regime where the system makes errors. It appears that the maximal capacity i_W is reached, as the number of stored patterns is increased, for any coding rate f smaller than some critical value f_c, and except at f_c it is reached in the error regime. Beyond this optimal point, if the number of stored patterns increases further, the information content decreases and eventually goes to zero - in contrast to what is observed for the Hebbian rule.

3.2.2 A third look at the Willshaw model

Recently the maximal capacity for {+1,-1} couplings has been computed with the replica techniques [17], and the calculation has been extended to other choices of discrete couplings [18,19]. In the case of (0,1) couplings in the sparse coding limit, the maximal capacity is found [18] to be around .29, much smaller than ln2...

Leaving aside the question of the validity of the replica approach, I have considered two possible reasons for this discrepancy:

1) the critical capacities computed so far with the replica method are associated with the requirement of *exactly no error* , (p_2 =0)whereas the maximal capacity of the Willshaw model is reached in a regime where the number of errors is non zero, but the noise to signal ratio is *vanishing in the large N limit* (as N goes to infinity, p_2 / p vanishes).

2) an implicit assumption done in the analysis of the Willshaw model is that the number of active neurons is *exactly* the same in every pattern, whereas in the replica approach the number of active neurons is given *in average* only.

The main results are as follows [14].

1) If one requires exactly no errors in the Willshaw model, the maximal capacity is *half* the maximal capacity i_W:

$i_0 = i_W / 2$ = .346 (15)

However, this value is not associated with a transition in the behaviour of the net: as the number of stored patterns increases, the information content increases regularly without any discontinuity at the point where the system starts to make some errors. It is only when the number of errors becomes not negligeable with respect to the total number of patterns that there is a change of behaviour.

2) The maximal capacity i1 of the willshaw model is much smaller if the patterns are chosen at random,

each activity being 1 or 0 with probability f. Indeed, the maximal capcity i_1 is found to be:

$$i_1 = .236, \quad (16)$$

about one third of i_W.

4. CONCLUSION

I have given a short review of some results obtained recently on the associative capacity of a simple perceptron for sparsely encoded data *and* rare positive outputs. The analysis stress the need for considering information theory criteria in the case where the network makes errors. The particular limit of sparsely encoded data is particularly interesting, since very simple rules allow to obtain very good performances as compared with the maximal possible capacity of such network in that limit.
It was shown also that the best performance may be obtained in the error regime; in fact one can show that this is not specific of the Willshaw model, nor of to the sparse limit [20].
I have given an explanation for the apparent discrepancy between the results coming from the Gardner approach and those from a direct study of the Willshaw rule. The value .29 is indeed above the value i_1=.236 obtained in the Willshaw model when one allows for fluctuations in the number of active neurons. From the replica calculation at several values of f the numerical extrapolation at f=0 is very hard to get [18], (the value .29 is only an upper bound) and thus we do not know exactly how .236 compares with the theoretical limit.
It was remarkable that the ln2 capacity of the Willshaw model was so close from the optimal capacity 1/(2 ln2) as computed by E. Gardner for continuous couplings. I have shown that this comparison was in fact inadequate: ln2 should be compared with the (unknown) capacity for patterns with exactly the same number of active neurons. The capacity of the Willshaw model for a fluctuating number of active neurons remains however remarkably good, when compared to the relevant upper limit, the one for 0,1 couplings.
In fact,the effects of the choice of the patterns distribution on the capacity is likely to be generic. Clearly it would be interesting to find a general method to compute directly the maximal capacity for patterns with exactly the same number of active neurons.

ACKNOWLEDGMENTS

I thank Hanoch Gutfreund, Marc Mézard and Gérard Toulouse for many fruitful discussions on sparse coding.

REFERENCES

[1] Palm G., This conference
[15] Lewenstein M., This conference
[4] Frolov, A. A. and Murav'ev I. P. 1988, *Informational characteristics of neuronal networks* , Nauka Ed. (In Russian) ; Frolovv A. A., 1991, to appear in Network.
[2] Willshaw D. J., Buneman O.P. and Longuet-Higgins H.C., 1969, *Non-Holographic Associative Memory* , Nature 222 960
[3] Palm G. 1980 *On Associative Memory* , Biol. Cybern. 36 19
[5] Palm G. 1987 *Techical Comment on "Computing with Neural Networks* , Science 235 1227

[6] Palm G. 1988 *On the Asymptotic Information Storage Capacity of Neural Networks* , in Neural Computers, Eckmiller R. and von der Malsburg C. Ed. (Springer, Berlin) p. 271
[7] Horner H. 1989 *Neural networks with low levels of activity: Ising versus McCulloch-Pitts neurons* , Z. Phys. B75 133
[8] Buhmann J., Divko R. and Schulten K. 1989 *On Sparsely Coded Associative Memories* , in Neural Networks from Models to Applications, Personnaz L. and Dreyfus G. Ed. (I.D.S.E.T., Paris) p. 360
[9] Tsodyks M.V. and Feigel'man M.V. 1988 *The Enhanced Storage Capacity in Neural Networks with Low Activity Level* , Europhys. Lett. 6 101
[10] Tsodyks M.V. 1988 *Associative Memory in Asymmetric Diluted Network with Low Level of Activity* , Europhys. Lett. 7 203
[11] Perez Vicente C.J. and Amit D.J. 1989 *Optimised Network for Sparsely Coded Patterns* , J.Phys. A22 559
[12] Amari S. 1989 *Characteristics of sparsely encoded associative memory* , Neural Networks 2 451
[13] Nadal J.-P. and Toulouse G.T. 1990 *Information storage in sparsely coded memory nets* , Network 1 61
[14] Nadal J.-P. 1991 *Associative memory: on the (puzzling) sparse coding limit* J. Phys. A. 24 1093
[16] Gardner E. 1988 T *he Space of Interactions in Neural Network Models* , J. Phys. A21 257
[17] Krauth W. and Mézard M. 1989 *Storage capacity of memory networks with binary couplings* , J. Physique 50 3057
[18] Gutfreund H. and Stein Y. 1990 *Capacity of neural Networks with discrete synaptic couplings* J. Phys. A23 2613
[19] Bouten M., Komoda A. and Serneels R. 1990 *Storage capacity of a diluted neural network with Ising couplings* , preprint
[20] Brunel N., Nadal J.-P. & Toulouse G., 1991 *Information Capacity of a Perceptron* , in preparation