Statistical Mechanics of Unsupervised Learning.

M. BIEHL and A. MIETZNER

Physikalisches Institut, Julius-Maximilians-Universität Am Hubland
D-97074 Würzburg, Germany

(received 23 February 1993; accepted in final form 4 October 1993)

PACS. 87.10 – General, theoretical, and mathematical biophysics (inc. logic of biosystems, quantum biology and relevant aspects of thermodynamics, information theory, cybernetics, and bionics).

PACS. 02.50 – Probability theory, stochastic processes, and statistics.

PACS. 05.90 – Other topics in statistical physics and thermodynamics.

Abstract. – We study two different unsupervised learning strategies for a single-layer perceptron. The environment provides a set of unclassified training examples, which belong to two different classes, depending on their overlap with an N-dimensional concept vector. By means of a statistical-mechanics analysis, using the replica method, we investigate how well the perceptron infers the unknown structure from the input data.

Neural networks deserve attention for their capability to learn from examples. Supervised learning is based on the presentation of training inputs together with their corresponding correct outputs, which are available from a «teacher». The aim is to efficiently extract the unknown rule, which assigns an output to any possible input, from these classified examples. This is termed generalization. Supervised learning techniques and their generalization properties have been thoroughly studied by means of statistical mechanics (e.g., [1-3]).

On the contrary, in unsupervised learning only unclassified training inputs are provided by the environment and the task is to implement a classification of these, which is chosen according to some optimality criterion. Unsupervised learning might be useful in the construction of multilayer networks (e.g., [4]). It is closely related to the problems of feature detection and clustering, where inputs are grouped into certain similarity classes [1, 5].

An illustrative example is a little child that might sort all animals it meets into classes, even without being told their names. In an unsupervised (or self-organizing) manner the child will learn to recognize the relevant features and distinguish cats from dogs, for instance. Of course this is only possible if there are indeed features which more or less uniquely separate the classes.

In the following we study a simple model in which a single vector $B \in \mathbb{R}^N$ defines two classes of inputs in N-space. An example $\xi' \in \mathbb{R}^N$ is classified according to its overlap with this vector such that the value of

$$S_B = \text{sign}(h_B),$$ (1)
where \( h_B = (1/\sqrt{N}) \sum_{j=1}^{N} B_j \xi_j \), indicates to which class it belongs. In the following we take the vector \( B \) to be normalized: \((1/N) \sum (B_j)^2 = 1\).

We assume that the examples the environment provides are more or less typical for their classes. This is modelled by presenting random patterns \( \{ \xi^r \in R^N \}_{r=1,...,p} \) with a distribution of overlaps \( h_B \), which has a double-peak structure:

\[
P(h_B) = \frac{1}{2 \sqrt{2\pi}} \left\{ \exp \left[ -\frac{1}{2} (h_B - \varphi)^2 \right] + \exp \left[ -\frac{1}{2} (h_B + \varphi)^2 \right] \right\}.
\]

The input vectors are taken to consist of independent random components with zero mean and unit variance, such that \( \xi \cdot \xi = N \) is satisfied exactly in the thermodynamic limit. The effect of this normalization is non-trivial and will be discussed elsewhere.

In the case of a binary \( B \in \{ +1, -1 \}^N \), for instance, these properties can be realized by drawing a variable \( \sigma^r \) for each pattern according to the distribution

\[
P(\sigma^r) = \frac{1}{2} \left\{ \delta(\sigma^r - 1) + \delta(\sigma^r + 1) \right\},
\]

and then the binary components of the pattern according to

\[
P(\xi_j^r | \sigma^r) = \frac{1 + \varphi/\sqrt{N}}{2} \delta(\xi_j^r - B_j \sigma^r) + \frac{1 - \varphi/\sqrt{N}}{2} \delta(\xi_j^r + B_j \sigma^r).
\]

Obviously we could consider \( B_j = 1 \) for all \( j \) without loss of generality in the following. Note that \( \sigma^r \) is a dummy variable, only indicating which peak in (2) the pattern contributes to. It does not necessarily coincide with the output of classification (1).

In contrast to the case of biased patterns (e.g., \[6\]), where \((1/N) \sum 1 \cdot \xi_j = O(1)\), the correlations considered here are much weaker: \((1/\sqrt{N}) \sum B_j \cdot \xi_j = O(1)\).

The parameter \( \varphi \) is a measure of the separability of the clusters, the width of the single peaks is chosen to be 1 in the following. Note that the peaks are indistinguishable in every direction perpendicular to \( B \).

Given a number \( p \) of examples, the normalized weight vector

\[
J \in R^N, \quad \frac{1}{N} \sum_{j=1}^{N} (J_j)^2 = 1,
\]

of a single-layer perceptron \[7,8\] is chosen, which itself defines a classification scheme in input space \( \tilde{S}^r = \text{sign}(h_j) \) analogous to (1).

How well the perceptron \( J \) has inferred the underlying structure \( B \) from the given examples is measured by the quantity

\[
R = \frac{1}{N} \sum_{j=1}^{N} J_j B_j.
\]

This overlap will be calculated for two different unsupervised learning strategies:
A) Minimization of the cost function

\[ H_A = - \sum_{v=1}^{P} (h_j)^2. \]  

In a geometric interpretation this corresponds to maximizing the mean-square distance of the input patterns from the hyperplane perpendicular to the vector \( J \), which separates the two output classes. In our case of zero mean data this means maximizing the variance \([1]\) of the quantities \( h_j \), or equivalently, finding the normalized eigenvector \( J \) corresponding to the largest eigenvalue of the correlation matrix \( C = \sum_{v=1}^{P} \xi_v \cdot \xi_v^T \) (principal-component analysis \([1,9]\)). There exist iterative learning algorithms, \( e.g. \) Oja’s rule or its modifications \([1,9]\), which indeed solve this problem. For the simulations of fig. 1 we used the method of von Mises \([10]\) to calculate the largest eigenvalue and the corresponding eigenvector iteratively: \( J(k+1) \propto C J(k), k \to \infty \).

B) Maximization of the so-called stability \([6]\)

\[ \kappa = \min_{v} \{ |h_j| \} \]  

of the perceptron. A cost function associated with this criterion is

\[ H_B = \sum_{v=1}^{P} \theta(\kappa - |h_j|) \]

and the optimal stability is the largest value of \( \kappa \) for which \( H_B \) can be made to zero. Such a classification can be shown to be very robust against input noise \([11]\).

The geometric interpretation is to maximize the distance of the pattern closest to the separating plane. This strategy has proven to yield very good generalization in the case of supervised learning of a linearly separable rule \([3,12]\). Although there exist algorithms which find the vector of optimal stability for a given, fixed input-output relation \([11,13,14]\), it is not clear how to deal with the additional degrees of freedom (the outputs) in the unsupervised problem. Methods similar to simulated annealing, as well as more sophisticated strategies, are currently under investigation \([15]\).

We study the properties of the corresponding solutions in the thermodynamic limit.
The cost functions defined above can be interpreted as the energy $H(J)$ of an interacting system of $N$ degrees of freedom with a corresponding partition function

$$Z = \left( \prod_{j=1}^{N} \int dJ_j \right) \varphi \left( \sum_{j=1}^{N} (J_j)^2 - N \right) \exp \left[ -\beta H(J) \right].$$

(10)

The limit $\beta \to \infty$ corresponds to the ground state and thus to the minimum of the cost function [6].

Assuming that the free energy $F = -\left(1/\beta\right) \ln Z$ is self-averaging with respect to the distribution of inputs given in eq. (4), the order parameter $R$ can be obtained by means of a saddle-point integration using the replica method [6]. We assume replica symmetry in the following. The calculation is similar to the ones related to supervised learning with a perceptron [2, 3, 12] and will be outlined in detail elsewhere.

**Case A)** For this choice of cost function the saddle point equations can be solved analytically. Given a separation $\varphi$ of the peaks in $P(h_i)$ we find the following dependence of $R$ on the number of patterns presented:

$$|R|(\alpha) = \begin{cases} \frac{\alpha \varphi^4 - 1}{\alpha \varphi^4 + \varphi^2}, & \text{for } \alpha \geq \frac{1}{\varphi^4}, \\ 0, & \text{for } \alpha < \frac{1}{\varphi^4}. \end{cases}$$

(11)

Up to a critical number of examples $\alpha_c = 1/\varphi^4$ the system does not at all detect the relevant direction $B$ in feature space. For $\alpha > \alpha_c$ the symmetry is broken, and the perceptron vector has a finite overlap with the concept vector. Note that the two solution $\pm R$ are equivalent in this context, because they correspond to the same classification schemes with only the labels exchanged.

In the limit $\varphi \to \infty$ the perceptron vector $J$ approaches $B$ as $|R| \approx 1 - 1/2\varphi^2 \alpha$, perfectly recognizing the underlying structure. Thus, the typical number needed for successful learning scales like $1/\varphi^2$, which coincides with a recent result [17] for supervised learning from similar data in the large separation limit. In this limit a teacher provides no additional information, because the structure in the data is self-evident.

For a specific value of $\varphi$, fig. 1 shows $|R|$ vs. $\alpha$ according to eq. (11) together with the results of simulations. A careful finite-size scaling confirms our result very well.

**Case B)** Following closely the work of [12], we obtain the replica-symmetric saddle-point equations

$$1 - R^2 = \alpha \int_{-\kappa}^{\kappa} \frac{dt}{\sqrt{2\pi}} \exp \left[ -\left( t - \varphi R \right)^2 / 2 \right] (|t| - \kappa)^2,$$

(12)

$$-2R = \alpha \frac{\partial}{\partial R} \left[ \int_{-\kappa}^{\kappa} \frac{dt}{\sqrt{2\pi}} \exp \left[ -\left( t - \varphi R \right)^2 / 2 \right] (|t| - \kappa)^2 \right],$$

(13)

which have to be solved numerically and yield the optimal stability and the overlap $R$ for given $\alpha$. The results for $\kappa$ will be discussed elsewhere.

For $\varphi = 0$ the second equation implies $R = 0$ and we recover the result of Anlauf, Opper and Pöppel [18, 19] for structureless input data.

In fig. 2 we have plotted $|R|$ vs. $\alpha$, for a separation of $\varphi = 2$. Again, $R = 0$ is the stable
solution up to a critical value $\alpha_c$, which together with $\kappa$ satisfies the system

$$
\begin{align*}
1 &= 2\alpha_c \int_0^\kappa Dz (\kappa - z)^2, \\
1 &= 2\alpha_c \varphi^2 \left( \frac{\kappa}{\sqrt{2\pi}} - \int_0^\kappa Dz \right),
\end{align*}
$$

(14)

where $Dz = dz/\sqrt{2\pi} \exp[-z^2/2]$. This critical value is always higher than in case A), see fig. 3 for comparison.

Above this value, $|R|$ increases with $\alpha$. Unlike case A), the asymptotic value $R(\alpha \to \infty)$ depends on the separation $\varphi$:

$$
|R|(\alpha \to \infty) = \begin{cases} 
\sqrt{1 - 2/\varphi^2}, & \text{for } \varphi > \sqrt{2}, \\
0, & \text{for } \varphi \leq \sqrt{2}.
\end{cases}
$$

(15)

This result has two remarkable implications:

Even for $\alpha \to \infty$ the student vector does not approach the concept vector $B(|R| \leq 1)$. Obviously $B$ itself is not the best vector with respect to the stability. Only for an infinite separation $\varphi \to \infty$ the weights of optimal stability coincide with $B$.

Another surprising result is the existence of a critical separation $\varphi_c = \sqrt{2}$, below which the structure in pattern space cannot be detected, no matter how many examples are presented. This is also reflected in the fact that $\alpha_c$ resulting from (14) diverges at $\varphi_c$.

Preliminary simulations, yielding high (yet suboptimal) values of $\kappa$, confirm qualitatively that indeed the resulting overlap $R(\alpha)$ is smaller than the one obtained by minimizing $H_A$. We expect, however, that the volume of solutions will be disconnected and that replica symmetry breaking [16] must be considered here in order to obtain exact results.

![Fig. 2.](image)

Fig. 2. $|R|$ vs. $\alpha$ for a separation of $\varphi = 2$. The solid line corresponds to strategy A) and converges to 1 for $\alpha \to \infty$. The dashed line is the result of eq. (12) for maximal stability, here $|R|$ tends to $1/\sqrt{2}$.

![Fig. 3.](image)

Fig. 3. The critical value $\alpha_c$ as a function of the separation $\varphi$. The dashed line is for maximal stability and diverges at $\varphi_c = \sqrt{2}$, the solid line corresponds to the maximal-variance strategy $\alpha_c = 1/\varphi^4$. 
It is also not obvious that our results will still hold qualitatively if the patterns provided by
the environment result in a different distribution $P(h_{ij})$.

The maximal variance strategy seems to be the natural tool for learning from data
distributed according to (4). Yet, in any practical problem, the form of this distribution will
be unknown in general. For instance, if the data is not normalized, the direction of largest
variance might not coincide with $B$ anymore. This effect will be studied in a forthcoming
publication.

Our result does not indicate that the maximum stability criterion is not appropriate for
unsupervised learning in general. It might be a useful concept indeed, for example in the case
of a separating gap between the two «clouds of data».

***

We thank W. KINZEL, E. KORUTCHEVA, G. REENTS and A. SCHARNAGL for useful
discussions and a critical reading of the manuscript. This work was supported by the
Volkswagen-Stiftung and the Deutsche Forschungsgemeinschaft. It is part of the Ph.D.
thesis of A. MIETZNER. Simulations were performed on the Cray Y-MP-EL of the
Rechenzentrum der Universität Würzburg.

REFERENCES

(Addison-Wesley, Redwood City, CA) 1991.
(Bibliographisches Institut, Zürich) 1993.
Heidelberg) 1990.