Supervised Learning from Clustered Input Examples.

C. Marangi (*), M. Biehl (**) (§) and S. A. Solla (**) (§§)

(*) Dipartimento di Fisica dell'Università di Bari
INFN, Sezione di Bari - Via Orabona 4, 70126 Bari, Italy
(**) CONNECT, The Niels Bohr Institute
Blegdamsvej 17, 2100 Copenhagen Ø, Denmark

(received 19 July 1994; accepted in final form 17 March 1995)

PACS. 87.10 + e - General, theoretical, and mathematical biophysics (including logic of biosystems, quantum biology, and relevant aspects of thermodynamics, information theory, cybernetics, and bionics).
PACS. 02.50 - r - Probability theory, stochastic processes, and statistics.
PACS. 05.90 + m - Other topics in statistical physics and thermodynamics.

Abstract. - In this paper we analyse the effect of introducing a structure in the input distribution on the generalization ability of a simple perceptron. The simple case of two clusters of input data and a linearly separable rule is considered. We find that the generalization ability improves with the separation between the clusters, and is bounded from below by the result for the unstructured case, recovered as the separation between clusters vanishes. The asymptotic behaviour for large training sets, however, is the same for structured and unstructured input distributions. For small training sets, the dependence of the generalization error on the number of examples is observed to be non-monotonic for certain values of the model parameters.

The use of layered neural networks as learning machines has been most successful in its application to classification tasks [1, 2], in which an N-dimensional feature vector $\xi$ is mapped onto a scalar label $\xi_0 \in \{1, \ldots, K\}$ in order to assign the input $\xi$ to one out of $K$ possible classes. Such tasks are easily implemented by a neural network in which classification is the result of a winner-take-all operation performed on the activities of $K$ linear output units [3]. The labelling function

$$f_j(\xi) = \arg \max_{1 \leq k \leq K} (J_k \cdot \xi)$$  \hspace{1cm} (1)

is characterized by the N-dimensional weight vectors $\{J_k\}$, $1 \leq k \leq K$ connecting the inputs to the $k$-th output unit.

(§) Permanent address: Institut für Theoretische Physik der Universität, Am Hubland, D-97074, Würzburg.

(§§) On leave from AT&T Bell Laboratories, Holmdel, NJ 07733, USA.

© Les Editions de Physique
In the framework of supervised learning the internal parameters $\{J\}$ of the neural network are modified so as to incorporate the information conveyed by a set of input-output pairs $\{(\xi^\mu, \xi^\mu_0), 1 \leq \mu \leq P\}$ which are examples of the classification task to be learned. A full specification of the joint probability distribution from which examples are drawn requires knowledge of the input distribution $P(\xi)$, as well as the decision rule $P(\xi_0 | \xi)$. The process of learning through adjustment of the weight vector $\{J\}$ is usually guided by the minimization of a training error, which measures the disagreement between the actual output $f_j(\xi)$ and the target output $\xi_0$ on the training examples. The performance of the trained network is measured through the generalization error, which quantifies the disagreement between $f_j(\xi)$ and $\xi_0$ on an arbitrary input $\xi$. 

A geometrical approach has been used in [5] to investigate the case in which each input component $\xi_j, 1 \leq j \leq N$ is chosen independently and randomly from the set $\{-1, +1\}$. Class labels are assigned to this unstructured, isotropic, discrete input space through a teacher network of the form (1) specified by vectors $\{B_k\}, 1 \leq k \leq K$. The task is obviously realizable for the student network (1): it suffices to set $J_k = B_k$ for all $1 \leq k \leq K$ (modulo a trivial permutation of the labels $k$ of the student output units). But real classification tasks rarely take place on unstructured input distributions: feature vectors often form distinguishable clusters in input space. The need to incorporate non-trivial input distributions within the framework of supervised learning of classification tasks, although widely recognized, has received limited theoretical attention. Available statistical-mechanics calculations [6,7] have considered input distributions $P(\xi)$ obtained as a mixture of $K$ Gaussian clusters centred at $\{C_k\}, 1 \leq k \leq K$, with class labels specified by the cluster labels. Such models impose an unrealistically strict correlation between the input space clustering described by the $\{C_k\}$ and the classification described by the $\{B_k\}$. The purpose of our paper is to investigate a model in which these two sets of directions are independent, and demonstrate that data clustering results in gains in generalization ability over the unstructured case even when the sets $\{B_k\}$ and $\{C_k\}$ are only weakly correlated. Performance gains increase as the two sets become identical.

We restrict ourselves to dichotomies ($K = 2$), for which the winner-take-all classifier (1) is implementable [5] by a single-layer perceptron:

$$f_j(\xi) = \text{sign} \left( \frac{1}{\sqrt{N}} J \cdot \xi \right),$$

with $J \in R^N$, $|J|^2 = N$. A structure is imposed on the discrete $N$-dimensional input space $\{-1, +1\}^N$ through the choice of a specific vector $C$ and a separation $\rho$ along the direction $\hat{C} = C/\sqrt{N}$ so that the inputs are distributed according to the discrete equivalent of two Gaussian clusters [8,9] centred at $\pm \rho \hat{C}$. A cluster label $\sigma$ is chosen from

$$P(\sigma) = \frac{1}{2} [\rho(\sigma - 1) + \rho(\sigma + 1)],$$

and the components of $\xi$ follow from

$$P(\xi | \sigma) = \frac{1}{2} \left[ (1 + \rho/\sqrt{N}) \phi(\xi_i - \sigma C_i) + (1 - \rho/\sqrt{N}) \phi(\xi_i + \sigma C_i) \right].$$
The input projection \( h_i = \tilde{c} \cdot \xi \) along the direction that joins the cluster centres is the superposition of two unit variance Gaussian peaks centred at \( \pm \beta \). Projections along any direction \( \tilde{c}' \perp \tilde{c} \) are structureless: \( h_\perp = \tilde{c}' \cdot \xi \) is a Gaussian variable of unit variance and zero mean.

Training examples are of the form \((\xi^\mu, \xi_0^\mu)\), where

\[
\xi_0^\mu = \text{sign} \left( \frac{1}{\sqrt{N}} B \cdot \xi^\mu \right),
\]

with \( B \in \mathbb{R}^N \), \( |B|^2 = N \), and inputs \( \xi^\mu \) drawn from the distribution (3), (4) for \( 1 \leq \mu \leq P \). The discreteness of input space is not relevant; results reported here are also valid for the continuous version [8,9] of distribution (4). Class labels \( \xi_0^\mu \) determined through (5) are in general not identical to the cluster labels \( \tau^\mu \). The task is by construction learnable by the student network (2): the classes are linearly separable even though the Gaussian clusters overlap. In this scenario, as in most practical applications, the minimization of the training error is a natural learning strategy (1).

The object of our analysis is to monitor the generalization ability of the student network as a function of the normalized number of examples \( \alpha = P/N \), the alignment \( \gamma = B \cdot C/N \) between teacher and structure, and the separation \( \beta \) between the centres of the input clusters. Training is based on the maximization of the stability

\[
\kappa = \min_{\mu} \left\{ \frac{\xi_0^\mu}{\sqrt{N}} J \cdot \xi^\mu \right\},
\]

over the \( P \) training examples. This strategy is known to yield almost optimal generalization for \( \beta = 0 \) [4,10]. Deterministic algorithms available to obtain the weight vector of optimal stability for any linearly separable training set [11-13] allow for comparisons between our analytic results and numerical simulations.

We apply the standard statistical-mechanics formalism [14] and consider the components of the student weight vector \( J \) as a system of \( N \) degrees of freedom subject to the constraint \( |J|^2 = N \) and interacting through an energy

\[
E = \sum_{\mu=1}^{P} \theta \left( \kappa - \frac{\xi_0^\mu}{\sqrt{N}} \sum_{i=1}^{N} J_i \xi_i^\mu \right), \quad \text{with } \theta(x) = \begin{cases} 1, & \text{for } x > 0, \\ 0, & \text{else.} \end{cases}
\]

The training error (7) reduces to the number of misclassified examples for \( \kappa = 0 \). The replica trick within the symmetric ansatz [15] is used to perform the average over all the possible choices for the \( P \) input vectors \( \{\xi^\mu\} \). Thermodynamic properties are expressed in terms of three order parameters: \( q \), representing the typical overlap between two error-free student vectors \( J \), and the additional overlaps \( R = J \cdot B/N \) and \( D = J \cdot C/N \). The limit \( \beta \to \infty \) forces the system into its ground state, and the optimal stability is the maximal value of \( \kappa \) for which this ground state still has zero energy at fixed \( \alpha \). The corresponding \( J \) is unique and is selected in the \( q \to 1 \) limit [14]. The following system of saddle-point equations determines

\(\text{(1)}\) The classification task considered by [6] and [7] is unrealizable: class labels determined by the cluster labels result in overlapping, non-separable classes. In this scenario, strategies based on the minimization of the training error can result in overtraining, and are easily outperformed by simple Hebbian learning [7].
the values of $R$, $D$, and the optimal $\kappa$:

$$1 - \frac{R^2 + D^2 - 2\gamma RD}{1 - \gamma^2} = \alpha \int \int D t \frac{dh}{\sqrt{2\pi}} \exp \left[ -\frac{1}{2} (h - \varphi \gamma)^2 \right] F^2(h, t),$$

$$2 \frac{\gamma D - R}{1 - \gamma^2} = \alpha \frac{3}{3R} \int \int D t \frac{dh}{\sqrt{2\pi}} \exp \left[ -\frac{1}{2} (h - \varphi \gamma)^2 \right] F^2(h, t),$$

$$2 \frac{\gamma R - D}{1 - \gamma^2} = \alpha \frac{3}{3D} \int \int D t \frac{dh}{\sqrt{2\pi}} \exp \left[ -\frac{1}{2} (h - \varphi \gamma)^2 \right] F^2(h, t).$$

Integrals $Dt = dt \exp[-t^2/2]/\sqrt{2\pi}$ are over a two-dimensional domain $\Gamma$ determined by $F(h, t) > 0$, where

$$F(h, t) = \kappa - t \sqrt{1 - R^2 + \varphi (\gamma R - D)} \, \text{sign} \ h - |h| R.$$

The generalization error $\varepsilon_\varphi(x)$ follows from averaging the error function $\varepsilon = \Theta[-(J \cdot \xi) \cdot (B \cdot \xi)]$ over the input distribution $P(\xi)$. The result

$$\varepsilon_\varphi(x) = 1 - \sum_{\varphi = \pm 1} \frac{1}{2\pi \sqrt{1 - R^2}} \int_0^\infty \int_0^\infty dx dy \exp \left[ -\frac{1}{2(1 - R^2)} (x^2 + y^2 - 2Rxy) \right]$$

is fully determined by the external parameters $\varphi$ and $\gamma$, and the equilibrium values of the self-averaging order parameters $R$ and $D$ as follow from solving the saddle-point equations (8)-(10) at fixed $\alpha$. Results are summarized below.

For $\gamma = 0$ there is no correlation between the labelling direction $B$ and the direction $C$ which characterizes the clustering of the input data. In this regime the generalization error becomes independent of the cluster separation $\varphi$, and the resulting $\varepsilon_\varphi(x)$ is identical to the known result for unstructured data [10]. Improved performance at $\varphi > 0$ arises even for weak

---

**Fig. 1.**

Fig. 1. – The generalization error $\varepsilon_\varphi$ versus $x$ for $\gamma = 0.3$ and $\varphi = 0$ (solid), 2 (dashed), and 4 (dotted). Inset: $\varepsilon_\varphi$ for small values of $x$.

**Fig. 2.**

Fig. 2. – Order parameters $D (\square)$ and $R (\triangle)$ versus $x$ for $\varphi = 5$ and $\gamma = 0.3$. The solid lines correspond to the numerical solution of the saddle-point equations. The data points represent the results of simulations for a system with $N = 500$ input units, averaged over 100 independent training sets. Standard error bars would be approximately the size of the symbols.
correlation between $B$ and $C$, as shown for $\gamma = 0.3$ in fig. 1: the performance improves monotonically with increasing $\varphi$ at fixed $\alpha$. The $\varphi = 0$ curve provides a universal upper bound. The advantage of learning structured data increases monotonically as the alignment $\gamma$ between $B$ and $C$ increases. Performance improvement is a finite $\alpha$ effect which disappears asymptotically, as indicated by the merging of the curves in fig. 1 with increasing $\alpha$. In the $\alpha \to \infty$ limit, $R \to 1$, $D \to \gamma$, and $D \to \gamma R$. In this limit the generalization error becomes independent of both $\varphi$ and $\gamma$, and exhibits the same decay as in the $\varphi = 0$ case [4]:

$$\varepsilon_g(\alpha) \approx 0.50/\alpha.$$  

The competition between $B$ and $C$ results in a novel effect only observable for $\gamma < 1$: a non-monotonic dependence of the generalization error on $\alpha$, illustrated for $\varphi = 4$ in fig. 1. The mechanism for this small $\alpha$ anomaly is found on the $\alpha$ dependence of the order parameters $R$ and $D$, as shown in fig. 2. Numerical solutions to the saddle-point equations are found to be in very good agreement with simulation results, and reveal the following behaviour: $R$ increases monotonically towards 1, although at a slower rate than for $\varphi = 0$, while the rapid growth of $D$ at small $\alpha$ identifies a regime dominated by alignment of the student vector $J$ with $C$ instead of its target $B$. Such behaviour is possible at intermediate values of $\varphi$, and requires a small training set for which the class labels $\xi^C$ happen to be to a large extent consistent with the cluster labels $\alpha^C$. Among all possible hypotheses of $J$ compatible with these labels, the maximum-stability requirement favours a separating hyperplane perpendicular to $C$. The compatibility between $\gamma^C$ and the correct labels $\xi^C$ is broken as $\alpha$ increases, and $D$ decreases monotonically towards its asymptotic value $D = \gamma R$. Large values of $D$ at small $\alpha$ result in a detectable loss of generalization ability at intermediate and small values of $\gamma$. As opposed to the non-monotonic behaviour observed in [16] for a perceptron which implements a continuous bounded function on a discrete unstructured input space, the non-monotonicity observed here is a direct consequence of the misalignment between $C$ and $B$, which is likely to arise in realistic circumstances.

***

This work has been supported in part by the EU grant CHRX-CT92-0063. CM has been partially supported by the Istituto Nazionale di Fisica Nucleare; MB has been supported by the Deutsche Forschungsgemeinschaft. Simulations were done on the Cray-YMP-EL of the Rechenzentrum der Universität Würzburg. The authors gratefully acknowledge discussions with G. JUNG, B. LAUTRUP, M. OPPER, H. SCHWARZE, and O. WINTHER, and wish to express their appreciation for the hospitality extended by CONNECT and The Niels Bohr Institute while this work was being performed.

REFERENCES


