OFF–LINE LEARNING FROM CLUSTERED INPUT EXAMPLES

CARMELA MARANGI
Dipartimento di Fisica dell'Universita' di Bari and I.N.F.N., Sez. di Bari
Via Orabona 4, I–70126 Bari, Italy

SARA A. SOLLA
CONNECT, The Niels Bohr Institute
Blegdamsvej 17, DK–2100 Copenhagen Ø, Denmark

MICHAEL BIEHL and PETER RIEGLER
Institut für Theoretische Physik, Julius–Maximilians–Universität
Am Hubland, D–97074 Würzburg, Germany

ABSTRACT

We analyze the generalization ability of a simple perceptron acting on a structured input distribution for the simple case of two clusters of input data and a linearly separable rule. The generalization ability computed for three learning scenarios: maximal stability, Gibbs, and optimal learning, is found to improve 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 behavior for large training sets is the same for structured and unstructured input distributions. For small training sets, the generalization error of the maximally stable perceptron exhibits a nonmonotonic dependence on the number of examples for certain values of the model parameters.

1. Introduction

The problem of *supervised learning* is usually formulated[1–4] as that of a *student* network architecture being trained from examples in order to implement a target input-output relation. The task to be learned is defined through a specific configuration of a *teacher* network architecture which provides target outputs for randomly drawn inputs. Most cases analysed to date consider unstructured inputs: vectors of binary or continuous components are drawn from a uniform or isotropic distribution.

In many practical situations the input distribution is structured: there is a natural clustering reflecting correlations among the features associated with different components of the input vector. The detection of such structures in input space is usually addressed in the context of *unsupervised learning*,[1,5,6] with no reference to a functional relation to an output space. The question we address here is that of incorporating such nontrivial input distributions within the framework of supervised learning.[7,8] The goal is to investigate potential performance improvements due to some degree of consistency between the rule output and the clustering in input space.

2. The model

We consider a simple classification task in which N-dimensional vectors $\boldsymbol{\xi}$ are assigned to one of two classes according to the state $\xi_0 = \pm 1$ of a single output unit. A structure is imposed on the discrete N-dimensional input space $\{-1,+1\}^N$ through the choice of a specific vector $\mathbf{C}$ and a separation ρ along the direction $\hat{\mathbf{c}} = \mathbf{C}/\sqrt{N}$ so that the inputs are distributed according to the discrete equivalent of two Gaussian clusters[5,6] centered at $\pm\rho\hat{\mathbf{c}}$. A cluster label σ is chosen from $P(\sigma) = \frac{1}{2}[\delta(\sigma-1)+\delta(\sigma+1)]$, and the components of $\boldsymbol{\xi}$ follow from

$$P(\xi_i \mid \sigma) = \frac{1}{2}\left[(1+\rho/\sqrt{N})\delta(\xi_i - \sigma C_i) + (1-\rho/\sqrt{N})\delta(\xi_i+\sigma C_i)\right] \ . \qquad (1)$$

The input projection $h_{\parallel} = \hat{\mathbf{c}}\cdot\boldsymbol{\xi}$ along the direction that joins the cluster centers is the superposition of two unit variance Gaussian peaks centered at $\pm\rho$. Projections along any direction $\hat{\mathbf{c}}' \perp \hat{\mathbf{c}}$ are structureless: $h_{\perp} = \hat{\mathbf{c}}'\cdot\boldsymbol{\xi}$ is a Gaussian variable of unit variance and zero mean. The discreteness of input space is not relevant; results reported here are also valid for the continuous version[5,6] of distribution (1).

Training examples are of the form $(\boldsymbol{\xi}^{\mu}, \xi_0^{\mu})$, $1 \le \mu \le P$, with inputs $\boldsymbol{\xi}^{\mu}$ drawn from the distribution (1) and class labels ξ_0^{μ} determined through the perceptron rule $\xi_0 = \mathrm{sign}(\mathbf{B}\cdot\boldsymbol{\xi}\ /\sqrt{N})$. The resulting dichotomy[1] corresponds to separating the two classes with a hyperplane through the origin. The teacher vector $\mathbf{B} \in I\!R^N$ is perpendicular to the separating hyperplane. Class labels ξ_0^{μ} are in general not identical to the cluster labels σ^{μ}.

The student network is also a perceptron, with couplings $\mathbf{J} \in I\!R^N$ which are modified through the learning process. Both weight vectors are normalized: $\mid \mathbf{J} \mid^2 = \mid \mathbf{B} \mid^2 = N$. The task is by construction learnable by the student network: the classes are linearly separable even though the Gaussian clusters overlap.

In this realizable scenario, as in most practical applications, the minimization of a *training error* is a natural learning strategy. The object of our analysis is to investigate the *generalization ability* of the trained student network as a function of the normalized number of examples $\alpha = P/N$, the alignment $\eta = \mathbf{B}\cdot\mathbf{C}/N$ between teacher and structure, and the separation ρ between the centers of the input clusters.

Training is based on the minimization of a training error

$$E = \sum_{\mu=1}^{P} \Theta\left(\kappa - \frac{\xi_0^{\mu}}{\sqrt{N}}\ \mathbf{J}\cdot\boldsymbol{\xi}^{\mu}\right), \quad \text{with} \quad \Theta(x) = \begin{cases} 1 & \text{for } x > 0 \\ 0 & \text{else} \end{cases} \ , \qquad (2)$$

where $\kappa = \min_{\mu}\{\xi_0^{\mu}\ \mathbf{J}\cdot\boldsymbol{\xi}^{\mu}/\sqrt{N}\}$ is the *stability*[9] of the training set. The training error (2) reduces to the number of misclassified examples for $\kappa = 0$. We apply the standard statistical mechanics formalism[9] and consider the components of the student weight vector $\mathbf{J}$ as a system of N degrees of freedom subject to the normalization

constraint and interacting through the energy (2). The replica trick within the symmetric ansatz[9] is used to perform the average over the quenched disorder introduced by all the possible choices for a training set of size P. The zero temperature limit is taken so as to force the system into its groundstate and minimize the training error (2).

In addition to the external parameters α, κ, ρ, and η, the thermodynamic properties of the system depend on three *order parameters*: q, representing the typical overlap between two zero–energy student vectors $\mathbf{J}$, and the additional overlaps $R = \mathbf{J}\cdot\mathbf{B}/N$ and $D = \mathbf{J}\cdot\mathbf{C}/N$. Equilibrium values for the order parameters result from solving the system of saddle point equations:

$$\begin{aligned}
\frac{1}{\sqrt{1-q}}\left(q - \frac{R^2+D^2-2\eta RD}{1-\eta^2}\right) &= \frac{\alpha}{2\pi}\iint Dt\ dh\ e^{-\frac{1}{2}(h-\rho\eta)^2}\ \frac{e^{-\frac{1}{2}W^2}}{H(W)} \\
&\times \left(\kappa + t\frac{(1-R^2)}{\sqrt{q-R^2}} + \rho(\eta R - D)\,\mathrm{sign}\,h - |h|R\right)\ , \\
\frac{R-\eta D}{(1-\eta^2)\sqrt{1-q}} &= \frac{\alpha}{2\pi}\iint Dt\ dh\ e^{-\frac{1}{2}(h-\rho\eta)^2}\ \frac{e^{-\frac{1}{2}W^2}}{H(W)} \\
&\times \left(|h| - \rho\eta\ \ \mathrm{sign}\,h + \frac{tR}{\sqrt{q-R^2}}\right)\ , \\
\frac{D-\eta R}{(1-\eta^2)\sqrt{1-q}} &= \frac{\alpha}{2\pi}\iint Dt\ dh\ e^{-\frac{1}{2}(h-\rho\eta)^2}\ \frac{e^{-\frac{1}{2}W^2}}{H(W)} \\
&\times (\rho\ \mathrm{sign}\,h)\ , \qquad (3)
\end{aligned}$$

which follow from the minimization of the corresponding zero-temperature free energy. In writing (3) we have used $Dt = dt\ \exp(-t^2/2)/\sqrt{2\pi}$, $H(x) = \int_x^\infty Dt$, and $W \equiv (\ \kappa + t\sqrt{q-R^2} + \rho\ \mathrm{sign}\,h\ (\eta R - D) - |h|R\ )/\sqrt{1-q}$.

The performance of the trained student network is measured through the generalization error which follows from averaging the error function $\epsilon = \Theta\left[-(\mathbf{J}\cdot\boldsymbol{\xi}\ )(\mathbf{B}\cdot\boldsymbol{\xi}\ )\right]$ over the input distribution $P(\boldsymbol{\xi}\ )$. The result

$$\epsilon_g(\alpha) = 1 - \sum_{\sigma=\pm 1}\int_{-\sigma\rho\eta}^{\infty} Dy\ H\left(-\frac{Ry + \sigma\rho D}{\sqrt{1-R^2}}\right)\ , \qquad (4)$$

is fully determined by the external parameters ρ and η, and the equilibrium values of the selfaveraging order parameters R and D as follow from solving the saddle point equations (3) at fixed α.

3. Three Learning Scenarios

We first consider learning the perceptron of *maximal stability*,[9,10] defined as the maximal value of κ for which the groundstate solution still has zero energy at fixed

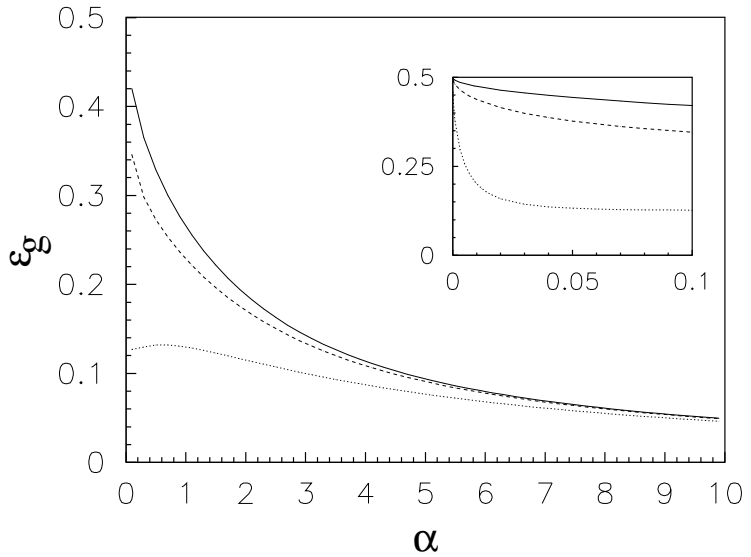

Fig. 1. The generalization error versus α for $\eta = 0.3$ and ρ= 0 (solid), 2 (dashed), and 4 (dotted). Inset: ϵ_g for small values of α.

α. The corresponding $\mathbf{J}$ is unique and is selected in the $q \to 1$ limit.[9] The saddle point equations (3) are solved in this limit to obtain the values of R, D, and κ. Deterministic algorithms available to obtain the weight vector of maximal stability for any linearly separable training set[11–13] allow for comparisons between our analytic results and numerical simulations. Results are summarized below.

For $\eta = 0$ there is no correlation between the labeling direction $\mathbf{B}$ and the direction $\mathbf{C}$ which characterizes the clustering of the input data. In this regime the generalization error becomes independent of the cluster separation ρ, and the resulting $\epsilon_g(\alpha)$ is identical to the known result for unstructured data.[10] Improved performance at $\rho > 0$ arises even for weak correlation between $\mathbf{B}$ and $\mathbf{C}$, as shown for $\eta = 0.3$ in Figure 1: performance improves monotonically with increasing ρ at fixed α. The $\rho = 0$ curve provides a universal bound. The advantage of learning structured data increases monotonically as the alignment η between $\mathbf{B}$ and $\mathbf{C}$ increases. Performance improvement is a finite α effect which disappears asymptotically, as indicated by the merging of the curves in Figure 1 with increasing α. In the $\alpha \to \infty$ limit, $R \to 1$, $D \to \eta$, and $D \to \eta R$. In this limit the generalization error becomes independent of both ρ and η, and exhibits the same decay[4] as in the $\rho = 0$ case: $\epsilon_g(\alpha) \approx 0.50/\alpha$.

The competition between $\mathbf{B}$ and $\mathbf{C}$ results in a novel effect only observable for $\eta < 1$: a nonmonotonic dependence of the generalization error on α, illustrated for $\rho = 4$ in Figure 1. The mechanism for this small α anomaly is found in the α dependence of the order parameters R and D, as shown in Figure 2. Numerical solutions to the saddle point equations are found to be in very good agreement with simulation results, and reveal the following behavior: R increases monotonically towards 1, although at a slower rate than for $\rho = 0$, while the rapid growth of D at small α identifies a

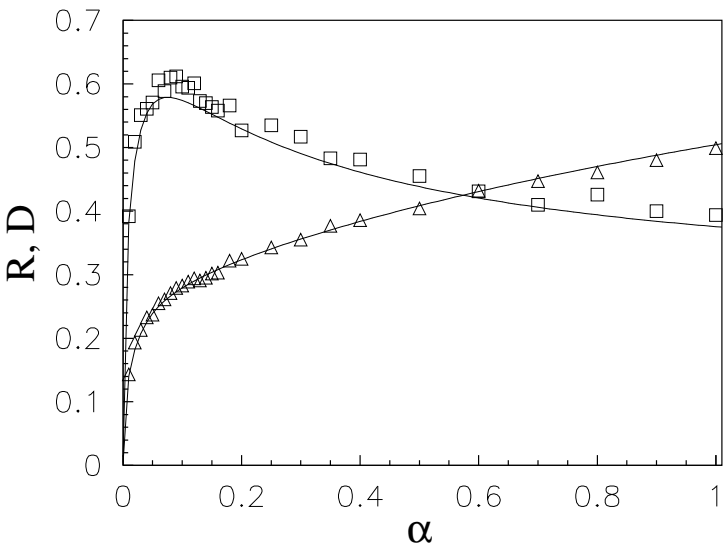

Fig. 2. Order parameters D ($\Box$) and R ($\triangle$) versus α for $\rho = 5$ and $\eta = 0.3$. The solid lines correspond to the numerical solution of the saddle point equations. The data points represent the results of simulations for a system with $N = 500$ input units, averaged over 100 independent training sets. Standard error bars would be approximately the size of the symbols.

regime dominated by alignment of the student vector $\mathbf{J}$ with $\mathbf{C}$ instead of its target $\mathbf{B}$. Such behavior is possible at intermediate values of ρ, and requires a small training set for which the class labels ξ_0^μ happen to be to a large extent consistent with the cluster labels σ^μ. Among all possible hypothesis $\mathbf{J}$ compatible with these labels, the maximum stability requirement favors a separating hyperplane perpendicular to $\mathbf{C}$. The compatibility between σ^μ and the correct labels ξ_0^μ is broken as α increases, and D decreases towards its asymptotic value $D = \eta R$. Large values of D at small α result in a detectable loss of generalization ability at intermediate and small values of η. This nonmonotonicity is a direct consequence of the misalignment between $\mathbf{C}$ and $\mathbf{B}$, which is likely to arise in realistic circumstances.

We now consider *Gibbs* learning. In this scenario, student vectors $\mathbf{J}$ which provide a correct classification for the training set define a *version space*[3] of equivalently good solutions to the learning problem, posed here as the minimization of the training error (2) at zero stability. The saddle point equations (3) are solved at $\kappa = 0$ to obtain the values of q_G, R_G, and D_G. Since the solution is not unique, the typical overlap q_G between two error–free student vectors is smaller than one. The symmetry $q = R$ is not satisfied, since possible teacher vectors $\mathbf{B}$ in version space are subject to the additional constraint $\mathbf{B} \cdot \mathbf{C}/N = \eta$.

Numerical solutions to the saddle point equations for zero-stability learning display a rather simple dependence of q_G, R_G, and D_G on α: for all values of ρ and η the order parameters increase monotonically towards their asymptotic values; both q_G and $R_G \to 1$, and $D_G \to \eta$, as $\alpha \to \infty$. The strictly monotonic increase of the

various order parameters with α results in a monotonic decrease of the generalization error with increasing α for all values of ρ and η. The generalization error at fixed α decreases monotonically with increasing ρ or η. The $\rho = 0$ curve provides here again a universal upper bound, and describes the asymptotic behavior: in the $\alpha \to \infty$ limit the generalization error becomes independent of ρ and η, and exhibits the same decay[2] as for $\rho = 0$: $\epsilon_g(\alpha) \approx 0.62/\alpha$.

Finally we consider the selection of a specific student vector: the *optimal* learner $\mathbf{J}^\star$ defined as the center of mass of the version space.[14] The corresponding order parameters are simply related to those obtained for a Gibbs learner: $R^\star = R_G/\sqrt{q_G}$ and $D^\star = D_G/\sqrt{q_G}$. $R^\star$ is found to increase monotonically with α for all values of ρ and η. $D^\star$ increases monotonically with α at small ρ, but exhibits nonmonotonic behavior resulting in a peak at small α for sufficiently large ρ. The nonmonotonicity of $D^\star$ is similar to the behavior of D for the maximally stable perceptron, but a slower decay to the $D^\star \to \eta$ asymptotic value results here in wider peaks. This softer effect is not sufficient to reverse the monotonic decrease of the generalization error with increasing α observed for all values of ρ and η. The generalization error at fixed α decreases with increasing ρ at fixed η; the same trend is observed when increasing η at fixed ρ. The $\rho = 0$ curve provides here again a universal upper bound. The dependence of ϵ_g on ρ and η disappears as $\alpha \to \infty$, resulting in the same asymptotic behavior[14,15] as for the $\rho = 0$ case: $\epsilon_g(\alpha) \approx 0.44/\alpha$.

References

1. J.A. Hertz, A. Krogh, and R.G. Palmer, *Introduction to the Theory of Neural Computation* (Addison-Wesley, California) 1991.
2. H.S. Seung, H. Sompolinsky, and N. Tishby, *Phys. Rev. A*, **45** (1992) 6056.
3. T.L.H. Watkin, A. Rau, and M. Biehl, *Rev. Mod. Phys.*, **65** (1993) 499.
4. M. Opper and W. Kinzel, in *Physics of Neural Networks III*, eds. E. Domany, J.L. van Hemmen, and K. Schulten (Springer, Berlin) in press.
5. M. Biehl and A. Mietzner, *J. Phys. A: Math. Gen.*, **27** (1994) 1885.
6. T.L.H. Watkin and J.–P. Nadal, *J. Phys. A: Math. Gen.*, **27** (1994) 1899.
7. N. Barkai, H.S. Seung , and H. Sompolinsky, *Phys. Rev. Lett.*, **70** (1993) 3167.
8. R. Meir, *Neural Computation*, **7(1)** (1995) 144.
9. E. Gardner, *J. Phys. A: Math. Gen.*, **21** (1988) 257.
10. M. Opper, W. Kinzel, J. Kleinz, and R. Nehl, *J. Phys. A: Math. Gen.*, **23** (1990) L581.
11. W. Krauth and M. Mezard, *J. Phys. A: Math. Gen.*, **20** (1987) L745.
12. J.K. Anlauf and M. Biehl, *Europhys. Lett.*, **10** (1990) 687.
13. P. Rujan, *J. Phys. I (Paris)*, **3** (1993) 277.
14. T.L.H. Watkin, *Europhys. Lett.*, **21** (1993) 871.
15. M. Opper and D. Haussler, *Phys. Rev. Lett.*, **66** (1991) 2677.