UNIVERSITÄT WÜRZBURG

INSTITUT FÜR THEORETISCHE PHYSIK

Am Hubland, D–97074 Würzburg, Germany

On–line learning from clustered input examples

Peter Riegler, Michael Biehl,
Sara A. Solla, and Carmela Marangi

to appear in:
Proceedings of the VII Italian Workshop on Neural Nets in Vietri s/m, 1995
World Scientific, in press

Ref.: WUE-ITP-95-013

e-mail : pr@physik.uni-wuerzburg.de
ftp : ftp.physik.uni-wuerzburg.de

ON–LINE LEARNING FROM CLUSTERED INPUT EXAMPLES

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

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

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

ABSTRACT

We analyse on–line learning of a linearly separable rule with a simple perceptron. Example inputs are taken from two overlapping clusters of data and the rule is defined through a teacher vector which is in general not aligned with the connection line of the cluster centers. We find that the Hebb algorithm cannot learn the rule perfectly in general. Moreover the dependence of the generalization error on the number of examples is nonmonotonic for certain choices of the model parameters. Perceptron and AdaTron training, however, approach perfect generalization with increasing size of the training set, and the asymptotic behavior is the same as for unstructured input data.

1. Introduction

In this work we study the problem of incorporating nontrivial input distributions within the framework of supervised learning in layered neural networks[1–4]. Structure in input space is expected to enhance the generalization ability whenever the rule output is consistent with the structure to a large extent. On the other hand, no improvement should arise when the classification to be learned is based on features completely different from the ones relevant to the clustering in input space.

We consider a simple classification task, in which points $\boldsymbol{\xi} \in I\!R^N$ are assigned to two categories according to the state of a single output unit $\xi_0 = \pm 1$. The dichotomy corresponds to placing a hyperplane through the origin to separate the two classes, and it is implemented through a single layer perceptron. Such a linearly separable target rule is defined by a vector $\mathbf{B} \in I\!R^N$ with $\mathbf{B}^2 = N$, which is perpendicular to the separating hyperplane. $\mathbf{B}$ can be interpreted as the weight vector of a *teacher perceptron* with output $\xi_0 = \mathrm{sign}\,(\mathbf{B} \cdot \boldsymbol{\xi})$. The student network is also a single layer perceptron, with couplings $\mathbf{J} \in I\!R^N$ chosen through the learning process.

A structure is imposed on the input space through the choice of a specific vector $\mathbf{C} \in \{+1, -1\}^N$ and a separation ρ along this direction so that the inputs are distributed according to the discrete equivalent of two Gaussian clouds centered at

$\pm\rho\mathbf{C}/\sqrt{N}$. The inputs are generated according to the following distribution[5]

$$P(\xi_i^\mu \mid \sigma^\mu) = \frac{1}{2}\left[(1+\rho/\sqrt{N})\delta(\xi_i^\mu - \sigma^\mu C_i) + (1-\rho/\sqrt{N})\delta(\xi_i^\mu + \sigma^\mu C_i)\right] \quad , \tag{1}$$

where the dummy variable σ^μ determines the cluster $\boldsymbol{\xi}^{\,\mu}$ belongs to. We assume $P(\sigma^\mu) = \frac{1}{2}\left[\delta(\sigma^\mu - 1) + \delta(\sigma^\mu + 1)\right]$. According to the central limit theorem the resulting distribution of overlaps $\mathbf{C}\cdot\boldsymbol{\xi}^{\,\mu}/\sqrt{N}$ is a superposition of two Gaussians with mean values $\pm\rho$ and unit width. In any arbitrary direction perpendicular to $\mathbf{C}$ the data appears structureless, as the corresponding distribution of overlaps is a single Gaussian with zero mean and unit variance. Note that the results reported here would also apply to a continuous version of distribution (2).

The object of our analysis is to investigate the generalization ability of the student as a function of the alignment $\eta = (\mathbf{B}\cdot\mathbf{C})/N$ between the teacher $\mathbf{B}$ and the vector $\mathbf{C}$, and the separation ρ between the centers of the input clusters.

By definition, the rule considered here is learnable for a student perceptron. The situation is different in a similar model recently studied by Meir[6], where the target outputs are defined by the labels σ^μ of the overlapping clusters (2).

2. The Formalism

Supervised learning is usually formulated as the extraction of information from a fixed set of examples through a learning process guided by the minimization of the training error. Here we will investigate *on–line learning*, where only the latest in a sequence of examples determines the change of the student weights in an iterative scheme.

On–line learning has recently been studied in a statistical mechanics framework in the context of perceptron learning[7–10]. This previous work considered only isotropic, unstructured input distributions. We will show in the following that the effect of introducing the more realistic clustered input distribution (2) leads to nontrivial effects including drastic changes of the generalization behavior.

The generic on–line perceptron scheme studied here is based on the following rule for the change of the student vector under the presentation of example μ:

$$\mathbf{J}^{\mu+1} = \mathbf{J}^\mu + f(h_J^\mu, \xi_o^\mu)\boldsymbol{\xi}^\mu \xi_o^\mu/\sqrt{N}. \tag{2}$$

Specific learning algorithms are defined through the choice of weight function f. This function can only depend on quantities which are available to the student, such as the teacher output ξ_o^μ, the student's current norm $Q^\mu = (\mathbf{J}^\mu)^2/N$, and its overlap with the μth example $h_J^\mu = \mathbf{J}^\mu\cdot\boldsymbol{\xi}^\mu/\sqrt{N}$. Note that no normalization is imposed on $\mathbf{J}$.

Eq. (2) can be interpreted as the evolution of the student weights in 'discrete time' μ. It is straightforward to derive recursion relations for the overlaps $\tilde{R}^\mu = \mathbf{J}^\mu\cdot\mathbf{B}/N$, $\tilde{D}^\mu = \mathbf{J}^\mu\cdot\mathbf{C}/N$, and Q^μ respectively.

The randomness of the input enters only through the overlaps h_J^μ, $h_B^\mu = \mathbf{B}\cdot\boldsymbol{\xi}^\mu/\sqrt{N}$ and $h_C^\mu = \mathbf{C}\cdot\boldsymbol{\xi}^\mu/\sqrt{N}$. If the inputs are drawn from distribution (2), the joint density of these quantities can be written as (omitting the indices μ) $P(h_J, h_B, h_C) = 1/2\sum_{\sigma=\pm1} P(h_J, h_B, h_C \mid \sigma)$. For large N these conditional densities become three–dimensional Gaussians with mean values $\langle h_J\rangle_c = \rho\tilde{D}\sigma$, $\langle h_B\rangle_c = \rho\eta\sigma$, $\langle h_C\rangle_c = \rho\sigma$ and correlations $\langle h_J h_B\rangle_c = \tilde{R} + \rho^2\eta\tilde{D}$, $\langle h_J h_C\rangle_c = \tilde{D}(1+\rho^2)$, $\langle h_B h_C\rangle_c = \eta(1+\rho^2)$, $\langle h_J^2\rangle_c = Q + \tilde{D}^2\rho^2$, $\langle h_B^2\rangle_c = 1+\rho^2\eta^2$, and $\langle h_C^2\rangle_c = 1+\rho^2$. Here $\langle\ldots\rangle_c$ denotes an average over the conditional probability $P(h_J, h_B, h_C \mid \sigma)$.

Thus the average over the sequence of uncorrelated training examples can be performed at every time step. In the limit $N\to\infty$ the order parameters are assumed to be selfaveraging with respect to the randomness of the inputs. Furthermore we interpret $\alpha = \mu/N$ as a 'continuous time' and obtain the first order differential equations

$$\frac{d\tilde{R}}{d\alpha} = \langle f h_B \xi_o\rangle, \qquad \frac{d\tilde{D}}{d\alpha} = \langle f h_C \xi_o\rangle, \qquad \frac{dQ}{d\alpha} = \left\langle 2 f h_J \xi_o + f^2\right\rangle. \tag{3}$$

The averages $\langle\ldots\rangle$ over the full distribution $P(h_J, h_B, h_C)$ are to be performed for a specific choice of the weight function f. The resulting system can be solved, at least numerically, yielding $R = \tilde{R}/\sqrt{Q}$ and $D = \tilde{D}/\sqrt{Q}$ and thus the generalization error $\epsilon_g = \langle\Theta(-h_J h_B)\rangle$ as a function of α.

We will consider initial conditions $\tilde{R}(0) = \tilde{D}(0) = 0$, and $Q(0) = 1$, corresponding to a normalized random initial student $\mathbf{J}^0$.

3. Three on–line algorithms

3.1. Hebbian Learning

The constant weight function $f = 1$ corresponds to the simple Hebb rule, which can be interpreted as an off–line training process constructing the weights $\mathbf{J}^p = \mathbf{J}^0 + \sum_{\mu=1}^{p}\boldsymbol{\xi}^\mu\xi_o^\mu/\sqrt{N}$ from a set of examples. Assuming an isotropic input distribution, Vallet[11] showed that a Hebbian student can learn a linearly separable rule perfectly, with a generalization error decreasing like $\epsilon_g(\alpha)\propto\alpha^{-1/2}$ as $\alpha\to\infty$.

The averages in eq.(5) can be performed analytically for $f = 1$ and the differential equations can be integrated to obtain

$$\tilde{R}(\alpha) = \left[\sqrt{\frac{2}{\pi}}e^{-\frac{1}{2}\eta^2\rho^2} + \eta\rho\,\mathrm{erf}\left(\frac{\eta\rho}{\sqrt{2}}\right)\right]\alpha, \qquad \tilde{D}(\alpha) = \left[\eta\sqrt{\frac{2}{\pi}}e^{-\frac{1}{2}\eta^2\rho^2} + \rho\,\mathrm{erf}\left(\frac{\eta\rho}{\sqrt{2}}\right)\right]\alpha,$$

$$Q(\alpha) = \left[1 + \rho\,\mathrm{erf}\left(\frac{\eta\rho}{2}\right)\tilde{D}(\alpha) + \sqrt{\frac{2}{\pi}}e^{-\frac{1}{2}\eta^2\rho^2}\tilde{R}(\alpha)\right]\alpha + 1. \tag{4}$$

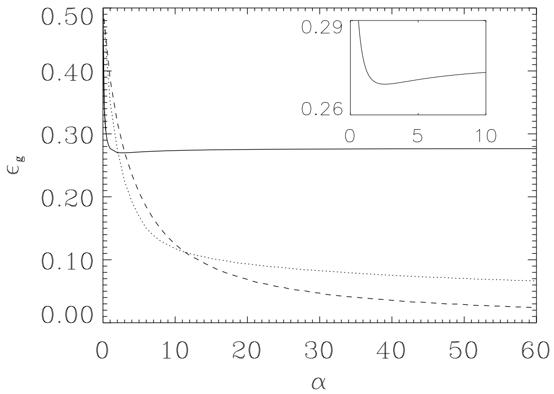

Fig. 1. The learning curves for $\eta = 0.2$ and $\rho = 2$. Solid line: Hebbian learning, dotted line: perceptron learning, and dashed line: on–line AdaTron. The inset shows the nonmonotinicity of the Hebbian learning curve.

Both normalized overlaps $R = \tilde{R}/\sqrt{Q}$ and $D = \tilde{D}/\sqrt{Q}$ increase monotonically with α for any non–negative values of ρ and η. Figure 1 shows the corresponding learning curve for $\eta = 0.2$ and $\rho = 2$.

We observe that in general the rule is not learnt perfectly in the limit $\alpha \to \infty$. Asymptotically, the Hebb student becomes a linear combination of the teacher $\mathbf{B}$ and the vector $\mathbf{C}$:

$$\mathbf{J}(\alpha \to \infty) \propto \sqrt{\frac{2}{\pi}} e^{-\eta^2\rho^2/2}\mathbf{B} + \rho \,\mathrm{erf}\left(\frac{\eta\rho}{\sqrt{2}}\right)\mathbf{C}. \tag{5}$$

Perfect generalization is only achieved for $\rho = 0$ and arbitrary η (unstructured data), for $\eta = 0$ and arbitrary ρ (learning in the subspace orthogonal to $\mathbf{C}$), and for $\eta = 1$ ($\mathbf{B}$ aligned with $\mathbf{C}$). In all other cases Hebbian learning fails to learn the linearly separable rule. We have found that for certain choices of η, ρ the function $\epsilon_g(\alpha)$ is nonmonotonic, indicating the existence of an optimal number of examples, for which the generalization error is minimal, see Fig. 1.

The residual error $\epsilon_g(\alpha \to \infty)$ is plotted vs. η for different values of ρ in Figure 2. This failure to learn the linearly separable rule is due to the fact that Hebbian learning assigns the same weight to all examples, whether correctly or incorrectly classified. More successful algorithms take into account whether the student disagrees with the teacher on the current example.

3.2. Perceptron learning

In the standard perceptron algorithm an example contributes to the student vector only when the teacher and the current student hypothesis provide different outputs: $f(h_J, \xi_o) = \Theta(-h_J\xi_o)$. This algorithm was studied in Ref. 9 for the case of unstructured inputs and the asymptotic decay of the generalization error was found to be $\epsilon_g \propto \alpha^{-1/3}$ as $\alpha \to \infty$, which is much slower than for Hebbian learning.

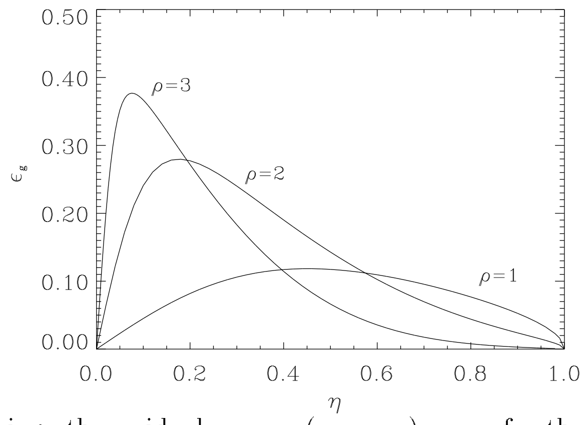

Fig. 2. Hebbian Learning: the residual error $\epsilon_g(\alpha \to \infty)$ vs. η for three different values of the separation ρ.

For the structured input distribution considered here, we have obtained differential equations which can be solved only numerically for general α, details will be published elsewhere. Figure 1 shows the learning curve for $\eta = 0.2$ and $\rho = 2$.

The normalized overlap with the teacher always approaches the asymptotic value $R = \tilde{R}/\sqrt{Q} = 1$; thus learning is always perfect in the limit $\alpha \to \infty$. This is in contrast to simple Hebbian learning. We find that

$$\epsilon_g = \frac{1}{\pi}\left(\frac{2}{3}e^{-\eta^2\rho^2}\right)^{1/3}\alpha^{-1/3} \quad \text{as} \quad \alpha \to \infty. \tag{6}$$

The decay is the same as for unstructured data[9], apart from an (η, ρ)–dependent prefactor. This observation is in agreement with Baum's prediction[12], that the generalization error should decrease like $\epsilon_g \propto \alpha^{-1/3}$ or faster for a perceptron algorithm applied to nonmalicious input distributions.

3.3. AdaTron learning

Here we consider training with a weight function [9] $f(h_J, \xi_o) = -h_J\xi_o\Theta(-h_J\xi_o)$. The differential equations resulting from performing the averages in Eq. (5) are to be solved numerically. Figure 1 displays the corresponding learning curve for $\eta = 0.2$ and $\rho = 2$. Learning is asymptotically perfect; the generalization error becomes independent of η and ρ, identical with the result for unstructured data[9] :

$$\epsilon_g = \frac{3}{2\alpha} \quad \text{as} \quad \alpha \to \infty. \tag{7}$$

This $1/\alpha$ decay is also found for off–line procedures considered in Ref. 13. The asymptotic behavior is obtained analytically by making the ansatz $\eta - D = \beta(\arccos\rho)^2$ for large α. For $1/\sqrt{3} < \rho\sqrt{1-\eta^2} < 1$ we find $\beta < 0$, indicating that D approaches

η from above. This implies a nonmonotonic dependence of the order parameter on α, since $D(0) = 0$. However, this rather weak effect does not lead to a nonmonotonic $\epsilon_g(\alpha)$.

4. Summary and outlook

We have investigated the effects of a nontrivial input distribution on the learning of a linearly separable rule by use of several on–line algorithms.

In particular we have found that the Hebb rule may fail to learn the linearly separable task completely. Moreover, the learning curve is nonmonotonic for certain choices of the model parameters, with a global minimum of the generalization error at an optimal number of examples. We will study this interesting example for overtraining in greater detail in a forthcoming publication.

For the perceptron and the on–line AdaTron algorithm the rule is learnt perfectly with the same asymptotic behavior as for unstructured data. It would be interesting to investigate the influence of more general types of input distributions on the learning curves.

It should also be possible to extend our studies to multiclass classification of data taken from a mixture of several Gaussian clusters [6,14].

References

1. J.A. Hertz , A. Krogh, R.G. Palmer, *Introduction to the Theory of Neural Computation*, (Addison–Wesley, Redwood City, Cal.) 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*, series 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. R. Meir, Neural Comp. **7**, (1995) 144.
7. O. Kinouchi and N. Caticha, J.Phys. **A25**, (1992) 6243.
8. M. Biehl and H. Schwarze, J.Phys. **A26**, (1993) 2651.
9. M. Biehl and P. Riegler, Europhys.Lett. **28**, (1994) 525.
10. N. Barkai, H.S. Seung, and H. Sompolinsky, *On–line Learning of Dichotomies*, preprint 1994.
11. F. Vallet, *Europhys. Lett.*, **8** (1989) 747
12. E.B. Baum, Neural Comp. **2**, (1990) 248.
13. C. Marangi, S.A. Solla, M. Biehl, and P. Riegler, this volume.
14. N. Barkai N., H.S. Seung, and H. Sompolinsky, *Phys. Rev. Lett.*, **70** (1993) 3167.