ESANN'1999 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 21-23 April 1999, D-Facto public., ISBN 2-600049-9-X, pp. 327-332

# Detection of two Gaussian clusters

Arnaud Buhot and Mirta B. Gordon *
Départment de Recherche Fondamentale sur la Matière Condensée
CEA-Grenoble, 17 rue des Martyrs,
38054 Grenoble Cedex 9, France
buhot@drfmc.ceng.cea.fr,gordon@drfmc.ceng.cea.fr

**Abstract.** We discuss the detection of two Gaussian clusters given a cloud of points. The optimal learning curve for this unsupervised learning scenario is determined with a replica calculation. A comparison with principal component analysis and supervised learning allows to understand the three different learning phases observed.

## 1. Introduction

The detection of structure underlying a set of randomly distributed points have been successfully studied in the context of neural networks with the tools of Statistical Mechanics [1, 2]. In the past few years, replica calculations allowed to obtain the optimal learning properties on such unsupervised tasks [3, 4]. In this article, we address the interesting problem in which the points are distributed in two Gaussian clusters. The optimal learning performance on this problem presents three consecutive phases as the number of points in the training set is increased: a first one, in which the information gathered from the data is not enough to detect any structure in the training set, an intermediate phase, similar to principal component analysis, and a last one where the structure of the two clusters is detected.

The paper is organized as follows: in section 2, we define our notations. Section 3 presents the properties of learning the principal component. In section 4, the optimal learning curve for the unsupervised detection of the double-Gaussian structure is determined. In section 5, we present the optimal learning performance in the case of a supervised formulation of the same problem. The discussion of the different learning phases observed is left to section 6.

## 2. Position of the problem

Let us consider a *training set* $\mathcal{L}_\alpha = \{\mathbf{x}_k\}_{k=1,\cdots,P}$ of $P = \alpha N$ points identically and independently distributed in a $N$-dimensional space. Hereafter, we refer to $\alpha$ as the (reduced) size of the training set. The detection of the probability

* also member of CNRS

ESANN'1999 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 21-23 April 1999, D-Facto public., ISBN 2-600049-9-X, pp. 327-332

distribution function (*pdf*) from which the points in the training set have been drawn is an example of *unsupervised learning.* If a class is associated to each point of the training set, learning is then called *supervised.* In the following, we consider the problem where the *pdf* consists of two Gaussian clusters, and may be written as:

$$P^* (\mathbf{x}|\mathbf{B}) \equiv (2\pi)^{-N/2} \exp\left\{-\frac{\mathbf{x}\cdot\mathbf{x}}{2} - V^*(\lambda)\right\} \tag{1}$$

where $\mathbf{x}$ represents a $N$-dimensional point and $\lambda = \mathbf{x}\cdot\mathbf{B} = \sum_i x_i B_i$ is the coordinate of $\mathbf{x}$ in the direction $\mathbf{B}$ ($\mathbf{B}\cdot\mathbf{B} = 1$). In the $N-1$ directions orthogonal to $\mathbf{B}$, the distributions are Gaussians with zero mean and unit variance. In the direction $\mathbf{B}$, the distribution is the sum of two Gaussians with variance $\sigma^2$ and mean $+\rho$ and $-\rho$ respectively. This defines the function $V^*$ introduced in (1):

$$P(\mathbf{x}\cdot\mathbf{B} = \lambda) = \frac{1}{2\sigma\sqrt{2\pi}}\left(\exp\left\{-\frac{(\lambda-\rho)^2}{2\sigma^2}\right\} + \exp\left\{-\frac{(\lambda+\rho)^2}{2\sigma^2}\right\}\right). \tag{2}$$

Such a *pdf* is the superimposition of two clusters, centered at $+\rho\mathbf{B}$ and $-\rho\mathbf{B}$ respectively. Thus, the *pdf* is assumed to be symmetric with respect to the origin. Generally this is not the case, and learning should be decomposed into two successive steps: learning of the mean, which is a relatively easy task [5], followed by the detection of the clusters. We are focusing here on the second step of learning. We assume that there are only two clusters. This restriction may be important but we expect that the different learning phases observed for two clusters also occurs for a larger number of clusters.

Whether the determination of the *pdf* is possible or not depends on the assumptions one is willing to accept. In the following, we assume that the variance $\sigma^2$ and the separation $\rho$ of the clusters are known. Thus, we restrict the learning problem to the determination of the direction $\mathbf{B}$ or $-\mathbf{B}$ (both directions are equivalent) for a given training set. Once this direction is known, the two clusters can be easily separated. The fact that the training set has finite (reduced) size implies that we cannot determine the direction $\mathbf{B}$ but only an estimator $\mathbf{J}$ (normalized to 1). In order to characterize it, we consider the overlap $R = |\mathbf{B}\cdot\mathbf{J}|$ between the true direction and the estimator. This overlap is zero when $\mathbf{J}$ contains no information about the true direction ($\mathbf{B}$ and $\mathbf{J}$ are orthogonal) and is one when learning is perfect ($\mathbf{J} = \pm\mathbf{B}$). For a given training set, the better the estimator the closer is the overlap to one. Different algorithms, in particular those based on the minimization of appropriate cost functions, allow to obtain such estimator.

## 3. Principal component analysis

A particular choice of cost functions consists in:

$$E(\mathbf{J};\mathcal{L}_\alpha,\epsilon) = \epsilon\sum_{k=1}^{P}(\mathbf{J}\cdot\mathbf{x}_k)^2 \tag{3}$$

where $\epsilon = \pm 1$. If $\epsilon = +1$, the minimum of (3) occurs for the direction $\mathbf{J}^*$ onto which the projections of the training-set points are as small as possible. More precisely, the variance of the training set is a minimized along this direction. Conversely, if $\epsilon = -1$, the direction $\mathbf{J}^*$ of minimal cost is the one with largest variance. Thus, the minimization of cost function (3) defines learning algorithms for principal component determination [1, 5, 6].

In our problem of two Gaussian clusters, the variances of the *pdf* in the directions orthogonal to $\mathbf{B}$ are one whereas the variance in the direction $\mathbf{B}$ is $\Delta^2 = \rho^2 + \sigma^2$. We expect that the direction $\mathbf{J}^*$ of minimal cost with $\epsilon = +1$ for $\Delta < 1$ and $\epsilon = -1$ for $\Delta > 1$ is a good estimator of the direction $\mathbf{B}$. The replica approach allows to calculate $R_{\text{pc}}(\alpha) = |\mathbf{B} \cdot \mathbf{J}^*|$ in the thermodynamic limit ($N, P \to +\infty$ with $\alpha = P/N$ constant), which turns out to be [3, 5]:

$$R_{\text{pc}}(\alpha) = \sqrt{\frac{\alpha - \alpha_c}{\alpha + 1/(\Delta^2 - 1)}} \tag{4}$$

with $\alpha_c = (\Delta^2 - 1)^{-2}$. First of all, no learning can succeed with too small training sets ($R_{\text{pc}} = 0$ for $\alpha \leq \alpha_c$), a fact called *retarded learning*. In the particular case of unit variance in the direction $\mathbf{B}$ ($\Delta^2 = 1$), $\alpha_c = +\infty$ so that $R_{\text{pc}} = 0$ for all values of $\alpha$: it is impossible to learn the direction $\mathbf{B}$ with principal component analysis. In the case where $\Delta^2 \neq 1$ then $R_{\text{pc}} \to 1$ when $\alpha \to +\infty$; perfect learning is possible asymptotically.

# 4. Optimal cost function

Let us now introduce a general cost function:

$$E\left(\mathbf{J}; \mathcal{L}_\alpha, V\right) = \sum_{k=1}^{P} V\left(\mathbf{J} \cdot \mathbf{x}_k\right). \tag{5}$$

For any *potential* $V$, replica calculations allow to determine the learning curve $R(\alpha; V)$ corresponding to the minimization of (5), in the thermodynamic limit. A functional maximization of $R(\alpha; V)$ allows to determine the optimal potential $V_{\text{opt}}$ [3, 4]. The determination of $V_{\text{opt}}$ may be done for any $\rho, \sigma$ and $\alpha$. The corresponding learning curve $R_{\text{opt}} = R(\alpha; V_{\text{opt}})$ is optimal by construction and satisfies the following equation:

$$\alpha = R_{\text{opt}}^2 \left\{ \int_{-\infty}^{+\infty} Dt \frac{\left[\int Dz\; z \exp\left(-V^*(\lambda)\right)\right]^2}{\int Dz \exp\left(-V^*(\lambda)\right)} \right\}^{-1} \tag{6}$$

where $\lambda = tR_{\text{opt}} + z\sqrt{1 - R_{\text{opt}}^2}$ and $Dz = \exp(-z^2/2)dz/\sqrt{2\pi}$. In order to obtain the *learning curve* $R_{\text{opt}}(\alpha)$, we need to invert Eq.(6). If no solution exists for a given $\alpha$, then $R_{\text{opt}}(\alpha) = 0$. If more than one solution exist, the one with largest $R$ has to be kept.

## 5. Supervised scenario

In this section, we suppose that not only the points, but also the clusters they come from, are given in the training set. This new information is coded as a binary variable $\tau$, such that $\tau_k = +1$ if the point $\mathbf{x}_k$ comes from the cluster centered at $+\rho\mathbf{B}$ and $\tau_k = -1$ otherwise. The properties of learning the direction $\mathbf{B}$ knowing the training set $\widetilde{\mathcal{L}}_\alpha = \{\mathbf{x}_k, \tau_k\}_{k=1,\cdots,P}$ containing points $\mathbf{x}_k$ drawn with the *pdf* (1) have been studied in [5]. The optimal supervised overlap $R_{\rm sup}(\alpha)$ is given by:

$$\begin{aligned} R_{\rm sup}(\alpha) &= \left\{ \frac{1-\alpha\left(1-2\Delta^2+\Delta^2\sigma^2\right)-\sqrt{Q}}{2\left(1-\sigma^2\right)\left(1-\alpha\left(1-\Delta^2\right)\right)} \right\}^{1/2}, \\ Q &= \left(1-\alpha\left(1-2\sigma^2+\sigma^2\Delta^2\right)\right)^2 + 4\alpha\rho^2\sigma^2. \end{aligned} \tag{7}$$

Since more information is available in the supervised scenario than in the unsupervised one, $R_{\rm sup}(\alpha)$ is an *upper bound* to the unsupervised learning curve.

## 6. Discussion of the learning phases

In this section, we discuss the three different kinds of optimal unsupervised learning curves that may arise in the process of detection of the two Gaussian clusters given a training set of points. Examples of these learning curves are shown on figures $a$, $b$ and $c$, which correspond to Gaussian clusters with the same value of $\sigma = 0.5$ but different separations $\rho = 1.4, 1.2$ and $1.1$ respectively. For comparison, we include in the same figures the curves corresponding to principal component learning and optimal supervised learning.

In all the cases, the supervised learning curves start increasing *at* $\alpha = 0$, in contrast with unsupervised learning. The retarded learning occurs as soon as there is a symmetry in the *pdf.* In this case, the directions $\pm\mathbf{B}$ are equivalent in the unsupervised scenario. In the supervised scenario this symmetry is absent: the average of the points $\mathbf{y}_k = \tau_k\mathbf{x}_k$ gives information on the direction $\mathbf{B}$, allowing to obtain a finite overlap $R$ as soon as $\alpha > 0$.

In figures $a$ and $b$, both the optimal and the principal component learning curves corresponding to the unsupervised scenario start increasing at the same value $\alpha_c$. They remain close to each other with increasing $\alpha$ in a range of $\alpha > \alpha_c$ whose width depends on the parameters of the Gaussian clusters. This phase, dominated by learning of the principal component, is absent in figure $c$.

Figure $a$ corresponds to a large variance $\Delta^2 = 2.21$. In this case, the optimal unsupervised learning curve presents a continuous cross-over between the principal component learning and the supervised learning.

In figure $b$, the optimal unsupervised learning curve jumps from a low-$R$ to a high-$R$ learning phase at $\alpha_1 > \alpha_c$. This discontinuity marks the boundary between two distinct learning phases. The first one is a learning phase similar to the principal component determination, and the corresponding optimal

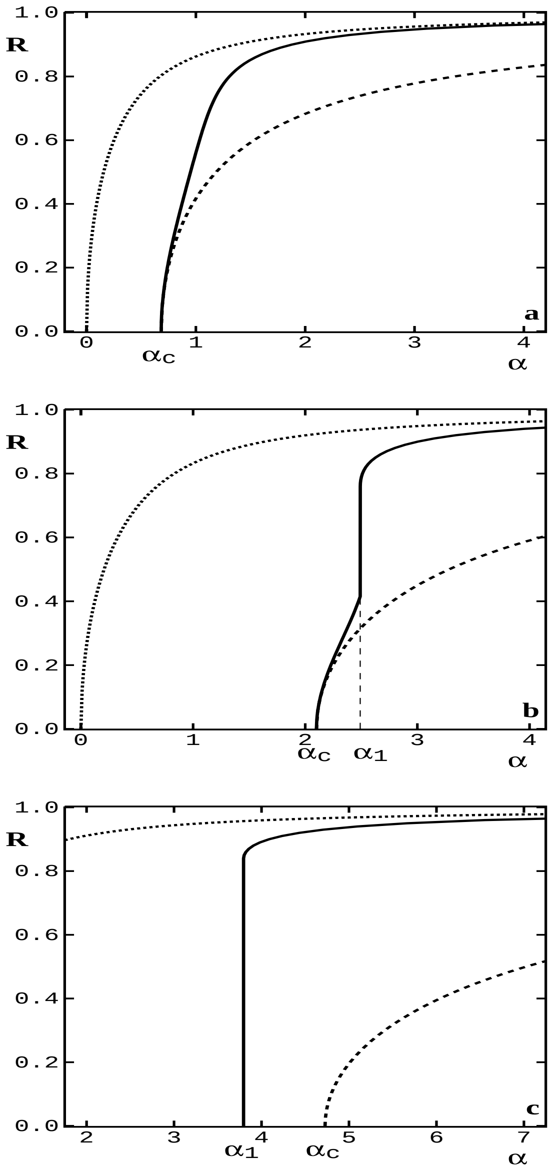

Figure 1: Learning curves. Figures $a$, $b$ and $c$ correspond to clusters with $\sigma = 0.5$, and three different separations: $\rho = 1.4, 1.2$ and $1.1$, respectively. In the three figures, the solid line is the optimal unsupervised learning curve $R_{\text{opt}}(\alpha)$, the upper (dotted) line is the supervised learning curve $R_{\text{sup}}(\alpha)$ and the lower (dashed) one is the principal component unsupervised learning curve $R_{\text{pc}}(\alpha)$.

potential turns out to be quite close to a quadratic function [4]. The second one corresponds to a phase where the double-Gaussian structure is detected. The learning performance of the latter is close to the one predicted for supervised learning. The corresponding optimal potential is a two-well function, whose minima are located close to $\pm\rho$, the positions of the *pdf*'s maxima in the direction $\mathbf{B}$.

Figure $c$ corresponds to a distribution whose variance in the direction $\mathbf{B}$ is close to the one in the other directions, as $\Delta^2 = 1.46$. In this case, the learning phase similar to the principal component one is absent from the optimal unsupervised learning scenario: $R_{\text{opt}}$ jumps from zero to a value close to the supervised learning curve at $\alpha = \alpha_1 < \alpha_c$, masking the principal component learning phase. The later needs a critical number of training patterns $\alpha_c > \alpha_1$ to develop.

In conclusion, the problem of the *unsupervised* detection of two Gaussian clusters given a cloud of points (the training set) has been analyzed within the Statistical Mechanics framework using replica calculations. We showed that the optimal unsupervised learning curves may present three different phases as the training set size $\alpha$ increases. First of all, a non-learning phase arises at small $\alpha$, due to the symmetry of the *pdf*. On increasing $\alpha$, a phase similar to principal component learning occurs if the variance in the direction of the two clusters is different enough from the variances in the other directions. Otherwise, this phase may not exist. The last phase corresponds to almost perfect learning. Its performance is close to the one of optimal *supervised* learning, which needs more information to be implemented, as the cluster from which the points have been drawn has to be included in the training set. We expect that these three different learning phases are characteristic to the detection of clusters and should also occur for larger numbers of clusters.

## References

[1] M. Biehl and A. Mietzner *Europhys. Lett.*, **24**, 421 (1993), *J. Phys. A: Math. Gen.*, **27**, 1885 (1994).

[2] T. L. H. Watkin and J.-P. Nadal *J. Phys. A: Math. Gen.*, **27**, 1899 (1994).

[3] C. Van den Broeck and P. Reimann *Phys. Rev. Lett.*, **76**, 2188 (1996), P. Reimann and C. Van den Broeck *Phys. Rev. E*, **53**, 3989 (1996).

[4] A. Buhot and M. B. Gordon *Phys. Rev. E*, **57**, 3326 (1998).

[5] P. Reimann, C. Van den Broeck and G. J. Bex *J. Phys. A: Math. Gen.*, **29**, 3521 (1996).

[6] E. Oja *Neural Networks*, **5**, 927 (1992).