Ultimate performance of QEM classifiers

Pierre COMON and Georges BIENVENU

Abstract—**Supervised learning of classifiers often resorts to the minimization of a quadratic error, even if this criterion is more especially matched to nonlinear regression problems. It is shown that the mapping built by a Quadratic Error Minimization (QEM) tends to output the Bayesian discriminating rules even with nonuniform losses, provided the desired responses are chosen accordingly. This property is for instance shared by the MultiLayer Perceptron (MLP). It is shown that their ultimate performance can be assessed with finite learning sets by establishing links with kernel estimators of density.**

I. Introduction

The classification problem consists of building a mapping ϕ from a set of patterns (observations), $\mathcal{E}$, to a set of classes. But in practice, ϕ often maps $\mathcal{E}$ to a set decision variables instead, $\mathcal{F}$. In *classification* problems, the set $\phi(\mathcal{E})$ is finite (and can be indexed by an integer i), and contains as many elements as classes. Denote y^i the variable encoding in $\mathcal{F}$ the i^{th} class, ω_i. With this formulation, any pattern x in $\mathcal{E}$ is wished to be associated with a variable y^i in $\phi(\mathcal{E}) \subset \mathcal{F}$.

In the context of *supervised* classification, a set of examples $\mathcal{A}(N) = \{(x^{(n)}, y^{i(n)}), 1 \leq n \leq N\}$ is given, so that mapping ϕ is apparently known at a finite number of points. This set of input-output pairs is subsequently referred to as the *learning set*. It is assumed throughout this paper that patterns are real valued and of dimension d, that is, $\mathcal{E} = \mathbb{R}^d$.

Next, let $\Phi(W, \cdot)$ be a mapping parametrized by a set of weights, W, that associates any vector x of $\mathcal{E}$ to an output vector $y = \Phi(W, x)$ in $\mathcal{F}$, from which the decision will be made; $\Phi(W, \cdot)$ is the *estimate* of ϕ.

Of course, regardless of the algorithm that will be used for this purpose, learning requires the existence of a link (generally of statistical nature) between the data present in the learning set, and the data to be classified [11] [13]. In the Bayesian context, it is assumed that any vector x of a given class ω_k that may be observed is drawn from a fixed (but a priori unknown) conditional density, $p(x|\omega_k)$. In addition, the occurence of any class ω_k has a constant probability denoted P_k. These notations will be subsequently assumed. The Bayesian approach is known to be able to detect new classes, but this will not be debated in the present letter. Also note that the true classes are not assumed to be disjoint, so that the ideal classifier may have a non zero misclassification rate (it does not bear overfitting).

In the classification context, the Quadratic Error Minimization (QEM) criterion consists of minimizing over the learning set a gap between desired responses and the outputs of the parametrized mapping, $\Phi(W, \cdot)$:

$$\epsilon(N) = \frac{1}{N} \sum_{n=1}^{N} \| y^{i(n)} - \Phi(W, x^{(n)}) \|^2. \tag{1}$$

The output space is assumed to be provided with a norm, and it is assumed throughout that $\mathcal{F} = \mathbb{R}^K$. Many neural networks dedicated to classification are proceeding this way, and the numerous algorithms proposed in the literature actually aim at reaching the same goal.

The matter presented in this letter has been already published in a French conference [5]. At the same time, results related to asymptotical performance of the MLP have been independently published in this journal [10]. One can also note that historically, asymptotical performance of the MLP have also been derived earlier [1], but the proof relied very much on the numerical algorithm utilized. It has been established in [9] that probabilities of misclassifications are minimized when data samples are infinite and when losses are uniform. The scope of the paper is to show that similar results hold true for non uniform losses, and for *finite* databases when noisy replicates are fed infinitly many times in the network. The statements presented are valid for general QEM classifiers independently of the exact form of the learning algorithm.

II. Notation

Assuming the existence of the above mentioned statistical links, the Bayesian solution minimizes a risk function, corresponding to probabilities of misclassification weighted by losses. More precisely the risk takes the form [6]:

$$R = \sum_{i,j=1}^{K} \kappa(i,j)\, P_i \int_{u \in D_j} p(u|\omega_i) du, \tag{2}$$

where $\kappa(i,j)$ denotes the loss associated with the classification in ω_j of a member of class ω_i, D_j is the domain in which patterns are assigned class ω_j, and K is the number of classes. In practice it is not very useful to assign a non-zero loss to patterns correctly classified. Therefore it can be set $\kappa(i,i) = 0$, and the minimization of (2) then simplifies. In this case a vector x will be assigned the class $\omega_{j(x)}$ which minimizes the expression $B_k(x)$ over index k:

$$\omega_{j(x)} \text{ assigned to } x \Leftrightarrow j(x) = Arg \operatorname*{Min}_k B_k(x), \tag{3}$$

$$B_k(x) = \sum_{\substack{1 \leq i \leq K \\ i \neq k}} \kappa(i,k)\, P_i\, p(x|\omega_i). \tag{4}$$

For instance, in the case of uniform losses, $\kappa(i,j) = 1 - \delta_{ij}$, and the minimization of $B_k(x)$ is equivalent to the maximization of:

$$b_k(x) = P_k\, p(x|\omega_k). \tag{5}$$

The Bayesian discriminating rule is generally better known in this latter form. See for instance [9] where Richard and Lippmann discuss this case in detail. In practice, a

Manuscript submitted May 19, 1995; revised April 23, 1996. Both authors are with Thomson-Sintra, BP 157, F-06903 Sophia-Antipolis Cedex, France; E-mail: comon@asm.thomson.fr. P. Comon is also with I3S-CNRS, Sophia-Antipolis, France.

finite learning set $A(N)$ containing N patterns is given. Let $A_k(N) = \omega_k \cap A(N)$ be the learning set for class k, and denote N_k the number of elements it contains. In order to calulate the Bayesian discriminating function (4), one can replace probabilities P_i by their relative frequency of occurrence:

$$\hat{P}_i = \frac{N_i}{N}. \quad (6)$$

Next, the conditional densities $p(x|\omega_i)$ can be estimated by resorting to kernel estimators of density, that have (among others) the remarkable property to be able to deliver estimates being positive, indefinitely differentiable, and of unit sum, regardless of the number of samples available, provided the kernel is appropriately chosen. The kernel estimator is defined as:

$$\hat{p}_x(u|\omega_i) = \frac{1}{N_i} \sum_{x^{(n)} \in A_i(N)} \frac{1}{h(N_i)^d} F(\frac{u - x^{(n)}}{h(N_i)}), \quad (7)$$

where d denotes the dimension of the space where vectors $x^{(n)}$ live, F is the kernel function, and $h(N_i)$ is a width parameter to be determined as a function of N_i. Under fairly mild conditions, it has been proved that this estimator is strongly consistent [3]. The kernel F may be chosen to be a radial function, that is, a function depending only on the norm of its argument, or may not.

This estimator was originally proposed in 1962 by Parzen [8], and Cacoullos [3] extended it a few years later to the multivariate case. The very useful suggestion of a variable width proposed independently by Wagner [12] and Breiman [2] will not be utilized in this paper. Because of their diversity, kernel estimators should not be generically called *Parzen estimators*, as is sometimes the case. Some links can also be emphasized with the so-called radial basis functions independently proposed twenty years later [7].

III. QEM classifiers

In this section, asymptotic performances of QEM-based learning algorithms are investigated and their convergence to Bayes general solution is stated; proofs are postponed to the last section.

The first lemma proves convergence to the Bayesian solution when the number of examples in each class, N_k, tends to infinity. It extends a theorem published by Ruck and Rogers [10], that applied only to uniform losses. Since the MLP is a particular QEM-based learning system, the analysis of its ultimate performance falls in the present framework. On the other hand, the second one proves convergence when the numbers N_k are fixed but when an increasing number of noisy replications are added to the learning set.

A. Infinite samples

Denote $G(u)$ the vector-valued function whose components are:

$$G_i(u) = B_i(u)/p(u), \quad (8)$$

with $B_i(u)$ defined as in (4), and $p(u)$ denoting the density of all observable patterns:

$$p(u) = \sum_{k=1}^{K} P_k \, p(u|\omega_k). \quad (9)$$

Such a $G(u)$ vector is also a Bayesian discriminating rule, since $p(u)$ is a scalar function.

Lemma III.1: If the ith component of the desired output, $y_i^{(n)}$, is chosen to be $\kappa(j,i)$ each time $x^{(n)} \in \omega_j$, then the error (1) converges to (10) when the numbers N_k tend all to infinity:

$$\epsilon(\infty) = \int p(v) \parallel \Phi(W,v) - G(v) \parallel^2 dv + constant. \quad (10)$$

In other words, the optimal value of W defines the best QEM estimate of $G(v)$ in the class of functions of the form $\Phi(W,\cdot)$. If this class is large enough, and if the optimization algorithm is able to reach an acceptable local minimum, the minimal component of $\Phi(W,x)$ will thus have the same index as the minimal one of $G(x)$, i.e., the same class will be assigned to x.

B. Finite samples

Let $p_z(u)$ be a probability density defined on $\mathcal{E}$. For finite N, define now the following estimates:

$$\hat{p}(v|\omega_k) = \frac{1}{N_k} \sum_{x^{(n)} \in A_k(N)} p_z(v - x^{(n)}), \quad (11)$$

$$\hat{p}(u) = \sum_{k=1}^{K} \hat{P}_k \, \hat{p}(u|\omega_k), \quad (12)$$

$$\hat{B}_k(v) = \sum_{j=1}^{K} \kappa(j,k) \hat{P}_j \; \hat{p}(v|\omega_j), \quad (13)$$

$$\hat{G}_k(v) = \frac{\hat{B}_k(v)}{\hat{p}(v)}. \quad (14)$$

Lemma III.2: Assume all N_k are strictly positive, and define a learning set of arbitrariliy large size obtained from $A(N)$ by duplication and noise addition: $A(N,R) = \{x^{(n)} + z^{(n,r)}\}$, where $z^{(n,r)}$ are independent random vectors drawn from a given distribution $p_z(u)$, with $1 \leq n \leq N$ and $1 \leq r \leq R$. Then, if again $y_k^j = \kappa(j,k)$, the error (15) converges to (16) when R tends to infinity:

$$\varepsilon(N,R) = \frac{1}{N}\frac{1}{R} \sum_{n=1}^{N} \sum_{r=1}^{R} \parallel y^{i(n)} - \Phi(W, x^{(n)} + z^{(n,r)}) \parallel^2, \quad (15)$$

$$\varepsilon(N,\infty) = \int \hat{p}(v) \parallel \Phi(W,v) - \hat{G}(v) \parallel^2 dv + constant. \quad (16)$$

C. Application

This last result shows that any QEM solution minimizing (15) tends to a kernel approximate of the Bayesian solution, where conditional densities have been replaced by kernel estimates (11), and with $p_z(u)$ as probability kernel.

This estimate is of the form (7). As a consequence, the noise pdf should satisfy the properties requested for kernel functions [3]. In addition, one should vary the width parameter $h(N_i)$ (which controls monotonically the variance) such that $h(N_i) \to 0$ and $N_i\, h(N_i)^d \to \infty$ for every i as $N_i \to \infty$ [3]. Of course every N_i is finite in practice, but as explained in [4], a good choice is to take $h(N_i)$ proportional to $N_i^{-1/(d+4)}$. This conclusion is important from the practical point of view.

Existing neural network training algorithms are used to minimize the unweighted quadratic error, so that loss terms are not taken into account. Our results show that *the same* algorithms can indeed take into account the losses provided the desired responses are chosen accordingly. For instance, the general Bayesian classifier can be implemented on a MLP and trained with any *standard* QEM algorithm.

D. A simpler case

If desired responses are set instead to $y_i^{(n)} = \delta_{ij}$ when $x^{(n)} \in \omega_j$, then it can be proved similarly [5] that QEM classifiers ultimately yield the same discriminating rule as in (5), where P_k and $p(x|\omega_k)$ are replaced by the same estimates as above. The reasoning is the same (though simpler): it suffices to interchange the roles played by $B_k(x)$ and $b_k(x)$, and to replace the minimization by a maximization. We do not repeat the statements. Note that this has been discussed in [9] in the infinite sample case.

IV. Proofs

A. Infinite samples

The error (1) can be written as a function of the K possible values of the desired response, y^k:

$$\epsilon(N) = \sum_{k=1}^{K} \frac{N_k}{N}\,\frac{1}{N_k} \sum_{x^{(n)} \in A_k(N)} ||y^k - \Phi(W, x^{(n)})||^2. \quad (17)$$

In fact, remind that y^k is the value of the desired response $y^{i(n)}$ when pattern $x^{(n)}$ is is class ω_k. Now let N tend to infinity and get:

$$\epsilon(\infty) = \sum_{k=1}^{K} P_k \int p(u|\omega_k)\, ||y^k - \Phi(W,u)||^2\, du. \quad (18)$$

Now by expanding the squared norm and using definition (9) yields:

$$\begin{aligned} \epsilon(\infty) &= \int p(u)\, ||\Phi(W,u)||^2\, du \\ &\quad -2 \sum_{k=1}^{K} P_k \int p(u|\omega_k) \Phi(W,u)^T y^k\, du + \epsilon_1, \end{aligned}$$

where $(^T)$ denotes transposition, and ϵ_1 is independent of $\Phi(W,\cdot)$. Now since $y_i^k = \kappa(k,i)$, the complete expanded form of the error turns out to be:

$$\begin{aligned} \epsilon(\infty) &= \int p(u)\, ||\Phi(W,u)||^2\, du \\ &\quad -2 \sum_{k=1}^{K}\sum_{i=1}^{K} P_k \int p(u|\omega_k)\Phi_i(W,u)\kappa(k,i)\, du + \epsilon_2, \end{aligned}$$

and finally utilizing the definition (8):

$$\begin{aligned} \epsilon(\infty) &= \int p(u)\, ||\Phi(W,u)||^2\, du \\ &\quad -2 \int p(u) G(u)^T \Phi(W,u)\, du + \epsilon_3, \end{aligned} \quad (19)$$

where ϵ_2 and ϵ_3 are independent of the $\Phi(W,\cdot)$'s. Thus, the error is eventually of the form:

$$\epsilon(\infty) = \int p(u)\, ||G(u) - \Phi(W,u)||^2\, du + \epsilon_4, \quad (20)$$

which proves the lemma.

B. Finite samples

Rewrite the Quadratic Error criterion (15), for finite N and R:

$$\begin{aligned} \varepsilon(N,R) = &\sum_{k=1}^{K} \frac{N_k}{N}\,\frac{1}{N_k} \sum_{x^{(n)} \in A_k(N)} \frac{1}{R}\,\cdot \\ &\sum_{r=1}^{R} ||y^{i(n)} - \Phi(W, x^{(n)} + z^{(n,r)})||^2. \end{aligned} \quad (21)$$

Now denote $\forall u \in \mathcal{E}$:

$$\xi_k(u) \;=\; ||y^k - \Phi(W,u)||^2 \quad \text{for } x^{(n)} \in \omega_k. \quad (22)$$

This is possible since the output $y^{i(n)} = y^k$ when $x^{(n)} \in \omega_k$, and thus depends only on the class label, k. Assume every N_i is non zero and let R tend to infinity. We get:

$$\varepsilon(N,\infty) = \sum_{k=1}^{K} \hat{P}_k \int p_z(u)\, \frac{1}{N_k} \sum_{x^{(n)} \in A_k(N)} \xi_k(x^{(n)} + u)\, du. \quad (23)$$

By making the change of variable $v = x^{(n)} + u$, and using (11), it can be obtained that:

$$\varepsilon(N,\infty) = \sum_{k=1}^{K} \hat{P}_k \int \hat{p}(v|\omega_k)\, \xi_k(v)\, dv. \quad (24)$$

Now with $\hat{G}(u)$ defined as in (14), it appears after a short manipulation that the error can be expressed as:

$$\begin{aligned} \varepsilon(N,\infty) &= \int \hat{p}(v)\, ||\Phi(W,v)||^2\, dv \\ &\quad -2 \int \sum_{k=1}^{K} \hat{G}_k(v)\, \hat{p}(v)\, \Phi_k(W,v)\, dv + \varepsilon_1, \end{aligned} \quad (25)$$

where ε_1 is independent of the Φ_k's. Another manipulation finally leads to:

$$\varepsilon(N,\infty) = \int \hat{p}(v)\, ||\Phi(W,v) - \hat{G}(v)||^2\, dv + \varepsilon_2, \quad (26)$$

where ε_2 is again independent of vector Φ. This last result shows that the mapping $\Phi(W, \cdot)$ obtained is the one closest to $\hat{G}(v)$. Yet, this is an estimate of the Bayesian discriminating functions $G_k(v)$ defined in (8). In other words, if the family of functions $\Phi(W, \cdot)$ is sufficiently large, the smallest $\Phi_k(W; v)$ will be reached for the same k as the smallest $\hat{G}_k(v)$, yielding the same decision.

References

[1] H. BOURLARD, C. WELLEKENS, "Links between Markov models and multilayer perceptrons", in *Advances in Neural Information Processing Systems I*. 1989, pp. 502–510, Morgan Kauffmann.

[2] L. BREIMAN, W. MEISEL, E. PURCELL, "Variable kernel estimates of multivariate densities", *Technometrics*, vol. 19, no. 2, pp. 135–144, May 1977.

[3] T. CACOULLOS, "Estimation of a multivariate density", *Annals of Inst. Stat. Math.*, vol. 18, pp. 178–189, 1966.

[4] P. COMON, "Supervised classification, a probabilistic approach", in *ESANN-European Symposium on Artificial Neural Networks*, Verleysen, Ed., Brussels, Apr 19-21 1995, pp. 111–128, D facto Publ., invited paper.

[5] P. COMON, G. BIENVENU, "Detection et estimation supervisees", in *XVième Colloque Gretsi*, Juan les Pins, 16–20 Sept 1991, pp. 277–280.

[6] R. O. DUDA, P. E. HART, *Pattern Classification and Scene Analysis*, Wiley, 1973.

[7] J. MOODY, C. J. DARKEN, "Fast learning in neural networks of locally-tune processing units", *Neural Computation*, vol. 1, pp. 281–294, 1989.

[8] E. PARZEN, "On the estimation of a probability density function and the mode", *Ann. Math. Stat.*, vol. 33, pp. 1065–1076, 1962.

[9] M. D. RICHARD, R. P. LIPPMANN, "Neural network classifiers estimate Bayesian a posteriori probabilities", *Neural Computation*, vol. 3, pp. 461–483, 1991.

[10] D. W. RUCK, S. K. ROGERS, "The multilayer perceptron as an approximation to Bayes optimal discriminant function", *IEEE Trans. Neural Networks*, vol. 1, pp. 296–298, Dec. 1990.

[11] V. VAPNIK, *Estimation of dependences based on empirical data*, Springer series in statics. Springer, Moscow, 1982.

[12] T. J. WAGNER, "Nonparametric estimates of probability densities", *IEEE Trans. on Inf. Theory*, vol. 21, no. 4, pp. 438–440, July 1975.

[13] H. WHITE, "Learning in artificial neural networks: A statistical perspective", *Neural Computation*, vol. 1, pp. 425–464, 1989.