International Journal of Pure and Applied Mathematics
Volume 59 No. 2 2010, 155-161

ON SUPERVISED AND UNSUPERVISED TRAINING SCHEMES FOR CLASSIFIERS

Eugen Grycko
Department of Mathematics and Computer Science
University of Hagen
125, Lützowstr., Hagen, D-58084, GERMANY
email: eugen.grycko@fernuni-hagen.de

Abstract: A stochastic model for the description of the classification problem is presented. Statistically motivated supervised and unsupervised training schemes for classifiers are considered; the resulting classifiers turn out to be asymptotically optimal. The rates of convergence of probability of successful classification to optimality are studied in a computer experiment. The supervised training scheme entails a sequence of classifiers whose quality converges faster to optimality than that in the unsupervised case.

AMS Subject Classification: 91E40, 62F12, 62G20
Key Words: Bayesian classifier, EM algorithm, consistent estimator

1. Introduction

Learning to classify is an ubiquitous process which can be modelled stochastically. The aim of this learning process can be formalized as the establishment of a good classifier whose quality is measured by the probability of successful guess.

A natural training scheme for a classifier is based on the statistical exploration of the mechanism generating the observations (measurements) as basis for classification; the so trained classifier imitates the optimal one as far as it is possible with the gained statistical information.

Received: December 20, 2009

The aim of the present contribution is the exemplification of a comparative evaluation of supervised/unsupervised training schemes for classifiers requiring the consideration of rates of convergence to optimality.

The paper is organized as follows. In Section 2 a stochastic model formalizing the classification problem is introduced; two training schemes for classifiers are motivated and described based on the introduced stochastic model. In Section 3 we specify a classification model and describe an evaluation strategy for the supervised/unsupervised training schemes introduced in Section 2. The role of the EM algorithm in unsupervised training is also pointed out. In Section 4 we describe and report a computer experiment yielding the rates of convergence of the quality of trained classifiers to optimality for increasing training sample size.

2. The Classification Problem

Let $(X, \mathcal{F})$ be a measurable space. For $j = 0, 1$ let Q_j be a probability measure on $(X, \mathcal{F})$ with density function $f_j : X \to \mathbb{R}_+$ w.r.t. a σ-finite measure μ.

We consider the following two-steps classification model:

Step 1. A non-observable class $J \in \{0, 1\}$ is generated where

$$\text{Prob}(J = 1) = 1 - \text{Prob}(J = 0) =: q \in (0, 1).$$

Step 2. An observable measurement $x \in X$ is generated according to Q_J.

The classification task is now to guess J. A classifier is a measurable function $c : X \to \{0, 1\}$ with the interpretation that class $c(x) \in \{0, 1\}$ is guessed if x is observed in Step 2.

A natural quality criterion for classifiers c is the probability of successful guess:

$$s(c) := (1 - q) \cdot Q_0(c = 0) + q \cdot Q_1(c = 1).$$

If the the stochastic mechanism generating observation x is completely known, then the optimal (Bayesian) classifier $c_B : X \to \{0, 1\}$ is given by

$$c_B(x) := \begin{cases} 0 & \text{if} \quad (1 - q) f_0(x) \geq q f_1(x), \\ 1 & \text{otherwise.} \end{cases}$$

In typical applications neither Q_0, Q_1 nor q are known but the past experience is represented by a training sample $(J_i, x_i)_{i=1}^N$ generated according to the introduced two-steps model.

A supervised training scheme for a classifier $\overline{c}^{(N)}$ is the estimation of f_j

by $\overline{f}_j^{(N)}$ based on the data set $\{x_i|J_i = j\}$ for $j = 0, 1$; parameter q can be estimated by

$$\overline{q}^{(N)} := \frac{1}{N} \cdot \#\{i|J_i = 1\}$$

where $\#S$ denotes the cardinality of set S.

The trained classifier $\overline{c}^{(N)} : X \to \{0, 1\}$ is given by

$$\overline{c}^{(N)}(x) := \begin{cases} 0 & \text{if} \quad (1 - \overline{q}^{(N)})\overline{f}_0^{(N)}(x) \geq \overline{q}^{(N)}\overline{f}_1(x) \\ 1 & \text{otherwise.} \end{cases} \tag{2.1}$$

Since variables J_i are observable in the training sample $(J_i, x_i)_{i=1}^N$, we speak of supervised training of classifier $\overline{c}^{(N)}$.

Remark 2.1. Let us assume in an asymptotic set-up that the applied estimators are strongly consistent:

$$\int_X |\overline{f}_j^{(N)} - f_j| d\mu \to 0 \qquad \text{almost surely for } N \to \infty \qquad (j = 0, 1), \tag{A1}$$

and

$$\overline{q}^{(N)} \to q \qquad \text{almost surely for } N \to \infty. \tag{A2}$$

It follows that the probability $s(\overline{c}^{(N)})$ of successful guess converges almost surely to the optimal value $s(c_B)$ for $N \to \infty$. Moreover,

$$\mathbb{E}(s(\overline{c}^{(N)})) \to s(c_B) \tag{2.2}$$

holds for $N \to \infty$ where $\mathbb{E}$ denotes the expected value.

Let us now suppose that the training sample is given by $(x_i)_{i=1}^N$ where x_i is generated according to the mixture

$$(1 - q)Q_0 + qQ_1. \tag{2.3}$$

Estimating f_0, f_1 and q in this situation is called unsupervised training of classifier $\widehat{c}^{(N)} : X \to \{0, 1\}$ which is given by

$$\widehat{c}^{(N)}(x) := \begin{cases} 0 & \text{if } (1 - \widehat{q}^{(N)})\widehat{f}_0^{(N)} \geq \widehat{q}\widehat{f_1}^{(N)}(x), \\ 1 & \text{otherwise,} \end{cases} \tag{2.4}$$

where $\widehat{f}_0^{(N)}, \widehat{f}_1^{(N)}, \widehat{q}^{(N)}$ are the corresponding estimators based on the sample $(x_i)_{i=1}^N$.

Analogously to Remark 2.1, if estimators $\widehat{f}_0^{(N)}, \widehat{f}_1^{(N)}, \widehat{q}^{(N)}$ are strongly consistent in the sense of (A1) and (A2), then

$$s(\widehat{c}^{(N)}) \to s(c_B) \text{ almost surely for } N \to \infty$$

and

$$\mathbb{E}(s(\widehat{c}^{(N)})) \to s(c_B) \qquad \text{for } N \to \infty. \tag{2.5}$$

At first sight, it might be quite surprising that a consistent unsupervised estimation is possible at all; in Section 3 we consider, however, an example where it really is if the class allowed for the (unknown) probability measures Q_0 and Q_1 is appropriately restricted. Note that strongly consistent estimation of densities f_0, f_1 is also possible in a nonparametric context (cf. [1] and [3]).

3. Two Training Schemes for Classifiers

Let us specify $(X, \mathcal{F}) := (\mathbb{R}, \mathcal{B})$, where $\mathcal{B}$ denotes the Borel σ-field. Let us, moreover, assume that

$$Q_j = N(a_j, 1) \qquad (j = 0, 1),$$

where $N(a, \sigma^2)$ denotes the normal distribution with mean a and variance σ^2 and $a_0 < a_1$.

We consider a training sample $(J_i, x_i)_{i=1}^N$, where

$$\text{Prob}(J_i = 1) =: q = \frac{1}{2} \qquad (i = 1, \ldots, N).$$

Put

$$N_j := \#\{i | J_i = j\} \qquad (j = 0, 1)$$

and

$$\overline{a}_j := \frac{1}{N_j} \cdot \sum_{J_i = j} x_i \qquad (j = 0, 1).$$

A natural construction of a classifier $\overline{c}^{(N)}$ based on sample $(J_i, x_i)_{i=1}^N$ is given in (2.1), where

$$\overline{f}_j^{(N)}(y) = \frac{1}{\sqrt{2\pi}} \cdot \exp\left(-\frac{1}{2}(y - \overline{a}_j)^2\right) \qquad (j = 0, 1)$$

and

$$\overline{q}^{(N)} := \frac{N_1}{N}.$$

The probability $s(\overline{c}^{(N)})$ of successful guess can be determined numerically by 1-dimensional integration. Note that $s(\overline{c}^{(N)})$ is in some sense random because it depends on the randomly generated training sample $(J_i, x_i)_{i=1}^n$; the expected value $\mathbb{E}(s(\overline{c}^{(N)}))$ can be interpreted as a characteristic number expressing the efficiency of the described supervised training scheme yielding classifier $\overline{c}^{(N)}$.

For the construction of a classifier $\widehat{c}^{(N)}$ based on sample $(x_i)_{i=1}^N$ (unsupervised case) we consider the family

$$g_{q,a_0,a_1}(x) = \prod_{i=1}^{N}((1-q)\varphi_{a_0}(x_i) + q\varphi_{a_1}(x_i)) \qquad (a_0 < a_1, q \in (0,1)) \tag{3.1}$$

of probability densities corresponding to the mixture model

$$\otimes_{i=1}^{N}((1-q)N(a_0,1) + qN(a_1,1)),$$

where

$$\varphi_a(y) = \frac{1}{\sqrt{2\pi}} \cdot \exp\left(-\frac{1}{2}(y-a)^2\right)$$

denotes the density of $N(a,1)$.

Let $(\widehat{q}, \widehat{a}_0, \widehat{a}_1)$ denote the maximum likelihood estimate of (q, a_0, a_1) in the model (3.1). This estimate can be efficiently computed by application of the EM algorithm (cf. [2], p. 238). A natural classifier $\widehat{c}^{(N)}$ based on $(x_i)_{i=1}^N$ is given by (2.4), where

$$\widehat{f}_j^{(N)}(y) = \varphi_{\widehat{a}_j}(y) \qquad (j = 0, 1)$$

and

$$\widehat{q}^{(N)} := \widehat{q}.$$

Analogously to the supervised case, it is possible to determine $s(\widehat{c}^{(N)})$ numerically by 1-dimensional integration. Again, the expectation $\mathbb{E}(s(\widehat{c}^{(N)}))$ expresses the quality of the unsupervised training scheme yielding classifier $\widehat{c}^{(N)}$.

Since all estimators involved in the construction of classifiers $\overline{c}^{(N)}$ and $\widehat{c}^{(N)}$ are consistent, we have

$$\lim_{N\to\infty} \mathbb{E}(s(\overline{c}^{(N)})) = \lim_{N\to\infty} \mathbb{E}(s(\widehat{c}^{(N)})) = s(c_B), \tag{3.2}$$

where c_B denotes the Bayesian classifier, cf. Remark 2.1.

In the sense of (3.2) both training schemes are asymptotically optimal; to compare the quality of the supervised/unsupervised training schemes we have to study the convergence rates in (3.2) which will be done in Section 4.

4. The Computer Experiment and its Outcome

We describe a long term computer experiment and report its outcome.

We consider the classification model specified in Section 3 where

$$a_0 := 0 \qquad \text{and} \qquad a_1 := 1$$

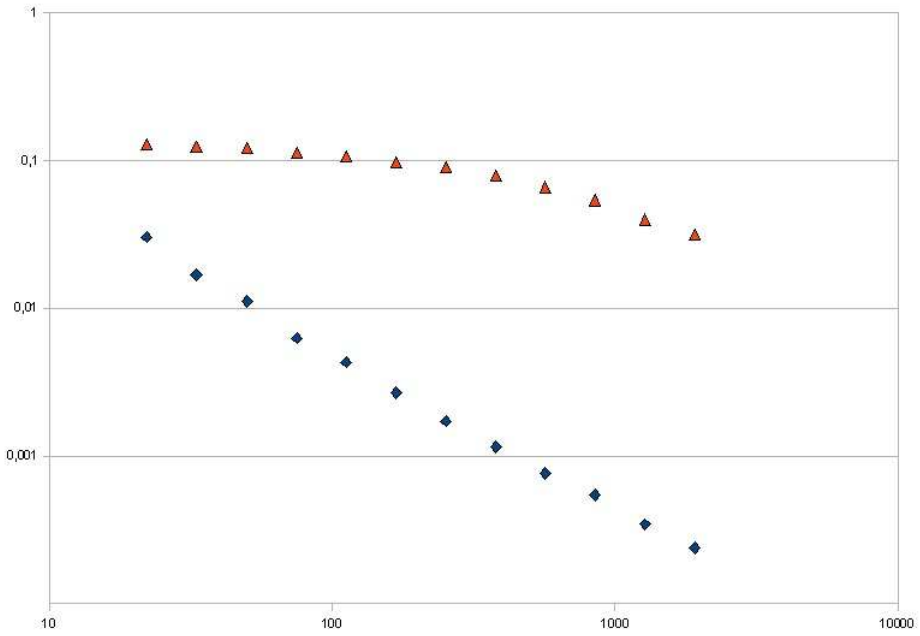

Figure 1: Convergence of the classifiers to optimality

are fixed. For a fixed training sample size N the following procedure is carried out.

Step 1. Generate a supervised training sample $(J_i, x_i)_{i=1}^N$ of size N distributed according to the mixture

$$\frac{1}{2}N(a_0, 1) + \frac{1}{2}N(a_1, 1).$$

Step 2. Estimate a_j by $\overline{a}_j$ and $q = 1/2$ by $\overline{q}^{(N)}$ from the sample according to Section 3.

Step 3. Construct classifier $\overline{c}^{(N)}$ and compute numerically probability $s(\overline{c}^{(N)})$ of successful guess.

Step 4. Forget the components J_i in the sample.

Step 5. Determine estimates $\widehat{a}_j$ and $\widehat{q}^{(N)}$ from the sample $(x_i)_{i=1}^N$ according to Section 3.

Step 6. Construct classifier $\widehat{c}^{(N)}$ and compute numerically probability $s(\widehat{c}^{(N)})$ of successful guess.

Step 7. Repeat Steps 1–6 1000 times to estimate the expectations $\mathbb{E}(s(\overline{c}^{(N)}))$

and $\mathbb{E}(s(\widehat{c}^{(N)}))$ by the averages of the probabilities of successful guess.

Step 8. Compute numerically $s(c_B)$.

In the long term computer experiment Steps 1–8 have been repeated where the sample size N has been varied.

In Figure 1 the horizontal axis corresponds to the sample size N and the vertical axis to $s(c_B)$ minus expected probability of successful guess. The diamonds and triangles represent the supervised and unsupervised training scheme, respectively. The logarithmic plots presented in Figure 1 confirm the convergence of the quality of the trained classifiers to optimality in the sense of (3.2). Figure 1 also reveals the fact that the convergence to optimality for the supervised training scheme is faster than for the unsupervised one.

Acknowledgments

The author would like to thank Professor Werner Kirsch from Hagen for encouragement concerning this contribution. The author also appreciates the technical support by Jens Rentmeister from Kierspe/Germany in connection with the graphical presentation of the computer experimental data.

References

[1] L. Devroye, L. Györfi, *Nonparametric Density Estimation*, John Wiley and Sons, Ney York (1985).

[2] T. Hastie, R. Tibshirani, J. Friedman, *The Elements of Statistical Learning*, Springer, New York (2001).

[3] B.W. Silverman, *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, London (1996).