LEARNING TO RECOGNIZE PATTERNS WITH A PROBABILISTIC TEACHER

Włodzimierz Greblicki
Institute of Engineering Cybernetics, Technical University of Wrocław,
Wyb. Wyspiańskiego 27, 50-370 Wrocław, Poland

printed in:
Pattern Recognition
Vol. 12, No. 3, pp. 154-164, 1980

Abstract - A problem of learning with a probabilistic teacher is considered. Neither prior class probabilities nor class densities are assumed to be known and pattern recognition procedures are derived from nonparametric density and regression estimates. Weak and strong Bayes risk consistency of the procedures is shown. Examples of procedures using the kernel, the nearest neighbor and the orthogonal series estimates are given.

Pattern recognition Classification Learning Probabilistic teacher Imperfect teacher Density estimation Regression estimation Nonparametric estimation.

1. INTRODUCTION

We consider a problem of learning with the help of a probabilistic teacher who labels observations coming from M classes. As the teacher is probabilistic his labels can differ from true classes. Learning with a probabilistic teacher was studied by a few authors. Parametric problems were examined by Shanmugam$^{(14)}$ and Gimlin,$^{(5)}$ while Agrawala$^{(1)}$ as well as Imai and Shimura$^{(10)}$ applied the idea of probabilistic labelling in unsupervised learning. A problem close to that considered in this paper, i.e. nonparametric learning, was studied by Shanmugam and Breipol$^{(15)}$. These authors employed the Rosenblatt-Parzen density estimate in order to estimate nonoverlapping class densities, whereas in our paper any constraint on class densities is imposed except for either continuity almost everywhere or square integrability. The usage of nonparametric density estimators in supervised learning was suggested by Van Ryzin$^{(19,20)}$ and developed by Wolverton and Wagner,$^{(21)}$ Glick$^{(6)}$ and Greblicki$^{(7)}$ who gave general theorems on Bayes risk consistency. In the present paper, however, pattern recognition procedures are derived not only from density but also from nonparametric regression estimators. It appears that under some very general conditions procedures obtained in this way are consistent in Bayes risk.

2. PRELIMINARIES

Let (ω, X) be a pair of random variables, where ω is $\Omega = \{1, \ldots, M\}$ valued and X is R^P valued. Elements of Ω will be called classes. Let $P\{\omega = i\} = p_i$ and let f_i be the class conditional density of X, i.e., the appropriate Radon-Nikodym derivative with respect to the Lebesgue measure μ on R^P. Let $L(i,j)$ be the loss incurred by taking action $i \in \Omega$. when the class is j. In this paper $L(i,j) = 0$ for $i = j$ and $L(i,j) = 1$ for $i \neq j$. For a classification function ψ, i.e. for a Borel function on R^P to Ω, the risk is

$$R(\psi) = \sum_{i=1}^{M} p_j \int L(\psi(x), j) f_j(x) d\mu(x)$$

and for our loss function it equals the probability of misclassification. It is well known that the Bayes classification function classifies almost every (μ) $x \in R^p$ to any class i for which

$$p_i f_i(x) = \max_j p_j f_j(x). \tag{1}$$

Denote by R_0 the Bayes risk, i.e. the risk for the Bayes classification function.

We assume that neither prior class probabilities p_i's nor class densities f_i's are known and we estimate them from observations accompanied by teacher labels. The teacher is represented by an Ω valued random variable Θ. The problem is described by a triple (ω, Θ, X), where ω is the true class and Θ is the teacher label. Let the teacher's errors be independent of observations, i.e., let

$$P\{\Theta = i/\omega = j, X \in B\} = P\{\Theta = i/\omega = j\} = p_{ij},$$

for all Borel sets $B \in R^p$. We assume that all p_{ij}'s are known. Moreover, the conditional density of X, given both $\omega = i$ and $\Theta = j$ is independent of j.

Let now

$$(\omega_1, \Theta_1, X_1), ..., (\omega_n, \Theta_n, X_n)$$

be a sequence of independent triples distributed like (ω, Θ, X). True classes are, however, unknown and we have only

$$(\Theta_1, X_1), ..., (\Theta_n, X_n). \tag{2}$$

Any Ω valued Borel function ψ_n defined for all realizations of sequence (2) is called an empirical decision function, while a sequence $\{\psi_n\}, n = 1, \ldots, n$, is called a pattern recognition procedure.

Definition

A pattern recognition procedure $\{\psi_n\}$ is said to be Bayes risk weakly or strongly consistent, if

$$R(\psi_n) \underset{a.s.}{\overset{p}{\rightarrow}} R_0$$

as $n \to \infty$, respectively.

3. BAYES RISK CONSISTENCY OF PROCEDURES USING DENSITY ESTIMATES

It will be useful to introduce the notions of the teacher class probabilities and the teacher class densities which are the appropriate distributions suggested by the teacher. $P\{\Theta = i\} = p_i^t$ and the conditional density of X, given $\Theta = i$, denoted by f_i^t will be called the teacher class probability and the teacher class density, respectively. It is easy to see that, for $i = 1, ..., M$,

$$f_i^t(x) = \sum_{j=1}^{M} p_{ij} p_j f_j(x) \Bigg/ \sum_{j=1}^{M} p_{ij} p_j \tag{3}$$

and

$$p_i^t = \sum_{j=1}^{M} p_{ij} p_j. \tag{4}$$

Denoting $[p_{ij}] = P$, assuming that P is nonsingular and using (3) and (4) we get

$$f_i(x) = \sum_{j=1}^{M} q_{ij} p_j^t f_j^t(x) \Bigg/ \sum_{j=1}^{M} q_{ij} p_j^t \tag{5}$$

and

$$p_i = \sum_{j=1}^{M} q_{ij} p_j^t, \tag{6}$$

for $i = 1, ..., M$, where $[q_{ij}] = P^{-1}$. Finally, by (5) and (6),

$$p_i f_i(x) = \sum_{j=1}^{M} q_{ij} p_j^t f_j^t(x). \tag{7}$$

The teacher class probability p_i^t is estimated by the $\hat{p}_i^t = N_i/n$, where N_i is the number of observations assigned by the teacher to the class i. Let

$$\hat{p}_i = \sum_{j=1}^{M} q_{ij} \hat{p}_i^t \tag{8}$$

be an estimate of the prior probability of the class i. Obviously

$$\hat{p}_i \overset{a.s.}{\rightarrow} p_i$$

as $n \to \infty$, for all i. In order to estimate the teacher class densities the observations $X_1, ..., X_n$ can be partitioned into M subsequences

$$(X_1^1, ..., X_{N_1}^1,), ..., (X_1^M, ..., X_{N_M}^M)$$

of observations recognized by the teacher as coming from particular classes. Let $\hat{f}_i^t(x)$ be an estimate of $f_i^t(x)$calculated from the ith subsequence, and let

$$\hat{f}_i(x) = \sum_{j=1}^{M} q_{ij} \hat{p}_j^t \hat{f}_j^t(x) \Bigg/ \sum_{j=1}^{M} q_{ij} \hat{p}_j^t \tag{9}$$

be an estimate of $f_i(x)$. Therefore

$$\hat{p}_i(x) \hat{f}_i(x) = \sum_{j=1}^{M} q_{ij} \hat{p}_j^t \hat{f}_j^t(x). \tag{10}$$

In this section we consider a class D of all pattern recognition procedures which, for $n = 1, 2, \ldots$, classify almost every (μ) $x \in R^p$ among any class i for which

$$\sum_{j=1}^{M} q_{ij} \hat{p}_j^t \hat{f}_j^t(x)$$

is maximal. In order to establish Bayes risk consistency of so defined procedures we shall use the following theorem:

Theorem 1

If, for $i = 1, \ldots, M$,

$$\hat{p}_i \underset{a.s.}{\overset{p}{\rightarrow}} p_i$$

as $n \to \infty$ and

$$\hat{f}_i(x) \underset{a.s.}{\overset{p}{\rightarrow}} f_i(x)$$

as $n \to \infty$ at almost every (μ) $x \in R^p$, then every D-procedure is Bayes risk weakly or strongly consistent, respectively.

The proof follows immediately from Greblicki's[7] Theorem 2.

By Theorem 1, (5) and (8) we get

Theorem 2

If, for $i = 1, \ldots, M$,

$$\hat{f}_i^t(x) \underset{a.s.}{\overset{p}{\rightarrow}} f_i^t(x)$$

as $n \to \infty$ at almost every (μ) $x \in R^p$, then every D-procedure is Bayes risk weakly or strongly consistent, respectively.

Observe that the estimate $\hat{f}_i^t$ is calculated from the random number N_i of observations. Using Greblicki's[8] arguments, we can show that if a density estimate employed by a D-procedure is weakly consistent at a point $x \in R^P$, i.e. if, for every $i = 1, \ldots, , M$, and every $\varepsilon > 0$,

$$P\{|\hat{f}_i^t(x) - f_i^t(x)| < \varepsilon / N_i > N\} \to 1$$

as $N \to \infty$, then

$$\hat{f}_i^t(x) \xrightarrow{p} f_i^t(x)$$

as $n \to \infty$. Hence, by virtue of Theorem 2, every D-procedure is Bayes risk weakly consistent, Similarly, if the density estimate is strongly consistent at a point $x \in R^P$, then

$$\hat{f}_i^t(x) \xrightarrow{a.s} f_i^t(x)$$

as $n \to \infty$, and consequently every D-procedure is Bayes risk strongly consistent. Thus, we have shown

Theorem 3

If the density estimate is weakly or strongly pointwise consistent at almost every (μ) $x \in R^p$, then every D-procedure is Bayes risk weakly or strongly consistent, respectively.

For an integrated uniformly consistent estimate, [for terminology see e.g. Wegman[18]] we employ the Wolverton-Wagner[21] Theorem 3.

Theorem 4

If, for $i = 1, ..., M$,

$$\hat{p}_i \underset{a.s.}{\xrightarrow{p}} p_i$$

and

$$\int [\hat{f}_i(x) - f_i(x)]^2 d\mu(x) \underset{a.s.}{\xrightarrow{p}} 0$$

as $n \to \infty$, then every D-procedure is Bayes risk weakly or strongly consistent, respectively.

The next two theorems follow from Theorem 4, (5) and (8).

Theorem 5

If, for $i = 1, ..., M$,

$$\int [\hat{f}_i^t(x) - f_i^t(x)]^2 d\mu(x) \underset{a.s.}{\xrightarrow{p}} 0$$

as $n \to \infty$, then every D-procedure is Bayes risk weakly or strongly consistent, respectively.

Theorem 6

If the density estimate is integrated uniformly weakly or strongly consistent, then every D-procedure is Bayes risk weakly or strongly consistent, respectively.

4. BAYES RISK CONSISTENCY OF PROCEDURES WITH REGRESSION ESTIMATES

Classical procedures for learning with the perfect teacher employ density estimates, Van Ryzin[19,20] used the Rosenblatt-Parzen as well as the orthogonal series estimates, whereas Greblicki[8] applied also the Loftsgaarden-Quesenberry one and studied, moreover, modifications of class density estimators. In this section we derive, like Stone,[l7] pattern recognition procedures from regression estimates.

Let λ_i be the indicator of class i, i.e. let

$$\lambda_i = \begin{cases} 1 & \text{for } \omega = i \\ 0 & \text{otherwise} \end{cases}$$

and let λ_i^t be the teacher indicator of class i, i.e., let

$$\lambda_i^t = \begin{cases} 1 & \text{for } \Theta = i \\ 0 & \text{otherwise.} \end{cases}$$

Define the following regressions:

$$r_i(x) = E\{\lambda_i / X = x\},$$

and

$$r_i^t(x) = E\{\lambda_i^t / X = x\},$$

$i = 1, \ldots, M$. Obviously

$$r_i^t(x) = \sum_{j=1}^{M} p_{ij} r_j(x), \tag{11}$$

and consequently

$$r_i(x) = \sum_{j=1}^{M} q_{ij} r_j^t(x), \tag{12}$$

if P is nonsingular.

Bayes classification function (1) can be rewritten in the following way: classify almost every (μ) $x \in R^P$ to any class i for which

$$r_i(x) = \max_j r_j(x). \tag{13}$$

That is why we study procedures derived from regression estimates.

We say that $\{\psi_n\}$ is an R-procedure if, for $n = 1, 2, \ldots$, recognizes almost every (μ) $x \in R^P$ as coming from any class i for which

$$\sum_{j=1}^{M} q_{ij} r_j^t(x) \tag{14}$$

is maximal.

The next theorem is a consequence of Greblicki's[7] Theorem 2.

Theorem 7

If, $i = 1, \dots, M$,

$$\hat{r}_i(x) \underset{a.s.}{\xrightarrow{p}} r_i(x)$$

as $n \to \infty$, almost everywhere (μ) in R^P, then every R-procedure is Bayes risk weakly or strongly consistent, respectively.

Hence, by (12) we get

Theorem 8

If the regression estimate is pointwise weakly or 1 strongly consistent almost everywhere (μ) in R^P, then , every R-procedure is Bayes risk weakly or strongly consistent,respectively.

Observe that $r_i(x) = g_i(x)/f(x)$ and $r_i^t(x) = g_i^t(x)/f(x)$, where $g_i(x) = p_i f_i(x)$, $g_i^t(x) = p_i f_i^t(x)$ and

$$f(x) = \sum_{i=1}^{M} p_i f_i(x)$$

For this reason regression estimates are usually of the following form:

$$\hat{r}_i^t(x) = \hat{g}_i^t(x)/\hat{f}(x),$$

where $\hat{g}_i^t(x)$ and $\hat{f}(x)$ are estimators of $g_i^t(x)$ and $f(x)$, respectively. Defining

$$\hat{g}_i(x) = \sum_{i=1}^{M} q_{ij} g_j^t(x), \tag{15}$$

we get the following, equivalent to (13), definition of R-procedures: classify almost every (μ) $x \in R^p$ among any class i for which

$$\sum_{i=1}^{M} q_{ij} \hat{g}_j^t(x) \tag{16}$$

is maximal. Now, using Wolverton-Wagner's Theorem 4 and (12) we easily get

Theorem 9

If, $i = 1, \dots, M$,

$$\int [\hat{g}_i(x) - g_i(x)]^2 \, d\mu(x) \underset{a.s.}{\xrightarrow{p}} 0$$

as $n \to \infty$, then every R-procedure is Bayes weakly or strongly consistent, respectively.

5. PATTERN RECOGNITION PROCEDURES

We shall give now a few examples of procedures derived from the most popular nonparametric estimators of density and regression. We shall apply the Rosenblatt-Parzen, the Loftsgaarden-Quesenberry and the orthogonal series density estimates as well as regression estimates associated with them. Rosenblatt[13] introduced the following density estimate:

$$\hat{f}_i^t(x) = \frac{1}{N_i h^p(N_i)} \sum_{k=1}^{N_i} K\left[\frac{x - X_k^i}{h(N_i)}\right], \tag{17}$$

while Parzen[12] studied its asymptotic properties. If a bounded and non-negative Borel kernel K satisfies the following conditions:

$$\int K(x) d\mu(x) = 1, \quad \lim_{||x|| \to \infty} ||x||^p K(x) = 0 \tag{18}$$

and, moreover,

$$h(n) \to 0, \; nh^p(n) \to \infty \tag{19}$$

as $n \to \infty$, then the estimator is weakly consistent at every continuity point of f_i^t, for $p = 1$ see Parzen,[l2] for multivariate case see Cacoullos.[2] Using this estimate we get a procedure which assigns almost every (μ) $x \in R^P$ to any class i for which

$$\sum_{j=1}^{M} q_{ij} \sum_{k=1}^{N_i} K\left[\frac{x - X_k^i}{h(N_i)}\right] \tag{20}$$

is maximal.

Nadaraya[11] introduced the following kernel estimate of regression:

$$\hat{r}_i^t(x) = \sum_{k=1}^{n} \lambda_i^t K\left[\frac{x - X_k}{h(n)}\right] \Bigg/ \sum_{k=1}^{n} K\left[\frac{x - X_k}{h(n)}\right].$$

Using the same arguments as Parzen[l2] and Cacoullos[2] it can be easily shown that, under (18) and (19), this estimate is weakly consistent at every point of continuity of g_i^t and f, i.e., at every point at which all the class densities are continuous. In this way we have obtained a procedure which recognizes almost every (μ) $x \in R^p$ as coming from any class i for which

$$\sum_{j=1}^{M} q_{ij} \sum_{k=1}^{n} \lambda_i^t K\left[\frac{x - X_k}{h(n)}\right]$$

i.e.

$$\sum_{j=1}^{M} q_{ij} \sum_{k=1}^{N_j} K\left[\frac{x - X_k^j}{h(n)}\right] \tag{21}$$

is maximal. By Theorems 6 and 8 the procedures (20) and (21) are Bayes risk weakly consistent if all the class densities are almost everywhere (μ) continuous.

Loftsgaarden and Quesenberry[9] gave the following density estimate:

$$\hat{f}_i^t(x) = k(N_i)/V N_i R_i^p[x, k(N_i)],$$

where V is the Lebesgue measure of a set of all $x \in R^P$ for which $||x|| \leq 1$ and $R_i(k, x)$ is the distance from x to the kth nearest observation of the ith subsequence, and have shown that

$$k(n) \to \infty,\ k(n)/n \to 0$$

as $n \to \infty$ implies its weak consistency at every continuity point of f_i^t. This estimate leads to a procedure which classifies almost every (μ) $x \in R^p$ among any class i for which

$$\sum_{j=1}^{M} q_{ij} k(N_j)/R_j^p[x, k(N_j)]$$

is maximal.

A nearest neighbor estimate of $r_i^t(x)$ is of the following form:

$$r_i^t(x) = N_i[x, k(n)]/k(n),$$

where $N_i(x, k)$ is the number of observations from the ith subsequence among k nearest neighbors to x, see Devroye.(4) An appropriate pattern recognition procedure classifies almost every (μ) $x \in R^p$ to any class i which maximizes

$$\sum_{j=1}^{M} q_{ij} N_j[x, k(n)].$$

Bayes risk consistency of such a procedure is due to Stone,(17) who, moreover, did not impose any constraints on class distributions.

The next type of procedure presented herein is based on orthogonal expansions. If all the class densities are square integrable, all the teacher class densities are also square integrable and can be estimated in the way suggested by Čencov.(3) Let

$$f_i^t(x) = N_i^{-1} \sum_{m=0}^{q(N_i)} \sum_{k=1}^{N_i} \varphi_m(X_k^i)\varphi_m(x)$$

where $\{\varphi_m\}$, $m = 0, 1, 2, \ldots$ is a complete set of orthonormal and jointly bounded functions. Schwartz(16) showed that, if

$$q(n) \to \infty,\ q(n)/n \to 0 \tag{22}$$

as $n \to \infty$, then the estimate is integrated uniformly weakly consistent. The above estimate leads to a procedure which uses the following discrimination functions:

$$\sum_{j=1}^{M} q_{ij} \sum_{m=0}^{q(N_j)} \sum_{k=1}^{N_j} \varphi_m(X_k^j)\varphi_m(x) \tag{23}$$

$j = 1, \ldots, M$.

Consider the following estimate of the nominator $g_i^t(x)$ of the regression $r_i^t(x)$:

$$\hat{g}_i^t(x) = n^{-1} \sum_{m=0}^{q(n)} \sum_{k=1}^{n} \lambda_{ik}^t \varphi_m(X_k)\varphi_m(x),$$

where

$$\lambda_{ik}^t = \begin{cases} 1 & \text{for } \Theta_k = i \\ 0 & \text{otherwise.} \end{cases}$$

Using the same method as Schwartz(16) one can show that

$$\int [\hat{g}_i^t(x) - g_i^t(x)]^2 d\mu(x) \overset{n}{\to} 0$$

as $n \to \infty$, provided that (22) is satisfied. Thus we have obtained a procedure which assigns almost every (μ) $x \in R^p$ to any class i which maximizes

$$\sum_{j=1}^{M} q_{ij} \sum_{m=0}^{q(n)} \sum_{k=1}^{n} \lambda_{ik}^t \varphi_m(X_k)\varphi_m(x). \tag{24}$$

By Theorems 6 and 9, the procedures (23) and (24) are Bayes risk consistent if all class densities are square integrable.

Notice that if a procedure $\{\psi_n\}$ is Bayes risk weakly consistent, then

$$ER(\psi_n) \to R_0$$

as $n \to \infty$. Obviously, the probability of the teacher error is

$$R_t = \sum_{i=1}^{M} (1 - p_{ii})p_i$$

and, if $R_t > R_0$, then, for sufficiently large n, every Bayes risk consistent procedure recognizes better than the teacher. This fact was observed empirically by Shanmugan and Breipol.(15)

6. SOME GENERALIZATIONS

Assuming that P is known and nonsingular we managed to calculate f_is and r_is from f_i^ts and r_i^ts, respectively. Generalizing Shanmugam-Breipol's(15) result for two-way classification it can be shown that in some cases Bayes risk consistent procedures may be obtained even if P is unknown. Let, for example, $p_{ii} = \alpha$ for $i = l, \ldots, M$, and let moreover the teacher's errors be distributed uniformly, i.e. let $p_{ij} = (1 - \alpha)/(M - 1)$ for $i \neq j$. Let α be unknown except for the fact that $\alpha > 1/M$. Now

$$p_i f_i(x) - p_j f_j(x) = c[p_i^t f_i^t(x) - p_j^t f_j^t(x)],$$

where c is a positive pointwise constant independent of both i and j. Thus, Bayes classification functions defined by (1) are equivalent to those which assign almost every (μ) $x \in R^P$ to any class i for which

$$p_i^t f_i^t(x) = \max_j p_j^t f_j^t(x).$$

In this case learning procedures given in the paper can be rewritten in considerably simpler forms, which, from computational viewpoint, are the same as the appropriate procedures for learning with the perfect teacher, see e.g., Greblicki.(8)

We would like also to remark that the results of the paper can be easily extended on learning with many probabilistic teachers. Moreover, Bayes risk consistent procedures can be obtained even if all the matrices, say $P_1, \ldots, P_N$, of particular teachers are singular, provided that the rank of $[P_1, \ldots, P_N]$ equals the number of classes.

7. CONCLUSIONS

In the paper we have assumed that the teacher is probabilistic, i.e. that his labels attached to the pat terns used for learning can be incorrect. We have derived pattern recognition procedures from nonparametric estimates of a density and a regression function. The procedures appear to be Bayes risk consistent despite the fact that the teacher is imperfect. This result has been obtained by using general theorems on Bayes risk consistency given by Wolverton and Wagner(21) and Greblicki(7) in the context of learning with the perfect teacher. Moreover, in some cases, the learning procedures can recognize better than the teacher.

SUMMARY

A multicategory problem of learning with a probabilistic teacher is considered. Neither prior class probabilities nor class conditional densities are known, They are estimated from observations labelled by the teacher. The labels can be different from true classes. Pattern recognition procedures are derived from nonparametric estimates of a density and a regression, Bayes risk consistency of the procedures is shown and examples of procedures using the kernel, the nearest neighbor and the orthogonal series estimates are given.

References

[1] A. K. Agrawala, Learning with a probabilistic teacher, *IEEE Trans. Information Theory* **IT-16**, 373 (1970).

[2] T. Cacoullos, Estimation of a multivariate density, *Ann. Inst. Statist. Math.* **18,** 179 (1965).

[3] N. N. Čencov, Evaluation of an unknown distribution density from observations, *Sovet. Math.* **3**, 1559 (1962).

[4] L. P. Devroye, The uniform convergence of nearest neighbor regression function estimators and their application in optimization, *IEEE Trans. Information Theory* **IT-24**, 142 (1978).

[5] D. R. Gimlin, A parametric procedure for imperfectly supervised learning with unknown class probabilities, *IEEE Trans. Information Theory* **IT-20**, 661 (1974).

[6] N. Glick, Sample-based classification procedures derived from density estimators, *J. Am. Statist. Ass.* **67**, 116 (1972).

[7] W. Greblicki Asymptotically optimal pattern recognition procedures with density estimates, *IEEE Trans. Information Theory* **IT-24**, 250 (1978).

[8] W. Greblicki, Pattern recognition procedures with nonparametric density estimates, *IEEE Trans. Systems, Man and Cybernet.* **SMC-8**, 809 (1978).

[9] O. Loftsgaarden, C. P. Quesenberry, A nonparametric estimation of multivariate density function, *Ann. Math. Statist.* **36**, 1049 (1965).

[10] T. Imai, M. Shimura, Learning with probabilistic labelling, *Pattern Recognition* **8**, 5 (1976).

[11] E. A. Nadaraya, On estimating regression, *Theory Prob. Appl.* **9**, 141 (1964).

[12] E. Parzen, On estimation of a probability density function and mode, *Ann. Math. Statist.* **33**, 1065 (1962).

[13] M. Rosenblatt, Remarks on some estimates of a density function, Ann. Math. Statist. **27**, 823 (1956).

[14] K. Shanmugam, A parametric procedure for learningwith an imperfect teacher, *IEEE Trans. Information Theory* **IT-18**, 300 (1972).

[15] K. Shanmugam, A. M. Breipol, An error correcting procedure for learning with an imperfect teacher, *IEEE Trans. Systems, Man and Cybernet.* **SMC-1**, 223 (1971).

[16] S. C. Schwartz, Estimation of probability density by an orthogonal series, *Ann. Math. Statist.* **38**, 1261 (1967).

[17] C. J. Stone, Consistent nonparametric regression, *Ann. Statist.* **5**, 595 (1977).

[18] E. J. Wegman, Nonparametric probability density esti mation; I. A summary of available methods, *Technometrics* **14**, 533 (1972).

[19] J. Van Ryzin, Non-parametric Bayesian decision procedures for (pattern) classification with stochastic learning, *Trans. of 4th Prague Conf. on Information Theory, Statistical Decision Functions and Random Processes*, Prague (1965).

[20] J. Van Ryzin, Bayes risk consistency of classification procedures using density estimation, *Sankhyā* **28**, 261 (1966).

[21] C. T. Wolverton, T. J. Wagner, Asymptotically optimal discriminant functions for pattern classification, *IEEE Trans. Information Theory* **IT-15**, 258 (1969).