Automatic Pattern Recognition: A Study of the Probability of Error

LUC DEVROYE

Abstract—**A test sequence is used to select the best rule from a rich class of discrimination rules defined in terms of the training sequence. The Vapnik–Chervonenkis and related inequalities are used to obtain distribution-free bounds on the difference between the probability of error of the selected rule and the probability of error of the best rule in the given class. The bounds are used to prove the consistency and asymptotic optimality for several popular classes, including linear discriminators, nearest neighbor rules, kernel-based rules, histogram rules, binary tree classifiers, and Fourier series classifiers. In particular, the method can be used to choose the smoothing parameter in kernel-based rules, to choose k in the k-nearest neighbor rule, and to choose between parametric and nonparametric rules.**

Index Terms—**Automatic parameter selection, empirical risk, error estimation, nonparametric rule, probability of error, statistical pattern recognition, Vapnik–Chervonenkis inequality.**

I. Introduction

In pattern recognition, we normally use the data, either directly (via formulas) or indirectly (by peeking), in the selection of a discrimination rule and/or its parameters. For example, a quick inspection of the data can convince us that a linear discriminator is appropriate in a given situation. The actual position of the discriminating hyperplane is usually determined from the data. In other words, we choose our discriminator from a class $\boldsymbol{D}$ of discriminators. This class can be small (e.g., "all k-nearest neighbor rules") or large (e.g., "all linear and quadratic discriminators, and all nonparametric discriminators of the kernel type with smoothing factor $h > 0$"). If we knew the underlying distribution of the data, then the selection process would be simple: we would pick the Bayes rule. Unfortunately, the Bayes rule is not in $\boldsymbol{D}$ unless we are incredibly lucky. Also, the underlying distribution is not known. Thus, it is important to know how close we are to the performance of the best discriminator in $\boldsymbol{D}$. If $\boldsymbol{D}$ is large enough, then hopefully, the performance of the best discriminator in it is close to that of the Bayes discriminator. There are two issues here which should be separated from each other.

1) The closeness of the best element of $\boldsymbol{D}$ to the Bayes rule.

Manuscript received May 30, 1986; revised July 29, 1987. Recommended for acceptance by J. Kittler. This work was supported by NSERC Grant A3456 and FCAR Grant EQ-1678.

The author is with the School of Computer Science, McGill University, Montreal, P.Q., Canada H3A 2K6.

IEEE Log Number 8718607.

2) The closeness of the actual element picked from $\boldsymbol{D}$ to the best element in $\boldsymbol{D}$.

The former issue is related to the consistency of the estimators in $\boldsymbol{D}$, and will only be dealt with briefly. Our main concern is with the second problem: to what extent can we let the data select the discriminator, and how much are we paying for this luxury? The paper is an exercise in compromises: on the one hand, $\boldsymbol{D}$ should be rich enough so that every Bayes rule can be asymptotically approached by a sequence of rules picked from a sequence of $\boldsymbol{D}$'s, and on the other hand, $\boldsymbol{D}$ should not be too rich because it would lead to trivial selections, as any data can be fit to some discriminator in such a class $\boldsymbol{D}$. One of the biggest advantages of the empirical selection is that the programmer does not have to worry about the choice of smoothing factors and design parameters.

Our statistical model is as follows. The *data* consists of a sequence of $n + m$ iid $R^d \times \{0, 1\}$-valued random vectors $(X_1, Y_1), \cdots, (X_{n+m}, Y_{n+m})$. The X_i's are called the *observations*, and the Y_i's are usually called the *classes*. The fact that we limit the number of classes to two should not take anything away from the main message of this paper. Note also that the data are artificially split by us into two independent sequences, one of length n, and one of length m. This will facilitate the discussion and the ensuing analysis immensely. We will call the n sequence the *training sequence*, and the m sequence the *testing sequence*. The testing sequence is used as an impartial judge in the selection process. A *discrimination rule* is a function $\psi: R^d \times (R^d \times \{0, 1\})^{n+m} \to \{0, 1\}$. It classifies a point $x \in R^d$ as coming from class $\psi(x, (X_1, Y_1), \cdots, (X_{n+m}, Y_{n+m}))$. We will write $\psi(x)$ for the sake of convenience.

The *probability of error* is

$$L_{n+m}(\psi) = L_{n+m} = P\big(\psi(X) \neq Y \,\big|\, (X_1, Y_1), \cdots, (X_{n+m}, Y_{n+m})\big)$$

where (X, Y) is independent of the data sequence and is distributed as (X_1, Y_1). Of course, we would like L_{n+m} to be small, although we know that L_{n+m} cannot be smaller than the *Bayes probability of error*

$$L_{\text{Bayes}} = \inf_{\psi: R^d \to \{0,1\}} P\big(\psi(X) \neq Y\big).$$

In the construction of a rule with small probability of error, we proceed as follows: $\boldsymbol{D}$ is a (possibly infinite) col-

0162-8828/88/0700-0530$01.00 © 1988 IEEE

lection of functions $\phi: R^d \times (R^d \times \{0, 1\})^n \to \{0, 1\}$, from which a particular function ϕ' is picked by minimizing the *empirical risk* based upon the testing sequence

$$\hat{L}_{n,m}(\phi') = \frac{1}{m} \sum_{i=n+1}^{n+m} I_{[\phi'(X_i) \neq Y_i]}$$

$$= \min_{\phi \in \boldsymbol{D}} \frac{1}{m} \sum_{i=n+1}^{n+m} I_{[\phi(X_i) \neq Y_i]}.$$

Here it is noted that

$$\phi(X_i) = \phi\big(X_i, (X_1, Y_1), \cdots, (X_n, Y_n)\big)$$

$$\phi'(X_i) = \phi'\big(X_i, (X_1, Y_1), \cdots, (X_n, Y_n)\big),$$

i.e., the discriminators themselves are based upon the training sequence. Let us formally write

$$\psi(x) = \psi\big(x, (X_1, Y_1), \cdots, (X_{n+m}, Y_{n+m})\big)$$

$$= \phi'\big(x, (X_1, Y_1), \cdots, (X_n, Y_n)\big), \qquad x \in R^d.$$

It is necessary to do this because ψ depends upon both the training sequence and the testing sequence. Since $\hat{L}_{n,m}(\phi)$ is an unbiased binomial estimate of $L_n(\phi)$, it is not unlikely that $L_{n+m}(\psi)$ is close to $\inf_{\phi \in \boldsymbol{D}} L_n(\phi)$, yet this has to be proven rigorously. It is this closeness that is under investigation here. We observe that the idea of minimizing the empirical risk in the construction of a rule goes back to Vapnik and Chervonenkis [130]–[134]. If we define our empirical risk entirely in terms of the training sequence, i.e., if we count the number of errors committed by a rule on the training sequence itself, then we can end up with strange rules. Consider, for example, the problem of the data-based choice of k in a k-NN rule. It is obvious that no errors are committed on the training sequence itself when $k = 1$, yet $k = 1$ can, but does not have to be the optimal choice in a given situation. Glick [53], [55] has shown, however, that for many nonparametric rules such as the kernel rule, counting the errors on the training sequence is essentially harmless provided that the nonparametric rule is consistent. Unfortunately, we want to choose the best discriminator from huge collections of discriminators from which it is possible to draw many nonconsistent sequences. The presence of nonconsistent rules is practically appealing (one can mix parametric and nonparametric discriminators; recall also that we can include all k-NN rules in $\boldsymbol{D}$ without restriction on k), but dangerous since we surely do not want our procedure to lead to nonconsistency.

Cover [22] suggested taking $m = 1$, and counting the number of errors committed by considering $n + 1$ training sets, each time leaving one of the observations (X_i, Y_i) out, and verifying whether the rule classifies the deleted X_i as Y_i. This, at least, reduces the anomaly observed when our collection of discriminators includes the 1-NN rule and no deletion is employed. Our approach is nothing more than an attempt to obtain an alternative to Cover's suggestion for which we can obtain good analytical guarantees of the performance. Not separating a training set from a testing set works in some cases, but it seems that good bounds on the probability of error can only be obtained when the collections are very nice or simple.

When $\boldsymbol{D}$ is large, $\inf_{\phi \in \boldsymbol{D}} L_n(\phi)$ is probably close to L_{Bayes}: this is the case when $\boldsymbol{D}$ contains all k-NN rules or when it contains all kernel-type rules. On the other hand, $\boldsymbol{D}$ can be so small that there is no hope of getting close to L_{Bayes}. A point in case is the class $\boldsymbol{D}$ of all linear discrimination rules. Having settled on a class $\boldsymbol{D}$, it is important to ensure that $L_{n+m}(\psi)$ is close to $\inf_{\phi \in \boldsymbol{D}} L_n(\phi)$. Often this is more important than actually knowing (even approximately) the value $L_{n+m}(\psi)$.

A last word about our split into a training sequence and a testing sequence. This split is primarily aimed at deriving results that are valid for many classes $\boldsymbol{D}$. There are well-known tricks of the trade such as cross validation (or leave-one-out) (Lunts and Brailovsky [85], Stone [121]), holdout, resubstitution, rotation, and bootstrap (Efron [39], [40]) which can be employed to construct an empirical risk from the training sequence, thus obviating the need for a testing sequence (see Kanal [74], Cover and Wagner [23], and Toussaint [126] for surveys, and Glick [56] for a discussion and empirical comparison). This works well in many important situations (see Vapnik and Chervonenkis [132]–[134], Vapnik [136], Devroye and Wagner [28]–[30]), but can fail miserably in other circumstances. This would then force us to restrict $\boldsymbol{D}$ to such an extent that our results would be less powerful. We will make a case for the split-data method by showing just how good the empirical choice is for most popular discrimination rules. This universality seems more difficult to obtain with other methods. In addition, we will argue that the testing sequence can often be taken much smaller than the training sequence ($m = o(n)$). It seems probable that more sophisticated methods such as cross validation would be equally good or better than the split-data method, but we have not been able to show this thus far.

Error estimation is used by us as a tool; we are not interested in actual values of error estimates per se, although it is always nice to have some good estimates of the probability of error. Thus, we will not address issues such as the bias and variance of error estimates, which have led to interesting discussions in the past (see Lachenbruch and Mickey [81], McLachlan [88]. Glick [54], [56], Lachenbruch *et al.* [82], and Lissack and Fu [84]). On the other hand, good automatic selection is impossible without good error estimates, and thus it should come as no surprise that the estimate on which the automatic selection is based can serve as an estimate of the probability of error of the selected rule. This relationship is captured in the following.

The Fundamental Inequalities:

$$L_{n+m}(\psi) - \inf_{\phi \in \boldsymbol{D}} L_n(\phi) \leq 2 \sup_{\phi \in \boldsymbol{D}} \left| \hat{L}_{n,m}(\phi) - L_n(\phi) \right|.$$

$$\left| \hat{L}_{n,m}(\phi') - L_{n+m}(\psi) \right| \leq \sup_{\phi \in \boldsymbol{D}} \left| \hat{L}_{n,m}(\phi) - L_n(\phi) \right|.$$

Proof: Everything is based upon the following observation:

$$\begin{aligned} L_{n+m}(\psi) - \inf_{\phi \in \boldsymbol{D}} L_n(\phi) &= L_{n+m}(\psi) - \hat{L}_{n,m}(\phi') + \hat{L}_{n,m}(\phi') - \inf_{\phi \in \boldsymbol{D}} L_n(\phi) \\ &\le L_{n+m}(\psi) - \hat{L}_{n,m}(\phi') + \sup_{\phi \in \boldsymbol{D}} \left(\hat{L}_{n,m}(\phi) - L_n(\phi)\right) \\ &\le 2 \sup_{\phi \in \boldsymbol{D}} \left|\hat{L}_{n,m}(\phi) - L_n(\phi)\right|. \end{aligned}$$

The second inequality is trivially true. ■

We see that upper bounds for $\sup_{\phi \in \boldsymbol{D}} |\hat{L}_{n,m}(\phi) - L_n(\phi)|$ provide us with upper bounds for two things simultaneously:

1) an upper bound for the suboptimality of ψ within $\boldsymbol{D}$, $L_{n+m}(\psi) - \inf_{\phi \in \boldsymbol{D}} L_n(\phi)$,

2) an upper bound for the error $|\hat{L}_{n,m}(\phi') - L_{n+m}(\psi)|$ committed when $\hat{L}_{n,m}(\phi')$ is used to estimate the probability of error $L_{n+m}(\psi)$.

In other words, by bounding $\sup_{\phi \in \boldsymbol{D}} |\hat{L}_{n,m}(\phi) - L_n(\phi)|$, we kill two flies at once. It is particularly useful to know that even though $\hat{L}_{n,m}(\phi')$ is usually optimistically biased, it is within given bounds of the unknown probability of error with ψ, and that no other test sample is needed to estimate this probability of error. Whenever our bounds indicate that we are close to the optimum in $\boldsymbol{D}$, we must at the same time have a good estimate of the probability of error, and vice versa.

All the probabilities and expected values written P_n and E_n are conditional on the training sequence of length n, whereas P and E refer to unconditional probabilities and expected values. The bounds derived below refer to conditional quantities, and they do not depend upon the training sequence. In other words, they are valid uniformly over all training sequences. The important consequence of this is that the testing sequence should have the right distribution and be iid, but the training sequence can, in fact, be arbitrary. In particular, annoying phenomena such as dependence between observations, noisy data, etc., become irrelevant for our bounds—they could have a negative impact on the actual value of the probability of error, though.

II. Finite Classes

We consider first finite classes $\boldsymbol{D}$, with cardinality bounded by N_n. We have the following.

Theorem 1: Let $\boldsymbol{D}$ be a finite class with cardinality bounded by N_n. For all $\epsilon > 0$,

$$P_n\left(\sup_{\phi \in \boldsymbol{D}} \left|\hat{L}_{n,m}(\phi) - L_n(\phi)\right| > \epsilon\right) \le 2N_n e^{-2m\epsilon^2}$$

and

$$E_n\left(\sup_{\phi \in \boldsymbol{D}} \left|\hat{L}_{n,m}(\phi) - L_n(\phi)\right|\right) \le \sqrt{\frac{\log(2N_n)}{2m}} + \frac{1}{\sqrt{8m \log(2N_n)}}.$$

In the proof of Theorem 1, we will need the following.

Lemma 1: If a nonnegative random variable Z satisfies the inequality $P(Z > t) \le ce^{-2nt^2}$ for all $t > 0$ and some $c > 0$, then

$$E(Z) \le \sqrt{\frac{\log(c)}{2n}} + \sqrt{\frac{1}{8n \log(c)}}. \quad ■$$

Proof of Lemma 1: For all $u > 0$,

$$\begin{aligned} E(Z) = \int_0^\infty P(Z > t)\, dt &\le \int_0^u dt + \int_u^\infty ce^{-2nt^2}\, dt \\ &\le u + c\sqrt{\frac{\pi}{2n}} \int_{2u\sqrt{n}}^\infty \frac{1}{\sqrt{2\pi}} e^{-t^2/2}\, dt \\ &\le u + c\sqrt{\frac{\pi}{2n}} \frac{1}{2u\sqrt{n}} \frac{1}{\sqrt{2\pi}} \\ &\quad \cdot \exp\left[-(2u\sqrt{n})^2/2\right] \\ &= u + \frac{c}{4un} e^{-2nu^2} \end{aligned}$$

where we used Gordon's inequality for the tail of the normal distribution (Gordon [59]; Mitrinovic [91, p. 177]). The last expression is approximately minimized for $u = \sqrt{\log(c)/2n}$. The corresponding value is

$$\sqrt{\frac{\log(c)}{2n}} + \sqrt{\frac{1}{8n \log(c)}}. \quad ■$$

Proof of Theorem 1:

$$\begin{aligned} P_n\left(\sup_{\phi \in \boldsymbol{D}} \left|\hat{L}_{n,m}(\phi) - L_n(\phi)\right| > \epsilon\right) &\le \sum_{\phi \in \boldsymbol{D}} P_n\left(\left|\hat{L}_{n,m}(\phi) - L_n(\phi)\right| > \epsilon\right) \\ &\le 2N_n e^{-2m\epsilon^2} \end{aligned}$$

where we used Hoeffding's inequality (Hoeffding [73]) and the fact that $m\hat{L}_{n,m}(\phi)$ is binomially distributed with parameters m and $L_n(\phi)$. The last part of Theorem 1 is a direct corollary of Lemma 1. ■

Remark 1—Size of the Error: If we take $m = n$ and assume that N_n is large, then Theorem 1 shows that on the average, we are within $\sqrt{\log(N_n)/(2n)}$ of the best possible error rate, whatever it is. Since most common error probabilities tend to the Bayes probability of error at a

rate much slower than $1/\sqrt{n}$, the loss in error rate studied here is asymptotically negligible in many cases relative to the difference between the probability of error and L_{Bayes}, at least when N_n increases at a polynomial rate in n. ■

Remark 2—Distribution-Free Properties: Theorem 1 shows that the problem studied here is purely combinatorial. The actual distribution of the data does not play a role at all in the upper bounds. ■

Remark 3—The k-Nearest Neighbor Rule: When $\boldsymbol{D}$ contains all k-nearest neighbor rules, then $N_n = n$ since there are only n possible values for k. It is easily seen that

$$E_n\left(L_{n+m}(\psi) - \inf_{\phi \in \boldsymbol{D}} L_n(\phi)\right) \leq \sqrt{\frac{\log(2n)}{2m}} + \frac{1}{\sqrt{8m \log(2n)}}.$$

Since $k/n \to 0$, $k \to \infty$ imply that $E(L_n) \to L_{\text{Bayes}}$ for the k-nearest neighbor with data-independent (deterministic) k for all possible distributions (Stone [119]), we see that our strategy leads to a universally consistent rule whenever $\log(n)/m \to 0$. Thus, we can take m equal to a small fraction of n without losing consistency. That we cannot take $m = 1$ and hope to obtain consistency should be obvious. It should also be noted that for $m = n$, we are roughly within $\sqrt{\log(n)/n}$ of the best possible probability of error within the given class. The same remark remains valid for k-nearest neighbor rules defined in terms of all L_p metrics or in terms of the transformation-invariant metric of Olshen (see Olshen [96], Devroye [27]). ■

III. Consistency

Although it was not our objective to discuss consistency of our rules, it is perhaps worth our while to present Theorem 2. Let us first recall the definition of a *consistent* rule (to be more precise, a consistent sequence of ϕ's): a rule is consistent if $E(L_n) \to L_{\text{Bayes}}$ as $n \to \infty$. Consistency may depend upon the distribution of the data. If it does not, then we say that the rule is universally consistent.

Theorem 2—Consistency: Assume that from each $\boldsymbol{D}$ (recall that $\boldsymbol{D}$ varies with n), we can pick one ϕ such that the sequence of ϕ's is consistent for a certain class of distributions. Then the automatic rule ψ defined above is consistent for the same class of distributions (i.e., $E(L_{n+m}(\psi)) \to L_{\text{Bayes}}$ as $n \to \infty$) if

$$\lim_{n \to \infty} \frac{m}{\log(1 + N_n)} = \infty.$$

Proof of Theorem 2: This is a direct corollary of Theorem 1. ■

If one is just worried about consistency, Theorem 2 reassures us that nothing is lost as long as we take m much larger than $\log(N_n)$. Often, this reduces to a very weak condition on the size m of the training set: recall Remark 3 for the k-nearest neighbor estimate.

Remark 4—Infinite N_n: Taking N_n very large, possibly infinite can be dangerous. One might even lose consistency in the process. Consider, for example, the class $\boldsymbol{D}$ of *all* measurable functions ϕ, i.e., all rules. This class contains by definition the Bayes rule, but it also contains many rules which agree completely with the training sequence, i.e., functions ϕ for which $\phi(X_i) = Y_i$ for all $n + 1 \leq i \leq n + m$. Our selection process selects one of the latter ϕ's. If it happens to select the function ϕ which takes the value 0 everywhere except possibly at the points X_i (where $\phi(X_i) = Y_i$), then there is no hope of obtaining universal consistency. The culprit here is the size of N_n. Note that consistency is lost regardless of how large m is picked. ■

IV. Asymptotic Optimality

Let us now introduce the notion of *asymptotic optimality*. A sequence of rules ψ is said to be asymptotically optimal for a given distribution of (X, Y) when

$$\lim_{n \to \infty} \frac{E(L_{n+m}(\psi)) - L_{\text{Bayes}}}{E\left(\inf_{\phi \in \boldsymbol{D}} L_n(\phi)\right) - L_{\text{Bayes}}} = 1.$$

Our definition is not entirely fair because ψ uses $n + m$ observations, whereas the class of rules in the denominator is restricted to using n observations. If ψ is not taken from the same $\boldsymbol{D}$, then it is possible to have a ratio which is smaller than one. But if $\psi = \phi' \in \boldsymbol{D}$, then the ratio always is at least one. That is why the definition makes sense in our setup.

When our selected rule is asymptotically optimal, we have achieved something very strong: we have, in effect, picked a rule (or better, a sequence of rules) which has a probability of error converging at the optimal rate attainable within the sequence of $\boldsymbol{D}$'s. And we do not even have to know what the optimal rate of convergence is. This is especially important in nonparametric rules where some researchers choose smoothing factors in function of theoretical results about the optimal attainable rate of convergence for certain classes of problems.

We are constantly faced with the problem of choosing between parametric and nonparametric discriminators. Parametric discriminators are based upon an underlying model in which a finite number of unknown parameters is estimated from the data. A point in case is the multivariate normal distribution, which leads to linear or quadratic discriminators. If the model is wrong, parametric methods can perform very poorly; when the model is right, their performance is difficult to beat. Our method chooses among the best discriminator depending upon which happens to be best for the given data. We can throw in $\boldsymbol{D}$ a variety of rules, including nearest neighbor rules, a few linear discriminators, a couple of tree classifiers, and perhaps a kernel-type rule. Theorems 1 and 2 should be used when the cardinality of $\boldsymbol{D}$ does not get out of hand.

To save space further on in the paper, we introduce here the notion of ϵ_n optimality where ϵ_n is a positive sequence

decreasing to 0 with n. A rule is said to be ϵ_n *optimal* when

$$\lim_{n\to\infty} \frac{E(L_{n+m}(\psi)) - L_{\text{Bayes}}}{E\left(\inf_{\phi\in D} L_n(\phi)\right) - L_{\text{Bayes}}} = 1$$

for all distributions of (X, Y) for which

$$\lim_{n\to\infty} \frac{E\left(\inf_{\phi\in D} L_n(\phi)\right) - L_{\text{Bayes}}}{\epsilon_n} = \infty.$$

For finite classes, we see that the empirical selection rule is $\sqrt{\log(1 + N_n)/m}$ optimal.

Remark 5: Let us now treat $n + m$ as the sample size. Assume that

$$\lim_{n\to\infty} \frac{E\left(\inf_{\phi\in D} L_n(\phi)\right) - L_{\text{Bayes}}}{\sqrt{\log(1 + N_n)/n}} = \infty$$

where we draw attention to the fact that this condition does not involve m. Then, it is possible to find a sequence m such that

$$\lim_{n\to\infty} \frac{E(L_{n+m}(\psi)) - L_{\text{Bayes}}}{E\left(\inf_{\phi\in D} L_{n+m}(\phi)\right) - L_{\text{Bayes}}} = 1$$

where D is appropriately enlarged to accommodate training sequences of size $n + m$. The condition under which this is true is very mild:

$E(\inf_{\phi\in D} L_n(\phi)) - L_{\text{Bayes}}$ is regularly varying at ∞ with parameter $\rho \in (-\infty, 0]$ (a sequence a_n is regularly varying at ∞ with parameter ρ if $\lim_{n\to\infty} a_n/a_{\lfloor cn \rfloor} = c^{-\rho}$ for all positive c).

The proof uses the following decomposition:

$$\frac{E(L_{n+m}(\psi)) - L_{\text{Bayes}}}{E\left(\inf_{\phi\in D} L_{n+m}(\phi)\right) - L_{\text{Bayes}}}$$
$$= \frac{E(L_{n+m}(\psi)) - L_{\text{Bayes}}}{E\left(\inf_{\phi\in D} L_n(\phi)\right) - L_{\text{Bayes}}}$$
$$\times \frac{E\left(\inf_{\phi\in D} L_n(\phi)\right) - L_{\text{Bayes}}}{E\left(\inf_{\phi\in D} L_{n+m}(\phi)\right) - L_{\text{Bayes}}}$$

where the second factor tends to $\lim(n/(n + m))^{\rho}$ if this limit exists and is finite. Choose $m = \lfloor \epsilon n \rfloor$ for ϵ positive, but very small. This shows that the second factor can be made arbitrarily close to 1. The first factor is less than

$$1 + O\left(\sqrt{\log(1 + N_n)/m}\right)/\left(M\sqrt{\log(1 + N_n)/n}\right)$$

for some large M and all n large enough (apply Theorem 1 and our assumption). This, in turn, is less than one plus a constant times M^{-1} times $\sqrt{1/\epsilon}$, which can be made arbitrarily small by choice of M. ■

V. Infinite Classes

Theorem 1 is useless when $N_n = \infty$. It is here that we can apply the inequality of Vapnik and Chervonenkis [130], [131] or one of its modifications. We will need some new notation. Let μ be the probability measure of (X, Y) on $\Omega = R^d \times \{0, 1\}$, and let μ_m be the empirical measure based upon $(X_{n+1}, Y_{n+1}), \cdots, (X_{n+m}, Y_{n+m})$. Then

$$\sup_{\phi\in D} \left|\hat{L}_{n,m}(\phi) - L_n(\phi)\right| = \sup_{C\in \mathcal{C}} \left|\mu_m(C) - \mu(C)\right|$$

where $\mathcal{C}$ is the collection of all sets

$$\{\{x: \phi = 1\} \times \{0\}\} \cup \{\{x: \phi = 0\} \times \{1\}\},$$
$$\phi \in D.$$

At this point, we recall the Vapnik–Chervonenkis inequality.

Theorem 3: Let $N_{\mathcal{C}}(x_1, \cdots, x_m)$, $(x_1, \cdots, x_m) \in \Omega^m$ be the number of different sets in

$$\{\{x_1, \cdots, x_m\} \cap C \mid C \in \mathcal{C}\},$$

and define the *shatter coefficient* as

$$s(\mathcal{C}, m) = \max_{(x_1,\cdots,x_m)\in\Omega^m} N_{\mathcal{C}}(x_1, \cdots, x_m).$$

Then, for all $\epsilon > 0$,

$$P\left(\sup_{C\in\mathcal{C}} \left|\mu_m(C) - \mu(C)\right| > \epsilon\right)$$
$$\le 4s(\mathcal{C}, 2m)\, e^{-(m\epsilon^2/8)} \qquad (m\epsilon^2 \ge 2)$$

(Vapnik and Chervonenkis [130], [131]) and

$$P\left(\sup_{C\in\mathcal{C}} \left|\mu_m(C) - \mu(C)\right| > \epsilon\right)$$
$$\le cs(\mathcal{C}, m^2)\, e^{-2m\epsilon^2}$$

where the constant c does not exceed $4e^{4\epsilon + 4\epsilon^2}$ (Devroye [33]). Also,

$$E\left(\sup_{C\in\mathcal{C}} \left|\mu_m(C) - \mu(C)\right|\right)$$
$$\le \sqrt{\frac{\log\left(4e^8 s(\mathcal{C}, m^2)\right)}{2m}}$$
$$+ \frac{1}{\sqrt{8m \log\left(4e^8 s(\mathcal{C}, m^2)\right)}}.$$

For more information about these inequalities, see also Vapnik [136], Gaenssler [48], Gaenssler and Stute [47], and Massart [87]. Devroye's bound provides competitive

values when $m\epsilon^2$ is large. It cannot compete for medium range values of $m\epsilon^2$ and relatively small values of $s(\mathbf{C}, m^2)$ with Massart's [87] and Alexander's [5] inequalities. For example, Alexander's bound is

$$P\left(\sup_{C\in\mathbf{C}}|\mu_m(C)-\mu(C)|>\epsilon\right) \le 16\left(\sqrt{m}\,\epsilon\right)^{4096V} e^{-2m\epsilon^2} \quad (m\epsilon^2 \ge 64)$$

where V is the *index* of the class $\mathbf{C}$, i.e., the least integer $k \ge 1$ for which $s(\mathbf{C}, k) < 2^k$. V can be considered as the complexity or size of $\mathbf{C}$. A set $\mathbf{C}$ for which $V < \infty$ is called a *Vapnik–Chervonenkis* (or VC) set. The bounds of Theorem 3 are useful when the shatter coefficients do not increase too quickly with m. For example, if $\mathbf{C}$ contains all Borel sets of Ω, then we can shatter any collection of m different points at will, and obtain $s(\mathbf{C}, m) = 2^m$. This would be useless, of course. The smaller is $\mathbf{C}$, the smaller is the shatter coefficient. It suffices now to compute a few shatter coefficients for certain classes of discrimination rules. For examples, see Cover [21], Vapnik and Chervonenkis [131], Devroye and Wagner [28]–[30], Feinholz [41] Devroye [33], Massart [87], and Dudley [38]. Note that every $\mathbf{D}$ yields one class $\mathbf{C}$ for every fixed training sequence. Thus, collecting results, we have the following.

Theorem 4: For fixed training sequence $(x_1, y_1), \cdots, (x_n, y_n)$, let $\mathbf{C}$ be the collection of all sets

$$\{\{x: \phi = 1\} \times \{0\}\} \cup \{\{x: \phi = 0\} \times \{1\}\}, \quad \phi \in \mathbf{D}.$$

Define

$$S(n, m) = \sup_{(x_1, y_1), \cdots, (x_n, y_n)} s(\mathbf{C}, m).$$

Then

$$P_n\left(\sup_{\phi\in\mathbf{D}}|\hat{L}_{n,m}(\phi)-L_n(\phi)|>\epsilon\right) \le 4S(n, 2m)\, e^{-(m\epsilon^2/8)} \quad (m\epsilon^2 \ge 2),$$

$$E_n\left(\sup_{\phi\in\mathbf{D}}|\hat{L}_{n,m}(\phi)-L_n(\phi)|\right) \le \sqrt{\frac{8\log(4S(n, 2m))}{m}} + \frac{1}{\sqrt{(m/2)\log(4S(n, 2m))}},$$

and

$$P_n\left(\sup_{\phi\in\mathbf{D}}|\hat{L}_{n,m}(\phi)-L_n(\phi)|>\epsilon\right) \le cS(n, m^2)\, e^{-2m\epsilon^2}$$

where the constant c can be taken equal to $4e^{4\epsilon+4\epsilon^2}$ and

$$E_n\left(\sup_{\phi\in\mathbf{D}}|\hat{L}_{n,m}(\phi)-L_n(\phi)|\right) \le \sqrt{\frac{\log(4e^8S(n, m^2))}{2m}} + \frac{1}{\sqrt{8m\log(4e^8S(n, m^2))}}.$$

From Alexander's bound, we derive the following bound.

Theorem 5: Let $\mathbf{C}$ be a Vapnik–Chervonenkis class with index V. Then

$$E\left(\sup_{\phi\in\mathbf{C}}|\mu_m(C)-\mu(C)|\right) \le \frac{1}{\sqrt{n}}\left(64\sqrt{\frac{V}{2}} + \frac{\sqrt{2}}{8\sqrt{V}}\left(\frac{64\sqrt{V}}{e\sqrt{2}}\right)^{4096V}\right).$$

Proof of Theorem 5: We will need the auxiliary inequality

$$\int_a^\infty t^b e^{-2t^2}dt \le \frac{a^b e^{-2a^2}}{4a - b/a} \quad (4a^2 > b > 0, a > 0),$$

which can easily be obtained by a standard analytical argument. Now,

$$E\left(\sup_{C\in\mathbf{C}}|\mu_m(C)-\mu(C)|\right) = \int_0^\infty P\left(\sup_{C\in\mathbf{C}}|\mu_m(C)-\mu(C)|>t\right)dt \le 64\sqrt{V/2m} + m^{-1/2}\int_{64\sqrt{V/2}}^\infty 16t^{4096V}e^{-2t^2}\,dt$$

where we used Alexander's inequality and the fact that $V \ge 1$. The threshold value $64\sqrt{V/2}$ is an approximation of the optimal threshold value. The last integral in the upper bound does not exceed

$$\frac{16(64\sqrt{V/2})^{4096V}e^{-4096V}}{4(64\sqrt{V/2}) - 64\sqrt{2V}} = \frac{(64\sqrt{V/2})^{4096V}e^{-4096V}}{4(4\sqrt{V/2}) - 4\sqrt{2V}} = \frac{\sqrt{2}\left(\frac{64\sqrt{V}}{e\sqrt{2}}\right)}{8\sqrt{V}}.$$

■

The dominating factor in the bound of Theorem 5 is V^{2048V}. Even though the bound decreases as $1/\sqrt{m}$ for fixed V, it is not as useful as the bound of Theorem 3

when V is very large. In most of our examples, V increases with m due to the fact that consistency forces the class $\boldsymbol{D}$ to become richer and richer as n (and m) grow. In comparison, note that since $s(\boldsymbol{C}, m) \le m^V$ for $m \ge 2$ (Vapnik and Chervonenkis [130], [131]), we see that $\sqrt{\log (s(\boldsymbol{C}, m))/m}$ is at most $\sqrt{V \log (m)/m}$. This is in our applications often smaller than $V^{2048V}/\sqrt{m}$. For fixed V, Theorem 5 is usually preferable. We will compute the functions $s(\boldsymbol{C}, m)$ for several important discrimination rules. The index V of $\boldsymbol{C}$ cannot be determined directly from these computations, so some additional (and often nontrivial) work will be required of the users.

The suprema in Theorems 3 and 4 are not always measurable; the measurability must be verified for every class $\boldsymbol{C}$ (for all our examples, the quantities are indeed measurable). For more on the measurability question, see Dudley [37], [38], Massart [87], and Gaenssler [48]. Gine and Zinn [52] and Yukich [141] provide further work on suprema of the type shown in Theorems 3 and 4.

Theorem 6—Consistency and Asymptotic Optimality: Assume that from each $\boldsymbol{D}$ (recall that $\boldsymbol{D}$ varies with n), we can pick one ϕ such that the sequence of ϕ's is consistent for a certain class of distributions. Then the automatic rule ψ defined above is consistent for the same class of distributions (i.e., $E(L_{n+m}(\psi)) \to L_{\text{Bayes}}$ as $n \to \infty$) if

$$\lim_{n \to \infty} \frac{\log \left(1 + S(n, m^2)\right)}{m} = 0.$$

Furthermore, ψ is $\sqrt{\log (1 + S(n, m^2))/m}$ optimal.

VI. Computation of $S(n, m)$

For a collection $\boldsymbol{D}$ of the form $\boldsymbol{D} = \cup_{j=1}^{k} \boldsymbol{D}_j$, we have

$$S(n, m) \le \sum_{j=1}^{k} S_j(n, m)$$

where $S_j(n, m)$ is computed for $\boldsymbol{D}_j$ only. This allows us to treat each homogeneous subcollection of $\boldsymbol{D}$ separately.

A. Linear Discrimination

Consider all rules that split the space R^d in two by virtue of a half plane, and assign class 1 to one half space, and class 0 to the other. Points on the border are treated as belonging to the same half space. Because the training sequence is not even used in the definition of the collection, $S(n, m)$ cannot possibly depend upon n. In other words, $S(n, m) = s(\boldsymbol{C}, m)$.

There are at most

$$2 \sum_{k=0}^{d} \binom{m-1}{k} \le 2(m^d + 1)$$

ways of dichotomizing m points in R^d by hyperplanes (see, e.g., Cover [21]) (this takes into account that there are two ways of attaching 0's and 1's to the two half spaces). We see that

$$S(n, m) \le 2 \left(\sum_{k=0}^{d} \binom{m-1}{k} \right) \le 2(m^d + 1).$$

In this case, it is clearly indicated to take $n = 0$. The resulting discriminator picks the best separating hyperplane from all possible hyperplanes simply based upon the testing sample m. Glick [55] pointed out that for this estimator, $|\hat{L}_{n,m}(\phi') - L_n(\phi')| \to 0$ almost surely. Bounds of the type obtained above can also be found in Devroye and Wagner [26], [28]–[30].

When $n > 0$ and the training sequence is used to assign a class to each half space by a majority vote, then

$$S(n, m) \le \left(\sum_{k=0}^{d} \binom{n+m-1}{k} \right) \le \left((n+m)^d + 1\right).$$

B. Generalized Linear Discrimination Rules

A rule ϕ in which the set $\{x: \phi(x) = 1\}$ coincides with a set of the form

$$\left\{ x: a_0 + \sum_{j=1}^{d^*} a_j f_j(x) \ge 0 \right\}$$

for given fixed functions $f_1, \cdots, f_{d^*}$ and some real numbers $a_0, \cdots, a_{d^*}$ is called a *generalized linear discrimination rule* (see Duda and Hart [36]). These include, for example, all quadratic discrimination rules in R^d when we choose all functions that are either components of x, or squares of components of x, or products of two components of x. In all, $d^* = 2d + d(d-1)/2$. The argument of the previous section remains valid, and we obtain

$$S(n, m) \le 2 \left(\sum_{k=0}^{d^*} \binom{m-1}{k} \right) \le 2(m^{d^*} + 1).$$

Note, nevertheless, that unless d^* is allowed to increase with m, there is no hope of obtaining universal consistency.

C. k-NN Rules

In the k-NN rule (Fix and Hodges [42], Cover and Hart [20]), a majority vote decision is made based upon the k nearest neighbors of X in the training set. If $\boldsymbol{D}$ contains all NN rules (all values of k), then, unlike most of the collections of the previous sections, $\boldsymbol{D}$ increases with n, and depends very much on the training set. A trivial bound in this case is

$$S(n, m) \le n$$

because there are only n members in $\boldsymbol{D}$. For universal consistency, we need $m/\log(n) \to \infty$. The selected rule is $\sqrt{\log(n)/m}$ optimal.

D. Variable Metric NN Rules

Fukunaga has observed in a series of papers (Fukunaga and Hostetler [45], Short and Fukunaga [115], Fukunaga and Flick [46]) that it is perhaps better to first define a suitable metric in R^d based upon the data, and then use

this metric in the determination of near neighbors. Typically, the metric is an L_2 metric based upon a data-dependent positive definite scale-rotation matrix. If $\boldsymbol{D}$ contains all NN rules with metric dependent upon the training sequence only, then the bound $S(n, m) \leq n$ of the previous section remains valid. If, however, $\boldsymbol{D}$ includes all NN rules with all L_2 metrics, then the situation is very different. The collection no longer depends upon n (as in the linear discrimination case), and it contains an infinite number of rules. If k is not properly restricted, there is virtually no hope of obtaining a useful bound for $S(n, m)$.

E. NN Rule Based upon Reference Data

Hart [71], Gates [49], Wilson [139], Wagner [138], Ullmann [128], Ritter *et al.* [107], Tomek [125], and Devijver and Kittler [25] all study 1-NN rules based upon a subsequence of the training sequence. This sequence can be considered as representative of the whole training sequence. Its choice is based upon certain criteria, which do not concern us here. Certainly, it seems that the most impartial criterion is the one that uses the empirical risk computed from a testing sequence. Assume thus that $\boldsymbol{D}$ includes all *condensed* (or edited) NN rules with reduced training set of at most k members, i.e., all NN rules based upon a subset of $k(X_i, Y_i)$'s selected from the training sequence. A simple counting argument shows that $\boldsymbol{D}$ has at most

$$\sum_{j=1}^{k} \binom{n}{j} \leq n^k + 1$$

members. The rule is $\sqrt{k \log (n)/m}$ optimal. This implies, for example, that to obtain $n^{-2/5}$ optimality, we can take k no larger than $O(n^{1/5}/\log (n))$ when $m = n$. Thus, k needs to be restricted. We also note that $k \to \infty$ and $m/(k \log (n)) \to \infty$ are sufficient for universal consistency.

F. Weighted NN Rules

In the k-NN rule, each of the k nearest neighbors of a point x plays an equally important role in the decision. Royall [110] first suggested using rules in which the k nearest neighbors are given unequal voting powers in the decision: the ith nearest neighbor receives weight v_i where usually $v_1 \geq v_2 \geq \cdots \geq v_k \geq 0$ and the v's sum to one. For consistency, the integer k and the v_i's have to satisfy certain properties given by Stone [119] and Devroye and Wagner [31], [32]. It is possible to let the testing sequence choose the best weight vector for a fixed k. In that case, $\boldsymbol{D}$ contains all weighted k-NN rules. Its cardinality is infinite. To compute $S(n, m)$, note that each X_j in the testing set is classified as 1 or 0 according to whether the sign of the following expression is positive or nonpositive:

$$\sum_{i=1}^{k} a_{ij} v_i$$

where $a_{ij} \in \{-1, 1\}$ depends upon the class of the ith nearest neighbor of X_j in the training sequence (and does not depend upon the v_i's). Every weight vector $v = (v_1, \cdots, v_k)$ yields a vector of m classes to which the X_j's in the testing sequence are assigned. In the computation of $s(\boldsymbol{C}, m)$, we consider the a_{ij}'s as fixed numbers. Let V be the k-dimensional space of all weight vectors v. The collection of all v's for which X_j is assigned to class 1 is a linear half space of V. Therefore, $s(\boldsymbol{C}, m)$ is bounded from above by the cardinality of the partition of V defined by m linear hyperplanes. This is bounded by m^k. Thus,

$$S(n, m) \leq m^k.$$

We observe that even though $s(\boldsymbol{C}, m)$ depends upon the training sequence (and thus, n) via the a_{ij}'s, we used an argument that did not require the actual values of the a_{ij}'s. The bound on $S(n, m)$ does not depend upon n.

The rule is universally consistent when $k \to \infty$ and $m/(k \log (m)) \to \infty$. It is $\sqrt{k \log (m)/m}$ optimal. Automatic selection is useful here when $\sqrt{k \log (m)/m}$ is small compared to the difference between the actual probability of error of the rule and the Bayes rule, which is nearly always at least $1/\sqrt{n}$. In the extreme case $m = n$, our bound is not good enough, as I will now show. With $m = n$, the choice $k = n^{4/5}$ (which is optimal in certain ways) leads to $\sqrt{\log n}\, n^{-1/10}$ optimality. Taking k smaller is not something we would like to do because the rate of convergence of the best rule within $\boldsymbol{D}$ is likely to slow down. In other words, $\boldsymbol{D}$ is too rich for the interesting values of k to apply automatic selection.

G. Kernel-Based Rules

Kernel-based rules are derived from the kernel estimate in density estimation originally studied by Parzen [98], Rosenblatt [108], and Cacoullos [16]. A point x is assigned class 1 if

$$g(x) = \sum_{i=1}^{n} \left(Y_i - \frac{1}{2}\right) K\left(\frac{x - X_i}{h}\right) \geq 0$$

and to class 0 otherwise where K is a fixed function called the kernel and $h > 0$ is a smoothing factor. It is easy to verify that this is a voting scheme in which the ith observation carries weight $K(x - X_i/h)$. Thus, K is usually decreasing along rays. For particular choices of K, rules of this sort have been proposed by Fix and Hodges [42], [43], Sebestyen [112], Bashkirov *et al.* [11], Aizerman *et al.* [1]–[4], Braverman [13], Braverman and Pyatniskii [14], Van Ryzin [129], and Meisel [89]. Statistical analysis of these rules and/or the corresponding regression function estimate can be found in Nadaraya [93], [95], Rejto and Revesz [106], Devroye and Wagner [26], [31], [32], and Greblicki [60]–[62].

Hardle and Marron [70] proposed and studied a cross-validation method for choosing the optimal h for the kernel regression estimate. They obtain asymptotic optimality for the integrated square error. Although their method gives us a choice for h if we consider $P(Y = 1 \mid X = x)$ as the regression function, it is not clear that the thus obtained h is optimal for the probability of error. Consider

next two well-separated classes. Then a little thought shows that the optimal h is, in fact, constant, independent of n. This shows that we should not *a priori* exclude any values of h, as is commonly done in studies on regression and density estimation.

Devroye and Wagner [28]–[30] obtained distribution-free error bounds for the cross-validation estimate of the probability of error. Unfortunately, their results are only valid for fixed choices of K and h.

We begin by considering the collection $\boldsymbol{D}$ of all kernel rules for all values of h, but fixed kernel $K = I_A$ where I is the indicator function and A is any star-shaped set of unit Lebesgue measure (a set A is star shaped if $x \notin A$ implies that $cx \notin A$ for all $c \geq 1$). We vary h monotonically from 0 to ∞. For fixed X_j in the testing sequence, the function $g(X_j)$ on which the decision is based can at most take n values. Therefore,

$$s(\boldsymbol{C}, m) \leq mn + 1$$

and

$$S(m, n) \leq mn + 1.$$

If $K = \Sigma_{i=1}^{k} a_i I_{A_i}$ for some finite k, some numbers a_i, and some star-shaped sets A_i, then $S(m, n) \leq kmn + 1$.

We can generalize in several directions. First, there is the question of more general K. There is an interesting subclass of kernels K of the form

$$K(x) = \|x\|^{-r} I_A(x)$$

where A is star shaped and $r \geq 0$ is a constant. Observe that $K_h(x) = h^{r-d}\|x\|^{-r} I_A(x/h)$. Thus, for fixed X_j,

$$\begin{aligned} g(X_j) &= \sum_{i=1}^{n} \left(Y_i - \frac{1}{2}\right) K\left(\frac{X_j - X_i}{h}\right) \\ &= h^{r-d} \sum_{i=1}^{n} \left(Y_i - \frac{1}{2}\right) \\ &\quad \cdot \|X_j - X_i\|^{-r} I_A((X_j - X_i)/h) \end{aligned}$$

changes sign at most n times as h increases from 0 to ∞. For these kernels, we also have $S(m, n) \leq mn + 1$. The kernels have the desirable property that if a decision is based upon l points, then changing h does not change the decision unless one or more points become excluded or new points are considered in the decision. For $r = 0$, we obtain the uniform kernel discussed earlier. For $0 < r < d$, the kernel is integrable, but has an infinite peak at the origin. For $r \geq d$, the kernel is not integrable. It has been pointed out by several authors that the integrability of K is not necessary for the consistency of kernel rules in discrimination and regression: in fact, kernels with $r = d$ have been suggested as early as 1962 by Sebestyen [112].

In the class $\boldsymbol{D}$ considered above, only one parameter was varied. In d-dimensional pattern recognition, it is often necessary to adjust the scales of many component variables. Thus, it seems natural to classify $x = (x_1, \cdots, x_d)$ in class one if

$$g(x) = \sum_{i=1}^{n} \left(Y_i - \frac{1}{2}\right) \prod_{l=1}^{d} K\left(\frac{x_l - X_{il}}{h_l}\right) \geq 0$$

where now K is a one-dimensional kernel, $h_1, \cdots, h_d$ are d positive numbers, and X_{il} is the lth component of X_i. It should be noted that this is certainly not the only way of introducing d different smoothing factors, one for each component. Let $\boldsymbol{D}$ be the collection of all rules of this type considered over all possible values $h_1, \cdots, h_d$. For this class, we will now show that

$$S(m, n) \leq (mn)^d + 1$$

when K is the function $I_{[-1,1]}$. Note that the rule is a majority vote over centered rectangles with sides equal to $2h_1, 2h_2, \cdots, 2h_d$. To see this, consider the d-dimensional quadrant of mn points obtained by taking the absolute values of the vectors $X_j - X_i$, $n < j < n + m$, $1 \leq i \leq n$ (the absolute value of a vector is a vector whose components are the absolute values of the components of the vector). To compute $S(m, n)$, it suffices to count how many different subsets can be obtained from these mn points by considering all possible rectangles with one vertex at the origin and the diagonally opposite vertex in the quadrant. This is $1 + (mn)^d$.

The *consistency* of the class $\boldsymbol{D}$ is ensured when K is the uniform kernel on the unit hypercube, by applying the universal consistency theorem of Devroye and Wagner [31], [32] and Spiegelman and Sacks [117] (see also Greblicki *et al.* [67] or Krzyzak [76]) provided that

$$\lim_{n \to \infty} \frac{m}{d \log (mn)} = \infty,$$

i.e.,

$$\lim_{n \to \infty} \frac{m}{\log (n)} = \infty.$$

ψ is $\sqrt{d \log (mn)/2m}$ optimal when both $m, n \to \infty$. In particular, it is $\sqrt{\log (n)/n}$ optimal when $m = n$. This is good news since the standard bounds for relating the probability of error to the L_1 error in density estimation (see, e.g., Devroye and Gyorfi [35]), combined with well-known results about the best possible expected error with any kernel density estimate (i.e., the best possible expected L_1 error is about equal to a constant times $n^{-2/(4+d)}$, see Devroye and Gyorfi [35], give us upper bounds for $E(L_n(\phi) - L_{\text{Bayes}})$ that decrease as $n^{-2/(4+d)}$ where ϕ is the kernel discrimination rule in which the h is chosen in an optimal way for the underlying densities. Since this tends to 0 slower than $\sqrt{\log (n)/n}$, it seems plausible that the automatic selection rule with $m = n^{1-\epsilon}$ (with an appropriately picked small ϵ) is asymptotically optimal for large classes of distributions. There are distributions for which $\sqrt{\log (n)/n}$ optimality does not give us asymptotic optimality: consider, for example, discrete distributions putting all their mass on a finite number

of points. For such distributions, the optimal rule within $\boldsymbol{D}$ has an error rate roughly equal to L_{Bayes} plus a constant times $1/\sqrt{n}$. Hence, the error introduced by the selection process exceeds the error of the best rule when all errors are considered relative to L_{Bayes}. Even when X has a density, the best h for density estimation is usually not the optimal h for discrimination, and the bounds linking density estimation errors to classification errors are suboptimal. Thus, we cannot just claim that the automatic selection rule is asymptotically optimal for all densities. A case in point is when given $Y = 1$, X puts its mass on $[0, 1]$, and given $Y = 0$, it puts its mass on $[3, 4]$. The kernel rule with $h = 1$ and kernel uniform on $[-1, 1]$ has expected error tending to zero at an exponential rate, while the Bayes error is 0. Yet, this choice for h is far from optimal for the individual density estimates since it usually does not even imply consistency of the density estimates!

H. Histogram Rules

Histogram rules are simply rules in which R^d is partitioned into a countable (but usually finite) number of sets A_l, and the decision for $x \in A_l$ is based upon a majority vote among all pairs (X_i, Y_i) for which $X_i \in A_l$. In case of a voting tie, we arbitrarily classify x as coming from class one (our strategy in case of a tie does not matter much). The partitions can be ordinary rectangular grids in which all A_i's are translates of A_1. If $\boldsymbol{D}$ contains all partitions into k or fewer sets without restriction as to the shape of the sets, then

$$S(n, m) \leq \sum_{i=1}^{k} \left\{ \begin{matrix} n+m \\ i \end{matrix} \right\}$$

where $\left\{ \begin{matrix} n+m \\ i \end{matrix} \right\}$ denotes a Stirling number of the second kind, i.e., the number of ways of partitioning $n + m$ points into i nonempty subsets. Even for $k = 2$, this is much too large since

$$\sum_{i=1}^{2} \left\{ \begin{matrix} n+m \\ i \end{matrix} \right\} = 2^{n+m-1}.$$

Thus, we need to restrict $\boldsymbol{D}$ drastically. Consider first *ordinary histogram rules* on R^1, i.e., rules defined by a regular interval partition of the real line: all intervals are of the form $[a + hi, a + h(i + 1))$ where h is the interval width, i is an integer, and a is the position of a fixed point of the partition. There are two free parameters, h and a.

The ordinary histogram rules can be traced back to a histogram regression function estimate of Tukey [127]. The consistency is established by Glick [54] and Gordon and Olshen [57], [58]. Devroye and Gyorfi [34] showed that for all distributions of (X, Y), the simple histogram rule is strongly consistent, provided that $h \to 0$ and $nh \to \infty$ as $n \to \infty$.

When h is fixed, a quick but rather loose upper bound for $S(n, m)$ can be obtained as follows: start with $a = 0$; clearly, we need only count the number of different m vectors of decisions as a increases to h (because of periodicity). It is straightforward to see that the vector of decisions can only change when a histogram interval boundary point reaches one of the $n + m$ data points. Therefore,

$$S(n, m) \leq n + m + 1.$$

If $a = 0$ is fixed, but h varies, a much more common assumption, $S(n, m)$ becomes very large.

Rather than varying h over an infinite range, it is often computationally more attractive to restrict h to all possible values $|X_i - X_j|$, $1 \leq i < j \leq n$. In that case, we can use Theorem 1 with $N_n = n(n-1)/2$. The d-dimensional generalization in which we consider all rectangular partitions with two training points as extreme vertices has the same N_n.

I. Statistically Equivalent Blocks

Considerable attention has been paid over the years to *histogram rules based upon order statistics*. Basically, the order statistics of the components of the training data are used to construct a partition into rectangles. The great advantage of these rules is their invariance with respect to all strictly monotone transformations of the coordinate axes. For example, it has been suggested to partition the real line by using the kth, $2k$th, etc., order statistics (Mahalanobis [86]; see also Parthasarathy and Bhattacharya [97]): when k is the free design parameter in the class of rules, Theorem 1 can be used with $N_n = n$. The d-dimensional generalizations of these rules include rules based upon statistically equivalent blocks. The idea is to define rectangles containing k points each. For example, the kth smallest x coordinate among the training data could define the first cut. The (infinite) rectangle with $n - k$ points can be cut according to the y axis, isolating another k points. This can be repeated on a rotational basis for all coordinate axes. However, it is obvious that one can proceed in many other ways as well; see, e.g., Anderson [8], Patrick [99], Patrick and Fisher [100], Quesenberry and Gessaman [105], and Gessaman and Gessaman [51]. Consider now all rules in which the partition depends entirely upon the training data sequence and upon k, and let k be the free design parameter. Once again, regardless of the dimension, Theorem 1 is applicable with $N_n = n$. The universal consistency is guaranteed by the results of Gordon and Olshen [57] when $d = 1$ and the kth-order statistics are used with $k/n \to 0$ and $k/\sqrt{n} \to \infty$.

Rules have been developed in which the rectangular partition depends not only upon the X_i's in the training sequence, but also upon the Y_i's; see, e.g., Henrichon and Fu [72], Meisel and Michalopoulos [90], and Friedman [44]. For example, Friedman cuts the axes at the places where the absolute differences between the marginal empirical distribution functions are largest to ensure minimal empirical error after the cut. His procedure is based upon an observation of Stoller [118].

The rules described above are called distribution free since they remain invariant under monotone transformations of the coordinate axes. For a survey of such rules,

see Das Gupta [24]. The problem of designing automatic distribution-free rules was first attacked in the papers of Anderson and Benning [7] and Beakley and Tuteur [12].

In our setup, we should let the test data decide where cuts should be made. This leads very quickly to oversized classes of rules, so we will impose reasonable restrictions. We consider cuts into at most k rectangles where k is a number picked beforehand. Recall that for a fixed partition, the class assigned to every rectangle is decided upon by a majority vote among the training points. On the real line, choosing a partition into at most k sets is equivalent to choosing $k-1$ cut positions from $n+m+1$ spacings between all test and training points. Hence,

$$S(n, m) \le \sum_{j=1}^{k-1} \binom{n+m+1}{j} \le (n+m+1)^{k-1}.$$

For d-dimensional partitions defined by at most $k-1$ consecutive orthogonal cuts, we see that for the first cut, there are at most $1 + d(n+m)$ possible combinations of spacings–directions to choose from. This yields the loose upper bound

$$S(n, m) \le \left(1 + d(n+m)\right)^{k-1}.$$

This bound is also valid for all *grids* defined by at most $k-1$ cuts. The main difference here is that every cut defines two half spaces, so that we usually end up with many more than k rectangles in the partition.

Assume that $\boldsymbol{D}$ contains all histograms with partitions into at most k (possibly infinite) rectangles. Then, considering that a rectangle in R^d requires choosing $2d$ spacings between all test and training points, two per coordinate axis,

$$S(n, m) \le (n+m+1)^{2d(k-1)}.$$

See Feinholz [41] for more work on such partitions.

J. Binary Tree Classifiers

Binary tree classifiers have become increasingly important because of their conceptual simplicity and computational feasibility. The forefathers of these classifiers are the histogram rules based upon statistically equivalent blocks described in the previous section. Many strategies have been proposed for constructing the binary decision tree (in which each internal node corresponds to a cut, and each terminal node corresponds to a set in the partition: see, for example, You and Fu [140], Bartolucci *et al.* [10], Sethi and Chatterjee [113], Payne and Meisel [101], Swain and Hauska [122], Taylor *et al.* [124], Kulkarni [77], Kulkarni and Kanal [78], Anderson and Fu [6,] Mui and Fu [92], Gustafson *et al.* [69], Rounds [109], Sethi and Sarvarayudu [114], Argentiero *et al.* [9], Qing-Yun and Fu [104], Kurzynski [79], Lin and Fu [83], Breiman *et al.* [15], and Casey and Nagy [17].

If we consider all binary trees in which each internal node corresponds to a split perpendicular to one of the axes, then, as we have shown in the previous section,

$$S(n, m) \le \left(1 + d(n+m)\right)^{k-1}.$$

Frequently, researchers consider smaller classes of rules, with particular recipes (dependent upon the training sequence only) for computing the cuts. In those situations, the bound is pessimistic. Others have proposed to generalize orthogonal cuts by including linear cuts in any direction. Recall that there are at most

$$\sum_{j=0}^{d} \binom{n+m}{j} \le (n+m)^d + 1$$

ways of dichotomizing $n+m$ points in R^d by hyperplanes (see, e.g., Cover [21]). Thus, if we allow up to $k-1$ internal nodes (or linear cuts),

$$S(n, m) \le \left(1 + (n+m)^d\right)^{k-1}.$$

The restriction imposed on the number of internal nodes is rather unrealistic. For example, Breiman *et al.* [15] construct a tree with n leaves, one per training point. Then, the tree is trimmed from the bottom up by combining leaves. Yet, without some sort of condition on the size of $\boldsymbol{D}$, our results are not useful.

K. Boolean Classifiers

Pearl [102] has studied Boolean classifiers. These classifiers can only be used when X takes values in $\{0, 1\}^d$. Each discrimination rule, $\{0, 1\}$-valued itself, can be written as a Boolean expression involving NOT, OR, and AND bit operations. The complexity c of such a discrimination rule is the minimum number of such operations needed to describe the discrimination rule. Thus, c adequately represents the computation time needed to apply the discrimination rule. Let $\boldsymbol{D}$ be the collection of all discrimination rules with complexity not exceeding c. Then

$$S(n, m) \le \left(\frac{16(d+c)^2}{c}\right)^c$$

(Pippenger [103]). The use of this expression in the Vapnik–Chervonenkis inequality leads to an inequality of Pearl [102]. Again, we should take $n = 0$, as the training data are not used in the definition of $\boldsymbol{D}$.

L. Series Method

Some classifiers are derived from the Fourier series estimate or other series estimates of an unknown density. The density estimates go back to the work of Cencov [18], Schwartz [111], Kronmal and Tarter [75], Tarter and Kronmal [123], and Specht [116]. Their use in classification was considered by Greblicki [64] and Greblicki and Pawlak [63], [65], [66].

Nearly all these estimators can be put into the following form: classify x as belonging to class 1 if

$$\sum_{i=1}^{N} a_{i,n}\, g_i(x) \ge \tfrac{1}{2}$$

where the g_i's are fixed functions, forming a base for the series estimate, $a_{i,n}$ is a fixed function of the training data, and N controls the amount of smoothing. When the g_i's are the usual trigonometric base, then this leads to the

Fourier series classifier studied by Greblicki and Pawlak [63], [65]. When the g_i's form an orthonormal system based upon Hermite polynomials, we obtain the classifiers studied by Greblicki [64] and Greblicki and Pawlak [66], [68].

If we let $\boldsymbol{D}$ be the class of all classifiers of one type (i.e., one set of g_i's and corresponding functions $a_{i,n}$), with N restricted to $1, 2, \cdots, k$ where k is a large integer usually not exceeding n, then the class is finite, with k members. Hence, we can choose the amount of smoothing automatically, and have $\sqrt{\log(k)/m}$ optimality. The rates of convergence of the probability of error to the Bayes probability of error commonly found in the literature are typically $O(n^{-a})$ for some constant $0 < a < 1/2$ (see, e.g., Greblicki and Pawlak [65], [66]). When $k = m = n$, the selected rule is $\sqrt{\log(n)/n}$ optimal, and thus possibly asymptotically optimal. The difficulty in the verification of the asymptotic optimality is due to the fact that lower bounds on actual rates of convergence to L_{Bayes} are not available.

If the collection of g_i's is fixed, but $\boldsymbol{D}$ contains all classifiers for all $1 \leq N \leq k$ and all values of $a_{i,n}$, then we are back in the position of the generalized linear discrimination rules with dimension k. Thus,

$$S(n, m) \leq 2\left(\sum_{i=0}^{k}\binom{m-1}{i}\right) \leq 2(m^k + 1).$$

The selected rule ψ is $O(\sqrt{k \log(m)/m})$ optimal, and this puts a modest restriction on k. Obviously, we should take $n = 0$. Greblicki and Pawlak have pointed out that for the d-dimensional Fourier series classifier, $N \sim m^{1/(5d)}$ yields a classifier whose expected error rate is equal to L_{Bayes} plus $O(m^{-2/5})$ under appropriate smoothness conditions on the distribution of (X, Y). (Note that m temporarily plays the role of the size of the data since $n = 0$.) Taking $k = O(m^{1/10}/\log(m))$ yields a rule that is $m^{-2/5}$ optimal. The Greblicki–Pawlak choice for N is entirely within the range (i.e., $N \leq k$) when $d > 2$.

References

[1] M. A. Aizerman, E. M. Braverman, and L. I. Rozonoer, "Theoretical foundations of the potential function method in pattern recognition learning," *Automat. Remote Contr.*, vol. 25, pp. 917–936, 1964.

[2] ——, "The probability problem of pattern recognition learning and the method of potential functions," *Automat. Remote Contr.*, vol. 25, pp. 1307–1323, 1964.

[3] ——, "The method of potential functions for the problem of restoring the characteristic of a function converter from randomly observed points," *Automat. Remote Contr.*, vol. 25, pp. 1546–1556, 1964.

[4] ——, "Extrapolative problems in automatic control and the method of potential functions," *Amer. Math. Soc. Transl.*, vol. 87, pp. 281–303, 1970.

[5] K. S. Alexander, "Probability inequalities for empirical processes and a law of the iterated logarithm," *Ann. Prob.*, vol. 12, pp. 1041–1067, 1984.

[6] A. C. Anderson and K. S. Fu, "Design and development of a linear binary tree classifier for leukocytes," Purdue Univ., Lafayette, IN, Tech. Rep. TR-EE-79-31, 1979.

[7] M. W. Anderson and R. D. Benning, "A distribution-free discrimination procedure based on clustering," *IEEE Trans. Inform. Theory*, vol. IT-16, pp. 541–548, 1970.

[8] T. W. Anderson, "Some nonparametric multivariate procedures based on statistically equivalent blocks," in *Multivariate Analysis*, P. R. Krishnaiah, Ed. New York: Academic, 1966, pp. 5–27.

[9] P. Argentiero, R. Chin, and P. Beaudet, "An automated approach to the design of decision tree classifiers," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. PAMI-4, pp. 51–57, 1982.

[10] L. A. Bartolucci, P. H. Swain, and C. Wu, "Selective radiant temperature mapping using a layered classifier," *IEEE Trans. Geosci. Electron.*, vol. GE-14, pp. 101–106, 1976.

[11] O. Bashkirov, E. M. Braverman, and I. E. Muchnik, "Potential function algorithms for pattern recognition learning machines," *Automat. Remote Contr.*, vol. 25, pp. 692–695, 1964.

[12] G. W. Beakley and F. B. Tuteur, "Distribution-free pattern verification using statistically equivalent blocks," *IEEE Trans. Comput.*, vol. C-21, pp. 1337–1347, 1972.

[13] E. M. Braverman, "The method of potential functions," *Automat. Remote Contr.*, vol. 26, pp. 2130–2138, 1965.

[14] E. M. Braverman and E. S. Pyatniskii, "Estimation of the rate of convergence of algorithms based on the potential function method," *Automat. Remote Contr.*, vol. 27, pp. 80–100, 1966.

[15] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*. Belmont, CA: Wadsworth Int., 1984.

[16] T. Cacoullos, "Estimation of a multivariate density," *Ann. Inst. Statist. Math.*, vol. 18, pp. 179–190, 1965.

[17] R. G. Casey and G. Nagy, "Decision tree design using a probabilistic model," *IEEE Trans. Inform. Theory*, vol. IT-30, pp. 93–99, 1984.

[18] N. N. Cencov, "Evaluation of an unknown distribution density from observations," *Sov. Math. Dokl.*, vol. 3, pp. 1559–1562, 1962.

[19] J. T. Chu, "Some new error bounds and approximations for pattern recognition," *IEEE Trans. Comput.*, vol. C-23, pp. 194–198, 1974.

[20] T. M. Cover and P. E. Hart, "Nearest neighbor pattern classification," *Ann. Math. Statist.*, vol. 36, pp. 1049–1051, 1965.

[21] T. M. Cover, "Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition," *IEEE Trans. Electron. Comput.*, vol. EC-14, pp. 326–334, 1965.

[22] ——, "Learning in pattern recognition," in *Methodologies of Pattern Recognition*, S. Watanabe, Ed. New York: Academic, 1969, pp. 111–132.

[23] T. M. Cover and T. J. Wagner, "Topics in statistical pattern recognition," *Commun. Cybern.*, vol. 10, pp. 15–46, 1975.

[24] S. DasGupta, "Nonparametric classification rules," *Sankhya Ser. A*, vol. 26, pp. 25–30, 1964.

[25] P. A. Devijver and J. Kittler, "On the edited nearest neighbor rule," in *Proc. 5th Int. Conf. Pattern Recognition*, 1980, pp. 72–80.

[26] L. Devroye and T. J. Wagner, "Nonparametric discrimination and density estimation," Electron. Res. Cen., Univ. Texas, Austin, Tech. Rep. 183, 1976.

[27] L. Devroye, "A universal k-nearest neighbor procedure in discrimination," in *Proc. 1978 IEEE Comput. Soc. Conf. Pattern Recognition Image Processing*, 1978, pp. 142–147.

[28] L. Devroye and T. J. Wagner, "Distribution-free performance bounds for potential function rules," *IEEE Trans. Inform. Theory*, vol. IT-25, pp. 601–604, 1979.

[29] ——, "Distribution-free performance bounds with the resubstitution error estimate," *IEEE Trans. Inform. Theory*, vol. IT-25, pp. 208–210, 1979.

[30] ——, "Distribution-free inequalities for the deleted and holdout error estimates," *IEEE Trans. Inform. Theory*, vol. IT-25, pp. 202–207, 1979.

[31] ——, "On the $L1$ convergence of kernel estimators of regression functions with applications in discrimination," *Zeitschrift fur Wahrscheinlichleitstheorie und verwandte Gebiete*, vol. 51, pp. 15–25, 1980.

[32] ——, "Distribution-free consistency results in nonparametric discrimination and regression function estimation," *Ann. Statist.*, vol. 8, pp. 231–239, 1980.

[33] L. Devroye, "Bounds for the uniform deviation of empirical measures," *J. Multivariate Anal.*, vol. 12, pp. 72–79, 1982.

[34] L. Devroye and L. Gyorfi, "Distribution-free exponential bound on the L_1 error of partitioning estimates of a regression function," in *Proc. 4th Pannonian Symp. Math. Statist.*, F. Konecny, J. Mogyorodi, and W. Wertz, Eds. Budapest, Hungary: Akademiai Kiado, 1983, pp. 67–76.

[35] ——, *Nonparametric Density Estimation: The L1 View*. New York: Wiley, 1985.

[36] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis*. New York: 1973.
[37] R. M. Dudley, "Central limit theorems for empirical measures," *Ann. Prob.*, vol. 6, pp. 899–929, 1978.
[38] ——, "Empirical processes," in *Ecole de Probabilite de St.-Flour 1982, Lecture Notes Math. 1097*. New York: Springer-Verlag, 1984.
[39] B. Efron, "Bootstrap methods: Another look at the jackknife," *Ann. Statist.*, vol. 7, pp. 1–26, 1979.
[40] ——, "Estimating the error rate of a prediction rule: Improvement on cross validation," *J. Amer. Statist. Assoc.*, vol. 78, pp. 316–331, 1983.
[41] L. Feinholz, "Estimation of the performance of partitioning algorithms in pattern classification," M.Sc.thesis, Dep. Math., McGill Univ., Montreal, Canada, 1979.
[42] E. Fix and J. L. Hodges, "Discriminatory analysis, nonparametric discrimination, consistency properties," USAF School of Aviation Med., Randolph Field, TX, Rep. 21-49-004, 1951.
[43] ——, "Discriminatory analysis: Small sample performance," USAF School of Aviation Med., Randolph Field, TX, Rep. 21-49-004, 1952.
[44] J. H. Friedman, "A recursive partitioning decision rule for nonparametric classification," *IEEE Trans. Comput.*, vol. C-26, pp. 404–408, 1977.
[45] K. Fukunaga and L. D. Hostetler, "Optimization of k-nearest neighbor density estimates," *IEEE Trans. Inform. Theory*, vol. IT-19, pp. 320–326, 1973.
[46] K. Fukunaga and T. E. Flick, "An optimal global nearest neighbor metric," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. PAMI-6, pp. 314–318, 1984.
[47] P. Gaenssler and W. Stute, "Empirical processes: A survey of results for independent and identically distributed random variables," *Ann. Prob.*, vol. 7, pp. 193–243, 1979.
[48] P. Gaenssler, "Empirical processes," in *Lecture Notes-Monograph Ser.* Hayward, CA: Inst. Math. Statist., 1983.
[49] G. W. Gates, "The reduced nearest neighbor rule," *IEEE Trans. Inform. Theory*, vol. IT-18, pp. 431–433, 1972.
[50] M. P. Gessaman, "A consistent nonparametric multivariate density estimator based on statistically equivalent blocks," *Ann. Math. Statist.*, vol. 41, pp. 1344–1346, 1970.
[51] M. P. Gessaman and P. H. Gessaman, "A comparison of some multivariate discrimination procedures," *J. Amer. Statist. Ass.*, vol. 67, pp. 468–472, 1972.
[52] E. Gine and J. Zinn, "Some limit theorems for empirical processes," *Ann. Prob.*, vol. 12, pp. 929–989, 1984.
[53] N. Glick, "Sample-based classification procedures derived from density estimators," *J. Amer. Statist. Ass.*, vol. 67, pp. 116–122, 1972.
[54] ——, "Sample-based multinomial classification," *Biometrics*, vol. 29, pp. 241–256, 1973.
[55] ——, "Sample-based classification procedures related to empiric distributions," *IEEE Trans. Inform. Theory*, vol. IT-22, pp. 454–461, 1976.
[56] ——, "Additive estimators for probabilities of correct classification," *Pattern Recognition*, vol. 10, pp. 211–222, 1978.
[57] L. Gordon and R. A. Olshen, "Asymptotically efficient solutions to the classification problem," *Ann. Statist.*, vol. 6, pp. 515–533, 1978.
[58] ——, "Consistent nonparametric regression from recursive partitioning schemes," *J. Multivariate Anal.*, vol. 10, pp. 611–627, 1980.
[59] R. D. Gordon, "Values of Mills' ratio of area to bounding ordinate and of the normal probability integral for large values of the argument," *Ann. Math. Statist.*, vol. 12, pp. 364–366, 1941.
[60] W. Greblicki, "Asymptotically optimal probabilistic algorithms for pattern recognition and identification," monograph 3, Prace Naukowe Instytutu Cybernetyki Technicznej Politechniki Wroclawsjiej 18, Wroclaw, Poland, 1974.
[61] ——, "Pattern recognition procedures with nonparametric density estimates," *IEEE Trans. Syst. Man, Cybern.*, vol. SMC-8, pp. 809–812, 1978.
[62] ——, "Asymptotically optimal pattern recognition procedures with density estimates," *IEEE Trans. Inform. Theory*, vol. IT-24, pp. 250–251, 1978.
[63] W. Greblicki and M. Pawlak, "Classification using the Fourier series estimate of multivariate density functions," *IEEE Trans. Syst., Man, Cybern.*, vol. 11, pp. 726–730, 1981.
[64] W. Greblicki, "Asymptotic efficiency of classifying procedures using the Hermite series estimate of multivariate probability densities," *IEEE Trans. Inform. Theory*, vol. IT-27, pp. 364–366, 1981.
[65] W. Greblicki and M. Pawlak, "A classification procedure using the multiple Fourier series," *Inform. Sci.*, vol. 26, pp. 115–126, 1982.
[66] ——, "Almost sure convergence of classification procedures using Hermite series density estimates," *Pattern Recognition Lett.*, vol. 2, pp. 13–17, 1983.
[67] W. Greblicki, A. Krzyzak, and M. Pawlak, "Distribution-free pointwise consistency of kernel regression estimate," *Ann. Statist.*, vol. 12, pp. 1570–1575, 1984.
[68] W. Greblicki and M. Pawlak, "Pointwise consistency of the Hermite series density estimate," *Statist. Prob. Lett.*, vol. 3, pp. 65–69, 1985.
[69] D. E. Gustafson, S. Gelfand, and S. K. Mitter, "A nonparametric multiclass partitioning method for classification," in *Proc. 5th Int. Conf. Pattern Recognition*, 1980, pp. 654–659.
[70] W. Hardle and J. S. Marron, "Optimal bandwidth selection in nonparametric regression function estimation," *Ann. Statist.*, vol. 13, pp. 1465–1481, 1985.
[71] P. E. Hart, "The condensed nearest neighbor rule," *IEEE Trans. Inform. Theory*, vol. IT-14, pp. 515–516, 1968.
[72] E. G. Henrichon and K. S. Fu, "A nonparametric partitioning procedure for pattern classification," *IEEE Trans. Comput.*, vol. C-18, pp. 614–624, 1969.
[73] W. Hoeffding, "Probability inequalities for sums of bounded random variables," *J. Amer. Statist. Ass.*, vol. 58, pp. 13–30, 1963.
[74] L. N. Kanal, "Patterns in pattern recognition," *IEEE Trans. Inform. Theory*, vol. IT-20, pp. 697–722, 1974.
[75] R. A. Kronmal and M. E. Tarter, "The estimation of probability densities and cumulatives by Fourier series methods," *J. Amer. Statist. Ass.*, vol. 63, pp. 925–952, 1968.
[76] A. Krzyzak, "The rates of convergence of kernel regression estimates and classification rules," *IEEE Trans. Inform. Theory*, vol. IT-32, pp. 668–679, 1986.
[77] A. V. Kulkarni, "On the mean accuracy of hierarchical classifiers," *IEEE Trans. Comput.*, vol. C-27, pp. 771–776, 1978.
[78] A. V. Kulkarni and L. N. Kanal, "Admissible search strategies for parametric and nonparametric hierarchical classifiers," in *Proc. 4th Int. Joint Conf. Pattern Recognition*, 1978, pp. 238–248.
[79] M. W. Kurzynski, "The optimal strategy of a tree classifier," *Pattern Recognition*, vol. 16, pp. 81–87, 1983.
[80] P. A. Lachenbruch, "An almost unbiased method of obtaining confidence intervals for the probability of misclassification in discriminant analysis," *Biometrics*, vol. 23, pp. 639–645, 1967.
[81] P. A. Lachenbruch and M. R. Mickey, "Estimation of error rates in discriminant analysis," *Technometrics*, vol. 10, pp. 1–11, 1968.
[82] P. A. Lachenbruch, C. Sneeringer, and L. T. Revo, "Robustness of the linear and quadratic discriminant functions to certain types of non-normality," *Commun. Statist.*, vol. 1, pp. 39–56, 1972.
[83] Y. K. Lin and K. S. Fu, "Automatic classification of cervical cells using a binary tree classifier," *Pattern Recognition*, vol. 16, pp. 69–80, 1983.
[84] T. Lissack and K. S. Fu, "Error estimation in pattern recognition via L^{α} distance between posterior density functions," *IEEE Trans. Inform. Theory*, vol. IT-22, pp. 34–45, 1976.
[85] A. L. Lunts and V. L. Brailovsky, "Evaluation of attributes obtained in statistical decision rules," *Eng. Cybern.*, vol. 0, pp. 98–109, 1967.
[86] P. C. Mahalanobis, "A method of fractile graphical analysis," *Sankhya Ser. A*, vol. 23, pp. 41–64, 1961.
[87] P. Massart, "Vitesse de convergence dans le theoreme de la limite centrale pour le processus empirique," Ph.D. dissertation, Univ. Paris-Sud, Orsay, France, 1983.
[88] G. J. McLachlan, "The bias of the apparent error rate in discriminant analysis," *Biometrika*, vol. 63, pp. 239–244, 1976.
[89] W. Meisel, "Potential functions in mathematical pattern recognition," *IEEE Trans. Comput.*, vol. C-18, pp. 911–918, 1969.
[90] W. S. Meisel and D. A. Michalopoulos, "A partitioning algorithm with application in pattern classification and the optimization of decision tree," *IEEE Trans. Comput.*, vol. C-22, pp. 93–103, 1973.
[91] D. S. Mitrinovic, *Analytic Inequalities*. New York: Springer-Verlag, 1970.
[92] J. K. Mui and K. S. Fu, "Automated classification of nucleated

blood cells using a binary tree classifier," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. PAMI-2, pp. 429-443, 1980.
[93] E. A. Nadaraya, "On estimating regression," *Theory Prob. Appl.*, vol. 9, pp. 141-142, 1964.
[94] ——, "On nonparametric estimates of density functions and regression curves," *Theory Prob. Appl.*, vol. 10, pp. 186-190, 1965.
[95] ——, "Remarks on nonparametric estimates for density functions and regression curves," *Theory Prob. Appl.*, vol. 15, pp. 134-137, 1970.
[96] R. A. Olshen, "Comments on a paper by C. J. Stone," *Ann. Statist.*, vol. 5, pp. 632-633, 1977.
[97] K. R. Parthasarathy and P. K. Bhattacharya, "Some limit theorems in regression theory," *Sankhya Ser.* A, vol. 23, pp. 91-102, 1961.
[98] E. Parzen, "On the estimation of a probability density function and the mode," *Ann. Math. Statist.*, vol. 33, pp. 1065-1076, 1962.
[99] E. A. Patrick, "Distribution-free minimum conditional risk learning systems," Purdue Univ., Lafayette, IN, Tech. Rep. TR-EE-66-18, 1966.
[100] E. A. Patrick and F. P. Fisher, II, "Introduction to the performance of distribution-free conditional risk learning systems," Purdue Univ., Lafayette, IN, Tech. Rep. TR-EE-67-12, 1967.
[101] H. J. Payne and W. S. Meisel, "An algorithm for constructing optimal binary decision trees," *IEEE Trans. Comput.*, vol. C-26, pp. 905-916, 1977.
[102] J. Pearl, "Capacity and error estimates for boolean classifiers with limited complexity," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. PAMI-1, pp. 350-355, 1979.
[103] N. Pippenger, "Information theory and the complexity of Boolean functions," *Math. Syst. Theory*, vol. 10, pp. 124-162, 1977.
[104] S. Qing-Yun and K. S. Fu, "A method for the design of binary tree classifiers," *Pattern Recognition*, vol. 16, pp. 593-603, 1983.
[105] C. P. Quesenberry and M. P. Gessaman, "Nonparametric discrimination using tolerance regions," *Ann. Math. Statist.*, vol. 39, pp. 664-673, 1968.
[106] L. Rejto and P. Revesz, "On empirical density function," *Probl. Contr. Inform. Theory*, vol. 2, pp. 67-80, 1973.
[107] G. L. Ritter, H. B. Woodruff, S. R. Lowry, and T. L. Isenhour, "An algorithm for a selective nearest neighbor decision rule," *IEEE Trans. Inform. Theory*, vol. IT-21, pp. 665-669, 1975.
[108] M. Rosenblatt, "Remarks on some nonparametric estimates of a density function," *Ann. Math. Statist.*, vol. 27, pp. 832-837, 1956.
[109] E. M. Rounds, "A combined nonparametric approach to feature selection and binary decision tree design," *Pattern Recognition*, vol. 12, pp. 313-317, 1980.
[110] R. M. Royall, "A class of nonparametric estimators of a smooth regression function," Ph.D. dissertation, Stanford Univ., Stanford, CA, 1966.
[111] S. C. Schwartz, "Estimation of probability density by an orthogonal series," *Ann. Math. Statist.*, vol. 38, pp. 1261-1265, 1967.
[112] G. Sebestyen, *Decision Making Processes in Pattern Recognition*. New York: Macmillan, 1962.
[113] I. K. Sethi and B. Chatterjee, "Efficient decision tree design for discrete variable pattern recognition problems," *Pattern Recognition*, vol. 9, pp. 197-206, 1977.
[114] I. K. Sethi and G. P. R. Sarvarayudu, "Hierarchical classifier design using mutual information," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. PAMI-4, pp. 441-445, 1981.
[115] R. D. Short and K. Fukunaga, "The optimal distance measure for nearest neighbor classification," *IEEE Trans. Inform. Theory*, vol. IT-27, pp. 622-627, 1981.
[116] D. F. Specht, "Series estimation of a probability density function," *Technometrics*, vol. 13, pp. 409-424, 1971.
[117] C. Spiegelman and J. Sacks, "Consistent window estimation in nonparametric regression," *Ann. Statist.*, vol. 8, pp. 240-246, 1980.
[118] D. S. Stoller, "Univariate two-population distribution-free discrimination," *J. Amer. Statist. Ass.*, vol. 49, pp. 770-777, 1954.
[119] C. J. Stone, "Consistent nonparametric regression," *Ann. Statist.*, vol. 8, pp. 1348-1360, 1977.
[120] ——, "Optimal global rates of convergence for nonparametric regression," *Ann. Statist.*, vol. 10, pp. 1040-1053, 1982.
[121] M. Stone, "Cross-validatory choice and assessment of statistical predictions," *J. Roy. Statist. Soc.*, vol. 36, pp. 111-147, 1974.
[122] P. H. Swain and H. Hauska, "The decision tree classifier: design and potential," *IEEE Trans. Geosc. Electron*, vol. GE-15, pp. 142-147, 1977.
[123] M. E. Tarter and R. A. Kronmal, "On multivariate density estimates based on orthogonal expansions," *Ann. Math. Statist.*, vol. 41, pp. 718-722, 1970.
[124] J. Taylor, P. H. Bartels, M. Bibbo, and G. L. Wied, "Automated hierarchic decision structures for multiple category cell classification by TICAS," *Acta Cytologica*, vol. 22, p. 4, 1978.
[125] I. Tomek, "Two modifications of CNN," *IEEE Trans. Syst. Man, Cybern.*, vol. SMC-6, pp. 769-772, 1976.
[126] G. T. Toussaint, "Bibliography on estimation of misclassification," *IEEE Trans. Inform. Theory*, vol. IT-20, pp. 472-479, 1974.
[127] J. W. Tukey, "Curves as parameters and touch estimation," in *Proc. 4th Berkeley Symp.*, 1961, pp. 681-694.
[128] J. R. Ullmann, "Automatic selection of reference data for use in a nearest-neighbor method of pattern classification," *IEEE Trans. Inform. Theory*, vol. IT-20, pp. 541-543, 1974.
[129] J. VanRyzin, "Bayes risk consistency of classification procedures using density estimation," *Sankhya Ser.* A, vol. 28, pp. 161-170, 1966.
[130] V. N. Vapnik and A. Ya. Chervonenkis, "Theory of uniform convergence of frequencies of events to their probabilities and problems of search for an optimal solution from empirical data," *Automat. Remote Contr.*, vol. 32, pp. 207-217, 1971.
[131] ——, "On the uniform convergence of relative frequencies of events to their probabilities," *Theory Prob. Appl.*, vol. 16, pp. 264-280, 1971.
[132] ——, "Ordered risk minimization. I," *Automat. Remote Contr.*, vol. 35, pp. 1226-1235, 1974.
[133] ——, "Ordered risk minimization. II," *Automat. Remote Contr.*, vol. 35, pp. 1403-1412, 1974.
[134] ——, *Theory of Pattern Recognition*. Moscow: Nauka, 1974.
[135] ——, "Necessary and sufficient conditions for the uniform convergence of means to their expectations," *Theory Prob. Appl.*, vol. 26, pp. 532-553, 1981.
[136] V. N. Vapnik, *Estimation of Dependences Based on Empirical Data*. New York: Springer-Verlag, 1982.
[137] T. J. Wagner, "Convergence of the nearest neighbor rule," *IEEE Trans. Inform. Theory*, vol. IT-17, pp. 566-571, 1971.
[138] ——, "Convergence of the edited nearest neighbor," *IEEE Trans. Inform. Theory*, vol. IT-19, pp. 696-699, 1973.
[139] D. L. Wilson, "Asymptotic properties of nearest neighbor rules using edited data," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-2, pp. 408-421, 1972.
[140] K. C. You and K. S. Fu, "An approach to the design of a linear binary tree classifier," in *Proc. Symp. Machine Processing of Remotely Sensed Data*, Purdue Univ., Lafayette, IN, 1976, pp. 3A-10.
[141] J. E. Yukich, "Laws of large numbers for classes of functions," *J. Multivariate Anal.*, vol. 17, pp. 245-260, 1985.

Luc Devroye received the Ph.D. degree from the University of Texas, Austin, in 1976.

He has been with the School of Computer Science, McGill University, Montreal, Canada, since 1977, where he is presently a Professor of Computer Science. He has mainly written on nonparametric estimation of densities, statistical pattern recognition, analysis of algorithms, probability theory, and random number generation.