AUTOMATIC SELECTION OF A DISCRIMINATION RULE BASED UPON MINIMIZATION OF THE EMPIRICAL RISK[1]

Luc Devroye

School of Computer Science, McGill University
805 Sherbrooke Street West, Montréal
Canada H3A 2K6

Abstract

A discrimination rule is chosen from a possibly infinite collection of discrimination rules based upon the minimization of the observed error in a test sample. For example, the collection could include all k nearest neighbor rules (for all k), all linear discriminators, and all kernel-based rules (for all possible choices of the smoothing parameter). We do not put any restrictions on the collection.

We study how close the probability of error of the selected rule is to the (unknown) minimal probability of error over the entire collection. If both training sample and test sample have n observations, the expected value of the difference is shown to be $O(\sqrt{\log(n)/n})$ for many reasonable collections, such as the one mentioned above. General inequalities governing this error are given which are of a combinatorial nature, *i.e.*, they are valid for all possible distributions of the data, and most practical collections of rules.

The theory is based in part on the work of Vapnik and Chervonenkis regarding minimization of the empirical risk. For all proofs, technical details, and additional examples, we refer to Devroye (1986).

As a by-product, we establish that for some nonparametric rules, the probability of error of the selected rule converges at the optimal rate (achievable within the given collection of non-parametric rules) to the Bayes probability of error, and this without actually knowing the optimal rate of convergence to the Bayes probability of error.

1. 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 $\mathbf{D}$ of discriminators. This class can be small (*e.g.*, "all k-nearest neighbor rules")

[1]Research of the author was sponsored by NSERC Grant A3456 and by FCAR Grant EQ-1678

NATO ASI Series, Vol. F30
Pattern Recognition Theory and Applications
Edited by P. A. Devijver and J. Kittler

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 $\mathbf{D}$ 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 $\mathbf{D}$. If $\mathbf{D}$ 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.

A. The closeness of the best element of $\mathbf{D}$ to the Bayes rule.

B. The closeness of the actual element picked from $\mathbf{D}$ to the best element in $\mathbf{D}$.

The former issue is related to the consistency of the estimators in $\mathbf{D}$, 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, $\mathbf{D}$ should be rich enough so that every Bayes rule can be asymptotically approached by a sequence of rules picked from a sequence of $\mathbf{D}$'s, and on the other hand, $\mathbf{D}$ 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 $\mathbf{D}$. 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.

Statistical model. Data-split technique

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 is 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

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

where (X, Y) is independent of the data sequence and 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*

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

Minimization of the empirical risk

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

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

Here it is noted that

$$\begin{aligned}\phi(X_i) &= \phi(X_i,(X_1,Y_1),\cdots,(X_n,Y_n)) \,,\\ \phi'(X_i) &= \phi'(X_i,(X_1,Y_1),\cdots,(X_n,Y_n)) \,,\end{aligned}$$

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

$$\begin{aligned}\psi(x) &= \psi(x,(X_1,Y_1),\cdots,(X_{n+m},Y_{n+m}))\\ &= \phi'(x,(X_1,Y_1),\cdots,(X_n,Y_n)), \quad x\in R^d.\end{aligned}$$

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 $\inf_{\phi\in D} L_n(\phi)$, 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 (1971, 1974).

Why split the data?

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 (1972, 1976) 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 $\mathbf{D}$ without restriction on k), but dangerous since we surely don't want our procedure to lead to nonconsistency.

Cover (1969) 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.

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 **D**. There are well-known tricks of the trade such as cross-validation (or leave-one-out) (Lunts and Brailovsky, 1967; Stone, 1974), holdout, resubstitution, rotation, and bootstrap (Efron, 1979, 1983) which can be employed to construct an empirical risk from the training sequence, thus obviating the need for a testing sequence (see Kanal (1974), Cover and Wagner (1975) and Toussaint (1976) for surveys, and Glick (1978) for a discussion and empirical comparison). This works well in many important situations (see Vapnik and Chervonenkis (1974), Vapnik (1982), Devroye and Wagner (1979)), but can fail miserably in other circumstances. This would then force us to restrict **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 haven't been able to show this thus far.

The size of the class of rules

When **D** contains all rules, selection is like a lottery. Even though the Bayes rule itself is in **D**, it is impossible to let the testing sequence pick a good rule, since the error committed on the testing sequence for the selected rule is zero, and thus has no relationship to the actual probability of error with that rule. When **D** is large but not gigantic, $\inf_{\phi \in D} L_n(\phi)$ is probably close to L_{Bayes}: this is the case when **D** contains all k-NN rules, or when it contains all kernel-type rules. On the other hand, **D** can be so small that there is no hope of getting close to L_{Bayes}. A point in case is the class **D** of all linear discrimination rules.

The selection error

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 Fundamental Inequalities

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

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

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

A. An upper bound for the suboptimality of ψ within **D**, $L_{n+m}(\psi) - \inf_{\phi \in D} L_n(\phi)$. We will call this difference the *selection error SE*.

B. 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 $L_{n+m}(\psi)$. This could be called the *error estimate's accuracy EEA*.

In other words, by bounding the *worst-case deviation* $W = \sup_{\phi \in D} |\hat{L}_{n,m}(\phi) - L_n(\phi)|$, 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 $\mathbf{D}$, we must at the the same time have a good estimate of the probability of error, and vice versa.

Conditional upper bounds

To make the random variable W small, m should be large so that we may benefit from the healthy averaging effect captured for example in the central limit theorem. Unfortunately, the size of $\mathbf{D}$ works against us. We will now derive bounds for $E(W|$ training data) that are functions of n, m and a quantity measuring the size of $\mathbf{D}$ only.

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 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, although the testing sequence should have the right distribution and be iid, the training sequence can in fact be arbitrary. In particular, annoying phenomena such as dependence between observations, noisy data, etcetera become irrelevant for our bounds — they could have a negative impact on the actual value of the probability of error though.

2. Finite Classes

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

Theorem 1. (Devroye, 1986) Let $\mathbf{D}$ be a finite class with cardinality bounded by N_n. Then

$$E_n W \leq \sqrt{\frac{\log(2N_n)}{2m}} + \frac{1}{\sqrt{8m \log(2N_n)}} .$$

Size of the bound

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.

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.

The k-nearest neighbor rule

When **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

$$E_n W \leq \sqrt{\frac{\log(2n)}{2m}} + \frac{1}{\sqrt{8m \log(2n)}} .$$

Since $k/n \to 0$, $k \to \infty$ imply that $E(L_n) \to L_{Bayes}$ for the k-nearest neighbor with data-independent (deterministic) k, for all possible distributions (Stone, 1977), 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 (Olshen, 1977; Devroye, 1978).

3. 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 $E(L_n) \to L_{Bayes}$ 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 **D** (recall that **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.*, $E(L_{n+m}(\psi)) \to L_{Bayes}$ as $n \to \infty$) if

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

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.

4. 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

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

Now, by the triangle inequality,

$$1 \leq \frac{E(L_{n+m}(\psi)) - L_{Bayes}}{E(\inf_{\phi\in D} L_n(\phi)) - L_{Bayes}} \leq 1 + \frac{E(SE)}{E(\inf_{\phi\in D} L_n(\phi)) - L_{Bayes}}$$

When the 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 **D**'s. And we don't 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. For the k-nearest neighbor rule with the best possible sequence of k's, the rate of convergence to the Bayes error is often of the order of $n^{-2/5}$ or worse. In those cases the selection rule is asymptotically optimal when $m = \varepsilon n$ for any $\varepsilon > 0$.

Mixing parametric and nonparametric rules

We are constantly faced with the problem of choosing between parametric discriminators 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 **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 **D** does not get out of hand.

5. Infinite Classes

Theorem 1 is useless when $N_n = \infty$. It is here that we can apply the inequality of Vapnik and Chervonenkis (1974) or one of it modifications. Fortunately, the bound remains formally valid if N_n, the bound on the cardinality of **D**, is replaced by $2e^8$ times the *shatter coefficient* S, which in turn depends upon n, m and the "richness" of **D** only. The shatter coefficient is always finite, regardless of the "size" of **D**.

The shatter coefficient

Let **C** be the collection of all sets

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

Thus, every ϕ in **D** can contribute at most one member to **C**. Then the shatter coefficient is defined as the maximum over all possible testing sequences of length m^2, and all possible training sequences of length n, of the number of possible misclassification vectors of the testing sequence. (The misclassification vector is a vector of m^2 zeroes and ones, a one occurring if and only if the corresponding testing point has been incorrectly classified.) Note that $S \leq N_n$. Also, S is not random by virtue of the definition in terms of maxima. Generally speaking, S increases with the size of **D**. It suffices now to compute a few shatter coefficients for certain classes of discrimination rules. For examples, see Cover (1965), Vapnik and Chervonenkis (1971), Devroye and Wagner (1979), Feinholz (1979), Devroye (1982), and Massart (1983).

Smorgasbords of rules

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

$$S \leq \sum_{j=1}^{k} S_j ,$$

where S_j is computed for $\mathbf{D}_j$ only. This allows us to treat each homogeneous subcollection of $\mathbf{D}$ separately.

Linear discrimination

Consider all rules that split the space R^d in two by virtue of a halfplane, and assign class 1 to one halfspace, and class 0 to the other. Points on the border are treated as belonging to the same halfspace. Because the training sequence is not even used in the definition of the collection, S can't possibly depend upon n.

There are at most

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

ways of dichotomizing m^2 points in R^d by hyperplanes (see *e.g.*, Cover, 1965) (this takes into account that there are two ways of attaching 0's and 1's to the two halfspaces). We see that

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

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

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

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, 1973). 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 counting argument of the previous paragraph remains valid, provided that d is replaced by d^*.

Kernel-based rules

Kernel-based rules are derived from the kernel estimate in density estimation originally studied by Parzen (1962), Rosenblatt (1956) and Cacoullos (1965). 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 i-th observation carries weight $K\left(\frac{x-X_i}{h}\right)$. Thus, K is usually decreasing along rays. For particular choices of K, rules of this sort have been proposed by Fix and Hodges (1951, 1952), Sebestyen (1962), Bashkirov, Braverman and Muchnik (1964), Van Ryzin (1966), and Meisel (1969).

We begin by considering the collection **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, $S \leq nm^2 + 1$.

A typical star-shaped set is the centered unit cube. In the class **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 l-th 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 **D** be the collection of all rules of this type considered over all possible values $h_1, \cdots, h_d$. For this class, we have $S \leq (nm^2)^d + 1$ for all kernels K that are indicators of centered hypercubes.

The *consistency* of the class **D** is insured when K is the uniform kernel on the unit hypercube, by applying the universal consistency theorem of Devroye and Wagner (1980) and Spiegelman and Sacks (1980) (see also Greblicki, Krzyzak and Pawlak, 1984), provided that $m/\log(n) \to \infty$. The standard bounds for relating the probability of error to the L_1 error in density estimation (see *e.g.*, Devroye and Gyorfi, 1985), 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, 1985), give us lower bounds for $E(L_n(\phi) - L_{Bayes})$ 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.

Binary tree classifiers

Binary tree classifiers have become increasingly important because of their conceptual simplicity and computational feasibility. 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 Sethi and Chatterjee (1977), Payne and Meisel (1977), Sethi and Sarvarayudu (1981), Lin and Fu (1983), Breiman, Friedman, Ohlsen and Stone (1983), and Casey and Nagy (1984).

If we consider all binary trees in which each internal node corresponds to a split perpendicular to one of the axes, and the space is partitioned into k hyperrectangles, then $S \leq (1 + d(n + m^2))^{k-1}$.

Here we have an example of a class that is too large to be practical for the present procedure, since k is typically a polynomially increasing function of n.

References

[1] O. Bashkirov, E.M. Braverman, and I.E. Muchnik, "Potential function algorithms for pattern recognition learning machines," *Automation and Remote Control*, vol. 25, pp. 692-695, 1964.

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

[3] T. Cacoullos, "Estimation of a multivariate density," *Annals of the Institute of Statistical Mathematics*, vol. 18, pp. 179-190, 1965.

[4] R.G. Casey and G. Nagy, "Decision tree design using a probabilistic model," *IEEE Transactions on Information Theory*, vol. IT-30, pp. 93-99, 1984.

[5] T.M. Cover, "Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition," *IEEE Transactions on Electronic Computers*, vol. EC-14, pp. 326-334, 1965.

[6] T.M. Cover, "Learning in pattern recognition," in *Methodologies of Pattern Recognition*, ed. S. Watanabe, pp. 111-132, Academic Press, New York, N.Y., 1969.

[7] T.M. Cover and T.J. Wagner, "Topics in statistical pattern recognition," *Communication and Cybernetics*, vol. 10, pp. 15-46, 1975.

[8] L. Devroye, "A universal k-nearest neighbor procedure in discrimination," in *Proceedings of the 1978 IEEE Computer Society Conference on Pattern Recognition and Image Processing*, pp. 142-147, 1978.

[9] L. Devroye and T.J. Wagner, "Distribution-free performance bounds for potential function rules," *IEEE Transactions on Information Theory*, vol. IT-25, pp. 601-604, 1979.

[10] L. Devroye and T.J. Wagner, "Distribution-free performance bounds with the resubstitution error estimate," *IEEE Transactions on Information Theory*, vol. IT-25, pp. 208-210, 1979.

[11] L. Devroye and T.J. Wagner, "Distribution-free inequalities for the deleted and holdout error estimates," *IEEE Transactions on Information Theory*, vol. IT-25, pp. 202-207, 1979.

[12] L. Devroye and T.J. Wagner, "Distribution-free consistency results in nonparametric discrimination and regression function estimation," *Annals of Statistics*, vol. 8, pp. 231-239, 1980.

[13] L. Devroye, "Bounds for the uniform deviation of empirical measures," *Journal of Multivariate Analysis*, vol. 12, pp. 72-79, 1982.

[14] L. Devroye and L. Gyorfi, *Nonparametric Density Estimation: the L_1 View*, John Wiley, New York, 1985.

[15] L. Devroye, "Automatic pattern recognition: a study of the probability of error," Technical Report, School of Computer Science, McGill University, 1986.

[16] R.O. Duda and P.E. Hart, *Pattern Classification and Scene Analysis*, John Wiley, New York, N.Y., 1973.

[17] B. Efron, "Bootstrap methods: another look at the jackknife," *Annals of Statistics*, vol. 7, pp. 1-26, 1979.

[18] B. Efron, "Estimating the error rate of a prediction rule: improvement on cross validation," *Journal of the American Statistical Association*, vol. 78, pp. 316-331, 1983.

[19] L. Feinholz, "Estimation of the performance of partitioning algorithms in pattern classification," M.Sc.Thesis, Department of Mathematics, McGill University, Montreal, 1979.

[20] E. Fix and J.L. Hodges, "Discriminating analysis, nonparametric discrimination, consistency properties," Report 21-49-004, USAF School of Aviation Medicine, Randolph Field, Texas, 1951.

[21] E. Fix and J.L. Hodges, "Discriminatory analysis: small sample performance," Report 21-49-004, USAF School of Aviation Medicine, Randolph Field, Texas, 1952.

[22] N. Glick, "Sample-based classification procedures derived from density estimators," *Journal of the American Statistical Association*, vol. 67, pp. 116-122, 1972.

[23] N. Glick, "Sample-based classification procedures related to empiric distributions," *Transactions on Information Theory*, vol. IT-22, pp. 454-461, 1976.

[24] N. Glick, "Additive estimators for probabilities of correct classification," *Pattern Recognition*, vol. 10, pp. 211-222, 1978.

[25] W. Greblicki, A. Krzyzak, and M. Pawlak, "Distribution-free pointwise consistency of kernel regression estimate," *Annals of Statistics*, vol. 12, pp. 1570-1575, 1984.

[26] L.N. Kanal, "Pattern in pattern recognition," *IEEE Transactions on Information Theory*, vol. IT-20, pp. 697-722, 1974.

[27] 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.

[28] A.L. Lunts and V.L. Brailosvsky, "Evaluation of attributes obtained in statistical decision rules," *Engineering Cybernetics*, vol. 5, pp. 98-109, 1967.

[29] P. Massart, "Vitesse de convergence dans le théorème de la limite centrale pour le processus empirique," Ph.D. Dissertation, Université de Paris-Sud, Orsay, France, 1983.

[30] W. Meisel, "Potential functions in mathematical pattern recognition," *IEEE Transactions on Computers*, vol. C-18, pp. 911-918, 1969.

[31] R.A. Olshen, "Comments on a paper by C.J. Stone," *Annals of Statistics*, vol. 5, pp. 632-633, 1977.

[32] E. Parzen, "On the estimation of a probability density function and the mode," *Annals of Mathematical Statistics*, vol. 33, pp. 1065-1076, 1962.

[33] H.J. Payne and W.S. Meisel, "An algorithm for constructing optimal binary decision trees," *IEEE Transactions on Computers*, vol. C-26, pp. 905-916, 1977.

[34] M. Rosenblatt, "Remark on some nonparametric estimates of a density function," *Annals of Mathematical Statistics*, vol. 27, pp. 832-837, 1956.

[35] G. Sebestyen, *Decision Making Processes in Pattern Recognition*, Macmillan, New York, N.Y., 1962.

[36] I.K. Sethi and B. Chatterjee, "Efficient decision tree design for discrete variable pattern recognition problems," *Pattern Recognition*, vol. 9, pp. 197-206, 1977.

[37] I.K. Sethi and G.P.R. Sarvarayudu, "Hierarchical classifier design using mutual Information," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-4, pp. 441-445, 1981.

[38] C. Spiegelman and J. Sacks, "Consistent window estimation in nonparametric regression," *Annals of Statistics*, vol. 8, pp. 240-246, 1980.

[39] C.J. Stone, "Consistent nonparametric regression," *Annals of Statistics*, vol. 8, pp. 1348-1360, 1977.

[40] M. Stone, "Cross-validatory choice and assessment of statistical predictions," *Journal of the Royal Statistical Society*, vol. 36, pp. 111-147, 1974.

[41] G.T. Toussaint, "Bibliography on estimation of misclassification," *IEEE Transactions on Information Theory*, vol. IT-20, pp. 474-479, 1974.

[42] J. VanRyzin, "Bayes risk consistency of classification procedures using density estimation," *Sankhya Series A*, vol. 28, pp. 161-170, 1966.

[43] V.N. Vapnik, *Estimation of Dependences Based on Empirical Data*, Springer-Verlag, 1982.

[44] 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," *Automation and Remote Control*, vol. 32, pp. 207-217, 1971.

[45] V.N. Vapnik and A. Ya. Chervonenkis, "On the uniform convergence of relative frequencies of events to their probabilities," *Theory of Probability and its Applications*, vol. 16, pp. 264-280, 1971.

[46] V.N. Vapnik and A. Ya. Chervonenkis, "Ordered risk minimization. I," *Automation and Remote Control*, vol. 35, pp. 1226-1235, 1974.

[47] V.N. Vapnik and A. Ya. Chervonenkis, "Ordered risk minimization. II," *Automation and Remote Control*, vol. 35, pp. 1043- 1412, 1974.

[48] V.N. Vapnik and A. Ya. Chervonenkis, *Theory of Pattern Recognition*, Nauka, Moscow, 1974.

[49] V. N. Vapnik and A. Ya. Chervonenkis, "Necessary and sufficient conditions for the uniform convergence of means to their expectations," *Theory of Probability and its Applications*, vol. 26, pp. 532-553, 1981.