PROCEEDINGS

AFOSR WORKSHOP in COMMUNICATION THEORY and APPLICATIONS

September 17-20, 1978 Provincetown, Massachusetts, U.S.A.

IEEE Library number: EHO 139-6

Sponsored by the Air Force Office of Scientific Research through Grant #AFOSR 78-3715 to The University of Connecticut, Storrs, Conn.

Also sponsored by the IEEE Communications and Information Theory Societies.
DISTRIBUTION-FREE CONSISTENCY RESULTS
IN DISCRIMINATION

Luc P. Devroye
School of Computer Science
Mc Gill University
P.O. Box 6070, Station A
Montreal H3C 3G1
Canada

Summary.

If one guesses at the value of a \{1,\ldots,M\}-valued random variable \(Y\) by using some function \(g(X)\) of an \(R^d\)-dimensional random vector \(X\), then no function \(g\) can be found for which the probability of a wrong guess is smaller than \(L^*\), the Bayes probability of error. The value \(L^*\) and the optimal mapping \(g\) are completely determined by the distribution of \((X,Y)\). Assume that the only information about this distribution is contained in a sample \(D_n=(X_1,Y_1),\ldots,(X_n,Y_n)\) of independent random vectors distributed as \((X,Y)\). The problem we are interested in is the one of the estimation of \(Y\) by some function \(g_n(X)\) of \(X\) and the data \(D_n\).

A sequence of such functions defines a discrimination rule. For given \(n\), the probability of error is \(L_n^*=\mathbb{P}(g_n(X)\neq Y|D_n)\). Ideally one would like to have some guarantee that at least for \(n\) large enough this probability of error is sufficiently close to the Bayes probability of error. This is necessarily tied to the concept of Bayes risk consistency. We say that a discrimination rule is weakly Bayes risk consistent (w.b.r.c.) if \(E(L_n^*)\rightarrow L^*\) as \(n\rightarrow\infty\) (or, equivalently, if \(L_n^*\rightarrow L^*\) in probability), and strongly Bayes risk consistent (s.b.r.c.) when \(L_n^*\rightarrow L^*\) with probability one. Until recently, it was generally believed that these properties were true for certain discrimination rules under certain conditions on the distribution of \((X,Y)\). Recent advances in the area of nonparametric discrimination seem to indicate that all "reasonable" discrimination rules are w.b.r.c. for all distributions of \((X,Y)\), that is, they are universally w.b.r.c.

We will attempt to sketch the development of some of these results and to indicate the key references.

Discrimination With Nearest Neighbor Rules.

In 1951, Fix and Hodges 12 showed that when the conditional distributions of \(X\) given \(Y=1\) all have almost everywhere continuous densities \(f_{1,j}\), then the \(k\)-nearest neighbor rule is w.b.r.c. under conditions on \(k\) and \(n\) that come very close to \(k=\infty\) and \(k/n=0\). The \(k\)-nearest neighbor rule picks for \(g_n(X)\) the class \(i\) with maximal representation among those \((X_{j,Y_{j}})\) for which \(X_{j}\) is among the \(k\) nearest neighbors to \(X\). Since local properties of the distribution of \((X,Y)\) are exploited here, one may find it very natural to have to impose some continuity condition on the regression functions \(P\{Y=1|X=x\}\) (a posteriori probabilities) as in the one proposed by Fix and Hodges 7.

To some people's surprise, Stone 31 was able to show in 1977 that under the same conditions on \(k\), the \(k\)-nearest neighbor rule is universally w.b.r.c.. His result is applicable to a larger class of rules which includes voting schemes of Royall 27 where the \(i\)-th nearest neighbor to \(X\) is given a vote \(v_{i}(n)\) while \(v_{i}(n)>v_{i+1}(n)>\ldots>v_{n}(n)>0\), (the \(k\)-nearest neighbor rule is a special case with \(v_{i}(n)=1\) for \(i<k\) and \(v_{i}(n)=0\) otherwise) and appropriate conditions are imposed on the votes.
Stone's very elegant proof is based on the observation that every Borel measurable function (such as \( P(Y \geq i | X=x) \)) is nearly continuous, a well-known fact from measure theory, and on a new nontrivial inequality to the effect that for all Borel measurable functions \( h \), after reordering \((X_1, \ldots, X_n)\) into \((X_{(1)}, \ldots, X_{(n)})\)

according to increasing values of \( |X_{(i)} - X| \),

\[
E \left( \sum_{i=1}^{n} V_i(n) \left|M(X_{(i)})\right| \right) \leq \alpha E\left( |h(X)| \right) \sum_{i=1}^{n} V_i(n),
\]

where \( \alpha \) is a parameter depending upon \( d \) only.

These results have been extended in two directions. First, if the nearest neighbor ordering is determined using an \( L_\infty \) norm on the ranks of the \( X_i \)'s, rather than through the Euclidean norm in \( \mathbb{R}^d \), then a discrimination rule results that is invariant to all strictly monotone transformations of the coordinate axes with regard to its probability of error, and that is universally w.b.r.c. \(^8\) \(^{21}\).

It has been noted that in low storage, fast computation situations the nearest neighbor rules become impractical, especially if a cheap supply of data is available. A sequential version of the nearest neighbor rule \(^{11}\) cuts the data sequence up into blocks of lengths \( f_1, f_2, \ldots \), finds the nearest neighbor to \( X \) in each block, and takes a vote with weights \( w_1, w_2, \ldots \) among the corresponding \( Y \)-values. This rule too is universally w.b.r.c. whenever \( \ell_n \to \infty \), \( \sum_{i=1}^{n} w_i = \infty \), and \( \frac{n}{\ell_n} \to 0 \). It is universally s.b.r.c. \(^7\) if also \( \sum_{i=1}^{n} (w_i / w_1)^2 < \infty \) and \( \ell_n / \log n \to \infty \).

**Discrimination Via Density Estimation.**

If conditional densities \( f_i \) exist, then one rule achieving \( L^* \) chooses \( g(X) \) according to

\[
\max_i f_i(X)
\]

when all classes \( i \) are equally probable. If in this formula \( f_i \) is replaced by an estimate \( \hat{f}_i \) then \( L_n \) is close to \( L^* \) whenever \( \hat{f}_i \) is close to \( f_i \) \(^{32}\) \(^7\). With the Parzen-Rosenblatt density estimate \(^{22}\) \(^{26}\) and taking into account unequal probabilities \( P(Y=1) \), rules are obtained of the following type: choose \( g_n(X) \) according to

\[
\max_i \sum_{j: Y_j = i} K((X_j - X)/h)
\]

where \( K \) is a given bounded density (kernel) and \( h>0 \) is a smoothing factor. The potential function method in pattern recognition was first formulated in this fashion by Sebestyen \(^{28}\) and Bashkirov et. al. \(^4\) (see also \(^1\) \(^2\)) for certain functions \( K \) that decrease with increasing values of \( |X_j - X| \). The consistency of \( K \)-rules under various conditions on the \( f_i \), \( K \) and \( h \) is treated in a series of papers \(^{32}\) \(^{14}\) \(^{24}\) \(^7\) \(^{15}\) \(^{17}\) \(^{18}\), all of which require \( \alpha \) a stronger than almost everywhere continuity condition on the \( f_i \)'s. However, the fact that for every density \( f \) almost every \( x \) is a Lebesgue point \(^ {29}\) (that is,

\[
a^{-d} \int \frac{|f(y) - f(x)|}{|y - x|} \, dy \to 0 \text{ as } a \to \infty,
\]

leads to the conclusion that weak Bayes risk consistency follows from \( h \to 0 \), \( h^{-d} \to \infty \), and

\[
\int \sup_{|y| > |x|} K(y) \, dx < \infty,
\]

39
whenever $X$ has a density $^9$. The next question is obvious: if this property is true for all densities of $X$, shouldn't it also be true for all distributions of $(X,Y)$? The answer, left as an open problem by Stone $^{31}$, is affirmative. Using probability theoretical covering lemmas, it is possible to show that kernel rules are universal w.r.t. under the same conditions on $h$ and some regularity conditions on $K$ $^{10}$.

Consistency results for recursive versions of kernel rules as developed in $^{33}$ $^{34}$ $^{24}$ $^{25}$ are also generalizable to the case that $X$ has any density $^9$.

**Discrimination With Partitioning Rules.**

Friedman $^{13}$ offers conclusive evidence that some algorithms that partition the space up into rectangular-shaped boxes and use majority rules on the member sets of the partition are computationally very attractive. These rules are rooted in the work of Stoller $^{30}$ and Anderson $^3$, and were studied, e.g., in $^{23}$ $^{24}$ $^{20}$. They are included in this survey because Gordon and Olshen $^{16}$ recently showed that under appropriate conditions on the construction of the partition these rules are w.r.t. for all distributions of $(X,Y)$.

**Open Problems.**

Stone's technique yields distribution-free weak Bayes risk consistency results. With minor modifications, strong consistency results can be obtained for sequential discrimination rules $^1$, but it is unknown whether conditions on the sequence of nearest neighbor weights can be found that ensure strong Bayes risk consistency for all such rules under no conditions on the distribution of $(X,Y)$.

All the discrimination rules mentioned above are oversimplifications: in this description, we do not take into account that data are preprocessed (scaling, dimensionality reduction, etc.), and that most parameters (window width $h$, neighborhood parameter $k$, etc.) are chosen as a function of the data and are therefore random variables not independent of $D$. I feel that for "reasonable" schemes to determine these parameters one probably does not have to reach to another level of sophistication in the proofs to show the universal Bayes risk consistency of practical versions of the basic nonparametric discrimination rules.

**References.**


whenever $X$ has a density $^9$. The next question is obvious: if this property is true for all densities of $X$, shouldn't it also be true for all distributions of $(X,Y)$? The answer, left as an open problem by Stone $^{31}$, is affirmative. Using probability theoretical covering lemmas, it is possible to show that kernel rules are universally w.b.r.c. under the same conditions on $h$ and some regularity conditions on $K$ $^{10}$.

Consistency results for recursive versions of kernel rules as developed in $^{32} 34 24 25$ are also generalizable to the case that $X$ has any density $^9$.

**Discrimination With Partitioning Rules.**

Friedman $^{13}$ offers conclusive evidence that some algorithms that recursively partition the space up into rectangular-shaped boxes and use majority rules on the member sets of the partition are computationally very attractive. These rules are rooted in the work of Stoller $^{30}$ and Anderson $^3$, and were studied, e.g., in $^{23} 19 20$. They are included in this survey because Gordon and Olshen $^{16}$ recently showed that under appropriate conditions on the construction of the partition these rules too are w.b.r.c. for all distributions of $(X,Y)$.

**Open Problems.**

Stone's technique yields distribution-free weak Bayes risk consistency results. With minor modifications, strong consistency results can be obtained for sequential discrimination rules $^{11}$, but it is unknown whether conditions on the sequence of nearest neighbor weights can be found that insure strong Bayes risk consistency for all such rules under no conditions on the distribution of $(X,Y)$.

All the discrimination rules mentioned above are oversimplifications: in their description, we do not take into account that data are preprocessed (scaling, dimensionality reduction, etc.), and that most parameters (window width $h$, the neighborhood parameter $k$, etc.) are chosen as a function of the data and are therefore random variables not independent of $D$. I feel that for "reasonable" schemes to determine these parameters one probably does not have to reach to another level of sophistication in the proofs to show the universal Bayes risk consistency of these practical versions of the basic nonparametric discrimination rules.

**References.**


