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 $\{1,\dots,M\}$-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 $D_n=(X_1,Y_1),\dots,(X_n,Y_n)$ 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 $L_n=P\{g_n(X)\neq Y|D_n\}$. 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 $E(L_n)\to L^*$ as $n\to\infty$ (or, equivalently, if $L_n\to L^*$ in probability), and strongly Bayes risk consistent (s.b.r.c.) when $L_n\to L^*$ 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=i all have almost everywhere continuous densities f_i, then the k-nearest neighbor rule is w.b.r.c. under conditions on k and n that come very close to $k\to\infty$ and $k/n\to 0$. 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 [5 6]. 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 $P\{Y=i|X=x\}$ (also called a posteriori probabilities) such as 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 a voting scheme of Royall [27] where the i-th nearest neighbor to X is given a vote $v_i(n)$ while $v_1(n)\geq v_2(n)\geq v_3(n)\geq\dots\geq v_n(n)\geq 0$, (the k-nearest neighbor rule is a special case with $v_i(n)=1$ for $i\leq k$ 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 $P\{Y=i|X=x\}$) 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 tc increasing values of $||X_i-X||$,

$$E\{\sum_{i=1}^{n} v_i(n)|h(X_{(i)})|\} \leq \alpha\, E\{|h(X)|\}\sum_{i=1}^{n} v_i(n) ,$$

where α 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_∞ norm on the ranks of the X_i's, rather than through the Euclidean norm in 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 computarion 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 $\ell_1,\ell_2,\ldots$, 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 $\ell_n\to\infty$, $\sum w_n=\infty$ and $w_n/\sum_1^n w_i\to 0$. It is universally s.b.r.c. if also $\sum(w_n/\sum_1^n w_i)^2<\infty$ 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 f_{ni} then L_n is close to L* whenever f_{ni} is close to f_i [32 7]. With the Parzen-Rosenblatt density estimate [22 26] and taking into account unequal probabilities $P\{Y=i\}$, 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 kernel 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 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,

$$a^{-d}\int_{||y-x||<a} |f(y)-f(x)|dy \to 0 \text{ as } a\to 0),$$

leads to the conclusion that weak Bayes risk consistency follows from $h\to 0$, $nh^d\to\infty$, and

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

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 [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 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_n. 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.

1 M.A.AIZERMAN,E.M.BRAVERMAN,L.I.ROZONOER:"Theoretical foundations of the potential function method in pattern recognition learning",Automation and Remote Control, vol. 25, pp. 821-837, 1964.

2 M.A.AIZERMAN,E.M.BRAVERMAN,L.I.ROZONOER:"The method of potential functions for the problem of restoring the characteristic of a function converter from randomly observed points",Automation and Remote Control, vol. 25, pp. 1546-1556, 1964.

3 T.W.ANDERSON:"Some nonparametric multivariate procedures based on statistically equivalent blocks",in: Multivariate Analysis, P.R.Krishnaiah Ed., Academic Press, New York, pp. 5-27, 1966.

4 O.BASHKIROV,E.M.BRAVERMAN,I.B.MUCHNIK:"Potential function algorithms for pattern recognition learning machines",Automation and Remote Control, vol. 25, pp. 629-631, 1964.

5 T.M.COVER,P.E.HART:"Nearest neighbor pattern classification",IEEE Transactions on Information Theory, vol. IT-13, pp. 21-27, 1967.

6 T.M.COVER:"Estimation by the nearest neighbor rule",IEEE Transactions on Information Theory, vol. IT-14, pp. 50-55, 1968.

7 L.P.DEVROYE,T.J.WAGNER:Nonparametric Discrimination And Density Estimation, Electronics Research Center, Univ. of Texas, Technical Report 183, 1976.

8 L.P.DEVROYE:"A universal k-nearest neighbor procedure in discrimination", Proceedings of the IEEE Computer Society Conference on Pattern Recognition and Image Processing, pp. 142-147, Chicago, 1978.

9 L.P.DEVROYE,T.J.WAGNER:"On the L1 convergence of kernel regression function estimators with applications in discrimination",submitted to Zeitschrift fur Wahrscheinlichkeitstheorie und Verwandte Gebiete, 1978.

10 L.P.DEVROYE,T.J.WAGNER:"Distribution-free consistency results in nonparametric discrimination and regression function estimation", submitted to Annals of Statistics, 1978.

11 L.P.DEVROYE,G.L.WISE:"Consistency of a sequential nearest neighbor regression function estimate",submitted to the Journal of Multivariate Analysis, 1978.

12 E.FIX,J.L.HODGES:Discriminatory Analysis.Nonparametric Discrimination:Consistency Properties, USAF School of Aviation Medicine, Randolph Field, Texas, Report 4, Project No. 21-49-004, 1951.

13 J.H.FRIEDMAN:"A recursive partitioning decision rule for nonparametric classification",IEEE Transactions on Computers, vol. C-26, pp. 404-408, 1977.

14 N.GLICK:"Sample-based classification procedures derived from density estimators", Journal of the American Statistical Association, vol. 67, pp. 116-122, 1972.

15 N.GLICK:"Sample-based classification procedures related to empiric distributions", IEEE Transactions on Information Theory, vol. IT-22, pp. 454-461, 1976.

16 L.GORDON,R.A.OLSHEN:"Asymptotically efficient solutions to the classification problem",Annals of Statistics, vol. 6, pp. 515-533, 1978.

17 W.GREBLICKI:"Pattern recognition procedures with nonparametric density estimates", IEEE Transactions an Systems,Man and Cybernetics, submitted in 1977.

18 W.GREBLICKI:"Asymptotically optimal pattern recognition procedures with density estimates",IEEE Transactions on Information Theory, vol. IT-24, pp. 250-251, 1978.

19 E.G.HENRICHON,K.S.FU:"A nonparametric partitioning procedure for pattern classification",IEEE Transactions on Computers, vol. C-18, pp. 614-624, 1969.

20 W.S.MEISEL,D.A.MICHALOPOULOS:"A partitioning algorithm with application in pattern classification and th optimization of decision trees",IEEE Transactions on Computers, vol. C-22, pp. 93-103, 1973.

21 R.A.OLSHEN:Comment on a paper by Stone [31],Annals of Statistics, vol. 5, pp. 620-621, 1977.

22 E.PARZEN:"On the estimation of a probability density function and the mode",Annals of Mathematical Statistics, vol. 33, pp. 1065-1076, 1962.

23 C.P.QUESENBERRY,M.P.GESSAMAN:"Nonparametric discrimination using tolerance regions", Annals of Mathematical Statistics, vol. 39, pp. 664-673, 1968.

24 L.REJTO,P.REVESZ:"Density estimation and pattern classification",Problems of Control and Information Theory, vol. 2, pp. 67-80, 1973.

25 P.REVESZ:"How to apply the method of stochastic approximation in the nonparametric estimation of a regression function", Mathematische Operationsforschung und Statistik, Series Statistics, vol. 8, pp. 119-126, 1977.

26 M.ROSENBLATT:"Remarks on some nonparametric estimates of a density function", Annals of Mathematical Statistics, vol. 27, pp. 832-837, 1957.

27 R.M.ROYALL:A Class Of Nonparametric Estimators Of A Smooth Regression Function, Ph.D.Dissertation, Stanford University, 1966.

28 G.SEBESTYEN:Decision Making Processes In Pattern Recognition,Macmillan,N.Y., 1962.

29 E.M.STEIN:Singular Integrals And Differentiability Properties Of Functions, Princeton University Press, Princeton, New Jersey, 1970.

30 D.S.STOLLER:"Univariate two-population distribution-free discrimination",Journal of the American Statistical Association, vol. 49, pp. 770-777, 1954.

31 C.J.STONE:"Consistent nonparametric regression",Annals of Statistics, vol. 5, pp. 595-645, 1977.

32 J.VAN RYZIN:"Bayes risk consistency of classification procedures using density estimation",Sankhya Series A, vol. 28, pp. 161-170, 1966.

33 J.VAN RYZIN:"A stochastic a posteriori updating algorithm for pattern recognition", Journal of Mathematical Analysis and Applications, vol. 20, pp. 359-379, 1967.

34 C.T.WOLVERTON,T.J.WAGNER:"Asymptotically optimal discriminant functions for pattern classification",IEEE Transactions on Information Theory, vol. IT-15, pp. 258-265, 1969.