LOWER BOUNDS FOR BAYES ERROR ESTIMATION

András Antos
Informatics Laboratory
Computer and Automation Research Institute of the
Hungarian Academy of Sciences
Lágymányosi u. 11
Budapest, Hungary H-1518

and

Luc Devroye
School of Computer Science
McGill University
Montreal, Canada H3A 2K6

and

László Györfi
Department of Computer Science and Information Theory
Techical University of Budapest
Stoczek u. 2
Budapest, Hungary H-1521

January 27, 1999

ABSTRACT. We give a short proof of the following result. Let (X,Y) be any distribution on $\mathcal{N} \times \{0,1\}$, and let $(X_1,Y_1),\ldots,(X_n,Y_n)$ be an i.i.d. sample drawn from this distribution. In discrimination, the Bayes error $L^* = \inf_g \mathbf{P}\{g(X) \neq Y\}$ is of crucial importance. Here we show that without further conditions on the distribution of (X,Y), no rate-of-convergence results can be obtained. Let $\phi_n(X_1,Y_1,\ldots,X_n,Y_n)$ be an estimate of the Bayes error, and let $\{\phi_n(.)\}$ be a sequence of such estimates. For any sequence $\{a_n\}$ of positive numbers converging to zero, a distribution of (X,Y) may be found such that

$$\mathbf{E}\left\{|L^* - \phi_n(X_1,Y_1,\ldots,X_n,Y_n)|\right\} \geq a_n$$

infinitely often.

INDEX TERMS. Discrimination, statistical pattern recognition, nonparametric estimation, Bayes error, lower bounds, rates of convergence.

The second author's work was supported by NSERC Grant A3456 and by FCAR Grant 90-ER-0291. The first and the third author was supported by a grant from the Hungarian Academy of Sciences (MTA SZTAKI).

1. Introduction.

The pattern recognition problem may be formulated as follows: we are given n i.i.d. observations $D_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$, drawn from the common unknown distribution of (X, Y) on $\mathbb{R}^d \times \{0, 1\}$. Given X, one must estimate Y as best as possible by a function $g_n(X)$ of X and the observations. The best one can hope for is to make an error equal to the Bayes error, L^*:

$$L_n \stackrel{\text{def}}{=} \mathbf{P}\{g_n(X) \neq Y | D_n\} \geq L^* \stackrel{\text{def}}{=} \inf_{g:\mathbb{R}^d \to \{0,1\}} \mathbf{P}\{g(X) \neq Y\} .$$

It is thus of great importance to be able to estimate L^* accurately, even before pattern recognition is attempted. Also, a comparison of estimates of L_n and L^* gives us an idea how much room is left for improvement.

In a first group of methods, L^* is estimated by an estimate $\widehat{L}_n$ of the error probability L_n of some consistent classification rule g_n. As such, this problem has been attempted by Fukunaga and Kessel (1971), Chen and Fu (1973), Fukunaga and Hummels (1987) and Garnett and Yau (1977), to cite just the early contributions. Concerning the error estimation of specific classification rules see Chapter 10 in McLachlan (1992). Clearly, if the estimate $\widehat{L}_n$ we use is consistent in the sense that $\widehat{L}_n - L_n \to 0$ with probability one as $n \to \infty$, and the rule is strongly consistent, then $\widehat{L}_n \to L^*$ with probability one. In other words, we have a consistent estimate of the Bayes error probability. The problem is that even though for many classifiers, $\widehat{L}_n - L_n$ can be guaranteed to converge to zero rapidly, regardless what the distribution of (X, Y) is (see Chapters 8, 23, 24 and 31 of Devroye, Györfi and Lugosi, 1996), in view of Cover (1968) and Devroye (1982), the rate of convergence of L_n to L^* using such a method may be arbitrarily slow. Thus, we can not expect a good performance for all distributions from such a method. The question thus is whether it is possible to come up with another method of estimating L^* (by $\phi_n(X_1, Y_1, \ldots, X_n, Y_n)$) such that the difference $\phi_n(X_1, Y_1, \ldots, X_n, Y_n) - L^*$ converges to zero rapidly for all distributions. Unfortunately, there is no method that guarantees a certain finite sample performance for all distributions. This disappointing fact is reflected in the following negative result (Theorem 8.5 of Devroye, Györfi and Lugosi, 1996):

THEOREM 1. *For every n, for any estimate $\phi_n(X_1, Y_1, \ldots, X_n, Y_n)$ of the Bayes error probability L^*, and for every $\epsilon > 0$, there exists a distribution of (X, Y), such that*

$$\mathbf{E}\{|\phi_n(X_1, Y_1, \ldots, X_n, Y_n) - L^*|\} \geq 1/4 - \epsilon .$$

The counterexamples in Theorem 1 vary with n, so it may still be possible that for every fixed distribution for (X, Y), there exists a universal rate of convergence to zero for $\mathbf{E}\{|\phi_n(X_1, Y_1, \ldots, X_n, Y_n) - L^*|\}$. The purpose of this note is to show that this too is impossible. We show the following:

THEOREM 2. *For any sequence $\{a_n\}$ of positive numbers converging to zero, a distribution of (X,Y) on $\{1,2,3,\ldots\}\times\{0,1\}$ may be found such that*

$$\mathbf{E}\left\{|\phi_n(X_1,Y_1,\ldots,X_n,Y_n)-L^*|\right\}\geq a_n$$

infinitely often.

We note that for the L_1 error in density estimation, similar global lower bounds were obtained by Devroye (1983, 1995) and Birgé (1986). We also note that the phrase "infinitely often" cannot be dropped from Theorem 2. Indeed, there exist deterministic sequences b_n with $|b_n - L^*| \leq c/n$ infinitely often for some constant c: just consider the dyadic sequence $0/2^0, 1/2^0, 0/2^1, 1/2^1, 2/2^1, 0/2^2, 1/2^2, 2/2^2, 3/2^2, 4/2^2, \ldots$. With $\phi_n \equiv b_n$, we thus obtain a very good estimate along an (unknown) subsequence.

2. Proof of Theorem 2.

Given $\{a_n\}$, we find a sequence of positive integers ℓ_n with a given property to be specified later. Then, we partition the positive integers into consecutive blocks of cardinality $\ell_1, \ell_2, \ell_3, \ldots$. Let $z=(z_1,z_2,\ldots)$ be a vector assigning a bit to each integer, and let $u=(u_1,u_2,\ldots)$ be a vector assigning a bit to each block. Then the distribution of (X,Y) is described constructively as follows: first a block B is drawn from the geometric distribution:

$$\mathbf{P}\{B=i\}=\frac{1}{2^i}\ ,\ i\geq 1\ .$$

Then X is drawn uniformly over the ℓ_B integers in that block. If $u_B=0$, then $Y=z_X$, while if $u_B=1$, Y is Bernoulli $(1/2)$. For this distribution, it is easy to verify that

$$L^*=L^*(u)=\sum_{i=1}^{\infty}\frac{u_i}{2^{i+1}}$$

as the only problem blocks are those with $u_i=1$, where locally, the Bayes error conditioned on $X\in$ block i is $1/2$. Note in particular that L^* depends upon u only.

Assume that all samples have the same common $X_1, X_2, \ldots, X_n$ components, consisting of i.i.d. observations drawn from the distribution of X (which is the same for all (u,z)). Then let $W=(W_1,\ldots,W_n)$ be a bit vector consisting of i.i.d. Bernoulli $(1/2)$ random bits. Then, define

$$Y_i=\begin{cases} z_{X_i} & \text{if } u_b=0 \text{ where } b \text{ is the block number of } X_i\\ W_i & \text{otherwise.}\end{cases}$$

Put

$$D_n=D_n(u,z)=\{(X_1,Y_1),\ldots(X_n,Y_n)\}.$$

In the proof we use randomization such that (u,z) is replaced by the independent random sequences (U,Z), where $U=(U_1,U_2,\ldots)$ and $Z=(Z_1,Z_2,\ldots)$ are i.i.d. Bernoulli $(1/2)$ sequences.

Let U_+^k and U_-^k denote the vector U, with the difference that U_+^k forces the k-th bit to be 1 and U_-^k forces the k-th bit to be 0. The event that all $X_1,\ldots X_n$ belong to block k are disjoint is denoted by T_k.

Introduce the notation

$$R_n(u,z) = \mathbf{E}\left\{|\phi_n(D_n(u,z)) - L^*(u)|\right\} ,$$

$$\bar{R}_n(u,z) = \sup_{m>n} \frac{R_m(u,z)}{a_m}$$

and

$$A_n = \{(u,z) : \bar{R}_n(u,z) \le 1\}.$$

Observe that given $X_1,\ldots,X_n$ and U and the event T_k, the distributions of $(D_n(U_-^k,Z), I_{\{(U_+^k,Z)\in A_n\}})$ and $(D_n(U_+^k,Z), I_{\{(U_+^k,Z)\in A_n\}})$ are equal. In order to see this decompose $D_n(U_+^k,Z)$ and $D_n(U_-^k,Z)$ as follows: Let $X_{i_1},\ldots,X_{i_M}$ be the subsample of $X_1,\ldots,X_n$ such that X_{i_j}'s fall into block k. Put

$$E_n(U,Z) = \{(X_{i_j}, Y_{i_j}) : j = 1,\ldots,M\}$$

and

$$F_n(U,Z) = D_n(U,Z) \setminus E_n(U,Z).$$

Then given $X_1,\ldots,X_n$ and U

$$F_n(U_+^k,Z) \equiv F_n(U_-^k,Z) \equiv F_n(U,Z),$$

while, because of

$$E_n(U_+^k,Z) = \{(X_{i_j}, W_{i_j}) : j = 1,\ldots,M\}$$

and

$$E_n(U_-^k,Z) = \{(X_{i_j}, Z_{X_{i_j}}) : j = 1,\ldots,M\},$$

under the event T_k $E_n(U_+^k,Z)$ and $E_n(U_-^k,Z)$ have the same distribution. Since the event $\{(U_+^k,Z) \in A_n\} = \{\bar{R}_n(U_+^k,Z) \le 1\}$ also only depends on $U \setminus U_k$ and Z_i's falling out of block k, $(F_n(U,Z), I_{\{(U_+^k,Z)\in A_n\}})$ is independent from both $E_n(U_+^k,Z)$ and $E_n(U_-^k,Z)$. This completes our statement.

We proceed with an indirect proof. Suppose that for all (u,z), $\bar{R}_n(u,z) \to 0$. This implies

$$\mathbf{P}\{(U,Z) \in A_n^c\} \to 0.$$

Thus by Fatou's lemma

$$\begin{aligned}
0 = \sup_{u,z} \limsup_n \frac{R_n(u,z)}{a_n} &\ge \mathbf{E}\left\{\limsup_n \frac{R_n(U,Z)}{a_n}\right\} \\
&\ge \mathbf{E}\{\limsup_n a_n^{-1} R_n(U,Z) I_{\{(U,Z)\in A_n\}}\} \\
&\ge \limsup_n a_n^{-1} \mathbf{E}\{R_n(U,Z) I_{\{(U,Z)\in A_n\}}\} \\
&= \limsup_n a_n^{-1} \mathbf{E}\{|\phi_n(D_n(U,Z)) - L^*(U)| I_{\{(U,Z)\in A_n\}}\} .
\end{aligned}$$

For proper $k = k_n$,

$$\begin{aligned}
&\limsup_n a_n^{-1} \mathbf{E}\left\{|\phi_n(D_n(U,Z)) - L^*(U)| I_{\{(U,Z)\in A_n\}}\right\} \\
&\ge \limsup_n a_n^{-1} \mathbf{E}\left\{\mathbf{E}\{|\phi_n(D_n(U,Z)) - L^*(U)| I_{\{(U,Z)\in A_n\}} | X_1,\ldots X_n, U\} I_{T_k}\right\}
\end{aligned}$$

$$
\begin{aligned}
&= \limsup_n a_n^{-1}\mathbf{E}\left\{ \Big(\mathbf{E}\{|\phi_n(D_n(U_-^k, Z)) - L^*(U_-^k)|I_{\{(U_-^k, Z)\in A_n\}}|X_1, \dots X_n, U\} I_{U_k=0} \right.\\
&\quad \left. +\ \mathbf{E}\{|\phi_n(D_n(U_+^k, Z)) - L^*(U_+^k)|I_{\{(U_+^k, Z)\in A_n\}}|X_1, \dots X_n, U\} I_{U_k=1} \Big) I_{T_k} \right\}\\
&= \limsup_n (2a_n)^{-1}\mathbf{E}\left\{ \Big(\mathbf{E}\{|\phi_n(D_n(U_-^k, Z)) - L^*(U_-^k)|I_{\{(U_-^k, Z)\in A_n\}}|X_1, \dots X_n, U\} \right.\\
&\quad \left. +\ \mathbf{E}\{|\phi_n(D_n(U_+^k, Z)) - L^*(U_+^k)|I_{\{(U_+^k, Z)\in A_n\}}|X_1, \dots X_n, U\} \Big) I_{T_k} \right\}\\
&= \limsup_n (2a_n)^{-1}\mathbf{E}\left\{ \Big(\mathbf{E}\{|\phi_n(D_n(U_-^k, Z)) - L^*(U_-^k)|I_{\{(U_-^k, Z)\in A_n\}}|X_1, \dots X_n, U\} \right.\\
&\quad \left. +\ \mathbf{E}\{|\phi_n(D_n(U_-^k, Z)) - L^*(U_-^k) - 1/2^{k+1}|I_{\{(U_+^k, Z)\in A_n\}}|X_1, \dots X_n, U\} \Big) I_{T_k} \right\},
\end{aligned}
$$

where in the last step we used that given $X_1, \dots X_n$ and U and the event T_k, the distributions of $(D_n(U_-^k, Z), I_{\{(U_+^k, Z)\in A_n\}})$ and $(D_n(U_+^k, Z), I_{\{(U_+^k, Z)\in A_n\}})$ are equal, and because of $L^*(U_+^k) = L^*(U_-^k) + 1/2^{k+1}$.

$$
\begin{aligned}
&\limsup_n a_n^{-1}\mathbf{E}\left\{|\phi_n(D_n(U, Z)) - L^*(U)|I_{\{(U,Z)\in A_n\}}\right\}\\
&\geq \limsup_n (2a_n)^{-1}\mathbf{E}\left\{ \Big(\mathbf{E}\left\{|\phi_n(D_n(U_-^k, Z)) - L^*(U_-^k)|I_{\left\{(U_-^k, Z)\in A_n\right\}}|X_1, \dots X_n, U\right\} \right.\\
&\quad \left. +\ \mathbf{E}\left\{|\phi_n(D_n(U_-^k, Z)) - L^*(U_-^k) - 1/2^{k+1}|I_{\left\{(U_+^k, Z)\in A_n\right\}}|X_1, \dots X_n, U\right\} \Big) I_{T_k} \right\}\\
&\geq \limsup_n (2a_n)^{-1}\mathbf{E}\left\{\mathbf{E}\left\{1/2^{k+1} I_{\left\{(U_-^k, Z)\in A_n\right\}} I_{\left\{(U_+^k, Z)\in A_n\right\}}|X_1, \dots X_n, U\right\} I_{T_k}\right\}\\
&= \limsup_n (2^{k+2}a_n)^{-1}\mathbf{P}\left\{(U_-^k, Z)\in A_n, (U_+^k, Z)\in A_n, T_k\right\}\\
&= \limsup_n (2^{k+2}a_n)^{-1}\mathbf{P}\left\{(U_-^k, Z)\in A_n, (U_+^k, Z)\in A_n\right\}\mathbf{P}\left\{T_k\right\}\\
&\geq \limsup_n (2^{k+2}a_n)^{-1}\left(1 - 2\frac{\mathbf{P}\left\{(U_-^k, Z)\in A_n^c\right\} + \mathbf{P}\left\{(U_+^k, Z)\in A_n^c\right\}}{2}\right)\mathbf{P}\left\{T_k\right\}\\
&= \limsup_n (2^{k+2}a_n)^{-1}\left(1 - 2\mathbf{P}\left\{(U, Z)\in A_n^c\right\}\right)\mathbf{P}\left\{T_k\right\}\\
&\geq \limsup_n (2^{k+2}a_n)^{-1}\left(1 - 2\mathbf{P}\left\{(U, Z)\in A_n^c\right\}\right)\left(1 - \frac{n^2}{2^{2k+1}\ell_k}\right) \geq 1/2
\end{aligned}
$$

if $1/2^{k_n+2} \geq a_n$ (e.g., the choice $k_n = \lfloor \log_2(1/a_n)\rfloor - 2 \to \infty$ is fine) and $\ell_{k_n} \geq n^2$. This can be satisfied by the choice $\ell_i = \max^2\{n : k_n \leq i\}$.

This is a contradiction, therefore for all $\{a_n\}$, there is $u, z, \{\ell_k\}$, that $\limsup_n \frac{R_n(u,z)}{a_n} > 0$, which implies that for all $\{a_n\}$, there is $u, z, \{\ell_k\}$, that

$$
\mathbf{E}|\phi_n(D_n(u, z)) - L^*(u)| > ca_n
$$

infinitely often. Applying this to the original $\sqrt{a_n}$, this concludes the proof of the Theorem.

3. Conclusion.

In a standard pattern recognition design process, one takes a number of features, and evaluates whether these suffice or will do for discrimination. If so, a discrimination method is designed. If not, more or different features must be considered. The quality of a collection of features is measured by the Bayes probability of error, L^*. Thus, the first phase of any pattern recognition method is based on estimates of L^* (even *before* a discriminant is picked!). In this paper, we show that no one can trust any Bayes error estimate, and that it is futile to even let the sample size tend to infinity. It is only possible to give error bounds or confidence bands for L^* under assumptions on the distribution of the data. In practice, one can thus never claim to have a universally superior feature extraction or Bayes error estimation method, no matter how many simulations are performed and no matter how large the sample sizes are.

4. References.

[1] L. Birgé, "On estimating a density using Hellinger distance and some other strange facts," *Probability Theory and Related Fields*, vol. 71, pp. 271–291, 1986.

[2] Z. Chen and K. S. Fu, "Nonparametric Bayes risk estimation for pattern classification," in: *Proceedings of the IEEE Conference on Systems, Man, and Cybernetics*, Boston, MA, 1973.

[3] T. M. Cover, "Rates of convergence for nearest neighbor procedures," In *Proceedings of the Hawaii International Conference on Systems Sciences*, Honolulu pp. 413–415, 1968.

[4] L. Devroye, "Any discrimination rule can have an arbitrarily bad probability of error for finite sample size," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 4, pp. 154–157, 1982.

[5] L. Devroye, "On arbitrarily slow rates of global convergence in density estimation," *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, vol. 62, pp. 475–483, 1983.

[6] L. Devroye, "Another proof of a slow convergence result of Birgé," *Statistics and Probability Letters*, vol. 23, pp. 63–67, 1995.

[7] L. Devroye, L. Györfi, and G. Lugosi, *A Probabilistic Theory of Pattern Recognition*, Springer-Verlag, New York, 1996.

[8] K. Fukunaga and D. M. Hummels, "Bias of nearest neighbor error estimates," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-9, pp. 103–112, 1987.

[9] K. Fukunaga and D. L. Kessel, "Estimation of classification error," *IEEE Transactions on Computers*, vol. 20, pp. 1521—1527, 1971.

[10] J. M. Garnett and S. S. Yau, "Nonparametric estimation of the Bayes error of feature extractors using ordered nearest neighbor sets," *IEEE Transactions on Computers*, vol. C-26, pp. 46–54, 1977.

[11] G. McLachlan, *Discriminant Analysis and Statistical Pattern Recognition*, John Wiley, New York, 1992.