RATES OF CONVERGENCE FOR NEAREST NEIGHBOR PROCEDURES

Thomas M. Cover
Stanford University
25 September 1967

Introduction and Summary

In previous papers [1] and [2] we have considered the question of estimating the parameter θ associated with a given observation x on the basis of n independent identically distributed parameter-observation pairs $(x_1,\theta_1),(x_2,\theta_2),\ldots,(x_n,\theta_n)$ drawn from the same distribution as (x, θ). We have shown for a probability of error and squared error loss criterion that the nearest-neighbor (NN) procedure (which associates with x the parameter θ_n' of the nearest neighbor among the observations $x_1,\ldots,x_n$) yields a large sample risk which is less than twice the Bayes risk under suitable regularity conditions.

In this paper we investigate the obvious question of the finite sample behavior of nearest neighbor estimates. Our most general conclusion is that, under suitable smoothness conditions, the n-sample nearest neighbor risk R_n converges to its limit R on the order of $1/n^2$. We are also able to show the vastness of the nonparametric family of distributions by giving examples in which the best nonparametric procedure has an arbitrarily slow algebraic rate of learning. Thus there is no hope in general of establishing efficacies of nonparametric classification procedures until regularity conditions are placed on the family.

Theorems

We shall first consider a classification problem with the probability of error loss criterion. Let $\theta_1,\theta_2,\ldots,\theta_n$ be independent identically distributed Bernoulli random variables taking the values 1 and 2 with probabilities p and $1-p$ respectively. Let x_i be drawn according to $f_{\theta_i}(x)$ where $f_1(x)$ and $f_2(x)$ are probability densities defined on the real line. Let R_n be the probability of error of the nearest neighbor procedure based on n samples expressed by:

$$R_n = \Pr\{\theta_n' \neq \theta\}$$

Let $R_\infty = \lim R_n$, and let R^* be the Bayes probability of error against known p, f_1 and f_2. In [1] it is shown that $R^* \leq R_\infty \leq 2R^*(1-R^*)$.

We note that the NN classification of an observation in the (n+1)-sample case differs from the classification based on n samples only if the $(n+1)^{th}$ sample is closest. Since this occurs with probability $1/(n+1)$, we have the following fact:

Theorem 1: $|R_n - R_{n+1}| \leq 1/(n+1)$, and in general $|R_n - R_{n+k}| \leq k/(n+1)$.

Remark: Thus increasing the number of samples by 10% will decrease the probability of error by at most .1. This bound on the fluctuations of the small-sample risk is rather conservative. The actual difference $|R_n - R_{n+1}|$ for large n is of the order of c/n^3 as will be seen from Theorem 2.

The following theorem demonstrates the rate at which R_n approaches its limit in the typical case, where the observations are real valued.

Theorem 2: Let $f_1(x)$ and $f_2(x)$ have uniformly bounded third derivatives and be bounded away from zero on their (probability one) support sets. Then $R_n = R_\infty + O(1/n^2)$.

Outline of Proof: We must show that there exists a constant k such that $|R_n - R_\infty| \leq k/n^2$ for sufficiently large n. The key to our proof is the realization that for a.e. x, $E[(x - x'_n)|x] \to 0$, and $E[(x-x'_n)^2|x] \sim 1/2n^2\, f^2(x)$, where $f(x) = pf_1(x) + (1-p)f_2(x)$. This is established by noting that, for large n, the samples near x are essentially Poisson distributed with density $\lambda = nf(x)$. Inclusion of these results into an appropriate Taylor series expansion of R_n yields the theorem. Recall that the example in [1] in which x is drawn from one of two equilikely triangular distributions resulted in the exact expression $R_n = R_\infty + 1/(n+1)(n+2)$, in agreement with the behavior developed in this theorem.

We shall now show that for the countable observation space problem there is no hope of finding interesting rates of convergence. This state of affairs comes about because the class of nonparametric classification problems includes so many problems that even optimal procedures converge arbitrarily slowly. To illustrate this point we shall select a problem for which the single-NN rule is optimal but has a rate of convergence of $O(i^{-\delta})$ for arbitrarily small $\delta > 0$.

Let $X = \{1,2,3,\ldots.\}$, and let $Pr\{x = i\} = p_i = c/i^{1+\delta}$, where c is such that $\Sigma p = 1$. Let $\theta_1, \theta_2, \theta_3, \ldots.$ be independent, identically

distributed Bernoulli random variables with $\Pr\{\theta_i = 1\} = 1/2 = \Pr\{\theta_i = 2\}$. This establishes a measure on the set of possible distributions $f(x,\theta)$ on $X \times \Theta$. Clearly the Bayes risk R^* is zero for each possible distribution, because knowledge of x yields perfect knowledge of the corresponding classification θ. However, if the only knowledge of $f(x,\theta)$ is that which may be inferred from samples, then the probability of error of the optimal procedure is $R_n^* = \frac{1}{2} \Sigma p_i (1-p_i)^n$. This is precisely the probability of error for the NN procedure. Inspection of this relation reveals that $R_n^* \geq c(\delta) n^{-\delta/(1+\delta)}$; thus the algebraic convergence is arbitrarily slow. We remark that this example may also be made continuous. We see in this case that the NN procedure is slow, not because the procedure is inherently bad, but because there exist some nonparametric problems for which the optimal decision procedure has the same slow rate of convergence.

References

[1] Cover, T. M., and Hart, P. E., "Nearest Neighbor Pattern Classification", IEEE Transactions on Information Theory, Vol. IT-13, No. 1, January 1967, pp. 21-27.

[2] Cover, Thomas M., "Estimation by the Nearest-Neighbor Rule", to appear, IEEE Transactions on Information Theory (October 1967 or January 1968).