P. R. Krishnaiah and L. N. Kanal, eds., *Handbook of Statistics, Vol. 2*

Nearest Neighbor Methods in Discrimination

L. Devroye and T. J. Wagner

In the discrimination problem one makes an observation $X=(X_1,\ldots,X_d)$ on some object whose state θ is known to be in some finite set which we may take to be $\{1,\ldots,M\}$. Assuming that the object is picked at random from some population, (X,θ) is a random vector with an arbitrary probability distribution. All that is assumed known about this distribution is that which can be inferred from a sample $(X_1,\theta_1),\ldots,(X_n,\theta_n)$ of size n made from objects drawn from the same population used for (X,θ). This sample, called *data*, is assumed to be independent of (X,θ). Using X and the data one makes an estimate $\hat{\theta}$ for θ where the procedure used for making this estimate is called a *rule*.

The rule which is the standard example for the class of rules considered in this article is the k-nearest neighbor rule. Here $\hat{\theta}$ is taken to be the state which occurs most frequently among the states of the k closest measurements to X from $X_1,\ldots,X_n$. To break ties in determining which of the vectors $X_1,\ldots,X_n$ is among the k closest to X and to break ties in determining which state occurs most frequently among these k closest, the independent sequence $Z,Z_1,\ldots,Z_n$ of independent random variables, each with a uniform distribution on $[0,1]$, is generated. We will think of Z as being attached to X and Z_i as being attached to $X_i, 1\leq i\leq n$. Then X_i is closer to X than X_j if

(a) $\|X-X_i\|<\|X-X_j\|$ or

(b) $\|X-X_i\|=\|X-X_j\|$ and $|Z-Z_i|<|Z-Z_j|$ or

(c) $\|X-X_i\|=\|X-X_j\|$, $|Z-Z_i|=|Z-Z_j|$ and $i<j$.

The k closest vectors to X from $X_1,\ldots,X_n$ are now determined and $\hat{\theta}$ is taken as the state occurring most frequently among these vectors. If several states occur most frequently among the k closest, the state whose observation is closest to X from among those tied is chosen. If (X^j,θ^j,Z^j) represents the jth closest observation to X, its corresponding state, and attached random variable, then we see that $\hat{\theta}$ for the k-nearest neighbor rule can be written as

$$\hat{\theta}=g\big((X^1,Z^1,\theta^1),\ldots,(X^k,Z^k,\theta^k)\big) \tag{1}$$

for some function g. Rules which have the form

$$\hat{\theta} = g_n((X^1, Z^1, \theta^1), \ldots, (X^n, Z^n, \theta^n)) \tag{2}$$

for some function g_n are termed nearest neighbor rules, while rules which can be put in the form (1) for some g are called k-local.

The probability of error for a rule given the data and attached random variables is given by

$$L_n = \mathbf{P}[\hat{\theta} \neq \theta \mid D_n]$$

where

$$D_n = ((X_1, \theta_1, Z_1), \ldots, (X_n, \theta_n, Z_n)).$$

The frequency interpretation of L_n is that a large number of new observations, whose states are estimated with the rule and the given data, will produce a frequency of errors equal to the value of L_n. (Each of these new observations will have a new independent Z attached to it but the $Z_1, \ldots, Z_n$ stay fixed with the data.) The random variable L_n is important then because it measures the future performance of the rule with the given data.

Most of the results dealing with nearest neighbor rules are of the asymptotic variety, that is, results concerned with where L_n converges to and how it converges as n tends to infinity. If the limiting behavior of L_n compares favorably to L^*, the Bayes probability of error (the smallest possible probability of error if one knew the distribution of (X, θ)), then one has some hope that the rule will at least perform well with large amounts of data. For the k-nearest neighbor rule with fixed k the first result of this type, and certainly the best known, is that of Cover and Hart (1967) who showed that

$$\mathbf{E}L_n \overset{n}{\to} L \tag{3}$$

when $\mathbf{P}[\theta = i \mid X = x]$ has an almost everywhere continuous version, $1 \leq i \leq M$. In (3) L is a constant satisfying, for $k = 1$,

$$L^* \leq L \leq 2L^*(1 - L^*) \leq 2L^*. \tag{4}$$

For arbitrary k the "2" in (4) is replaced by α_k where $\alpha_k \downarrow 1$ as $k \to \infty$. For these same assumptions it is also known that

$$L_n \overset{n}{\to} L \text{ in probability} \tag{5}$$

(Wagner, 1971) with convergence in (5) actually being with probability one for $k = 1$ (Fritz, 1975).

If k is allowed to vary with n, then Stone (1977) showed that for *any* distribution of (X, θ)

$$L_n \overset{n}{\to} L^* \text{ in probability} \tag{6}$$

if

$$k = k_n \overset{n}{\to} \infty \quad \text{and} \quad k_n/n \overset{n}{\to} 0.$$

This distribution-free result extends to a large class of nearest neighbor rules, which are also discussed by Stone and, because of its sheer technical achievement, rivals the original accomplishment of Fix and Hodges (1951) who introduced k-nearest neighbor rules and proved (6) in a slightly different setting *with* analytic assumptions on the distribution of (X, θ). We should note here that Stone breaks ties differently than described earlier. For example, if $k_n = 5$ and if six vectors, with the attached Z's, have positions 4–9 in the distance ordering of $X_1, \ldots, X_n$ to X and all have the same distance to X, then each of the states of these six vectors gets a $2/6 = 1/3$ 'vote' for the estimate $\hat{\theta}$. By contrast, in the first way of breaking ties two of these six vectors would get one vote each and the other four would get 0. Devroye (1981a) has recently shown that if one also assumes that

$$k_n/(\log n) \overset{n}{\to} \infty,$$

then (6) holds with the convergence being with probability one.

In view of Stone's result, it might be expected that the asymptotic results of the k-nearest neighbor rule with k fixed are also distribution-free, that is, no conditions on the distribution of (X, θ) are needed for (5). In fact, using Stone's way of breaking ties, Devroye (1981b) has shown exactly that. Moreover, the constant L for the general case, which is the same as Cover and Hart's for their assumptions on the distribution of (X, θ), continues to obey the inequality (4).

As intellectually satisfying as these results are, one is still faced with the finite sample situation. You have data D_n and your immediate need is for a reliable estimate of L_n for your chosen rule. You may even wish to examine the data and then pick the rule. In this case reliable estimates of L_n for each rule may guide you in your choice. If one is using a *local* rule, then a natural estimate is the deleted estimate of L_n given by

$$\hat{L}_n = (1/n) \sum_{i=1}^{n} I_{[\hat{\theta}_i \neq \theta_i]}$$

where $\hat{\theta}_i$ is the estimate of θ_i from X_i, Z_i, and D_n with (X_i, θ_i, Z_i) deleted. This definition requires, of course, that $k \leq n-1$. Deleted estimates are not easy to compute but, in cases like the k-nearest neighbor rule, the computation is reasonable and the intuitively appealing use of the data can be taken advantage of. Rogers and Wagner (1977) have shown that for *all* distributions of (X, θ) and any k-local rule

$$\mathbf{E}\left(\hat{L}_n - L_n\right)^2 \leq \frac{2k + 1/4}{n} + \frac{2k(2k + 1/4)^{1/2}}{n^{3/2}} + \frac{k^2}{n^2}. \tag{7}$$

Using Chebychev's inequality and (7), distribution-free upper bounds for $\mathbf{P}[|\hat{L}_n - L_n| \geqslant \varepsilon]$ can be obtained which are $O(1/n)$. In Devroye and Wagner (1979a) distribution-free upper bounds for $\mathbf{P}[|L_n - \hat{L}_n| \geqslant \varepsilon]$ of the form Ae^{-nB} are also given where A and B are positive constants which depend only on d, M, and ε. In these bounds, however, the rate of decrease of B to 0 with d is quite rapid. In contrast, the right-hand side of (7) does not depend on d at all. Finally, simulations carried out by Penrod and Wagner (1979) suggest that $2e^{-2n\varepsilon^2}$ is generally an upper bound for $\mathbf{P}[|\hat{L}_n - L_n| \geqslant \varepsilon]$. Other estimates of L_n are discussed in the references mentioned above.

If one considers just the single nearest neighbor rule for the finite sample case, two features stand out. The first is that one must store and search all of the data for each of the future estimates. The second point is that the nearest neighbor rule performance deteriorates from the Bayes rule (e.g., the rule used to achieve L^* when the distribution of (X, θ) is known), because in the region of $\mathbb{R}^d$ where $\mathbf{P}[\theta = m | X = x]$ is maximal (which is where $\hat{\theta}(x) = m$ in the Bayes rule) all of the samples X_i which fall there 'carve' out a subset where $\hat{\theta} = \theta_i$, regardless of whether $i = m$ or not. To reduce one or both of these effects, many authors have suggested condensing or editing the data before the nearest neighbor rule is applied (e.g., see Ritter et al. (1975) for recent references). There are no really general asymptotic results for condensing methods at this writing, but it seems clear that condensing, properly done, will definitely reduce computation for future estimates *and* improve performance. Devroye and Wagner (1979b) have also shown that if the original data is condensed in *any* way to J points,

$$(Y_1, \xi_1), \ldots, (Y_J, \xi_J), \tag{8}$$

and if the single nearest neighbor rule is used with these J points, then

$$\mathbf{P}\left[|L_n - \hat{\hat{L}}_n| \geqslant \varepsilon\right] \leqslant 4(4n)^{dJ(J-1)} e^{-n\varepsilon^2/8} \tag{9}$$

where $\hat{\hat{L}}_n$ is the frequency of errors one gets on the original data with the single nearest neighbor rule now using (8) as data. The right-hand side of (9) is, of course, distribution-free, but requires that J be small to be useful.

References

Cover, T. and Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Trans. Inform. Theory* **11**, 21–27.

Devroye, L. (1981a). On the almost everywhere convergence of nonparametric regression function estimates. *Ann. Statist.* **9**, 1310–1319.

Devroye, L. P. (1981b). On the inequality of Cover and Hart in nearest neighbor discrimination. *IEEE Trans. Pattern Analysis Machine Intelligence* **3**, 75–78.

Devroye, L. P. and Wagner, T. J. (1979a). Distribution-free inequalities for the deleted and holdout error estimates. *IEEE Trans. Inform. Theory* **25**, 202–207.

Devroye, L. P. and Wagner, T. J. (1979b). Distribution-free performance bounds with the resubstitution error estimate. *IEEE Trans. Inform. Theory* **25**, 208–210.

Fix, E. and Hodges, J. (1951). Discriminatory analysis: Nonparametric discrimination: consistency properties. Rept. No. 4, USAF School of Aviation Medicine, Randolph Field, TX.

Fritz, J. (1975). Distribution-free exponential error bound for nearest neighbor pattern classification. *IEEE Trans. Inform. Theory* **21**, 552–557.

Penrod, C. S. and Wagner, T. J. (1979). Risk estimation for nonparametric discrimination and estimation rules: A simulation study. *IEEE Trans. Inform. Theory* [to appear].

Ritter, G. L., Woodruff, H. B., Lowry, S. R., and Isenhour, T. L. (1975). An algorithm for a selective nearest neighbor rule. *IEEE Trans. Inform. Theory* **21**, 665–669.

Rogers, W. H. and Wagner, T. J. (1977). A finite sample distribution-free performance bound for local discrimination rules. *Ann. Statist.* **6**, 506–514.

Stone, C. J. (1977). Consistent nonparametric regression. *Ann. Statist.* **5**, 595–645.

Wagner, T. J. (1971). Convergence of the nearest neighbor rule. *IEEE Trans. Inform. Theory* **17**, 566–571.