k Nearest Neighbors in Search of a Metric

Robert R. Snapp[1]
Computer Science and Electrical Engineering Department
University of Vermont
Burlington, VT 05405 USA
snapp@emba.uvm.edu

Santosh S. Venkatesh
Department of Electrical Engineering
University of Pennsylvania
Philadelphia, PA 19104 USA
venkates@ee.upenn.edu

Abstract — **The finite-sample risk of the k-nearest neighbor classifier that uses a weighted L_p metric as a measure of class similarity is examined. For a family of multiclass, classification problems with smooth distributions in $\mathbb{R}^n$, the risk is represented as an asymptotic expansion in decreasing fractional powers of the reference-sample size. An analysis of the leading coefficients reveals that the optimal metric (i.e., the metric that minimizes the risk) tends to a weighted Euclidean (i.e., L_2) metric as the sample size is increased. Numerical calculations corroborate this finding.**

I. The k-Nearest-Neighbor Classifier

Let the elements of $\mathbb{L} = \{1, \ldots, C\}$ denote C states of nature, or pattern classes, and let $P_1, \ldots, P_C$ denote their corresponding stationary prior probabilities. Each pattern is represented by a feature vector $\mathbf{x}$, drawn at random from $\mathbb{R}^n$. Specifically, patterns originating from class $\ell \in \mathbb{L}$ are generated by the stationary conditional distribution F_ℓ.

Labeled feature vectors are generated by a two-step process. First, a class $\ell \in \mathbb{L}$ is chosen at random so that $\Pr[\ell = j] = P_j$; then a random feature vector is drawn according to F_ℓ. After m independent repetitions of this process, we obtain the labeled reference sample,

$$\mathcal{X}_m = \{(\mathbf{x}^1, \ell^1), \ldots, (\mathbf{x}^m, \ell^m)\}.$$

Given a weighted L_p metric, $d(\mathbf{x}, \mathbf{y}) = \|A(\mathbf{x}-\mathbf{y})\|_p$, where A denotes an n-by-n, positive-definite, symmetric matrix with $\det A = 1$, and an arbitrary point $\mathbf{x} \in \mathbb{R}^n$, the indices of the labeled feature vectors in $\mathcal{X}_m$ can be permuted so that

$$d(\mathbf{x}, \mathbf{x}^1) \le d(\mathbf{x}, \mathbf{x}^2) \le \cdots \le d(\mathbf{x}, \mathbf{x}^m). \tag{1}$$

Here $\|\mathbf{x}\|_p = (|x_1|^p + \cdots + |x_n|^p)^{1/p}$ for $1 \le p < \infty$, and $\|\mathbf{x}\|_\infty = \max_{1 \le i \le n} |x_i|$, denote the L_p norm. The k nearest neighbors of $\mathbf{x}$ then form the subset $\{(\mathbf{x}^1, \ell^1), \ldots, (\mathbf{x}^k, \ell^k)\}$; and the k-nearest-neighbor classifier assigns $\mathbf{x}$ to class $L'(\mathbf{x}) = \mathrm{maj}(\ell^1, \ldots, \ell^k)$, viz., the most frequently appearing class label in the subset. (Ties, and degeneracies in (1), can be resolved by an arbitrary procedure.) Using this algorithm every point in $\mathbb{R}^n$ can be assigned to a class in $\mathbb{L}$.

II. The Finite-Sample Risk

Given a positive integer k, an L_p metric, and a finite random reference sample $\mathcal{X}_m$, a single test vector $(\mathbf{X}, L)$, drawn independently by the same random process, is assigned to class $L' = L'(\mathbf{X})$ by the k-nearest-neighbor classifier. We now consider the m-sample risk,

$$R_m = \sum_{i=1}^{C} \sum_{j=1}^{C} \Lambda_{ij} \Pr[L' = i, L = j],$$

with the zero-one cost matrix $\Lambda_{ij} = 1 - \delta_{ij}$.

For a family of classification problems, $\mathcal{F}_N$, described by class-conditional probability densities f_ℓ with uniformly bounded partial derivative up through order $N+1$, and a mixture density $f = \sum_{\ell=1}^{C} P_\ell f_\ell$ that is bounded away from zero a.e. on its probability-one support $S \subset \mathbb{R}^n$, we obtain the following:

Theorem 1 *There exist constants c_j, for $j = 2, 3, \ldots, N$, such that*

$$R_m = R_\infty + \sum_{j=2}^{N} c_j m^{-j/n} + O(m^{-(N+1)/n})$$

where R_∞ is the infinite-sample risk derived by Cover and Hart [1].

(A version of this theorem, restricted to the case $k = 1$, $p = 2$, $A = I$, and $C = 2$, appears in a recent paper [2].) The coefficient c_2 evaluates to

$$c_2 = D_n(p) \frac{\Gamma(k + 1 + \frac{2}{n})}{24 \left[\Gamma\left(\frac{k+1}{2}\right)\right]^2} \mathrm{tr}\left\{(A^{-1})^T H A^{-1}\right\},$$

where,

$$D_n(p) = \frac{\Gamma\left(\frac{3}{p}+1\right)\Gamma\left(\frac{n}{p}+1\right)^{1+(2/n)}}{\Gamma\left(\frac{n+2}{p}+1\right)\Gamma\left(\frac{1}{p}+1\right)^3},$$

A^{-1} denotes the inverse of the metric weight matrix A, and H is an n-by-n matrix, independent of p. For the two-class problem ($C = 2$),

$$H_{ij} = \int_S d\mathbf{x} f^{1-\frac{2}{n}} (\hat{P}_1 \hat{P}_2)^{\frac{k+1}{2}} (\hat{P}_2 - \hat{P}_1) \left(\frac{1}{f_1} \frac{\partial^2 f_1}{\partial x_i \partial x_j} - \frac{1}{f_2} \frac{\partial^2 f_2}{\partial x_i \partial x_j} \right).$$

Here, $\hat{P}_\ell = P_\ell f_\ell(\mathbf{x}) / f(\mathbf{x})$ denotes the posterior probability that a feature vector with value $\mathbf{x}$ originates from class ℓ.

III. A Desirable Metric

Since R_∞ does not depend upon the chosen metric, Theorem 1 suggests that the finite-sample risk of the k-nearest-neighbor may be reduced, for large values of m, by selecting a metric that minimizes c_2. It can be shown that $D_n(p)$ has a global minimum at $p = 2$ for fixed $n > 1$. Using the Euler-Lagrange multiplier theorem, the trace in c_2 is minimized if the weight matrix A satisfies $A^T A = H/(\det H)^{1/n}$. Although it may be difficult to determine H, and consequently the optimal matrix A, in practice, this analysis and corroborating numerical simulations motivate the use of a weighted Euclidean metric for large reference samples.

[1] This work was supported in part by Rome Laboratory, Air Force Material Command, USAF, under grant number F30602-94-1-0010.

References

[1] T. M. Cover and P. E. Hart, "Nearest neighbor pattern classification," *IEEE Trans. Inform. Theory*, vol. IT-13, pp. 21-27, 1967.

[2] D. Psaltis, R. R. Snapp, and S. S. Venkatesh, "On the finite sample performance of the nearest neighbor classifier," *IEEE Trans. Inform. Theory*, vol. IT-40, pp. 820-837, 1994.