Asymptotic Predictions of the Finite-Sample Risk of the k-Nearest-Neighbor Classifier

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

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

Abstract

The finite-sample risk of the k*-nearest-neighbor classifier is analyzed for a family of two-class problems in which patterns are randomly generated from smooth probability distributions in an* n*-dimensional Euclidean feature space. First, an exact integral expression for the* m*-sample risk is obtained for a* k*-nearest-neighbor classifier that uses a reference sample of* m *labeled feature vectors. Using a multidimensional application of Laplace's method of integration, this integral can be represented as an asymptotic expansion in negative rational powers of* m*. The leading terms of this asymptotic expansion elucidate the curse of dimensionality and other properties of the finite-sample risk.*

1 Introduction

Since its introduction over forty years ago, the k-nearest-neighbor classifier continues to serve as an important pattern recognition paradigm. Using this algorithm, an input pattern (represented as a feature vector in an n-dimensional metric space) is assigned to a pattern class (i.e., a discrete state of nature) by means of a reference sample of m previously classified patterns. In its classic form, each input vector is assigned to the class represented by the majority of the k reference vectors that are closest to the input vector. (Ties can be broken by an arbitrary procedure.)

In practical applications, the accuracy of this algorithm is often competitive with leading pattern classifiers. Its favorable performance in the infinite-sample limit ($m \to \infty$) was revealed in the seminal analysis of Cover and Hart [2]. Specifically, if the reference sample and input patterns are independently selected at random by arbitrary, stationary distributions, then the statistical risk of the k-nearest neighbor classifier (e.g., the probability that a random input vector is misclassified) tends to a value that is close to the Bayes risk (viz., the optimal misclassification probability). In particular, as $m \to \infty$:

1. If $k = 1$, then the statistical risk tends to a value that does not exceed twice the Bayes risk [2]; and
2. If $k \to \infty$ such that $k/m \to 0$, then the statistical risk tends to the Bayes risk. [2, 9].

To complement these findings, we present new results that elucidate the performance of this classifier for realizable sample sizes. In particular, for a family of two-class problems with sufficiently smooth distributions in $\mathbb{R}^n$, we obtain an asymptotic description of the *finite-sample risk*, R_m, in the form

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

where R_∞ is the infinite-sample risk (cf. [2]), and N is an integer greater than one. We also give an explicit formula for the leading coefficient c_2 in terms of k and the underlying distributions, and describe how expressions for the higher-ordered coefficients can be obtained. The leading terms in (1) analytically underline the desire to represent patterns in low-dimensional feature spaces for effective convergence. For problems with known distributions that satisfy the requisite smoothness hypotheses, these leading terms can be used to estimate the finite-sample risk, or conversely, the size of a sample that yields an acceptable risk. More importantly, this asymptotic expansion provides a useful framework for analyzing properties of the finite-sample risk.

To our knowledge, Eqn. (1) represents the most complete description yet attained of the finite-sample risk of the k-nearest-neighbor algorithm for a broad family of nontrivial classification problems. It is obtained by Laplace's method of integration [3, 5] in direct analogy with a recent analysis of the nearest-neighbor classifier [6]. Other studies of finite-sample risk include Cover's one-dimensional analysis of the nearest-neighbor classifier [1] and the heuristic n-dimensional generalization of Fukunaga and Hummels [4].

1051-4651/94 $04.00

In Section 2, we formally describe the k-nearest-neighbor classifier. After a definition of the finite-sample risk, Section 3 presents our main theorem and the smoothness hypotheses under which it holds. An outline of a constructive proof appears in Section 4. Two examples, given in Section 5, illustrate the relevance of (1). The first example, a two-class problem with radial (i.e., spherically symmetric) distributions, clarifies how the leading expansion coefficients depend on k, n, and local properties of the underlying probability distributions. The second example — an empirical study of a two-class problem with multidimensional normal densities — suggests that the theorem holds for a larger family of distributions than is captured by our hypotheses. Finally we present our conclusions in Section 6.

Before proceeding to the main body, we list some of our notational conventions. Let $\mathbf{x}, \mathbf{x}^1, \ldots, \mathbf{x}^m$ denote points in $\mathbb{R}^n$, and let $\mathbf{X}, \mathbf{X}^1, \ldots, \mathbf{X}^m$ denote random vectors in that same space. For simplicity, we shall consider only two-class decision problems, with the set of class labels $\mathbb{L} = \{1, 2\}$. Let $\ell, \ell^1, \ldots, \ell^m$ denote (deterministic) class labels, and let $L, L^1, \ldots, L^m$ denote random class labels.

In the following, we assume that k is a predefined *odd* positive integer. The special symbol $\mathbf{l}$ denotes a k-dimensional vector with components $\ell^1, \ell^2, \ldots, \ell^k$. We define the *class majority function* as the mapping $\mathrm{maj} : \mathbb{L}^k \to \mathbb{L}$ that assigns a k-component class vector to the class represented by the majority of its components. (Note that ties are forbidden if k is an odd integer.) We define the *majority-j set* as $\mathcal{L}_j = \{\mathbf{l} : \mathrm{maj}(\mathbf{l}) = j\}$; thus, $\mathcal{L}_1$, and $\mathcal{L}_2$ form a complete, disjoint, partition of the space $\mathbb{L}^k$, for odd k.

We use the symbol

$$(m)_k = m(m-1)\cdots(m-k+1)$$

to represent the number of ways of selecting k elements from a set of size m without replacement. We let Γ denote Euler's gamma function, $\Gamma(z) = \int_0^\infty t^{z-1}e^{-t}\,dt$.

The norm of a vector $\mathbf{x} \in \mathbb{R}^n$ is represented by

$$\|\mathbf{x}\| = \sqrt{\sum_{i=1}^{n} x_i^2}.$$

We let $B(\rho, \mathbf{x})$ denote the set of points that lie inside the n-dimensional ball of radius ρ, and center at $\mathbf{x}$. We also let $S_{n-1} = \partial B(1, 0)$ denote the points that lie on the boundary of the n-dimensional, unit ball; and

$$V_n = \frac{\pi^{n/2}}{\Gamma(1 + n/2)},$$

denotes the volume of the n-dimensional unit ball.

Finally, we use the usual notation $\nabla^2 \equiv \sum_{i=1}^{n} \frac{\partial^2}{\partial x_i^2}$ for the n-dimensional Laplacian derivative.

2 The k-nearest-neighbor classifier

Let the elements of $\mathbb{L} = \{1, 2\}$ denote two states of nature corresponding to two pattern classes, and let P_1 and $P_2 = 1 - P_1$ denote their corresponding prior probabilities. Each pattern is represented by a random feature vector $\mathbf{X}$, drawn at random from $\mathbb{R}^n$. Specifically, patterns originating from class $\ell \in \mathbb{L}$ are generated by the conditional distribution F_ℓ. We let $F = P_1F_1 + P_2F_2$ denote the (unconditional) mixture distribution of the feature vector $\mathbf{X}$.

Labeled feature vectors are obtained by a two-step process: first the class $L \in \mathbb{L}$ is chosen at random, so that $\mathbf{P}[L = \ell] = P_\ell$ for $\ell \in \mathbb{L}$; then a random feature vector is drawn according to F_L. After m independent repetitions of this process, the labeled reference sample

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

is obtained.

Given a point $\mathbf{x} \in \mathbb{R}^n$, the ordered reference sample, $\mathcal{X}'_m$ contains every element of $\mathcal{X}_m$, with indices permuted so that

$$\|\mathbf{x} - \mathbf{X}^1\| \le \|\mathbf{x} - \mathbf{X}^2\| \le \cdots \le \|\mathbf{x} - \mathbf{X}^m\|. \quad (2)$$

(Equalities can be resolved by an arbitrary procedure; however, as we will assume that F_1 and F_2 are absolutely continuous, an equality in (2) is a zero-probability event.) The k-nearest neighbors of $\mathbf{x}$ form the subset $\{(\mathbf{X}^1, L^1), \ldots, (\mathbf{X}^k, L^k)\} \subset \mathcal{X}'_m$; and the k-nearest-neighbor classifier assigns $\mathbf{x}$ to the class $L' = \mathrm{maj}(\mathbf{L})$, where $\mathbf{L} = (L^1, \ldots, L^k)$. Using this algorithm every point in $\mathbb{R}^n$ can be assigned to a class in $\mathbb{L}$.

3 Finite-sample risk

A single test vector $(\mathbf{X}, L)$ is drawn by the same two-step process, independent of $\mathcal{X}_m$. The finite-sample risk, is

$$R_m = \lambda_{12}\mathbf{P}[L' = 1, L = 2] + \lambda_{21}\mathbf{P}[L' = 2, L = 1],$$

where $\lambda_{i,j}$ is the cost of assigning a pattern from class j to class i. (We assume that $\lambda_{11} = \lambda_{22} = 0$.) If $\lambda_{12} = \lambda_{21} = 1$, then $R_m = \mathbf{P}[L' \neq L]$, the probability that the input vector is misclassified. For simplicity, we assume this *zero-one* loss function. In addition, we assume that the two-class problems under consideration satisfy the following smoothness hypotheses.

3.1 Smoothness hypotheses

H1. For $\ell \in \{1, 2\}$, the class-conditional distributions F_ℓ are absolutely continuous over $\mathbb{R}^n$ and have corresponding densities f_ℓ.

H2. The mixture density, $f = P_1f_1 + P_2f_2$ is bounded away from zero a.e.[1] over its probability-one support $\mathcal{S} \subset \mathbb{R}^n$. (We can thus assume, without loss

[1] Here and elsewhere, with respect to Lebesgue measure.

of generality, that $\mathcal{S}$ is compact.) We also assume that the probability-one supports of f_1 and f_2 intersect over a set of positive measure.

H3. The class-conditional densities f_j possess uniformly bounded partial derivatives up to order $N+1$ almost everywhere on their probability one support.

H4. One or the other of the class-conditional densities vanishes close to the boundary of $\mathcal{S}$. More precisely, let $\partial\mathcal{S} = \text{closure}(\mathcal{S}) \cap \text{closure}(\mathbb{R}^n \setminus \mathcal{S})$ denote the boundary of $\mathcal{S}$, and for $t \geq 0$, let $\bar{\mathcal{S}}_t \subset \mathcal{S}$ denote the set of points in $\mathcal{S}$ of distance no more than t from the boundary:

$$\bar{\mathcal{S}}_t = \{\mathbf{x} \in \mathcal{S} : \|\mathbf{x} - \partial\mathcal{S}\| \leq t\}.$$

(Note, for example, $\bar{\mathcal{S}}_0 = \partial\mathcal{S}$.) Then there exists a $t_0 > 0$ such that for a.e. $\mathbf{x} \in \bar{\mathcal{S}}_{t_0}$, either $f_1(\mathbf{x}) = 0$ or $f_2(\mathbf{x}) = 0$.

REMARKS: Cover [1] has shown that Hypothesis H1 is necessary; otherwise the rate of convergence can be arbitrarily slow, even in one dimension. Hypothesis H2 satisfies a uniformity condition required by our application of Laplace's method of integration in Section 4. Hypothesis H3 permits each class-conditional density to be represented as a truncated Taylor series. Actually, it is possible to generalize this hypothesis slightly by requiring each class-conditional density to have a uniform asymptotic expansion in the form

$$f_j(\mathbf{x}') = f_j(\mathbf{x}) + \sum_{i=1}^{N} f_{j,i}(\Omega, \mathbf{x})\rho^i + o(\rho^N) \qquad (3)$$

about almost every point in $\mathcal{S}$. Here, $\rho = \|\mathbf{x}' - \mathbf{x}\|$ and $\Omega = (\mathbf{x}' - \mathbf{x})/\rho$ (cf. [6]). Hypothesis H4 is introduced to eliminate boundary contributions to the risk integral. After analyzing some specific examples, e.g., the triangular distribution in [2], it is evident that this condition is not necessary.

3.2 The main theorem

Theorem 3.1 *Under Hypotheses H1 through H4 (stated above), 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

$$R_\infty = \int_{\mathcal{S}} d\mathbf{x}\, f(\mathbf{x}) \sum_{j=0}^{(k-1)/2} \binom{k}{j} \left(\widehat{P}_1(\mathbf{x})^{j+1}\, \widehat{P}_2(\mathbf{x})^{k-j} + \widehat{P}_1(\mathbf{x})^{k-j}\, \widehat{P}_2(\mathbf{x})^{j+1} \right). \qquad (4)$$

is the infinite-sample risk [2].

In the above,

$$\widehat{P}_\ell(\mathbf{x}) = \frac{P_\ell f_\ell(\mathbf{x})}{f(\mathbf{x})}$$

denotes the posterior probability that a feature vector with value $\mathbf{x}$ originates from class ℓ.

REMARKS: The leading finite-sample term $c_2\, m^{-2/n}$, reveals how the infinite-sample limit is approached, as it dominates the remaining terms in the series for sufficiently large m. Clearly, convergence is enhanced when n is small. For the coefficient c_2, we obtain

$$c_2 = \frac{\Gamma(k + 1 + \frac{2}{n})}{2\, V_n^{2/n}\, (n+2)\, \left(\left(\frac{k-1}{2}\right)!\right)^2} \int_{\mathcal{S}} d\mathbf{x}\, f(\mathbf{x})^{1-2/n} \times \left(\widehat{P}_1(\mathbf{x})\, \widehat{P}_2(\mathbf{x})\right)^{\frac{k+1}{2}} \left(\frac{1}{f_1}\nabla^2 f_1 + \frac{1}{f_2}\nabla^2 f_2 - \frac{2}{f}\nabla^2 f \right). \qquad (5)$$

In Section 5, we discuss how c_2 depends upon k and n for two specific examples. Note that if $k = 1$, (5) reduces to the corresponding coefficient for the nearest-neighbor classifier [6, 7].

4 Outline of the proof

Using the zero-one loss function, R_m represents the expected probability that the k-nearest-neighbor algorithm misclassifies the random feature vector $\mathbf{X}$. In computing this probability, we average over all possible test vectors, $\mathbf{X}$, as well as over all possible labeled reference samples of size m. Thus,

$$R_m = \int_{\mathcal{S}} d\mathbf{x}\, \mathbf{P}(L' \neq L \,|\, \mathbf{x})\, f(\mathbf{x}), \qquad (6)$$

where,

$$\mathbf{P}(L' \neq L \,|\, \mathbf{x}) = \int_{\mathcal{S}} d\mathbf{x}^k \int_{B_k} d\mathbf{x}^{k-1} \cdots \int_{B_2} d\mathbf{x}^1 \; \mathbf{P}(L' \neq L \,|\, \mathbf{x}, \mathbf{x}^1, \ldots, \mathbf{x}^k) f_m(\mathbf{x}^1, \ldots, \mathbf{x}^k \,|\, \mathbf{x}). \qquad (7)$$

Here,

$$f_m(\mathbf{x}^1, \ldots, \mathbf{x}^k \,|\, \mathbf{x}) = (m)_k\, f(\mathbf{x}^1) \cdots f(\mathbf{x}^k) \times \left(1 - \psi(\|\mathbf{x}^k - \mathbf{x}\|, \mathbf{x})\right)^{m-k}$$

represents the probability density associated with the event that the k-nearest neighbors of $\mathbf{x}$ assume the values $\mathbf{x}^1, \ldots, \mathbf{x}^k$, and

$$\psi(\rho, \mathbf{x}) = \int_{B(\rho, \mathbf{x})} f(\mathbf{x}')\, d\mathbf{x}' \qquad (8)$$

equals the probability that a feature vector falls within a ball or radius ρ centered at $\mathbf{x}$. Our indexing convention (2) enforces the inequalities

$$\|\mathbf{x}-\mathbf{x}^1\| \le \|\mathbf{x}-\mathbf{x}^2\| \le \cdots \le \|\mathbf{x}-\mathbf{x}^k\|. \quad (9)$$

The combinatorial factor $(m)_k$ is introduced to count the number of ways that (9) can be realized, given that the k-nearest neighbors of $\mathbf{x}$ in an m-sample, assume the values $\mathbf{x}^1,\ldots,\mathbf{x}^k$. The inequalities (9) also constrain the domains of integration in (7) to concentric balls, centered at $\mathbf{x}$, with radii bounded by the distance between the next nearest neighbor and $\mathbf{x}$. These domains are expressed by the sets $B_j = B(\|\mathbf{x}^j-\mathbf{x}\|,\mathbf{x}) \cap \mathcal{S}$.

By statistical independence, we obtain

$$R_m = (m)_k \int_{\mathcal{S}} d\mathbf{x} \int_{\mathcal{S}} d\mathbf{x}^k \int_{B_k} d\mathbf{x}^{k-1} \cdots \int_{B_3} d\mathbf{x}^2 \int_{B_2} d\mathbf{x}^1 \; g(\mathbf{x},\mathbf{x}^1,\ldots,\mathbf{x}^k)\, e^{-mh(\|\mathbf{x}^k-\mathbf{x}\|,\mathbf{x})} \quad (10)$$

where

$$g(\mathbf{x},\mathbf{x}^1,\ldots,\mathbf{x}^k) = \frac{\sum_{\ell=1}^{2} P_\ell f_\ell(\mathbf{x}) \sum_{\mathbf{l}\notin\mathcal{L}_\ell} \prod_{j=1}^{k} P_{\ell j} f_{\ell j}(\mathbf{x}^j)}{\big(1-\psi(\|\mathbf{x}^k-\mathbf{x}\|,\mathbf{x})\big)^k} \quad (11)$$

$$h(\rho,\mathbf{x}) = -\log\big(1-\psi(\rho,\mathbf{x})\big).$$

We now proceed to evaluate (10) for large, but finite values of m. As in [6], we introduce the family of "cylinder" sets

$$C_t = \{(\mathbf{x},\mathbf{x}^k) \in \mathbb{R}^{2n} : \|\mathbf{x}^k-\mathbf{x}\| = \rho_k \le t\} \cap \mathcal{S}^2.$$

where $t>0$. R_m can then be represented as the sum

$$R_m = (m)_k \iint_{C_t} d\mathbf{x}\, d\mathbf{x}^k \int_{B_k} d\mathbf{x}^{k-1} \cdots \int_{B_2} d\mathbf{x}^1 \; g\, e^{-mh} + (m)_k \iint_{\mathcal{S}^2\setminus C_t} d\mathbf{x}\, d\mathbf{x}^k \int_{B_k} d\mathbf{x}^{k-1} \cdots \int_{B_2} d\mathbf{x}^1 \; g\, e^{-mh}$$

As $h(\rho,\mathbf{x})$ is bounded from below if $\rho \ge t$ (Hypothesis H2), the second term is exponentially subdominant for fixed k as $m\to\infty$, and can thus be neglected for sufficiently large m. Letting $\mathcal{S}_t = \mathcal{S}\setminus\bar{\mathcal{S}}_t = \{\mathbf{x}\in\mathcal{S} : \|\mathbf{x}-\partial\mathcal{S}\| > t\}$, we partition C_t into an interior set $Q_t = \{(\mathbf{x},\mathbf{x}^k)\in C_t : \mathbf{x}\in S_t\}$ and a boundary set $\bar{Q}_t = \{(\mathbf{x},\mathbf{x}^k)\in C_t : \mathbf{x}\in\bar{S}_t\}$. We thus obtain,

$$R_m \sim I_m + J_m$$

where

$$I_m = (m)_k \iint_{Q_t} d\mathbf{x}\, d\mathbf{x}^k \int_{B_k} d\mathbf{x}^{k-1} \cdots \int_{B_2} d\mathbf{x}^1 \; g\, e^{-mh}, \quad (12)$$

$$J_m = (m)_k \iint_{\bar{Q}_t} d\mathbf{x}\, d\mathbf{x}^k \int_{B_k} d\mathbf{x}^{k-1} \cdots \int_{B_2} d\mathbf{x}^1 \; g\, e^{-mh}.$$

(The "$\sim$" symbol indicates that we have neglected exponentially subdominant terms.) If t is chosen so that $0<t<t_0/2$, where t_0 is defined in Hypothesis H4, then $J_m = 0$ as $(\mathbf{x},\mathbf{x}^k)\in\bar{Q}_t \Rightarrow g = 0$.

To evaluate I_m, we use Hypothesis H3 to expand each class-conditional density $f_{\ell j}(\mathbf{x}^j)$ in (11) as a truncated Taylor series about $\mathbf{x}$. Letting, $\rho_j = \|\mathbf{x}^j-\mathbf{x}\|$, and $\Omega_j = (\mathbf{x}^j-\mathbf{x})/\rho_j$, for $j=1,\ldots,k$, we obtain,

$$f_{\ell j}(\mathbf{x}^j) \sim f_{\ell j}(\mathbf{x}) + \sum_{i=1}^{N} f_{\ell j,i}(\Omega_j,\mathbf{x})\rho_j^i + O(\rho_j^{N+1}).$$

We first integrate over the variables $\mathbf{x}^1,\ldots,\mathbf{x}^{k-1}$ in (12). As $d\mathbf{x}^j = \rho_j^{n-1} d\rho_j d\Omega_j$, we obtain

$$\int_{B(\rho_k,\mathbf{x})} d\mathbf{x}^{k-1}\, f_{\ell^{k-1}}(\mathbf{x}^{k-1}) \cdots \int_{B(\rho_2,\mathbf{x})} d\mathbf{x}^1\, f_{\ell^1}(\mathbf{x}^1) = \sum_{\substack{j_1+\cdots+j_{k-1}\le N \\ j_1\ge 0,\ldots,j_{k-1}\ge 0}} \frac{\rho_k^{(k-1)n+j_1+\cdots+j_{k-1}} \Phi_{\ell^1,j_1}(\mathbf{x})\cdots\Phi_{\ell^{k-1},j_{k-1}}(\mathbf{x})}{(n+j_1)\cdots\big((k-1)n+j_1+\cdots+j_{k-1}\big)} + O\Big(\rho_k^{(k-1)n+N+1}\Big), \quad (13)$$

where

$$\Phi_{\ell,i}(\mathbf{x}) = \int_{S_{n-1}} d\Omega_\ell\, f_{\ell,i}(\Omega_\ell,\mathbf{x}).$$

In particular, Hypothesis H3 demands that $\Phi_{\ell,0} = nV_n f_\ell$, and $\Phi_{\ell,2} = (V_n/2)\nabla^2 f_\ell$, while $\Phi_{\ell,i}(\mathbf{x}) = 0$ if i is odd. Similarly, H3 enables us to expand the mixture density in (8) as a Taylor series about $\mathbf{x}$. Whence,

$$\psi(\rho,\mathbf{x}) = \rho^n\big(\psi_0(\mathbf{x}) + \rho^2\psi_2(\mathbf{x}) + \cdots\big)$$

where, $\psi_0 = V_n f$, $\psi_2 = (V_n/(2(n+2))\nabla^2 f$, and again by parity $\psi_i = 0$ for odd values of i. After applying the Binomial Theorem to the denominator of g, and using (13), we obtain an expression of the form

$$R_m = (m)_k \int_{Q_t} d\mathbf{x}\, d\mathbf{x}^k\; e^{-mh(\rho_k,\mathbf{x})}\rho_k^{kn-1} \times \left(\sum_{i=0}^{N} g_i(\mathbf{x})\rho_k^i + O\big(\rho_k^{N+1}\big)\right). \quad (14)$$

Following a rigorous application of Laplace's method to multidimensional integrals [5], Eqn. (14) evaluates to an expression in the form

$$R_m = (m)_k \sum_{i=0}^{N} \Gamma\left(k+1+\frac{i}{n}\right) \int_{\mathcal{S}} d\mathbf{x}\; \lambda_i(\mathbf{x})\, m^{-k-i/n} + O\Big(m^{-(N+1)/n}\Big). \quad (15)$$

In obtaining the above, we have omitted several pages of analysis. (An analogous calculation for the nearest-neighbor classifier appears in [6].) We now observe that

$$(m)_k = m^k \sum_{i=0}^{k} (-1)^j \begin{bmatrix} k \\ k-i \end{bmatrix} \frac{1}{m^i}$$

where $\left[{p \atop q}\right]$ denotes a Stirling Number of the First Kind. Finally, we obtain (1) where c_2 evaluates to the right-hand side of (5).

5 Examples

5.1 Radial distributions

If each class-conditional density, $f_j(\mathbf{x})$, is spherically symmetric in n dimensions about a common origin, and the classification problem admits a single Bayesian decision surface within the interior of $\mathcal{S}$ at radius $r = \|\mathbf{x}\| = r_0$, then for $1 \ll k \ll m < \infty$ we can apply Laplace's method to the integrals in (4) and (5) to obtain,

$$R_\infty = R_B + \frac{n}{8\sqrt{-\widehat{P}_1'(r_0)\,\widehat{P}_2'(r_0)}} \left(\frac{r_0}{2}\right)^{n-1} \frac{1}{k} + O\left(\frac{1}{k^2}\right)$$

and

$$c_2 = \frac{n\Gamma\left(1+\frac{n}{2}\right)^{(2/n)-1}}{64(n+2)\pi^{1-(n/2)}} \frac{\gamma''(r_0)\,k^{(2/n)-1}}{\left[-\widehat{P}_1'(r_0)\,\widehat{P}_2'(r_0)\right]^{3/2}} + O\left(k^{(2/n)-2}\right). \quad (16)$$

Here, R_B is the Bayes risk, and

$$\gamma(r) = r^{n-1} f(r)^{1-(2/n)} \left(\frac{1}{f_1}\nabla^2 f_1 + \frac{1}{f_2}\nabla^2 f_2 - \frac{2}{f}\nabla^2 f\right)$$

with,

$$\nabla^2 f = \frac{n-1}{r}\frac{df}{dr} + \frac{d^2 f}{dr^2}.$$

The above equations reveal several interesting trends: If $n = 1$, c_2 grows linearly with k. If $n = 2$, then c_2 is asymptotically independent of k. Finally, if $n > 2$, then c_2 is inversely related to k. (These features are also qualitatively evident in the following nonradial example.)

5.2 Normal distributions

Here we consider the two-class problem described by the class-conditional densities

$$f_j(\mathbf{x}) = \frac{1}{(2\pi)^{n/2}} e^{-\frac{1}{2}\left((x_1+(-1)^j)^2 + \sum_{i=2}^n x_i^2\right)},$$

for $j = 1$ and 2, with $P_1 = P_2 = 1/2$. This two-class problem violates Hypotheses H2 and H4, so Theorem 3.1 may not directly apply. Nevertheless, the following empirical study suggests that the expected finite-sample risk converges to its infinite-sample limit in a fashion consistent with our theoretical findings. For this problem R_B evaluates numerically to 0.15865, while the values of R_∞ form the decreasing sequence 0.224800, 0.191539, 0.174598, 0.166439, 0.162493, and 0.160560 for $k = 1$, 3, 7, 15, 31, and 63, respectively. These values are valid for all n.

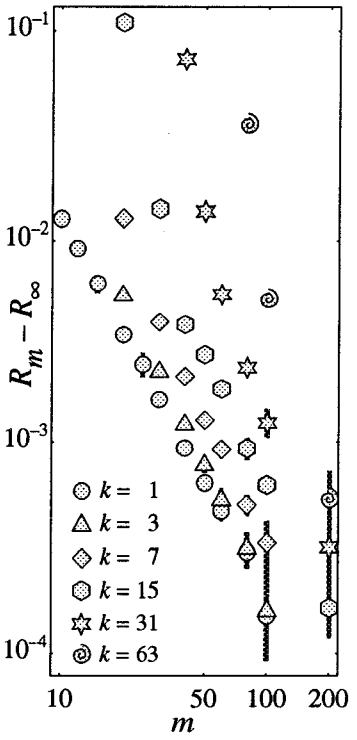

Figure 1: For $n = 1$, numerical simulations suggest that the risk for a two-class problem with normal distributions is asymptotically consistent with formula (1). Note also that the empirical value of c_2 appears to increase with k, in qualitative agreement with formula (16).

Figures 1, 2, and 3 respectively reveal how R_m approaches R_∞ for $n = 1$, 2, and 5. Each data point reflects the relative frequency of classification errors from a large number of independent Bernoulli trials. For each trial, an i.i.d. reference sample of m vectors and an independent test vector are constructed. For each data point, a 95% confidence interval was inferred. (Only those confidence intervals that exceed the marker size are shown.) On these log-log plots, graphs of $R_m - R_\infty = c_2 m^{-2/n}$ will appear as straight lines with slope $-2/n$. Note that before plotting each symbol, the computed value of R_∞ corresponding to the symbol's k value was subtracted from our empirical estimate of R_m. These results suggest that it may be possible to extend Theorem 3.1 to a larger family of distributions. Moreover, it is evident that this asymptotic behavior dominates for realizable sample sizes.

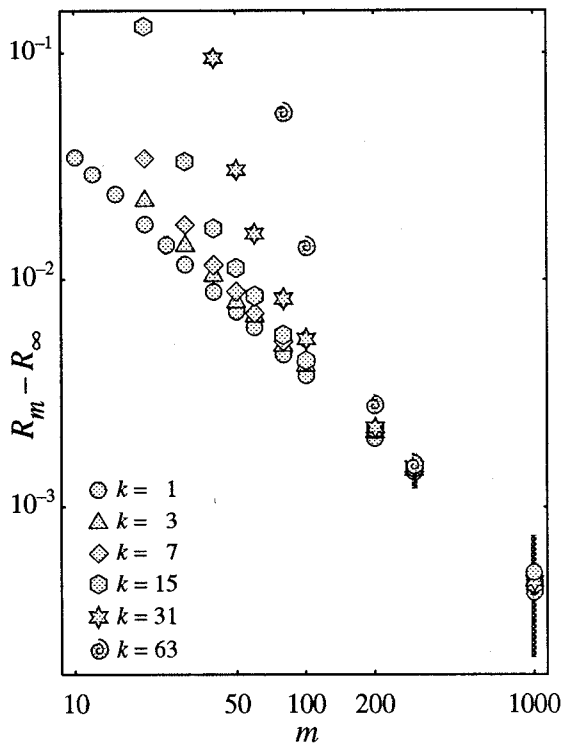

Figure 2: For $n = 2$, numerical estimates of the risk of the two-class, normal-distribution problem support formula (1). As in the radial-distribution example, c_2 is insensitive to variations in k for this dimensionality.

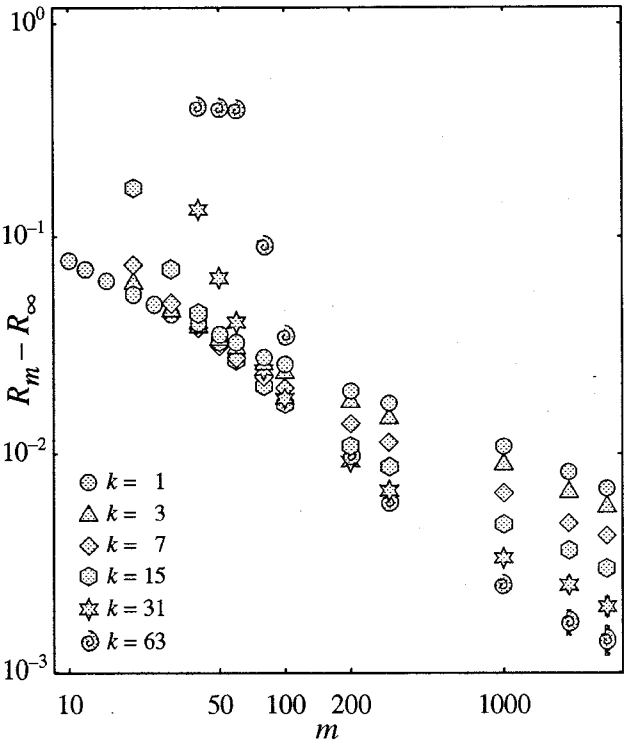

Figure 3: For $n = 5$, numerical estimates of the risk of the two-class, normal-distribution problem again support (1). As in the radial-distribution problem with $n > 2$ (cf. (16)), the empirical value of c_2 decreases as k increases.

6 Concluding remarks

In the preceding, we have demonstrated how Laplace's method yields asymptotic predictions of the finite-sample risk for the k-nearest-neighbor algorithm. As Hypotheses H2 and H4 are not necessary conditions for Eqn. (1), Theorem 3.1 may apply for a large family of smooth classification problems. These results may help estimate optimal values of k in terms of m and n. In a future article, we shall describe how these results can be extended to k-nearest-neighbor classifiers that use more general L_p distance functions [8].

References

[1] T. M. Cover, "Rates of convergence of nearest neighbor decision procedures," *Proc. First Annual Hawaii Conf. on Systems Theory*, 1968, pp. 413–415.

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

[3] A. Erdélyi, *Asymptotic Expansions*, Dover: New York, NY, 1956.

[4] K. Fukunaga and D. M. Hummels, "Bias of nearest neighbor estimates," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI–9, 1987, pp. 103–112.

[5] W. Fulks and J. O. Sather, "Asymptotics II: Laplace's method for multiple integrals," *Pacific J. Math.*, vol. 11, 1961, pp. 185–192.

[6] 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, 1994, pp. 820–837.

[7] R. R. Snapp, D. Psaltis, and S. S. Venkatesh, "Asymptotic slowing down of the nearest-neighbor classifier," in *Advances in Neural Information Processing Systems, 3* (ed. R. Lippmann, et al.). San Mateo, California: Morgan Kaufmann, 1991, pp. 932–938.

[8] R. R. Snapp and S. S. Venkatesh, "Finite-sample risk of the k-nearest-neighbor classifier under the L_p metric," (in preparation).

[9] C. J. Stone, "Consistent nonparametric regression," *Annals of Statistics*, vol. 5, 1977, pp. 595–645.