# $k$-NN CLASSIFIERS: INVESTIGATING THE $k = k(n)$ RELATIONSHIP

*Cesare Alippi, Manuel Roveri*
DEI, Politecnico di Milano, Piazza L. Da Vinci 32, 20133 Milano, Italy
Alippi, Roveri@elet.polimi.it

A $k$-NN classifier [1][2] associates a classification label to an input as the majority of its $k$ nearest training samples. This classification rule, which does not require a proper training phase, tends asymptotically to the optimal Bayes's classifier provided that $k$ grows less than linearly w.r.t. *n* [3] but no indications are given regarding the optimal selection of $k$ having a *n*-sized sample training set.

In general, identification of the optimal $k$, i.e., the $k$ leading to the optimal classification accuracy for a given *n*, is carried out with a Leave-One-Out (LOO) technique [4], which, even if it does not require a-priori information, is computationally expensive. Existing theoretical results (e.g., [1]-[3]) focus on the asymptotic relationships between $k$ and *n* but do not address the nonasymptotical case.

In the research we investigate the issues leading to the identification of the optimal $k$ in $k$-NN classifiers in correspondence of a sufficiently large *n*. Results are *pdf*-dependent but can be easily used to derive, at the numerical-experimental level, indications about the relationships among the optimal $k$, the robustness of the generalization error w.r.t. $k$ and the number of training samples, hence shedding light on the intrinsic behavior of $k$-NN classifiers.

Determination of the optimal $k$ requires two subsequent logical steps:

1. identifying a neighborhood of an input point containing, in probability, $k$ samples
2. estimating the classification error probability and relate it to $k$; the optimal $k$ can then be derived by minimizing the classification error.

Validity of results, here applied to the two-class classifications problem, can be extended to cover multi-class applications.

*Step 1: neighborhood identification*

Let $x \in T \subset R^d$ be the input vector and $y \in \{\omega_1, \omega_2\}$ the associated binary classification output, respectively. $p(x)$, $p(\omega_1)$ and $p(\omega_2)$ refer to the pdfs of inputs, and outputs (with $p(\omega_1) = 1 - p(\omega_2)$); $p(x \mid \omega_1)$ and $p(x \mid \omega_2)$ are their respective conditional probability distributions. $f(x,k)$ represents the classification function carried out by the $k$-NN classifier.

Define $\Omega(\bar{x}, r)$ to be the *d*-dimensional hypersphere centered at $\bar{x}$ of radius *r* and $q(\bar{x}, r)$ the integral of $p(x)$ over $\Omega(\bar{x}, r)$. In probability, $\Omega(\bar{x}, r)$ contains $k$ training samples, when radius $r^*$ satisfies $q(\bar{x}, r^*) = k/n$. The reason for considering a spherical neighborhood derives from the operational nature of the traditional $k$-NN classifier, which identifies the $k$ nearest training samples of point $\bar{x}$ through the Euclidean distance.

*Step 2: minimizing the classification error*

Classification $f(\bar{x}, k)$ of sample $\bar{x}$ in $\Omega(\bar{x}, r)$ introduces an error whenever the correct output value *y* differs from $f(\bar{x}, k)$. The error probability $p_e(\bar{x}, k)$ is hence the probability of having more than ($k$-1)/2 training samples in $\Omega(\bar{x}, r)$ associated with the wrong class $\omega_{i' \neq i}$, i.e.,

$$p_e(\bar{x}, k) = \sum_{j=1}^{2} \left( \sum_{m=\frac{k-1}{2}+1}^{k} \binom{k_j}{m} \left(q_{\omega_j}(\bar{x}, r^*)\right)^m \left(1 - q_{\omega_j}(\bar{x}, r^*)\right)^{k_j - m} \right) p(\bar{x} \mid \omega_j)\, p(\omega_j) \tag{1}$$

where $\lceil k_1 \rceil = kp(\omega_1)$ represents the number of training samples of class $\omega_1$, $k_2 = k - k_1$ is the number of training samples of class $\omega_2$ and $q_{\omega_j}(\bar{x}, r^*) = \int_{\Omega(\bar{x}, r^*)} p(x \mid \omega_j) dx$.

(1) can be structurally simplified when $n \to \infty$. In such a case, the Binomial distribution can be approximated with a Poisson distribution [6] and (1) simplifies as

$$p_e(\bar{x}, k) \cong \sum_{j=1}^{2} \left( \frac{\Gamma\left(k_j + 1, k_j q_{\omega_j}(\bar{x}, r^*)\right)}{\Gamma\left(k_j + 1\right)} - \frac{\Gamma\left(\frac{k+1}{2}, k_j q_{\omega_j}(\bar{x}, r^*)\right)}{\Gamma\left(\frac{k+1}{2}\right)} \right) p(\bar{x} \mid \omega_j) p(\omega_j) \tag{2}$$

By integrating either non-asymptotical (1) or asymptotical (2) over the input domain we obtain the error probability $P_e(k) = \int_T p_e(x, k) dx$. The optimal value of $k$ given *n* training samples is the one minimizing $P_e(k)$:

$$k^o = \min_k \{P_e(k)\} \tag{3}$$

with the dependence of *n* hidden in the $q(\bar{x}, r)$s.

In general, (3) cannot be solved in a closed form for any application; yet it can be solved numerically provided that the required probability density functions or an estimate are available or can be derived from the data. The numerical procedure for solving (3) is finally given in Algorithm 1.

| **Algorithm 1: Determining the optimal $k$ in $k$-NNs classifiers** |
|---|
| 1. Determine $r^*(x, r)$ from $q(\bar{x}, r^*) = k / n$; |
| 2. Compute $q_{\omega_i}(x, r^*)$, *i*=1,2; |
| 3. Estimate $p_e(x, k)$ with (1) or (2); |
| 4. Estimate $P_e(k) = \int_T p_e(x, k) dx$; |
| 5. Minimize $P_e(k)$ and obtain $k^o$. |

From simulation it appears that (3) is robust w.r.t. $k$. As such, any estimate close to $k^o$ will practically work well. Results, together with a critical view of the methodology and details will be presented at the workshop.

**References:**


[1] T. M. Cover and R. E. Hart, "Nearest neighbor pattern classification," *IEEE Trans. on Information Theory*, vol. 13, no. 1, pp. 21-27, January 1967.

[2] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, Academic Press, 1972.

[3] C. Stone, "Consistent nonparametric regression," *Annals of Statistics*, vol. 8, pp. 1348-1360, 1977.

[4] P. A. Lachenbruch and M. R. Mickey, "Estimation of error rates in discriminant analysis," *Technometrics*, vol. 10, pp. 1-10, 1968.

[5] M. Keans and D. Ron. "Algorithmic stability and sanity check bounds for leave-one-out cross validation bounds", *Neural Computation,* vol. 11, no. 6, pp. 1427-1453, 1999.

[6] A.M. Mood and F.A. Graybill, *Introduction to the theory of statistics*, McGraw Hill, 1963.