Statistical Science
2006, Vol. 21, No. 3, 347–351
DOI: 10.1214/088342306000000457
Main article DOI: 10.1214/088342306000000493

Comment

Grace Wahba

1. INTRODUCTION

The authors are to be commended for jumping in to describe support vector machines (SVMs), not an easy thing to do since the the literature for SVMs has grown at least exponentially in the last few years. A Google search for "support vector machines" gave "about 1,180,000" hits as of this writing. The authors have nevertheless made a nice selection of important points to emphasize. As noted, SVMs were proposed for classification in the early 1990s by arguments like those behind Figure 1 in their paper. The use of SVMs grew rapidly among computer scientists, as it was found that they worked very well in all kinds of practical applications. The theoretical underpinnings that went with the original proposals were different than those in the classical statistical literature, for example, those related to Bayes risk, and so had less impact in the statistical literature. The convergence of SVMs and regularization methods (or, rather the convergence of the "SVM community" and the "regularization community") was a major impetus in the study of the (classical) statistical properties of the SVM. One point at which this convergence took place was at an American Mathematical Society meeting at Mt. Holyoke in 1996. The speaker was describing the SVM with the so-called kernel trick when an anonymous person at the back of the room remarked that the SVM with the kernel trick was the solution to an optimization problem in a reproducing kernel Hilbert space (RKHS). Once it was clear to statisticians that the SVM can be obtained as the result of an optimization/regularization problem in a RKHS, tools known to statisticians in this context were rapidly employed to show how the SVM could be modified to take into account nonrepresentative sample sizes, unequal misclassification costs and more than two classes, and to show in each case that it directly targets the Bayes risk under very general circumstances (see also [5, 8]). Thus, a "classical" explanation of why they work so well was provided.

Grace Wahba is the IJ Schoenberg-Hilldale Professor of Statistics, Department of Statistics, University of Wisconsin, 1300 University Avenue, Madison, Wisconsin 53706, USA (e-mail: wahba@stat.wisc.edu), and is also a member of the Computer Sciences Department and the Biostatistics and Medical Informatics Department.

2. MERCER'S KERNELS AND POSITIVE DEFINITE FUNCTIONS

Let $\mathcal{T}$ be a.d.o. (any dirty old) domain and let $K(s,t), s,t \in \mathcal{T}$, be a symmetric, positive definite function of two variables; K is said to be positive definite if for any n, and any $t_1, \ldots, t_n \in \mathcal{T}$, the $n \times n$ matrix with ijth element $K(t_i, t_j)$ is nonnegative definite. In the early SVM literature, as well as in the present paper, the kernel is described as having a representation $K(s,t) = \sum_{\nu=1}^{\infty} \lambda_\nu \Phi_\nu(s)\Phi_\nu(t)$. Here the (nonnegative) λ_ν and the Φ_ν are the eigenvalues and eigenvectors of K. A representation as in this sum is sufficient for K to be positive definite (see [13] on the Mercer Hilbert–Schmidt theorem), but the so-called radial basis functions (RBF) popular in machine learning, of the form $K(s,t) = k(\|s-t\|)$, s,t in Euclidean d-space E^d, do not have a countable sequence of eigenvalues and eigenvectors—complex exponentials play the role of eigenvectors (see [3]). The Gaussian kernel $K_c(x,y) = e^{-\|x-y\|^2/c}$ is such an example. Although the notion of a countable expansion was used in uncoupling the linear SVM from its linearity restriction (and seems to be repeated over and over), the lack of a countable set of eigenvectors and eigenvalues does not affect the use of the Gaussian kernel or any other positive definite function in an SVM; as the authors note, only values of K are needed. The RBF probably just do not want to be called "Mercer's kernels" (!). Positive definite functions are sometimes called reproducing kernels, relating to their association with RKHS [1].

Given a collection of objects (which could be vectors, images, sounds, graphs, texts, trees, ...) in a.d.o. domain $\mathcal{T}$, a positive definite matrix with ij entry $K(i,j)$ defines a (squared) distance d_{ij} between the ith and jth object as

$$d_{ij} = K(i,i) + K(j,j) - 2K(i,j)$$

(and, in addition, this distance comes with an inner product). It can be argued that using distance between objects, defined in some way, is truly fundamental to classification and, therefore, positive definite kernels, since they provide a distance, play a fundamental role.