Statistical Science
2006, Vol. 21, No. 3, 341–346
DOI: 10.1214/088342306000000475
Main article DOI: 10.1214/088342306000000493
© Institute of Mathematical Statistics, 2006

Comment

Peter L. Bartlett, Michael I. Jordan and Jon D. McAuliffe

INTRODUCTION

The support vector machine (SVM) has played an important role in bringing certain themes to the fore in computationally oriented statistics. However, it is important to place the SVM in context as but one member of a class of closely related algorithms for nonlinear classification. As we discuss, several of the "open problems" identified by the authors have in fact been the subject of a significant literature, a literature that may have been missed because it has been aimed not only at the SVM but at a broader family of algorithms. Keeping the broader class of algorithms in mind also helps to make clear that the SVM involves certain specific algorithmic choices, some of which have favorable consequences and others of which have unfavorable consequences—both in theory and in practice. The broader context helps to clarify the ties of the SVM to the surrounding statistical literature.

We have at least two broader contexts in mind for the SVM. The first is the family of "large-margin" classification algorithms—a class that includes boosting and logistic regression. All of these algorithms involve the minimization of a convex contrast or loss function that upper bounds the 0–1 loss function. The SVM makes a specific choice of convex loss function—the so-called hinge loss. Hinge loss has some potentially desirable properties (e.g., sparseness) and some potentially undesirable properties (e.g., lack of calibration to posterior probabilities). As we discuss, much of the theoretical analysis of the SVM is best carried out by focusing on convexity and abstracting away from the details of specific loss functions.

Second, as the authors note, the SVM is an instance of the broader family of statistical procedures based on reproducing kernel Hilbert spaces (RKHSs). The authors' emphasis is on the use of RKHS methods to provide basis expansions for discriminant functions and regression functions. RKHS ideas have, however, been carried significantly further in recent years, enlivening areas of computationally oriented statistics beyond classification and regression. We wish to convey some of the reasons for this broader interest in RKHS-based approaches.

There are both computational and statistical motivations for focusing on methods based on convexity and reproducing kernel Hilbert spaces. In the remainder of this discussion we attempt to disentangle some of these motivations, but we wish to emphasize at the outset that it is precisely because these methods bring computational and statistical considerations together that they are so interesting.

CONVEXITY

The SVM is one example of a general strategy for solving the binary classification problem via a "convex surrogate loss function." To develop this perspective, let us define binary classification as the problem of choosing a discriminant function $f : \mathcal{X} \to \mathbb{R}$ that minimizes misclassification risk,

$$R(f) = P\big(Y \neq \operatorname{sgn}(f(X))\big) = \mathbf{E}\ell(Yf(X)),$$

where $X \in \mathcal{X}$ is the covariate, $Y \in \{\pm 1\}$ is the binary response, and $\ell(\alpha) = 1$ for $\alpha \leq 0$ and $= 0$ otherwise. The family of large-margin classification algorithms attacks this problem indirectly by minimizing a quantity known as the *ϕ-risk*,

$$R_\phi(f) = \mathbf{E}\phi(Yf(X)),$$

where $\phi : \mathbb{R} \to \mathbb{R}$ is a surrogate for the loss function ℓ, and $Yf(X)$ is called the *margin* of f on the observation (X, Y). The margin indicates not only whether the observation is correctly classified by f, but how close f comes to choosing the opposite label. The surrogate loss function ϕ is chosen so that large margins correspond to small losses.

Given a data set $(X_1, Y_1), \ldots, (X_n, Y_n)$, we can form the *empirical ϕ-risk*

$$\hat{R}_\phi(f) = \frac{1}{n}\sum_{i=1}^{n} \phi(Y_i f(X_i))$$

Peter L. Bartlett and Michael I. Jordan are Professors, Computer Science Division and Department of Statistics, University of California, Berkeley, California 94720, USA (e-mail: bartlett@stat.berkeley.edu; jordan@stat.berkeley.edu). Jon D. McAuliffe is Assistant Professor, Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, Pennsylvania 19104-6340, USA (e-mail: mcjon@wharton.upenn.edu).