Statistical Science
2006, Vol. 21, No. 3, 352–357
DOI: 10.1214/088342306000000466
Main article DOI: 10.1214/088342306000000493

Comment

Trevor Hastie and Ji Zhu

We congratulate the authors for a well written and thoughtful survey of some of the literature in this area. They are mainly concerned with the geometry and the computational learning aspects of the support vector machine (SVM). We will therefore complement their review by discussing from the statistical function estimation perspective. In particular, we will elaborate on the following points:

- Kernel regularization is essentially a generalized ridge penalty in a certain feature space.
- In practice, the effective dimension of the data kernel matrix is not always equal to n, even when the implicit dimension of the feature space is infinite; hence, the training data are not always perfectly separable.
- Appropriate regularization plays an important role in the success of the SVM.
- The SVM is not fundamentally different from many statistical tools that our statisticians are familiar with, for example, penalized logistic regression.

We acknowledge that many of the comments are based on our earlier paper Hastie, Rosset, Tibshirani and Zhu (2004).

KERNEL REGULARIZATION AND THE GENERALIZED RIDGE PENALTY

Given a positive definite kernel $K(\mathbf{x}, \mathbf{x}')$, where $\mathbf{x}, \mathbf{x}'$ belong to a certain domain $\mathcal{X}$, we consider the general function estimation problem

$$\min_{\beta_0, f} \sum_{i=1}^{n} \ell(y_i, \beta_0 + f(\mathbf{x}_i)) + \frac{\lambda}{2} \| f(\mathbf{x}) \|_{\mathcal{H}_K}^2. \tag{1}$$

Here $\ell(\cdot, \cdot)$ is a convex loss function that describes the "closeness" between the observed data and the fitted model, and f is an element in the span of $\{K(\cdot, \mathbf{x}'), \mathbf{x}' \in \mathcal{X}\}$. More precisely, $f \in \mathcal{H}_K$ is a function in the reproducing kernel Hilbert space $\mathcal{H}_K$ (RKHS) generated by $K(\cdot, \cdot)$ (see Burges, 1998; Evgeniou, Pontil and Poggio, 2000; and Wahba, 1999, for details).

Suppose the positive definite kernel $K(\cdot, \cdot)$ has a (possibly finite) eigenexpansion,

$$K(\mathbf{x}, \mathbf{x}') = \sum_{j=1}^{\infty} \delta_j \phi_j(\mathbf{x}) \phi_j(\mathbf{x}'),$$

where $\delta_1 \geq \delta_2 \geq \cdots \geq 0$ are the eigenvalues and $\phi_j(\mathbf{x})$'s are the corresponding eigenfunctions. Elements of $\mathcal{H}_K$ have an expansion in terms of these eigenfunctions

$$f(\mathbf{x}) = \sum_{j=1}^{\infty} \beta_j \phi_j(\mathbf{x}), \tag{2}$$

with the constraint that

$$\| f \|_{\mathcal{H}_K}^2 \overset{\text{def}}{=} \sum_{j=1}^{\infty} \beta_j^2 / \delta_j < \infty,$$

where $\| f \|_{\mathcal{H}_K}$ is the norm induced by $K(\cdot, \cdot)$.

Then we can rewrite (1) as

$$\min_{\beta_0, \boldsymbol{\beta}} \sum_{i=1}^{n} \ell\left(y_i, \beta_0 + \sum_{j=1}^{\infty} \beta_j \phi_j(\mathbf{x}_i) \right) + \lambda \sum_{j=1}^{\infty} \frac{\beta_j^2}{\delta_j}, \tag{3}$$

and we can see that the regularization term $\| f \|_{\mathcal{H}_K}^2$ in (1) can be interpreted as a generalized ridge penalty, where eigenfunctions with small eigenvalues in the expansion (2) get penalized more and vice versa.

Formulation (3) seems to be an infinite dimensional optimization problem, but according to the representer theorem (Kimeldorf and Wahba, 1971; Wahba 1990), the solution is finite dimensional and has the form

$$f(\mathbf{x}) = \sum_{i=1}^{n} \alpha_i K(\mathbf{x}, \mathbf{x}_i).$$

Using the reproducing property of $\mathcal{H}_K$, that is, $\langle K(\cdot, \mathbf{x}_i), K(\cdot, \mathbf{x}_{i'}) \rangle = K(\mathbf{x}_i, \mathbf{x}_{i'})$, (3) also reduces to a finite-dimensional criterion,

$$\min_{\beta_0, \alpha} L(\mathbf{y}, \beta_0 + \mathbf{K}\alpha) + \lambda \alpha^{\mathrm{T}} \mathbf{K} \alpha. \tag{4}$$

Here we use vector notation, $\mathbf{K}$ is the $n \times n$ data kernel matrix with elements equal to $K(\mathbf{x}_i, \mathbf{x}_{i'}), i, i' =$

Trevor Hastie is Professor, Department of Statistics, Stanford University, Stanford, California 94305, USA (e-mail: hastie@stanford.edu). Ji Zhu is Assistant Professor, Department of Statistics, University of Michigan, Ann Arbor, Michigan 48109, USA (e-mail: jizhu@umich.edu).