DEPARTMENT OF STATISTICS
University of Wisconsin
1210 West Dayton St.
Madison, WI 53706

TECHNICAL REPORT NO. 984rr

July 2, 1998

Support Vector Machines, Reproducing Kernel Hilbert Spaces and the Randomized GACV [1]

by

Grace Wahba

[1] Prepared for the NIPS 97 Workshop on Support Vector Machines. Research sponsored in part by NSF under Grant DMS-9704758 and in part by NEI under Grant R01 EY09946. This is a second revised and corrected version of a report of the same number and title dated November 29, 1997

1 Support Vector Machines, Reproducing Kernel Hilbert Spaces and the Randomized GACV

Grace Wahba
Department of Statistics, University of Wisconsin-Madison
wahba@stat.wisc.edu http://www.stat.wisc.edu/~wahba
July 1, 1998 Revised version, pls destroy earlier versions.

This chapter is an expanded version of a talk presented in the NIPS 97 Workshop on Support Vector Machines. It consists of three parts: (1) A brief review of some old but relevant results on constrained optimization in Reproducing Kernel Hilbert Spaces (RKHS), and a review of the relationship between zero-mean Gaussian processes and RKHS. Application of tensor sums and products of RKHS including smoothing spline ANOVA spaces in the context of SVM's is also described. (2) A discussion of the relationship between penalized likelihood methods in RKHS for Bernoulli data when the goal is risk factor estimation, and SVM methods in RKHS when the goal is classification. When the goal is classification it is noted that replacing the likelihood functional of the logit (log odds ratio) with an appropriate SVM functional is a natural method for concentrating computational effort on estimating the logit near the classification boundary and ignoring data far away. Remarks concerning the potential of SVM's for variable selection as an efficient preprocessor for risk factor estimation are made. (3) A discussion of how the the GACV (Generalized Approximate Cross Validation) for choosing smoothing parameters proposed in Xiang and Wahba (1996, 1997) may be adapted and implemented in the context of certain convex SVM's.

Introduction

Several old results in Reproducing Kernel Hilbert Spaces (RKHS) and Gaussian processes are proving to be very useful in the application of support vector machine

(SVM) methods in classification. In Section 1.1 of this paper we very briefly review some of these results. RKHS can be chosen tailored to the problem at hand in many ways, and we review a few of them, including radial basis function and smoothing spline ANOVA spaces.

Girosi (1997), Smola and Schölkopf (1997), Schölkopf *et al* (1997) and others have noted the relationship between SVM's and penalty methods as used in the statistical theory of nonparametric regression. In Section 1.2 we elaborate on this, and show how replacing the likelihood functional of the logit (log odds ratio) in penalized likelihood methods for Bernoulli [yes-no] data, with certain other functionals of the logit (to be called SVM functionals) results in several of the SVM's that are of modern research interest. The SVM functionals we consider more closely resemble a "goodness-of-fit" measured by classification error than a "goodness-of-fit" measured by the comparative Kullback-Liebler distance, which is frequently associated with likelihood functionals. This observation is not new or profound, but it is hoped that the discussion here will help to bridge the conceptual gap between classical nonparametric regression via penalized likelihood methods, and SVM's in RKHS. Furthermore, since SVM's can be expected to provide more compact representations of the desired classification boundaries than boundaries based on estimating the logit by penalized likelihood methods, they have potential as a prescreening or model selection tool in sifting through many variables or regions of attribute space to find influential quantities, even when the ultimate goal is not classification, but to understand how the logit varies as the important variables change throughout their range. This is potentially applicable to the variable/model selection problem in demographic medical risk factor studies as described, for example in Wahba, Wang, Gu, Klein and Klein (1995).

When using SVM functionals to produce classification boundaries, typically a tradeoff must be made between the size of the SVM functional and the 'smoothness' or complexity of the logit function. This tradeoff is in the first instance embodied in smoothing parameters. In Section 1.3 we discuss how the GACV for choosing smoothing parameters proposed in Xiang and Wahba (1996, 1997) may be adapted to some support vector machines.

1.1 Some facts about RKHS

1.1.1 The Moore-Aronszajn Theorem

Let $\mathcal{T}$ be a set, for example, $\mathcal{T} = \{1, 2, \cdots, N\}, \mathcal{T} = [0, 1]$, or $\mathcal{T} = E^d$, (Euclidean d-space), or $\mathcal{T} = \mathcal{S}_d$, (the d-dimensional sphere). A real symmetric function $K(s,t), s,t \in \mathcal{T}$ is said to be positive definite on $\mathcal{T} \times \mathcal{T}$ if for every $n = 1, 2, \cdots$, and every set of real numbers $\{a_1, a_2, \cdots, a_n\}$ and $t_1, t_2, \cdots t_n,\ t_i \in \mathcal{T}$, we have $\sum_{i,j=1}^n a_i a_j K(t_i, t_j) \geq 0$. We have the famous

Moore-Aronszajn Theorem: (Aronszajn 1950)

`To every positive definite function` K `on` $\mathcal{T}\times\mathcal{T}$ `there corresponds a unique RKHS` $\mathcal{H}_K$ `of real valued functions on` $\mathcal{T}$ `and vice versa.`

The proof is trivial. We just suggest how to construct $\mathcal{H}_K$ given K. Let $K_t(s)$ be the function of s obtained by fixing t and letting $K_t(s) \doteq K(s,t)$. $\mathcal{H}_K$ consists of all finite linear combinations of the form $\sum_{\ell=1}^{L} a_\ell K_{t_\ell}$ with $t_\ell \in \mathcal{T}$ and limits of such functions as the t_ℓ become dense in $\mathcal{T}$, in the norm induced by the inner product

$$< K_s, K_t >_{\mathcal{H}_K} = K(s,t). \tag{1.1}$$

See Wahba (1990) (W) for further details on most of the material in this Section. The positive definiteness of K guarantees that (1.1) defines a bona fide inner product. (Furthermore strong limits here imply pointwise limits [1] .) The function $K_t(\cdot)$ is the so-called representer of evaluation at t in $\mathcal{H}_K$ - this means: For any $f \in \mathcal{H}_K$ and fixed t

$$< f, K_t >_{\mathcal{H}_K} = f(t), \tag{1.2}$$

where $< \cdot,\cdot >_{\mathcal{H}_K}$ is the inner product in $\mathcal{H}_K$. If $K(s,t)$ has a representation of the form

$$K(s,t) = \sum_\nu \lambda_\nu \Psi_\nu(s)\Psi_\nu(t) \tag{1.3}$$

with $\int_{\mathcal{T}} \Psi_\xi(s)\Psi_\eta(s)d\mu(s) = 1$ if $\xi = \eta$, and 0 otherwise, where μ is some measure on $\mathcal{T}$, then $< f, g >_{\mathcal{H}_K} = \sum_\nu \frac{f_\nu g_\nu}{\lambda_\nu}$ where $f_\nu = \int \Psi_\nu(s) f(s) d\mu(s)$ and similarly for g_ν. In particular $< \Psi_\xi, \Psi_\eta >_{\mathcal{H}_K} = \frac{1}{\lambda_\xi}$ if $\xi = \eta$ and 0 otherwise. Examples of μ include Lebesgue measure on [0,1] and counting measure on $\{1, 2, \cdots, N\}$. $K(\cdot,\cdot)$ is known as the reproducing kernel (RK) for $\mathcal{H}_K$, due to the 'reproducing property' (1.1).

1.1.2 The Representer Theorem

Let $\mathcal{T}$ be an index set, $\mathcal{H}_K$ be an RKHS of real valued functions on $\mathcal{T}$ with RK $K(\cdot,\cdot)$. Let $\{y_i, t_i, i = 1, 2, \cdots n\}$ be given (the "training set"), with t_i (the "attribute vector") $\in \mathcal{T}$. y_i is the "response" (usually a real number, but may be more general, see Wahba (1992)). Let $\{\phi_\nu\}_{\nu=1}^{M}$ be M functions on $\mathcal{T}$ [2] with the property that the $n \times M$ matrix T with $i\nu$th entry $\phi_\nu(t_i)$ is of rank M. ("Least squares regression on *span* $\{\phi_\nu\}$ is unique.") Let $g_i(y_i, f)$ be a functional of f which depends on f only through $f(t_i) \doteq f_i$, that is, $g_i(y_i, f) \doteq g_i(y_i, f_i)$. Then we have

The Representer Theorem: (Kimeldorf and Wahba 1971 (KW))

1. i. e. $\|f_n - f\| \to 0 \Rightarrow |f_n(t) - f(t)|$, every $t \in \mathcal{T}$.
2. Sufficient conditions on the $\{t_i\}$ for existence are being assumed.

Any solution to the problem: find $f \in span\ \{\phi_\nu\} + h$ with $h \in \mathcal{H}_K$ to minimize

$$\frac{1}{n}\sum_{i=1}^{n} g_i(y_i, f_i) + \lambda \|h\|^2_{\mathcal{H}_K} \tag{1.4}$$

has a representation of the form

$$f(\cdot) = \sum_{\nu=1}^{M} d_\nu \phi_\nu(\cdot) + \sum_{i=1}^{n} c_i K(t_i, \cdot). \tag{1.5}$$

1.1.3 Remarks

Remark 1. This theorem is explicitly stated in KW only for $g_i(y_i, f_i) = (y_i - f_i)^2$ and (letting $y_i = (y_{i1}, y_{i2})$) for $g_i(y_i, f_i) = 0, y_{i1} \le f_i \le y_{i2}, = \infty$ otherwise. However, the extension to general g_i is obvious from the argument there (and has appeared in various places, see, for example Cox and O'Sullivan (1990)). One of the most popular support vector machines corresponds to the case $M = 1, \phi_1(t) \equiv 1$ and $g_i(y_i, f_i) = V_\epsilon(y_i - f_i)$, where V_ϵ, Vapnik's ϵ-insensitive loss function, is given by $V_\epsilon(u) = max\{0, |u| - \epsilon\}$. f of the form (1.5) is substituted back into (1.4), resulting in an optimization problem in the unknown d_1 and $c = (c_1, \cdots, c_n)'$. Details concerning how this optimization problem is converted to the familiar SVM QP may be found, e. g. in Girosi (1997), see also Vapnik (1995).

Remark 2. Probably the best known example of this problem is the case $\mathcal{T} = [0,1], M = 2, \|h\|^2_{\mathcal{H}_K} = \int_0^1 (h''(u))^2 du, \phi_1(t) = 1, \phi_2(t) = t$. Then f is a cubic spline with knots at the data points, see KW, Wahba (1990) (W) for details. Reproducing kernels for $\|h\|^2_{\mathcal{H}_K} = \int_0^1 [(L_m f)(u)]^2 du$ where L_m is a differential operator with a null space spanned by a Tchebychev system are found in KW and involve Green's functions for $L_m * L_m$. Typically ϕ_1 is a constant function and the ϕ_ν's are linear or low degree polynomials. Under certain circumstances a large λ in (1.4) will force the minimizer into $span\{\phi_\nu\}$. In KW and W this theorem is stated for $f \in \mathcal{H}_{\tilde{K}}$ where $\mathcal{H}_K$ is a subspace of $\mathcal{H}_{\tilde{K}}$ of codimension M orthogonal to $span\{\phi_\nu\}$.

Remark 3. Let $'$ denote transpose. If we make some assumptions a simple proof exists that the coefficient vector c of any minimizer satisfies $T'c = 0$. First, note that $\begin{pmatrix} f_1 \\ \vdots \\ f_n \end{pmatrix} = Kc + Td$, where $d = (d_1, \cdots d_M)'$ and (with some abuse of notation) we are letting K be the $n \times n$ matrix with i,jth entry $K(t_i, t_j)$ (where it will be clear from the context that we mean K is an $n \times n$ matrix rather than an RK). Similarly, note that $\| \sum_{i=1}^n c_i K_{t_i} \|^2_{\mathcal{H}_K} = c'Kc$. The vectors c and d are found as the minimizers of

$$\frac{1}{n}\sum_{i=1}^{n} g_i(y_i, f_i) + \lambda c' K c. \tag{1.6}$$

Assuming that we can differentiate g_i with respect to f_i, differentiating (1.6)with respect to c and d gives

$$\frac{1}{n} K \frac{\partial g}{\partial f} = -2\lambda K c \tag{1.7}$$

$$\frac{1}{n} T' \frac{\partial g}{\partial f} = 0 \tag{1.8}$$

where $\frac{\partial g}{\partial f} = (\frac{\partial g_1}{\partial f_1}, \cdots, \frac{\partial g_n}{\partial f_n})'$, and, assuming K is of full rank, and multiplying (1.7) by K^{-1} and substituting the result into (1.8) gives the result.

Remark 4. If the matrix K is not of full rank, as would happen if, for example, $K(\cdot,\cdot)$ is of the form

$$K(s,t) = \sum_{\mu=1}^{N} \Psi_\mu(s)\Psi_\mu(t) \tag{1.9}$$

with $N < n$ then c is not uniquely determined by the setup in (1.7), (1.8). Here $\mathcal{H}_K$ contains at most N linearly independent functions. Letting X be the $n \times N$ matrix with i, μ th entry $\Psi_\mu(t_i)$ then $K = XX'$, and if c is a minimizer of (1.6), then $c + \delta$, where δ is orthogonal to the column span of X will also be a minimizer. We may substitute $c = X\gamma$ where γ is an N vector into (1.6), then Kc becomes $X\tilde{\gamma}$ and $c'Kc$ becomes $\tilde{\gamma}'\tilde{\gamma}$, where $\tilde{\gamma} = X'X\gamma$. For uniqueness we also need that if $f(t) = \sum_{\nu=1}^{M} d_\nu \phi_\nu(t)$, then $argmin_d \sum_{i=1}^{n} g_i(y_i, f_i)$ is unique. If the g_i are strictly convex functions of f_i this will be true whenever T is of full column rank. However, the strict convexity will be violated in some of the cases we consider later.

Remark 5. Characterization of isotropic RK's on E^d may be found in Skorkohod and Yadrenko (1973) and some examples along with their RKHS norms are given in the slides for my NIPS 96 workshop talk available via my home page. Characterization of isotropic RK's on the sphere may be found in Schoenberg (1942) and some examples along with their RKHS norms may be found in Wahba (1981, 1982b). $K(s,t)$ of the form $\int_{u \in \mathcal{U}} G(t,u)G(s,u)du$ will always be positive definite if the integral exists.

Remark 6. If $R_1(u_1, v_1), u_1, v_1 \in \mathcal{T}^{(1)}$ and $R_2(u_2, v_2), u_2, v_2 \in \mathcal{T}^{(2)}$ are positive definite functions on $\mathcal{T}^{(1)} \otimes \mathcal{T}^{(1)}$ and $\mathcal{T}^{(2)} \otimes \mathcal{T}^{(2)}$ respectively, then both the tensor product and the tensor sum of R_1 and R_2 are positive definite. That is, letting $\mathcal{T} = \mathcal{T}^{(1)} \otimes \mathcal{T}^{(2)}$, $s = (u_1, u_2) \in \mathcal{T}, t = (v_1, v_2) \in \mathcal{T}$, we have that $K(s,t) = R_1(u_1, v_1)R_2(u_2, v_2)$ and $K(s,t) = R_1(u_1, v_1) + R_2(u_2, v_2)$ are both positive definite on $\mathcal{T} \otimes \mathcal{T}$.

Remark 7. Re: Smoothing Spline ANOVA Spaces: Let $\mathcal{H}_K^\alpha$ be an RKHS of functions on $\mathcal{T}^{(\alpha)}$, for $\alpha = 1, \cdots, d$, and suppose $\mathcal{H}_K^\alpha$ has an orthogonal decomposition

$$\mathcal{H}_K^\alpha = [1^{(\alpha)}] \oplus \mathcal{H}_K^{(\alpha)} \tag{1.10}$$

where $[1^{(\alpha)}]$ is the one-dimensional space of constants on $\mathcal{T}^{(\alpha)}$, and let $R_\alpha(s^\alpha, t^\alpha)$ be the RK for $\mathcal{H}_K^{(\alpha)}$. Examples may be found in Wahba, Wang, Gu, Klein and Klein (1995) (WWGKK) and Gu and Wahba (1993). A Smoothing Spline ANOVA space $\mathcal{H}_K$ of functions on $\mathcal{T} = \mathcal{T}^{(1)} \otimes \cdots \otimes \mathcal{T}^{(d)}$ may be constructed by defining $\mathcal{H}_K$ as

$$\mathcal{H}_K = \prod_{\alpha=1}^{d} [[1^{(\alpha)}] \oplus \mathcal{H}_K^{(\alpha)}] \tag{1.11}$$

which then has the RK

$$K(s,t) = \prod_{\alpha=1}^{d} [1 + R_\alpha(s^\alpha, t^\alpha)] \tag{1.12}$$

$$= 1 + \sum_{\alpha=1}^{d} R_\alpha(s^\alpha, t^\alpha) + \sum_{\alpha<\beta} R_\alpha(s^\alpha, t^\alpha) R_\beta(s^\beta, t^\beta) + .. + \prod_{\alpha=1}^{d} R_\alpha(s^\alpha, t^\alpha). \tag{1.13}$$

Ordinarily the series in (1.13) is truncated somewhere and the direct sum of the corresponding subspaces in the corresponding expansion in (1.11) (which are orthogonal in this construction) constitute the 'model space'. Multiple smoothing parameters can be arranged by multiplying each of the individual RK's which remain in (1.13) after truncation, by $\theta_\alpha, \theta_{\alpha\beta}, \cdots$, and so forth. See W and WWGKK for details. The so-called main effects spaces, which involve only one t^α at a time are particularly popular, see Hastie and Tibshirani (1990).

Remark 8. The Smoothing Spline ANOVA spaces can be built up including conditionally positive definite functions (Micchelli 1986), leading to thin plate spline components (Gu and Wahba 1993), we omit the details.

1.1.4 Gaussian Processes, The Isometric Isomorphism Theorem

The relationship between conditional expectations on Gaussian processes and solutions to variational problems in RKHS has been known for a long time, see W, KW, Kimeldorf and Wahba (1970, 1971), Wahba (1978). This is not a coincidence. Let $X(t), t \in \mathcal{T}$ be a zero mean Gaussian stochastic process with $EX(s)X(t) = K(s,t)$. The Hilbert space $\mathcal{X}_K$ spanned by this stochastic process can be defined as all finite linear combinations of all random variables of the form $\sum_{\ell=1}^{L} a_\ell X(t_\ell)$ with $t_\ell \in \mathcal{T}$ and limits of such functions in the norm induced by the inner product

$$EX(s)X(t) = K(s,t). \tag{1.14}$$

Then we have

The Isometric Isomorphism Theorem (Parzen 1962, 1970)

`To every RKHS` $\mathcal{H}_K$ `there corresponds a zero mean Gaussian stochastic process` $X(t), t \in \mathcal{T}$ `with covariance` $K(s,t)$. `There is an isometric isomorphism [one-one inner product preserving map] between` $\mathcal{X}_K$, `the Hilbert space spanned by this stochastic process, and` $\mathcal{H}_K$, `whereby the random variable` $X(t) \in \mathcal{X}_K$ `corresponds to the representer` $K_t \in \mathcal{H}_K$.

The proof is trivial, details may be found in W. We note that sample functions of $X(t), t \in \mathcal{T}$ are not in $\mathcal{H}_K$ (with probability 1) if $\mathcal{H}_K$ is infinite dimensional. One may understand why this should be true by considering the case where K has a representation of the form (1.3). Then X has a Karhunen-Loeve expansion, namely

$$X(t) = \sum_{\nu} \xi_\nu \Psi_\nu(t) \tag{1.15}$$

where the ξ's are independent, zero mean Gaussian random variables with variance λ_ν, and a little algebra shows that $EX(s)X(t) = K(s,t)$ and also that the expected value of the RKHS norm if it exists, would be

$$E\|X(\cdot)\|^2_{\mathcal{H}_K} = \sum_{\nu} E\frac{\xi_\nu^2}{\lambda_\nu} \tag{1.16}$$

but this will be ∞ if $\mathcal{H}_K$ is infinite dimensional. This has consequences for how one might choose smoothing and other parameters, see, for example Wahba (1985a).

1.2 From Soft Classification to Hard Classification to SVM's

1.2.1 Hard Classification

Let $\mathcal{T}$ be a set as before, one observes n instances, $\{y_i, t_i\}, i = 1, \cdots, n, y_i \in \{+1, -1\}$ [the training set], where $t_i \in \mathcal{T}$ and $y_i = +1$ if the ith instance is member of class $\mathcal{A}$ and $y_i = -1$ if it is in class $\mathcal{B}$. Consider a random model for $\{y, t\}$:

$$Prob\{y = +1|t\} = p(t) \tag{1.17}$$

$$Prob\{y = -1|t\} = 1 - p(t) \tag{1.18}$$

Let $f(t) = \ln(p(t)/(1 - p(t))$ be the logit [also called the log odds ratio]. Assuming that the cost of misclassification is the same for both kinds of misclassification, then the optimal strategy for generalization, [minimization of expected loss], if one knew f, would be to classify as $\mathcal{A}$ if $f(t) > 0$ and $\mathcal{B}$ if $f(t) < 0$. Thus, letting $[f]_* = 1$ if $f > 0$ and 0 otherwise, one really wants to know sign f, equivalently it is desired to estimate $[-f(t)]_*$ from the training set $\{y_i, t_i\}, i = 1, \cdots, n$. This particular formulation is convenient, because we note that if $\hat{f}$ is used for classification, then the number of misclassifications on the training set will just be $\sum_{i=1}^n [-y_i \hat{f}(t_i)]_*$.

1.2.2 Soft Classification

If, on the other hand one's goal is not simply classification, but to understand how the relative risk $[e^{f(t)}]$ of $\mathcal{A}$ to $\mathcal{B}$, varies with t, as is frequently the case in demographic and environmental studies, then one is interested in estimating the actual value of f for all t in a region of $\mathcal{T}$, for which one is likely to have future observations. See, for example, WWGKK (1994, 1995). In this latter case one might estimate f from the training set by the methods of penalized log likelihood, that is, one finds f in $\{span\ \phi_\nu\} \oplus \mathcal{H}_K$ to minimize

$$\frac{1}{n}\sum_{i=1}^{n}\mathcal{L}(y_i, f_i) + \lambda\|h\|^2_{\mathcal{H}_K}. \tag{1.19}$$

Here, $f_i \doteq f(t_i)$ and $\mathcal{L}(y_i, f_i)$ is the negative log likelihood function[3]. In this example the likelihood that $y_i = 1$ is $p(t_i)$, and the likelihood that $y_i = -1$ is $(1 - p(t_i))$. Thus $\mathcal{L}(y_i, f_i) \doteq l(y_i f_i)$ where $l(\tau) = ln(1 + e^{-\tau})$. To see this, let $p_i \doteq p(t_i)$ and note that

$$\begin{array}{rcccl} \mathcal{L}(1, f_i) & = & -\ln(\frac{e^{f_i}}{1+e^{f_i}}) & = & -\ln p_i \\ \mathcal{L}(-1, f_i) & = & -\ln(\frac{1}{1+e^{f_i}}) & = & -\ln(1 - p_i) \end{array} \tag{1.20}$$

Thus, we may rewrite (1.19) as

$$\frac{1}{n}\sum_{i=1}^{n} l(y_i f_i) + \lambda\|h\|^2_{\mathcal{H}_K} \tag{1.21}$$

where $l(\tau) = \ln(1 + e^{-\tau})$. $l(\tau)$ is plotted in Figure 1.1 as ln(1+exp(-tau)).

Note that $l(\tau)$ is strictly convex. We know that $h = \sum_{i=1}^{n} c_i K_{t_i}$, $\|h\|^2_{\mathcal{H}_K} = c'Kc, f = (f_1, \cdots, f_n)' = Kc + Td$ and, if K is of full rank, $T'c = 0$, with the modifications noted if K is not of full rank. c and d are substituted into (1.21) and, if K is of full rank or the dimension of c is reduced appropriately, a strictly convex optimization problem with readily accessible gradient and Hessian results. A natural target for choosing λ is then to minimize the comparative Kullback-Liebler distance $CKL(\lambda) \doteq CKL(f_{true}, f_\lambda)$ between f_{true} and f_λ[4]. Here f_λ is the minimizer of (1.21) and f_{true} is the logit of the 'true' distribution p_{true} which generated the data. $CKL(\lambda)$ in this case becomes $E_{true}\sum_{i=1}^{n} l(y_i f_{\lambda i})$, see Xiang and Wahba (1996) for more details. Later we will turn to the randomized GACV method for estimating a computable proxy for $CKL(\lambda)$ (Xiang and Wahba 1996,1997, Lin and Wahba, in preparation).

---

3. In the statistics literature the usual log likelihood functional is formulated for $y = 1$ or 0.
4. Recall that the Kullback-Liebler distance is not really a distance.

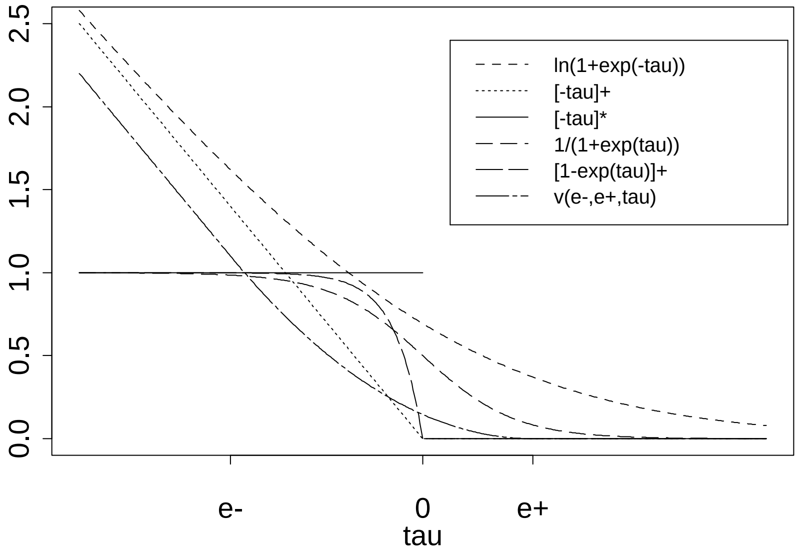

Figure 1.1 Pseudo-distance functions of tau (τ) mentioned in the text.

1.2.3 Back to Hard Classification

Section 1.2.1 suggests that we choose $f \in \{span\ \phi_\nu\} + h$ with $h \in \mathcal{H}_K$ to minimize

$$\frac{1}{n}\sum_{i=1}^{n}[\epsilon - y_i f_i]_* + \lambda \|h\|^2_{\mathcal{H}_K}, \tag{1.22}$$

for some fixed $\epsilon > 0$, thereby penalizing the misclassification rate rather than the log likelihood. $[-\tau]_*$ is plotted in Figure 1.1 as [-tau]* for comparison with $l(\tau)$. Substituting c and d into (1.22) as before, one seeks to find the minimizers, while choosing λ. It appears that a large λ will force f into $span\{\phi_\nu\}$, thus making f and (hence the boundaries of the different classification regions) less flexible, while a small λ will allow the classification boundaries to follow the training set more closely. However, if the attribute data is well separated by class, then the minimizer over d of $\sum_{i=1}^{n}[\epsilon - y_i \sum_{\nu=1}^{M} d_\nu \phi_\nu(t_i)]_*$ may not be unique or bounded, so that it will be necessary to further constrain the d_ν's. Supposing $\phi_1(t) = 1$, and letting $\phi(t) = \sum_{\nu=2}^{M} d_\nu \phi_\nu(t)$, in what follows we could replace the penalty $\lambda\|h\|^2_{\mathcal{H}_K}$ by $J_\lambda(f)$ where $J_\lambda(f) = \lambda_0\|\phi\|_0^2 + \lambda_1\|h\|^2_{\mathcal{H}_K}$, where $\|\phi\|_0^2$ is some appropriate positive definite quadratic form in $d_2, \cdots, d_M$, for example $\sum_{\nu=2}^{M} d_\nu^2$. Alternatively, the $\{\phi_\nu\}$ could retain their special role by being absorbed into K. In this case, $K(s,t)$ is replaced by $\theta \sum_{\nu=2}^{M} \phi_\nu(s)\phi_\nu(t) + K(s,t)$. Increasing θ forces more of the solution into the $\{\phi_\nu\}$.

Unfortunately the use of $[\epsilon - \tau]_*$ in (1.22) results in a nonconvex optimization problem, with its attendant pitfalls. However, Mangasarian (1994) has recently proposed numerical algorithms in the $\lambda_1 = \infty$ case with $t = (t^1, \cdots, t^d) \in E^d$, $\phi_1(t) = 1, \phi_\alpha(t) = t^\alpha, \alpha = 1, \cdots, d$. Bradley, Mangasarian and Street (1997) recently considered problems where the rather nasty function $[-\tau]_*$ is replaced with other more tractable functions including the sigmoidal approximation $1/(1+e^{a\tau})$ and the function $[1 - e^{a\tau}]_+$, concave for $\tau < 0$. Here, $[x]_+ = x, x > 0, = 0$ otherwise. For comparison, these two functions are also plotted in Figure 1.1 with $a = 1$. Bradley *et al* considered examples with a large number of variables, where the goal was to screen out some non-informative variables for deletion. They penalized the number of variables included and used 10-fold cross validation on the misclassification rate to choose a penalty parameter on the number of variables. See also Bennett and Blue (1997).

1.2.4 Convex Compromises with SVM's

Let $v(\tau) \doteq v_{\epsilon_+,\epsilon_-}(\tau)$ be defined by

$$\begin{array}{rclcrcl} v_{\epsilon_+,\epsilon_-}(\tau) & = & [-(\tau - \frac{\epsilon_+ + \epsilon_-}{2})]_+ & & \tau & < & \epsilon_- \\ & = & \frac{(\tau-\epsilon_+)^2}{2(\epsilon_+ - \epsilon_-)} & \epsilon_- \leq & \tau & \leq & \epsilon_+ \\ & = & 0 & \epsilon_+ \leq & \tau. & & \end{array} \tag{1.23}$$

For fixed $\epsilon_- < \epsilon_+$, $v_{\epsilon_+,\epsilon_-}(\tau)$ is convex and possesses a continuous first derivative, and a non-negative second derivative everywhere except at ϵ_- and $\epsilon+$, where the second derivative could be defined by assigning it to be continuous from the left, say. $v_{\epsilon_+,\epsilon_-}(\tau)$ is plotted in Figure 1.1 as v(e+,e-,tau), along with $v_{0,0}(\tau) \doteq [-\tau]_+$. $v_{\epsilon_+,\epsilon_+}(\tau) \doteq [\epsilon - \tau]_+$. The v's may be thought of as (in some sense) convex approximations to $[\epsilon - \tau]_*$, which for $\epsilon_- < \epsilon_+$ possess a continuous first derivative and non-negative second derivative which could be defined everywhere.

1.3 The Randomized GACV for Choosing λ

So far our discussion has been a relatively straightforward description of bridges between well known results in optimization in RKHS, Gaussian processes, penalized likelihood methods in soft classification (more commonly known as risk factor estimation) and SVM methods. This section is more heuristic and in the nature of work in progress. The goal is to explore to what extent the randomized GACV method in Xiang and Wahba (1996,1997) for choosing λ n the case $g(\tau) = ln(1 + e^{-\tau})$ may be extended to apply in the context of SVM's. Minimization of the generalized comparative Kullback-Liebler distance (GCKL) of f_λ to the '*true*' f as a function of λ is the target of the GACV. We first describe the GCKL and how it relates in some cases to the expected misclassification rate. Then we describe how the (computable) minimizer of the GACV should be a good estimate of the minimizer of the (not computable) GCKL. The randomized trace method for computing the GACV relatively efficiently is described and the details worked out for a simple case. Finally relations between the the GACV here and its versions in other contexts is noted.

1.3.1 The Generalized Comparative Kullback-Liebler Distance

Suppose unobserved y_i's will be generated according to an (unknown) probability model with $p(t) = p_{true}(t)$ being the probability that an instance with attribute vector t is in class $\mathcal{A}$. Let y_j be an (unobserved) value of y associated with t_j. Given f_λ, define the generalized comparative Kullback-Liebler distance (GCKL distance) with respect to g as

$$GCKL(f_{true}, f_\lambda) \doteq GCKL(\lambda) = E_{true}\frac{1}{n}\sum_{j=1}^{n} g(y_j f_{\lambda j}). \tag{1.24}$$

If $g(\tau) = ln(1 + e^{-\tau})$, then $GCKL(\lambda)$ reduces to the usual CKL [5] , averaged over the attribute vectors of the training set. If $g(\tau) = [\epsilon - \tau]_*$, then

$$\begin{aligned} E_{true}[\epsilon - y_j f_{\lambda j}]_* &= p_{[true]j}[\epsilon - f_{\lambda j}]_* + (1 - p_{[true]j})[\epsilon + f_{\lambda j}]_* & (1.25)\\ &= p_{[true]j}, \quad f_{\lambda j} < -\epsilon & (1.26)\\ &= 1, \quad -\epsilon \le f_{\lambda j} \le \epsilon & (1.27)\\ &= (1 - p_{[true]j}), \quad f_{\lambda j} > \epsilon, & (1.28) \end{aligned}$$

where $p_{[true]j} = p_{[true]}(t_j)$, so that the $GCKL(\lambda)$ is (a slight over estimate of) the expected misclassification rate for f_λ on unobserved instances if they have the same distribution of t_j as the training set (since the $GCKL$ is assigning 'misclassified' to all $f_{\lambda j} \in [-\epsilon, \epsilon]$.) Similarly, if $g(\tau) = [\epsilon - \tau]_+$, then

$$\begin{aligned} E_{true}[\epsilon - y_j f_{\lambda j}]_+ &= p_{[true]j}(\epsilon - f_{\lambda j}), \quad f_{\lambda j} < -\epsilon & (1.29)\\ &= \epsilon + (1 - 2p_{[true]j})f_{\lambda j}, \quad -\epsilon \le f_{\lambda j} \le \epsilon & (1.30)\\ &= (1 - p_{[true]j})(\epsilon + f_{\lambda j}), \quad f_{\lambda j} > \epsilon, & (1.31) \end{aligned}$$

not quite the misclassification rate, but related to it. The misclassification rate would be small if the large negative $f_{\lambda j}$ go with small $p_{[true]j}$ and the large positive $f_{\lambda j}$ go with small $(1 - p_{[true]j})$. We do not, of course, know $p_{[true]}$, so we cannot calculate $GCKL(\lambda)$ directly. However if it were cheap and easy to obtain an estimate of the minimizer of $GCKL(\lambda)$ it would be an appealing method for choosing λ.

5. The usual CKL (comparative Kullback-Liebler distance) is the Kullback-Liebler distance plus a term which depends only on $p_{[true]}$.

Since for $y_i = \pm 1$ and $0 < \epsilon < 1$, $V_\epsilon(y_i, f_i) = max\{0, |1 - y_i f_i| - \epsilon\}$, $g(\tau) = [|1-\tau| - \epsilon]_+$ corresponds to the 'usual' SVM. Note that this $g(\tau)$ is not monotonic in τ, but is 0 for $\tau \in [1-\epsilon, 1+\epsilon]$ and increases outside of this interval linearly as τ goes away from the interval in either direction. The relation of the GCKL to the misclassification rate in this example is not quite so direct, but it still may still be useful.

1.3.2 A Computable Proxy for the $GCKL$

1.3.2.1 Approximate Cross Validation

Xiang and Wahba (1996,1997) proposed the randomized GACV method for estimating a proxy for $CKL(\lambda)$. By a proxy for $CKL(\lambda)$ is meant a computable function whose minimizer is a good estimate for the minimizer of $CKL(\lambda)$. Define

$$I_\lambda(f, Y) = \frac{1}{n}\sum_{i=1}^{n} g(y_i f_i) + J_\lambda(f), \tag{1.32}$$

where $J_\lambda(f)$ is a quadratic penalty on f depending on λ. In this section we follow the derivation in Xiang and Wahba (1996) to find a computable proxy for $GCKL(\lambda)$, in the case that I_λ is strictly convex. In the SVM cases we are interested in, I_λ is generally convex but not strictly convex. However, the end result, below at (1.63) is well defined and plausible, even though some of the steps to get there are heuristic. The derivation proceeds by describing a leaving-out-one cross validation procedure for the GCKL and a series of approximations to get an approximate proxy for the GCKL. Then we describe a randomization procedure for computing this proxy efficiently. We emphasize that we do not actually do leaving-out-one, the randomization technique is a Monte Carlo estimate of a quantity approximating what we would expect to get if we actually did leaving-out-one.

Let $f_\lambda^{[-i]}$ be the solution to the variational problem: find $f \in \{span\ \phi_\nu\} \oplus \mathcal{H}_K$ to minimize

$$\frac{1}{n}\sum_{\substack{j=1\\ j\neq i}}^{n} g(y_j f_j) + J_\lambda(f). \tag{1.33}$$

Then the leaving-out-one function $V_0(\lambda)$ is defined as

$$V_0(\lambda) = \frac{1}{n}\sum_{i=1}^{n} g(y_i f_{\lambda i}^{[-i]}). \tag{1.34}$$

Since $f_{\lambda i}^{[-i]}$ does not depend on y_i but is (presumably) on average close to $f_{\lambda i}$, we may consider $V_0(\lambda)$ a proxy for $GCKL(\lambda)$, albeit one that is not generally feasible to compute in large data sets. Now let

$$V_0(\lambda) = OBS(\lambda) + D(\lambda), \tag{1.35}$$

where $OBS(\lambda)$ is the observed match of f_λ to the data,

$$OBS(\lambda) = \frac{1}{n}\sum_{i=1}^{n} g(y_i f_{\lambda i}) \tag{1.36}$$

and

$$D(\lambda) = \frac{1}{n}\sum_{i=1}^{n}[g(y_i f_{\lambda i}^{[-i]}) - g(y_i f_{\lambda i})]. \tag{1.37}$$

Using a first order Taylor series expansion gives

$$D(\lambda) \approx -\frac{1}{n}\sum_{i=1}^{n}\frac{\partial g}{\partial f_{\lambda i}}(f_{\lambda i} - f_{\lambda i}^{[-i]}). \tag{1.38}$$

Next we let $\mu(f)$ be a 'prediction' of y given f. Here we let

$$\mu_i = \mu(f_i) = \sum_{y \in \{+1,-1\}} \frac{\partial}{\partial f_i} g(y_i f_i). \tag{1.39}$$

When $g(\tau) = ln(1 + e^{-\tau})$ then $\mu(f) = 2p - 1 = E\{y|p\}$. For $g(\tau) = v_{\epsilon_+,\epsilon_-}(\tau)$, $\mu(f) = -1, f < min\{\epsilon_+, -\epsilon_-\}$, $\mu(f) = +1, f > max\{\epsilon_+, -\epsilon_-\}$, and varies in a non-decreasing piecewise linear fashion in between. $\mu(f)$ is plotted in Figure 1.2 for both these cases. For $g(\tau) = \frac{1}{2}[|1-\tau| - \epsilon]_+$, $\mu(f)$ is a nondecreasing step function with $\mu(-(1+\epsilon)) = -1$, $\mu(1+\epsilon) = +1$.

Letting $\mu_{\lambda i} = \mu(f_{\lambda i})$ and $\mu_{\lambda i}^{[-i]} = \mu(f_{\lambda i}^{[-i]})$, we may write (ignoring, for the moment, the possibility of dividing by 0),

$$D(\lambda) \approx -\frac{1}{n}\sum_{i=1}^{n}\frac{\partial g}{\partial f_{\lambda i}}\frac{(f_{\lambda i} - f_{\lambda i}^{[-i]})}{(y_i - \mu_{\lambda i}^{[-i]})}(y_i - \mu_{\lambda i}^{[-i]}) \tag{1.40}$$

$$\equiv -\frac{1}{n}\sum_{i=1}^{n}\frac{\partial g}{\partial f_{\lambda i}}\frac{(f_{\lambda i} - f_{\lambda i}^{[-i]})}{(y_i - \mu_{\lambda i}^{[-i]})}\frac{(y_i - \mu_{\lambda i})}{(1 - \frac{\mu_{\lambda i} - \mu_{\lambda i}^{[-i]}}{y_i - \mu_{\lambda i}^{[-i]}})}. \tag{1.41}$$

Next, approximate $\mu_{\lambda i} - \mu_{\lambda i}^{[-i]}$ as

$$\mu_{\lambda i} - \mu_{\lambda i}^{[-i]} = \mu(f_{\lambda i}) - \mu(f_{\lambda i}^{[-i]}) \approx \frac{\partial \mu}{\partial f_{\lambda i}}(f_{\lambda i} - f_{\lambda i}^{[-i]}). \tag{1.42}$$

Making the definitions

$$g_i' = \frac{\partial g}{\partial f_{\lambda i}} \tag{1.43}$$

$$\mu_i' = \frac{\partial \mu}{\partial f_{\lambda i}} \tag{1.44}$$

$$h_{ii} = \frac{f_{\lambda i} - f_{\lambda i}^{[-i]}}{y_i - \mu_{\lambda i}^{[-i]}} \tag{1.45}$$

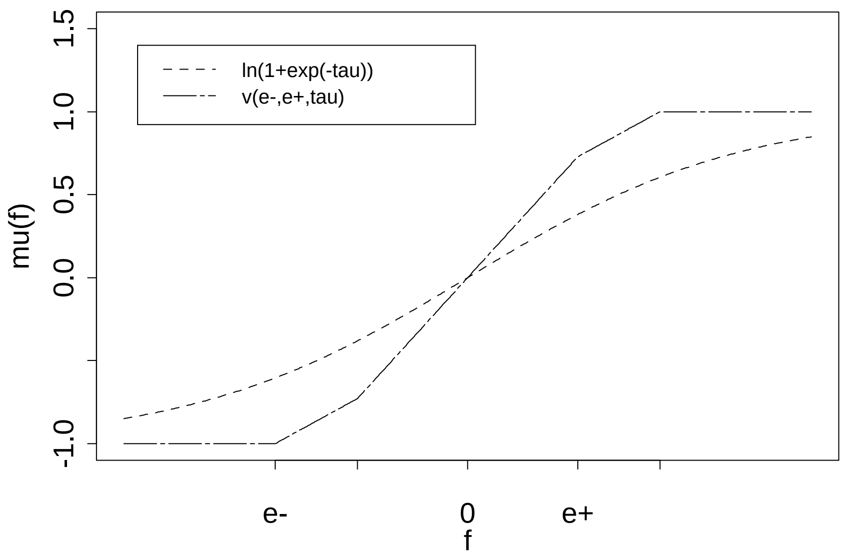

Figure 1.2 Prediction functions $\mu(f)$ defined in the text for $g(\tau) = ln(1 + e^{-\tau})$ and $g(\tau) = v_{\epsilon_+,\epsilon_-}(\tau)$.

in (1.41) gives, finally

$$D(\lambda) \approx -\frac{1}{n}\sum_{i=1}^{n} g_i' h_{ii} \frac{y_i - \mu_{\lambda i}}{1 - \mu_i' h_{ii}}. \tag{1.46}$$

Now $g_i, \mu_{\lambda i}$ and μ_i' depend only on $f_{\lambda i}$, which presumably, is at hand if the original variational problem has been solved. It remains to find an approximation for the right hand side of (1.45) to use for h_{ii}. The Leaving Out One Lemma will help.

1.3.3 The Leaving Out One Lemma

Let $g(\tau)$ be convex, and let $\mu = \mu(f)$ be a nondecreasing function of f with $\mu(-\infty) = -1$ and $\mu(\infty) = +1$, and let $f^\dagger(\mu)$ be any function of μ such that $\mu(f^\dagger(\mu)) = \mu$. Thus if $\mu(f)$ is a strictly monotone function of f then $f^\dagger$ is uniquely defined and $f^\dagger(\mu) = f(\mu)$. Suppose that

$$g(\mu(f)f^\dagger(\mu(f))) \leq g(\mu(f)f*) \tag{1.47}$$

for any $f*$ for which $\mu(f*) \neq \mu(f)$. It can be shown that $g = l$ and $g = v_{\epsilon_+,\epsilon_-}$ with the 'prediction functions' μ as in Figure 1.2 have property (1.47). Then we have the

Leaving-Out-One Lemma (Craven and Wahba 1979, Xiang and Wahba 1996)

Let $f_\lambda^{[-i]}$ be the minimizer of (1.33) as before. Then $f_\lambda^{[-i]}$ is also the minimizer of $\frac{1}{n}\sum_{i=1}^n g(y_j f_j) + J_\lambda(f)$ given the data $Y^{[-i]} = \{y_1, \cdots, y_{i-1}, \mu(f_{\lambda i}^{[-i]}), y_{i+1}, \cdots, y_n\}$.

A proof is in the Appendix.

This lemma says that if we leave out the ith data point, use the prediction function $\mu(f_\lambda^{[-i]})$ to 'predict' (or impute) y_i, and then solve the variational problem with $\mu(f_{\lambda i}^{[-i]})$ substituted in for y_i, we will get $f_\lambda^{[-i]}$ back for the solution.

1.3.4 An Approximation for $h_{ii} = \frac{f_{\lambda i} - f_{\lambda i}^{[-i]}}{y_i - \mu_{\lambda i}^{[-i]}}$

With some abuse of notation, in this subsection we let f stand for a vector of function values, at $t_1, \cdots, t_n$ rather than a function, as we have been doing. That is, here

$$f_\lambda = (f_{\lambda 1}, \cdots, f_{\lambda n})' \tag{1.48}$$
$$f_\lambda^{[-i]} = (f_{\lambda 1}^{[-i]}, \cdots f_{\lambda n}^{[-i]})' \tag{1.49}$$

and let

$$Y = (y_1, \cdots, y_n)' \tag{1.50}$$
$$Y^{[-i]} = (y_1, \cdots, y_{i-1}, \mu(f_{\lambda i}^{[-i]}), y_{i+1}, \cdots, y_n)'. \tag{1.51}$$

Recalling the definition of I_λ from (1.32), it can be shown that I_λ depends only on f through the values of f at (some of) the data points $t_1, \cdots, t_n$. We don't need to know this relationship explicitly here, however. Details have been worked out for the strictly convex case in Xiang and Wahba (1996). Expanding the vector $\frac{\partial I_\lambda}{\partial f}$ in a Taylor series about f_λ and Y gives

$$\frac{\partial I_\lambda}{\partial f}(f_\lambda^{[-i]}, Y^{[-i]}) = \frac{\partial I_\lambda}{\partial f}(f_\lambda, Y) + \frac{\partial^2 I_\lambda}{\partial f \partial f'}(f_\lambda, Y)(f_\lambda^{[-i]} - f_\lambda) + \frac{\partial^2 I_\lambda}{\partial f_\lambda \partial Y'}(Y^{[-i]} - Y) + .. \tag{1.52}$$

Now since f_λ is a minimizer of $I_\lambda(f, Y)$ and, by the Leaving Out One Lemma, $f_\lambda^{[-i]}$ is a minimizer of $I_\lambda(f, Y^{[-i]})$, we have

$$\frac{\partial I_\lambda}{\partial f}(f_\lambda^{[-i]}, Y^{[-i]}) = \frac{\partial I_\lambda}{\partial f}(f_\lambda, Y) = 0 \tag{1.53}$$

and, from (1.52),

$$H_{ff}(f_\lambda - f_\lambda^{[-i]}) \approx -H_{fY}(Y - Y^{[-i]}) \tag{1.54}$$

where $H_{ff} = \frac{\partial^2 I_\lambda}{\partial f \partial f'}$ and $H_{fy} = \frac{\partial^2 I_\lambda}{\partial f \partial Y'}$. If H_{ff} were invertible, we could write

$$h_{ii} \equiv \frac{f_{\lambda i} - f_{\lambda i}^{[-i]}}{y_i - \mu_{\lambda i}^{[-i]}} \approx -(H_{ff}^{-1} H_{fY})_{ii} = \tilde{h_{ii}}, \tag{1.55}$$

say, where $(H_{ff}^{-1} H_{fY})_{ii}$ is the iith entry of $H_{ff}^{-1} H_{fY}$. Here $H_{ff} = W + \Sigma_\lambda$ where W is the diagonal matrix with iith entry $w_{ii} = \frac{\partial^2}{\partial f_{\lambda i}^2} g(y_i f_{\lambda i})$, Σ_λ is the Hessian matrix of J_λ with respect to the $f_{\lambda i}$ and H_{fY} is the diagonal matrix with iith entry $\frac{\partial^2}{\partial f_{\lambda i} \partial y_i} g(y_i f_{\lambda i})$. Setting

$$\tilde{h_{ii}} = -(H_{ff}^{-1} H_{fY})_{ii} \tag{1.56}$$

gives our approximate cross validation function $ACV(\lambda)$ as an approximation to $V_0(\lambda)$, the leaving-out-one function of (1.34):

$$ACV(\lambda) = \frac{1}{n} \sum_{i=1}^{n} g(y_i f_{\lambda i}) - \frac{1}{n} \sum_{i=1}^{n} g_i' \tilde{h_{ii}} \frac{y_i - \mu_{\lambda i}}{1 - \mu_i' \tilde{h_{ii}}}. \tag{1.57}$$

$ACV(\lambda)$ can be shown to be equivalent to the $ACV_2(\lambda)$ of Xiang and Wahba (1996), p 689, after suitable modification for the setup here.

1.3.4.1 The Randomized Trace Estimate of $GACV(\lambda)$ *for* $g(\tau) = [\epsilon - \tau]_+$

Next, we consider $g(\tau) = v_{\epsilon, \epsilon-\delta}(\tau)$ and $\lim_{\delta \to 0} v_{\epsilon,\epsilon-\delta}(\tau) = [\epsilon - \tau]_+$. Table 1.3.4.1 gives the ingredients of $D(\lambda)$ other than $\tilde{h_{ii}}$ for $g(\tau) = [\epsilon - \tau]_+$. Note that as we take the above limit certain derivatives used in the derivation of (1.57) do not exist at $\tau = \pm\epsilon$. Nevertheless we proceed. Assuming that we need not be concerned at exactly the degenerate points $y_i f_i = \pm\epsilon$, we have, substituting the entries from Table 1.3.4.1 into (1.46),

$$D(\lambda) \approx \frac{1}{n} \sum_{y_i f_{\lambda i} - < \epsilon} 2\tilde{h_{ii}} + \frac{1}{n} \sum_{-\epsilon < y_i f_{\lambda i} < \epsilon} \tilde{h_{ii}}. \tag{1.58}$$

If H_{ff} were invertible, we would have the simple expression

$$D(\lambda) \approx -\frac{2}{n} trace E_\epsilon H_{ff}^{-1} H_{fY}, \tag{1.59}$$

where E_ϵ is the diagonal matrix with 1 in the iith position if $y_i f_i < -\epsilon$, with $\frac{1}{2}$ in the iith position if $-\epsilon < y_i f_i < \epsilon$, and 0 otherwise. We now give a heuristic argument for the randomized trace estimation of $D(\lambda)$ for $g(\tau) = [\epsilon - \tau]_+$, based on a perturbation of the data, and not requiring H_{ff} strictly positive definite. Let $Z = (z_1, \cdots, z_n)'$, where the z_i's will be generated by a random number generator with $Ez_i = 0$ and $Ez_i z_j = \sigma_Z^2, i = j, = 0$ otherwise. Let $f_\lambda \equiv f_\lambda^Y$ [6] be the minimizer of $I_\lambda(f, Y)$ as before and let f_λ^{Y+Z} be the minimizer of $I_\lambda(f, Y + Z)$. That is, we

6. Again, with some abuse of notation, we are letting f stand for a function, when convenient, and for the vector of its values at $t_1, \cdots, t_n$ when convenient.

	$y_i f_i < -\epsilon$	$-\epsilon < y_i f_i < \epsilon$	$\epsilon < y_i f_i$
$g(y_i f_i)$	$\epsilon - y_i f_i$		0
$\frac{\partial g}{\partial f_i}$	$-y_i$		0
$\frac{\partial^2 g}{\partial f_i \partial y_i}$	-1		0
$\frac{\partial g}{\partial f_i}(y_i - \mu_i)$	-2	-1	0
	$f_i < -\epsilon$	$-\epsilon < f_i < \epsilon$	$\epsilon < f_i$
$\mu(f_i)$	-1	0	1
$\frac{d\mu}{df_i}$	0	0	0

Table 1.1 Ingredients of $D(\lambda)$ for $g(\tau) = [\epsilon - \tau]_+$.

are perturbing the response vector Y by adding a (small) random perturbation Z. Note that in what follows $y_i + z_i$ does not have to be in $\{-1, 1\}$, and in general the variational problems here do not require the responses to be in that set. Using the Taylor series expansion (1.52) with $(f_\lambda^{Y+Z}, Y + Z)$ replacing $(f_\lambda^{[-i]}, Y^{[-i]})$ gives, assuming that Z is a small perturbation,

$$H_{ff}(f_\lambda^{Y+Z} - f_\lambda^Y) \approx -H_{fY} Z. \tag{1.60}$$

If H_{ff} were invertible we could write

$$f_\lambda^{Y+Z} - f_\lambda^Y \approx -H_{ff}^{-1} H_{fY} Z. \tag{1.61}$$

Then, observing that for any $n \times n$ matrix A, that $EZ'AZ = \sigma_Z^2 traceA$, we would have that

$$\frac{2}{n}\frac{1}{\sigma_z^2} Z' E_\epsilon (f_\lambda^{Y+Z} - f_\lambda^Y) \tag{1.62}$$

is an estimate of $-\frac{2}{n} traceE_\epsilon H_{ff}^{-1} H_{fY}$. For $g(\tau) = [\epsilon - \tau]_+$ generally H_{ff} will not be invertible at f_λ since f_λ does not depend on the inactive data points, that is, those y_i for which $y_i f_{\lambda i} > \epsilon$. However, we argue heuristically that the restriction of the argument above (and thus the restriction of H_{ff}) to just the active data points does make sense. (Recall that we will be limiting ourselves to I_λ with unique solutions.) Thus we conjecture that (1.62) will provide a reasonable randomized estimate of $D(\lambda)$ of (1.58). The end result is the randomized GACV function for $g(\tau) = [\epsilon - \tau]_+$ defined as

$$ranGACV(\lambda) = \frac{1}{n} \sum_{y_i f_{\lambda i} < \epsilon} [\epsilon - y_i f_{\lambda i}^Y]_+ + \frac{2}{n}\frac{1}{\sigma_z^2} Z' E_\epsilon (f_\lambda^{Y+Z} - f_\lambda^Y). \tag{1.63}$$

We conjecture that the minimizer of $ranGACV(\lambda)$, under suitable assumptions (in particular, fairly large sample sizes) should be a good estimate of the minimizer of $GCKL(\lambda)$ of (1.24). Note that the reasonableness of the result (1.63) is independent of *how* the solutions to the variational problem are found. The minimizer of I_λ with $g(\tau) = [\epsilon - \tau]_+$ will be found via a mathematical programming algorithm, whereas the case with $g(\tau) = l(\tau)$ (with its corresponding $GACV$ function) is typically found using a descent algorithm which uses the Hessian.

Note that the same Z should be used for all λ, however it is possible to compute several replicates of $D(\lambda)$ and take a suitable average, see Xiang (1996) who examined this question in the log likelihood case. It is to be expected that $D(\lambda)$ in the $[\epsilon - \tau]_+$ case will be 'bumpy' considered as a function of λ as instances move in and out of the active constraint set as λ varies, see Wahba(1982a), Villalobos and Wahba (1987) for related examples involving linear inequality constraints.

1.3.5 Discussion of $ranGACV$

Using the fact that for $g(\tau) = l(\tau)$ it can be shown that $-g'_i = \frac{1}{2}(y_i - \mu_i)$, then, (as noted before) (1.57) corresponds to the formula ACV_2 in Xiang and Wahba (1996) p. 689. In that paper a slightly different leaving out one was used in the setup $y_i = 1$ or $y_i = 0$ with probability p_i and $(1 - p_i)$ respectively. Then the negative log likelihood can be written as $-yf + b(f)$ where $b(f) = \ln(1 + e^f)$[7]. In that paper we used the same argument as described here starting with the leaving out one in the form $-y_i f_{\lambda i}^{[-i]} + b(f_{\lambda i})$, that is, we did not leave out one in the b term. The end result, called ACV there (which is based on $y_i \in \{0, 1\}$) resulted in

$$D(\lambda) = \sum_{i=1}^{n} \frac{y_i \tilde{h_{ii}}(y_i - \mu_{\lambda i})}{(1 - \mu'_i \tilde{h_{ii}})} \tag{1.64}$$

In that paper we replaced D by D_{GACV} defined by

$$D_{GACV}(\lambda) = \bar{h} \frac{\sum_{i=1}^{n} y_i (y_i - \mu_{\lambda i})}{1 - (\bar{\mu' h})} \tag{1.65}$$

where $\bar{h} = \frac{1}{n}\sum_{i=1}^{n} \tilde{h_{ii}}$, $\bar{\mu' h} = \frac{1}{n}\sum_{i=1}^{n} \mu'_i \tilde{h_{ii}}$ and demonstrated that the resulting $GACV(\lambda)$ provided an excellent proxy for the CKL in the examples tried. In Xiang and Wahba (1997) and Lin and Wahba, in preparation randomized versions of the GACV were tried and proved to be essentially as good as the exact version calculated via matrix decompositions.

The problem of choosing smoothing parameters in the log likelihood case with Gaussian response data with unknown noise variance has been extensively studied. In that case $f_i = Ey_i$ and $l(f) = \sum_{i=1}^{n}(y_i - f_i)^2$. In that case it can be shown

7. This is the usual formulation in the Statistics literature for the log likelihood for a member of an exponential family.

(Craven and Wahba 1979) that

$$(y_i - f_{\lambda i}^{[-i]})^2 = \frac{(y_i - f_{\lambda i})^2}{(1-h_{ii})} \tag{1.66}$$

which gave rise to the GCV estimate $V(\lambda)$

$$V(\lambda) = \frac{\frac{1}{n}\sum_{i=1}^{n}(y_i - f_{\lambda i})^2}{(1-\bar{h})}, \tag{1.67}$$

which is known to have various theoretically optimum properties (See Li (1986), W). The randomized trace version of it can be found in Girard (1991,1998), who showed that the randomized version was essentially as good as the exact version for large data sets, Hutchinson (1989), who used Bernoulli data for the perturbations, Wahba, Johnson, Gao and Gong (1995) who further compared exact and randomized versions of GCV, Gong, Wahba, Johnson and Tribbia (1998) where it was applied to a complex variational problem with multiple smoothing parameters, Golub and vonMatt (1997), who did extensive simulations. Wahba (1982a, 1985b) and Villalobos and Wahba (1987) considered variational problems in RKHS with Gaussian data and linear inequality constraints as side conditions, where a GCV function adapted to inequality constraints was used. It can be seen in Wahba (1982a) how $GCV(\lambda)$ has jumps as data points move in and out of the active constraint set as λ varies.

Acknowledgements

I would like to thank the organizers of the NIPS 97 SVM Workshop for inviting me to speak. I would especially like to thank Chris Burges for graciously providing me with some unpublished numerical results of his on estimating the misclassification rate, and for Olvi Mangasarian for some very helpful comments and for providing pointers to some important references. A lively dinner discussion with Trevor Hastie and Jerry Friedman at NIPS 96 contributed to the ideas in this paper. I am grateful to David Callan for his patient help as a sounding board and in producing this document, and to Fangyu Gao for the plots. This work was partly supported by NSF under Grant DMS-9704758 and NIH under Grant R01 EY09946.

Appendix: Proof of the Leaving Out One Lemma

The hypotheses of the Lemma give

$$g(\mu(f_{\lambda i}^{[-i]})f^{\dagger}(\mu(f_{\lambda i}^{[-i]}))) \le g(\mu(f_{\lambda i}^{[-i]})f_*)$$

for any f_* for which $\mu(f_*) \neq \mu(f_{\lambda i}^{[-i]})$. and, in particular

$$g(\mu(f_{\lambda i}^{[-i]})f_{\lambda i}^{[-i]}) \le g(\mu(f_{\lambda i}^{[-i]})f_*)$$

for any f_* for which $\mu(f_*) \neq \mu(f_{\lambda i}^{[-i]})$. Thus, letting $J(f) = \|h\|^2_{\mathcal{H}_K}$ (in an obvious notation), we have

$$
\begin{aligned}
g(\mu(f_{\lambda i}^{[-i]})f_i) + \sum_{j\neq i} g(y_j f_j) + n\lambda J(f) &\geq g(\mu(f_{\lambda i}^{[-i]})f_{\lambda i}^{[-i]}) + \sum_{j\neq i} g(y_j f_j) + n\lambda J(f) \\
&\geq g(\mu(f_{\lambda i}^{[-i]})f_{\lambda i}^{[-i]}) + \sum_{j\neq i} g(y_j f_{\lambda j}^{[-i]}) + n\lambda J(f_\lambda^{[-i]})
\end{aligned}
$$

giving the result.

References

1. Aronszajn, N. (1950), 'Theory of reproducing kernels', *Trans. Am. Math. Soc.* **68**, 337–404.
2. Bennett, K. & Blue, J. (1997), A support vector machine approach to decision trees, Technical report, Mathematical Sciences Department, RPI, Troy NY.
3. Bradley, P., Mangasarian, O. & Street, N. (1997), 'Feature selection via mathematical programming', *INFORMS J. Complexity, to appear.*
4. Cox, D. & O'Sullivan, F. (1990), 'Asymptotic analysis of penalized likelihood and related estimators', *Ann. Statist.* **18**, 1676–1695.
5. Craven, P. & Wahba, G. (1979), 'Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation', *Numer. Math.* **31**, 377–403.
6. Girard, D. (1991), 'Asymptotic optimality of the fast randomized versions of *GCV* and C_L in ridge regression and regularization', *Ann. Statist.* **19**, 1950–1963.
7. Girard, D. (1998), 'Asymptotic comparison of (partial) cross-validation, GCV and randomized GCV in nonparametric regression', *Ann. Statist.* **126**, 315–334.
8. Girosi, F. (1997), An equivalence between sparse approximation and support vector machines, Technical Report A. I. 1606, MIT artificial Intelligence Laboratory, Boston MA.
9. Golub, G. & vonMatt, U. (1997), 'Generalized cross-validation for large-scale problems', *J. Comput. Graph. Statist.* **6**, 1–34.
10. Gong, J., Wahba, G., Johnson, D. & Tribbia, J. (1998), 'Adaptive tuning of numerical weather prediction models: simultaneous estimation of weighting, smoothing and physical parameters', *Monthly Weather Review* **125**, 210–231.
11. Gu, C. & Wahba, G. (1993), 'Semiparametric analysis of variance with tensor product thin plate splines', *J. Royal Statistical Soc. Ser. B* **55**, 353–368.
12. Hastie, T. & Tibshirani, R. (1990), *Generalized Additive Models*, Chapman and Hall, 335pp.
13. Hutchinson, M. (1989), 'A stochastic estimator for the trace of the influence matrix for Laplacian smoothing splines', *Commun. Statist.-Simula.* **18**, 1059–1076.
14. Kimeldorf, G. & Wahba, G. (1970), 'A correspondence between Bayesian estimation of stochastic processes and smoothing by splines', *Ann. Math. Statist.* **41**, 495–502.
15. Kimeldorf, G. & Wahba, G. (1971), 'Some results on Tchebycheffian spline functions', *J. Math. Anal. Applic.* **33**, 82–95.

16. Li, K. C. (1986), 'Asymptotic optimality of C_L and generalized cross validation in ridge regression with application to spline smoothing', *Ann. Statist.* **14**, 1101–1112.
17. Mangasarian, O. (1994), 'Misclassification minimization', *J. Global Optimization* **5**, 309–323.
18. Micchelli, C. (1986), 'Interpolation of scattered data: distance matrices and conditionally positive definite functions', *Constructive Approximation* **2**, 11–22.
19. Parzen, E. (1962), 'An approach to time series analysis', *Ann. Math. Statist.* **32**, 951–989.
20. Parzen, E. (1970), Statistical inference on time series by rkhs methods, *in* R. Pyke, ed., 'Proceedings 12th Biennial Seminar', Canadian Mathematical Congress, Montreal. 1-37.
21. Schoenberg, I. (1942), 'Positive definite functions on spheres', *Duke Math. J.* **9**, 96–108.
22. Schölkopf, B., Sung, K., Burges, C., Girosi, F., Niyogi, P., Poggio, T. & Vapnik, V. (1997), Comparing support vector machines with gaussian kernels to radial basis function classifiers, Technical Report 1599, Center for biological and Computational Learning, MIT, Boston MA.
23. Skorokhod, A. & Yadrenko, M. (1973), 'On absolute continuity of measures corresponding to homogeneous Gaussian fields', *Theory of Probability and its Applications* **XVIII**, 27–40.
24. Smola, A. & Schölkopf, B. (1997), 'On a kernel-based method for pattern recognition, regression, approximation, and operator inversion', ms.
25. Vapnik, V. (1995), *The Nature of Statistical Learning Theory*, Springer.
26. Villalobos, M. & Wahba, G. (1987), 'Inequality constrained multivariate smoothing splines with application to the estimation of posterior probabilities', *J. Am. Statist. Assoc.* **82**, 239–248.
27. Wahba, G. (1978), 'Improper priors, spline smoothing and the problem of guarding against model errors in regression', *J. Roy. Stat. Soc. Ser. B* **40**, 364–372.
28. Wahba, G. (1981), 'Spline interpolation and smoothing on the sphere', *SIAM J. Sci. Stat. Comput.* **2**, 5–16.
29. Wahba, G. (1982*a*), Constrained regularization for ill posed linear operator equations, with applications in meteorology and medicine, *in* S. Gupta & J. Berger, eds, 'Statistical Decision Theory and Related Topics, III, Vol.2', Academic Press, pp. 383–418.
30. Wahba, G. (1982*b*), 'Erratum: Spline interpolation and smoothing on the sphere', *SIAM J. Sci. Stat. Comput.* **3**, 385–386.
31. Wahba, G. (1985*a*), 'A comparison of GCV and GML for choosing the smoothing parameter in the generalized spline smoothing problem', *Ann. Statist.* **13**, 1378–1402.
32. Wahba, G. (1985*b*), Multivariate thin plate spline smoothing with positivity and other linear inequality constraints, *in* E. Wegman & D. dePriest, eds, 'Statistical Image Processing and Graphics', Marcel Dekker, pp. 275–290.
33. Wahba, G. (1990), *Spline Models for Observational Data*, SIAM. CBMS-NSF Regional Conference Series in Applied Mathematics, v. 59.
34. Wahba, G. (1992), Multivariate function and operator estimation, based on smoothing splines and reproducing kernels, *in* M. Casdagli & S. Eubank, eds,

‘Nonlinear Modeling and Forecasting, SFI Studies in the Sciences of Complexity, Proc. Vol XII’, Addison-Wesley, pp. 95–112.

35. Wahba, G., Johnson, D., Gao, F. & Gong, J. (1995*a*), ‘Adaptive tuning of numerical weather prediction models: randomized GCV in three and four dimensional data assimilation’, *Mon. Wea. Rev.* **123**, 3358–3369.

36. Wahba, G., Wang, Y., Gu, C., Klein, R. & Klein, B. (1994), Structured machine learning for ‘soft’ classification with smoothing spline ANOVA and stacked tuning, testing and evaluation, *in* J. Cowan, G. Tesauro & J. Alspector, eds, ‘Advances in Neural Information Processing Systems 6’, Morgan Kauffman, pp. 415–422.

37. Wahba, G., Wang, Y., Gu, C., Klein, R. & Klein, B. (1995*b*), ‘Smoothing spline ANOVA for exponential families, with application to the Wisconsin Epidemiological Study of Diabetic Retinopathy’, *Ann. Statist.* **23**, 1865–1895.

38. Xiang, D. (1996), Model Fitting and Testing for Non-Gaussian Data with a Large Data Set, PhD thesis, Technical Report 957, University of Wisconsin-Madison, Madison WI.

39. Xiang, D. & Wahba, G. (1996), ‘A generalized approximate cross validation for smoothing splines with non-Gaussian data’, *Statistica Sinica* **6**, 675–692.

40. Xiang, D. & Wahba, G. (1997), Approximate smoothing spline methods for large data sets in the binary case, Technical Report 982, Department of Statistics, University of Wisconsin, Madison WI. To appear in the Proceedings of the 1997 ASA Joint Statistical Meetings, Biometrics Section, pp 94-98 (1998).