16 GACV for Support Vector Machines

Grace Wahba
Department of Statistics
University of Wisconsin
1210 West Dayton Street
Madison, WI 53706, USA
wahba@stat.wisc.edu

Yi Lin
Department of Statistics
University of Wisconsin
1210 West Dayton Street
Madison, WI 53706, USA
yilin@stat.wisc.edu

Hao Zhang
Department of Statistics
University of Wisconsin
1210 West Dayton Street
Madison, WI 53706, USA
hzhang@stat.wisc.edu

We introduce the Generalized Approximate Cross Validation (GACV) for estimating tuning parameter(s) in SVMs. The GACV has as its target the choice of parameters which will minimize the Generalized Comparative Kullback-Leibler Distance (GCKL). The GCKL is seen to be an upper bound on the expected misclassification rate. Some modest simulation examples suggest how it might work in practice. The GACV is the sum of a term which is the observed (sample) GCKL plus a margin-like quantity.

16.1 Introduction

reproducing kernel Hilbert space (RKHS)

It is now common knowledge that the support vector machine (SVM) paradigm, which has proved highly successful in a number of classification studies, can be cast as a variational/regularization problem in a reproducing kernel Hilbert space (RKHS), see [Kimeldorf and Wahba, 1971, Wahba, 1990, Girosi, 1998, Poggio and Girosi, 1998], the papers and references in [Schölkopf et al., 1999a], and elsewhere. In this note, which is a sequel to [Wahba, 1999b], we look at the SVM paradigm from the point of view of a regularization problem, which allows a comparison with penalized log likelihood methods, as well as the application of model selection and tuning approaches which have been used with those and other regularization-type algorithms to choose tuning parameters in nonparametric statistical models.

GCKL

GACV

We first note the connection between the SVM paradigm in RKHS and the (dual) mathematical programming problem traditional in SVM classification problems. We then review the Generalized Comparative Kullback-Leibler distance (GCKL) for the usual SVM paradigm, and observe that it is trivially a simple upper bound on the expected misclassification rate. Next we revisit the GACV (Generalized Approximate Cross Validation) as a proxy for the GCKL proposed by Wahba [1999b] and the argument that it is a reasonable estimate of the GCKL. We found that it is not necessary to do the randomization of the GACV in [Wahba, 1999b], because it can be replaced by an equally justifiable approximation which is readily computed exactly, along with the SVM solution to the dual mathematical programming problem. This estimate turns out interestingly, but not surprisingly to be simply related to what several authors have identified as the (observed) VC dimension of the estimated SVM. Some preliminary simulations are suggestive of the fact that the minimizer of the GACV is in fact a reasonable estimate of the minimizer of the GCKL, although further simulation and theoretical studies are warranted. It is hoped that this preliminary work will lead to better understanding of "tuning" issues in the optimization of SVM's and related classifiers.

16.2 The SVM Variational Problem

reproducing kernel

Let $\mathcal{T}$ be an index set, $t \in \mathcal{T}$. Usually $\mathcal{T} = E^d$, Euclidean d-space, but not necessarily. Let $K(s,t), s,t \in \mathcal{T}$, be a positive definite function on $\mathcal{T} \otimes \mathcal{T}$, and let $\mathcal{H}_K$ be the RKHS with reproducing kernel (RK) K. See [Wahba, 1990, 1982, Lin et al., 1998] for more on RKHS. RK's which are tensor sums and products of RK's are discussed there and elsewhere. K may contain one or more tuning parameters, to be chosen. A variety of RK's with success in practical applications have been proposed by various authors, see, e.g., the Publications list at `http://www.kernel-machines.org`. Recently [Poggio and Girosi, 1998] interestingly observed how different scales may be accommodated using RKHS methods. We are given a training set $\{y_i, t_i\}$, where the attribute vector $t_i \in \mathcal{T}$, and $y_i = \pm 1$ according as an example with attribute

vector t_i is in category $\mathcal{A}$ or $\mathcal{B}$. The classical SVM paradigm is equivalent to: find f_λ of the form $const + h$, where $h \in \mathcal{H}_K$ to minimize

regularization problem

$$\frac{1}{n}\sum_{i=1}^{n}(1 - y_i f_i)_+ + \lambda \|h\|^2_{\mathcal{H}_K}, \tag{16.1}$$

here $f_i = f(t_i)$, and $(\tau)_+ = \tau, \tau > 0; = 0$ otherwise. Similar regularization problems have a long history, see, for example [Kimeldorf and Wahba, 1971]. Once the minimizer, call it f_λ is found, then the decision rule for a new example with attribute vector t is: $\mathcal{A}$ if $f_\lambda(t) > 0$, $\mathcal{B}$ if $f_\lambda(t) < 0$.

We will assume for simplicity that K is strictly positive definite on $\mathcal{T} \otimes \mathcal{T}$, although this is not necessary. The minimizer of (16.1) is known to be in the span $\{K(\cdot, t_i), i = 1, \cdots n\}$, of representers of evaluation in $\mathcal{H}_K$. The function $K(\cdot, t_i)$ is $K(s, t_i)$ considered as a function of s with t_i fixed. The famous "reproducing" property gives the inner product in $\mathcal{H}_K$ of two representers as $< K(\cdot, t_i), K(\cdot, t_j) >_{\mathcal{H}_K} = K(t_i, t_j)$. Thus, if $h(\cdot) = \sum_{i=1}^{n} c_i K(\cdot, t_i)$, then $\|h\|^2_{\mathcal{H}_K} = \sum_{i,j=1}^{n} c_i c_j K(t_i, t_j)$. Letting $e = (1, \cdots, 1)', y = (y_1, \cdots, y_n)', c = (c_1, \cdots, c_n)', (f(t_1), \cdots f(t_n))' = (f_1, \cdots, f_n)'$, and with some abuse of notation, letting $f = (f_1, \cdots, f_n)'$ and K now be the $n \times n$ matrix with ijth entry $K(t_i, t_j)$, and noting that $f(t) = d + \sum_{i=1}^{n} c_i K(t, t_i)$ for some c, d, we have

$$f = Kc + ed \tag{16.2}$$

and the variational problem (16.1) becomes: find (c, d) to minimize

$$\frac{1}{n}\sum_{i=1}^{n}(1 - y_i f_i)_+ + \lambda c' K c. \tag{16.3}$$

16.3 The Dual Problem

Let Y be the $n \times n$ diagonal matrix with y_i in the iith position, and let $H = \frac{1}{2n\lambda} YKY$. By going to the dual form of (16.3), it can be shown that $c = \frac{1}{2n\lambda} Y\alpha$, where α is the solution to the problem

$$\text{maximize } L = -\frac{1}{2}\alpha' H \alpha + e'\alpha \tag{16.4}$$

$$\text{subject to } \begin{cases} 0 & \leq \alpha & \leq e \\ e'Y\alpha & = y'\alpha & = 0. \end{cases} \tag{16.5}$$

Assuming that there is an i for which $0 < \alpha_i < 1$, it can also be shown that $d = 1/y_i - \sum_{j=1}^{n} c_j K(t_i, t_j)$. This is the usual form in which the SVM is computed. In the experiments reported below, we used the MINOS [Murtagh and Saunders, 1998] optimization routine to find α, and hence c. The support vectors are those $K(\cdot, t_i)$ for which $\alpha_i \neq 0$, equivalently $c_i \neq 0$. d can be found from any of the support vectors for which $0 < \alpha_i < 1$.

For future reference we review the relation between the (hard) margin (γ) of the support vector machine classifier and $\sum_{y_i f_{\lambda i} \leq 1} \alpha_{\lambda i}$. In the situation where we can separate the training set points perfectly, γ is given by

$$\gamma^2 = 2n\lambda \left(\sum_{y_i f_{\lambda i} \leq 1} \alpha_{\lambda i} \right)^{-1} . \tag{16.6}$$

See [Cortes and Vapnik, 1995, Bartlett and Shawe-Taylor, 1999]. (Notice the notation is a bit different from ours in these papers.) By definition the margin of the (hard margin) support vector machine classifier is $\gamma = \frac{1}{\|h\|_{\mathcal{H}_K}} = (c'Kc)^{-1/2}$. The equality (16.6) can be seen from the following: In the perfectly separable case, where all members of the training set are classified correctly, $\alpha_{\lambda i}$ is the solution of the problem below:

margin of the SVM classifier

$$\text{maximize } L = -\frac{1}{2}\alpha' H\alpha + e'\alpha \tag{16.7}$$

$$\text{subject to } \alpha_i \geq 0 \text{ and } y'\alpha = 0. \tag{16.8}$$

Introducing the Lagrangian multipliers $\xi = (\xi_1, \ldots, \xi_n)'$ and β for the constraints, the Lagrangian for this problem is

$$L_P = -\frac{1}{2}\alpha' H\alpha + e'\alpha - \beta y'\alpha - \xi'\alpha$$

and $\alpha_{\lambda i}$ satisfies the Kuhn-Tucker conditions:

$$\begin{aligned} \frac{\partial}{\partial \alpha} L_P = -H\alpha + e - \beta y - \xi = 0 \\ \alpha_i \geq 0, \ i = 1, 2, ..., n \\ y'\alpha = 0 \\ \xi_i \geq 0, \ i = 1, 2, ..., n \\ \xi_i \alpha_i = 0, \ i = 1, 2, ..., n \end{aligned}$$

From these and the relation that $c = Y\alpha_\lambda/(2n\lambda)$, it is easy to get

$$c'Kc = \frac{1}{2n\lambda}\alpha'_\lambda H \alpha_\lambda = \frac{1}{2n\lambda}\left[\alpha'_\lambda e - \beta \alpha'_\lambda y - \alpha'_\lambda \xi\right] = \frac{1}{2n\lambda}\left[\alpha'_\lambda e\right]. \tag{16.9}$$

Since $\alpha_{\lambda i} = 0$ if $y_i f_i > 1$, we finally get

$$\gamma^2 = (c'Kc)^{-1} = 2n\lambda \left[\sum_{y_i f_{\lambda i} \leq 1} \alpha_{\lambda i} \right]^{-1} .$$

16.4 The Generalized Comparative Kullback-Leibler Distance

Suppose unobserved y_i's will be generated according to an (unknown) probability model with $p(t) = p_{\text{true}}(t)$ 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.

GCKL

Given f_λ, define the Generalized Comparative Kullback-Leibler distance (GCKL distance) with respect to g as

$$GCKL(p_{\text{true}}, f_\lambda) \doteq GCKL(\lambda) = E_{\text{true}} \frac{1}{n} \sum_{j=1}^{n} g(y_j f_{\lambda j}). \tag{16.10}$$

penalized log likelihood

Here f_λ is considered fixed and the expectation is taken over future, unobserved y_j. If $g(\tau) = \ln(1 + e^{-\tau})$, (which corresponds to classical penalized log likelihood estimation if it replaces $(1 - \tau)_+$ in (16.1)) $GCKL(\lambda)$ reduces to the usual CKL for Bernoulli data[1] averaged over the attribute vectors of the training set. More details may be found in [Wahba, 1999b]. Let $[\tau]_* = 1$ if $\tau > 0$ and 0 otherwise. If $g(\tau) = [-\tau]_*$, then

$$E_{\text{true}}[-y_j f_{\lambda j}]_* = p_{[\text{true}]j}[-f_{\lambda j}]_* + (1 - p_{[\text{true}]j})[f_{\lambda j}]_* \tag{16.11}$$

$$= p_{[\text{true}]j}, \quad f_{\lambda j} < 0 \tag{16.12}$$

$$= (1 - p_{[\text{true}]j}), \quad f_{\lambda j} > 0, \tag{16.13}$$

where $p_{[\text{true}]j} = p_{[true]}(t_j)$, so that the $GCKL(\lambda)$ is the expected misclassification rate for f_λ on unobserved instances if they have the same distribution of t_j as the training set. Similarly, if $g(\tau) = (1 - \tau)_+$, then

$$E_{\text{true}}(1 - y_j f_{\lambda j})_+ = p_{[\text{true}]j}(1 - f_{\lambda j}), \quad f_{\lambda j} < -1 \tag{16.14}$$

$$= 1 + (1 - 2p_{[\text{true}]j}) f_{\lambda j}, \quad -1 \le f_{\lambda j} \le 1 \tag{16.15}$$

$$= (1 - p_{[\text{true}]j})(1 + f_{\lambda j}), \quad f_{\lambda j} > 1. \tag{16.16}$$

Note that $[-y_i f_i]_* \le (1 - y_i f_i)_+$, so that the $GCKL$ for $(1 - y_i f_i)_+$ is an *upper bound* for the *expected misclassification rate* - see Figure 16.1.

16.5 Leaving-out-one and the GACV

Recently there has been much interest in choosing λ (or its equivalent, referred to in the literature as $\frac{1}{2nC}$), as well as other parameters inside K. See for example [Burges, 1998, Cristianini et al., 1999, Kearns et al., 1997], surely not a complete list. Important references in the statistics literature that are related include [Efron and Tibshirani, 1997, Ye and Wong, 1997]. Lin et al. [1998] consider in detail the case $g(\tau) = \ln(1+e^{-\tau})$. We now obtain the GACV estimate for λ and other tuning parameters.

1. The usual CKL (comparative Kullback-Leibler distance) is the Kullback-Leibler distance plus a term which depends only on $p_{[true]}$. In this case g is the negative log likelihood and f_λ plays the role of (an estimate of) the logit $ln[p/1-p]$. See also [Friedman et al., 1998].

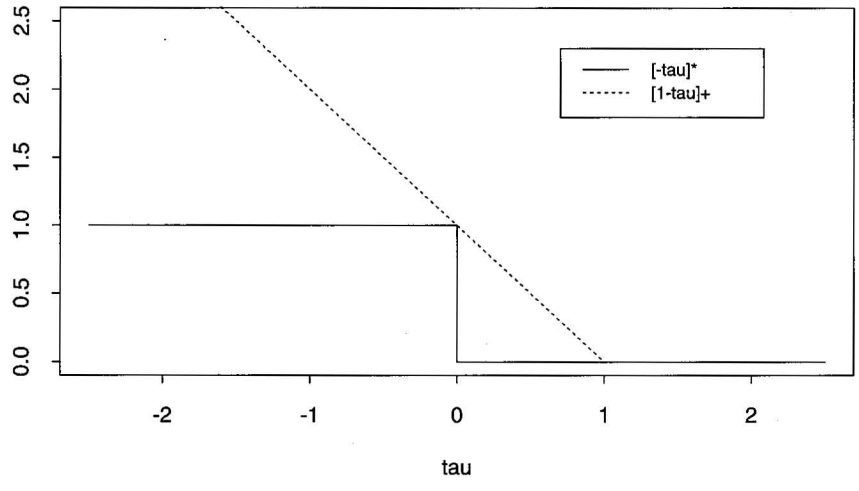

Figure 16.1 $g(\tau) = (1-\tau)_+$ and $g(\tau) = [-\tau]_*$ compared.

Let $f_\lambda^{[-i]}$ be the solution to the variational problem: find f of the form $f = const + h$ with $h \in \mathcal{H}_K$ to minimize

$$\frac{1}{n}\sum_{\substack{j=1\\ j\neq i}}^{n} g(y_j f_j) + \lambda \|h\|^2_{\mathcal{H}_K}. \tag{16.17}$$

leaving-out-one

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

$$V_0(\lambda) = \frac{1}{n}\sum_{i=1}^{n} g(y_i f_{\lambda i}^{[-i]}). \tag{16.18}$$

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

$$V_0(\lambda) = OBS(\lambda) + D(\lambda), \tag{16.19}$$

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

$$OBS(\lambda) = \frac{1}{n}\sum_{i=1}^{n} g(y_i f_{\lambda i}) \tag{16.20}$$

and

$$D(\lambda) = \frac{1}{n}\sum_{i=1}^{n} [g(y_i f_{\lambda i}^{[-i]}) - g(y_i f_{\lambda i})]. \tag{16.21}$$

Using a first order Taylor series expansion gives

$$D(\lambda) \approx -\frac{1}{n}\sum_{i=1}^{n}\frac{\partial g}{\partial f_{\lambda i}}(f_{\lambda i} - f_{\lambda i}^{[-i]}). \tag{16.22}$$

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

$$\mu_i = \mu(f_i) = \sum_{y\in\{+1,-1\}} \frac{\partial}{\partial f_i} g(y_i f_i). \tag{16.23}$$

When $g(\tau) = \ln(1+e^{-\tau})$ then $\mu(f) = 2p-1 = E\{y|p\}$. Since this $g(\tau)$ corresponds to the penalized log likelihood estimate, it is natural in this case to define the "prediction" of y given f as the expected value of y given f (equivalently, p). For $g(\tau) = (1-\tau)_+$, this definition results in $\mu(f) = -1, f < -1; \mu(f) = 0, -1 \le f \le 1$ and $\mu(f) = 1$ for $f > 1$. This might be considered a kind of all-or-nothing prediction of y, being, essentially, ± 1 outside of the margin and 0 inside it. 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),

$$D(\lambda) \approx -\frac{1}{n}\sum_{i=1}^{n}\frac{\partial g}{\partial f_{\lambda i}}\frac{(f_{\lambda i} - f_{\lambda i}^{[-i]})}{(y_i - \mu_{\lambda i}^{[-i]})}(y_i - \mu_{\lambda i}^{[-i]}) \tag{16.24}$$

This is equation (6.36) in [Wahba, 1999b]. We now provide somewhat different arguments than in [Wahba, 1999b] to obtain a similar result, which, however is easily computed as soon as the dual variational problem is solved.

Let $f_\lambda[i,x]$ be the solution of the variational problem (16.1) [2] given the data $\{y_1,\cdots,y_{i-1},x,y_{i+1},\cdots,y_n\}$. Note that the variational problem does not require that $x = \pm 1$. Thus $f_\lambda[i,y_i](t_i) \equiv f_{\lambda i}$. To simplify the notation, let $f_\lambda[i,x](t_i) = f_{\lambda i}[i,x] = f_{\lambda i}[x]$. In [Wahba, 1999b] it is shown, via a generalized leaving-out-one lemma, that $\mu(f)$ as we have defined it has the property that $f_{\lambda i}^{[-i]} = f_\lambda[i,\mu_{\lambda i}^{[-i]}](t_i)$. Letting $\mu_{\lambda i}^{[-i]} = x$, this justifies the approximation

$$\frac{f_{\lambda i} - f_{\lambda i}^{[-i]}}{y_i - \mu_{\lambda i}^{[-i]}} \equiv \frac{f_{\lambda i}[y_i] - f_{\lambda i}[x]}{y_i - x} \approx \frac{\partial f_{\lambda i}}{\partial y_i}. \tag{16.25}$$

Furthermore, $\mu_{\lambda i}^{[-i]} \equiv \mu(f_{\lambda i}^{[-i]}) = \mu(f_{\lambda i})$ whenever $f_{\lambda i}^{[-i]}$ and $f_{\lambda i}$ are both in the interval $(-\infty,-1)$, or $[-1,1]$, or $(1,\infty)$, which can be expected to happen with few exceptions. Thus, we make the further approximation $(y_i - \mu_{\lambda i}^{[-i]}) \approx (y_i - \mu_{\lambda i})$, and we replace (16.24) by

$$D(\lambda) \approx -\frac{1}{n}\sum_{i=1}^{n}\frac{\partial g}{\partial f_{\lambda i}}\frac{\partial f_{\lambda i}}{\partial y_i}(y_i - \mu_{\lambda i}). \tag{16.26}$$

---

2. d is not always uniquely determined; this however does not appear to be a problem in practice, and we shall ignore it.

Now, for $g(\tau) = (1-\tau)_+$

$$\begin{aligned}\frac{\partial g}{\partial f_{\lambda i}}(y_i - \mu_{\lambda i}) &= -2,\ y_i f_{\lambda i} < -1 \\ &= -1,\ y_i f_{\lambda i} \in [-1,1] \\ &= \ 0,\ y_i f_{\lambda i} > 1,\end{aligned}$$

giving finally

$$D(\lambda) \approx \frac{1}{n} \sum_{y_i f_{\lambda i} < -1} 2\frac{\partial f_{\lambda i}}{\partial y_i} + \frac{1}{n} \sum_{y_i f_{\lambda i} \in [-1,1]} \frac{\partial f_{\lambda i}}{\partial y_i}. \tag{16.27}$$

It is not hard to see how $\frac{\partial f_{\lambda i}}{\partial y_i}$ should be interpreted. Fixing λ and solving the variational problem for f_λ we obtain $\alpha = \alpha_\lambda$, $c = c_\lambda = \frac{1}{2n\lambda} Y\alpha_\lambda$ and for the moment letting f_λ be the column vector with ith component $f_{\lambda i}$, we have $f_\lambda = Kc_\lambda + ed = \frac{1}{2n\lambda} KY\alpha_\lambda + ed$. From this we may write

$$\frac{\partial f_{\lambda i}}{\partial y_i} = K(t_i, t_i)\frac{\alpha_{\lambda i}}{2n\lambda} \equiv \|K(\cdot, t_i)\|^2_{\mathcal{H}_K} \frac{\alpha_{\lambda i}}{2n\lambda}. \tag{16.28}$$

The resulting $GACV(\lambda)$, which is believed to be a reasonable proxy for $GCKL(\lambda)$, is, finally

GACV

$$GACV(\lambda) = \frac{1}{n}\sum_{i=1}^{n}(1 - y_i f_{\lambda i})_+ + \hat{D}(\lambda), \tag{16.29}$$

where

$$\hat{D}(\lambda) = \frac{1}{n}\left[2 \sum_{y_i f_{\lambda i} < -1} \frac{\alpha_{\lambda i}}{2n\lambda} \cdot \|K(\cdot, t_i)\|^2_{\mathcal{H}_K} + \sum_{y_i f_{\lambda i} \in [-1,1]} \frac{\alpha_{\lambda i}}{2n\lambda} \cdot \|K(\cdot, t_i)\|^2_{\mathcal{H}_K}\right]. \tag{16.30}$$

If $K = K_\theta$, where θ are some parameters inside K to which the result is sensitive, then we may let $GACV(\lambda) = GACV(\lambda, \theta)$. Note the relationship between $\hat{D}$ and $\sum_{y_i f_{\lambda i} \le 1} \alpha_{\lambda i}$ and the margin γ. If $K(\cdot,\cdot)$ is a radial basis function then $\|K(\cdot, t_i)\|^2_{\mathcal{H}_K} = K(0,0)$. Furthermore $\|K(\cdot, t_i) - K(\cdot, t_j)\|^2_{\mathcal{H}_K}$ is bounded above by $2K(0,0)$. If all members of the training set are classified correctly then $y_i f_i > 0$ and the sum following the 2 in (16.30) does not appear and $\hat{D}(\lambda) = K(0,0)/n\gamma^2$.

We note that Opper and Winther (Chapter 17) have obtained a different approximation for $f_{\lambda i} - f_{\lambda i}^{[-i]}$.

16.6 Numerical Results

We give two rather simple examples. For the first example, attribute vectors t were generated according to a uniform distribution on $\mathcal{T}$, the square depicted in Figure 16.2. The points outside the larger circle were randomly assigned $+1$ ("$+$") with probability $p_{[true]} = .95$ and -1 ("o") with probability .05. The points between the outer and inner circles were assigned $+1$ with probability $p_{[true]} = .50$, and the

points inside the inner circle were assigned +1 with probability $p_{[true]} = .05$. In this and the next example, $K(s,t) = e^{-\frac{1}{2\sigma^2}\|s-t\|^2}$, where σ is a tunable parameter to be chosen. Figure 16.3 gives a plot of $\log_{10}(GACV)$ of (16.29) and $\log_{10}(GCKL)$ of (16.10) as a function of $\log_{10}\lambda$, for $\log_{10}\sigma = -1$. Figure 16.4 gives the corresponding plot as a function of $\log_{10}\sigma$ for $\log_{10}\lambda = -2.5$, which was the minimizer of $\log_{10}(GACV)$ in Figure 16.3. Figure 16.5 shows the level curve for $f_\lambda = 0$ for $\log_{10}\lambda = -2.5$ and $\log_{10}\sigma = -1.0$, which was the minimizer of $\log_{10}(GACV)$ over the two plots. This can be compared to the theoretically optimal classifier, which the Neyman-Pearson Lemma says would be any curve between the inner and outer circles, where the theoretical log-odds ratio is 0. For the second example, Figure 16.6 corresponds to Figure 16.2, with $p_{[true]} = .95, .5$ and $.05$ respectively in the three regions, starting from the top. Figure 16.7 gives a plot of $\log_{10}(GACV)$ and $\log_{10}(GCKL)$ as a function of $\log_{10}\lambda$ for $\log_{10}\sigma = -1.25$ and Figure 16.8 gives $\log_{10}(GACV)$ and $\log_{10}(GCKL)$ as a function of $\log_{10}\sigma$ for $\log_{10}\lambda = -2.5$, which was the minimizer of Figure 16.7. Figure 16.9 gives the level curves for f_λ at 0 for $\log_{10}\lambda = -2.5$, $\log_{10}\sigma = -1.25$, which was the minimizer of $\log_{10}(GACV)$ over Figures 16.7 and 16.8. This can also be compared to the theoretically optimal classifier, which would be any curve falling between the two sine waves of Figure 16.7.

It can be seen that $\log_{10} GACV$ tracks $\log_{10} GCKL$ very well in Figures 16.3, 16.4, 16.7 and 16.8, more precisely, the minimizer of $\log_{10} GACV$ is a good estimate of the minimizer of $\log_{10} GCKL$.

A number of cross-sectional curves were plotted, first in $\log_{10}\lambda$ for a trial value of $\log_{10}\sigma$ and then in $\log_{10}\sigma$ for the minimizing value of $\log_{10}\lambda$ (in the GACV curve), and so forth, to get to the plots shown. A more serious effort to obtain the global minimizers over of $\log_{10}(GACV)$ over $\log_{10}\lambda$ and $\log_{10}\sigma$ is hard to do since both the $GACV$ and the $GCKL$ curves are quite rough. The curves have been obtained by evaluating the functions at increments on a log scale of .25 and joining the points by straight line segments. However, these curves (or surfaces) are not actually continuous, since they may have a jump (or tear) whenever the active constraint set changes. This is apparently a characteristic of generalized cross validation functions for constrained optimization problems when the solution is not a continuously differentiable function of the observations, see, for example [Wahba, 1982, Figure 7]. In practice, something reasonably close to the minimizer can be expected to be adequate.

Work is continuing on examining the $GACV$ and the $GCKL$ in more complex situations.

Acknowledgments

The authors thank Fangyu Gao and David Callan for important suggestions in this project. This work was partly supported by NSF under Grant DMS-9704758 and NIH under Grant R01 EY09946.

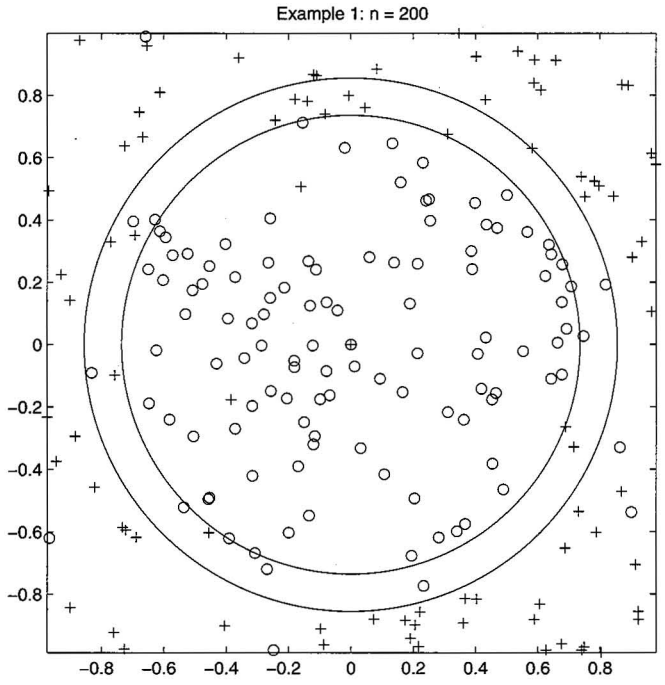

Figure 16.2 Data for Example 1, With Regions of Constant (Generating) Probability.

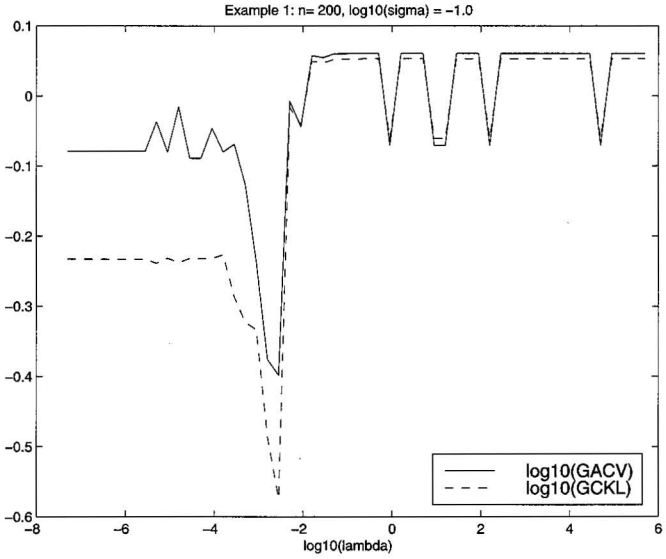

Figure 16.3 Plot of $\log_{10} GACV$ and $\log_{10} GCKL$ as a function of $\log_{10} \lambda$ for $\log_{10} \sigma = -1.0$.

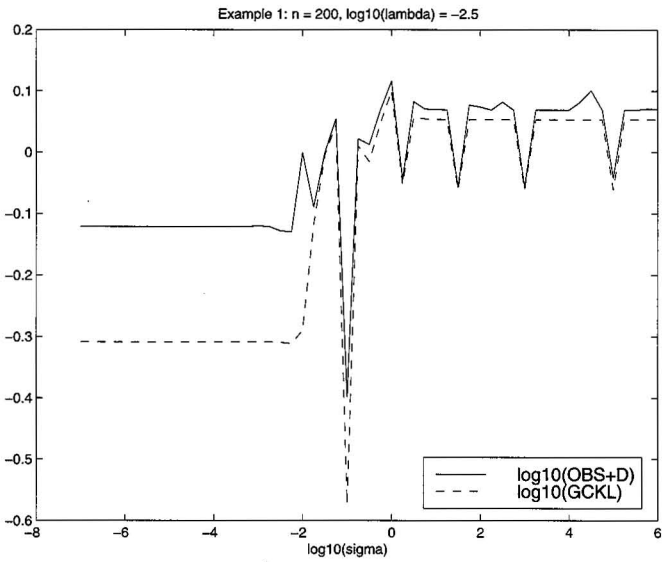

Figure 16.4 Plot of $\log_{10} GACV$ and $\log_{10} GCKL$ as a function of $\log_{10} \sigma$ for $\log_{10} \lambda = -2.5$.

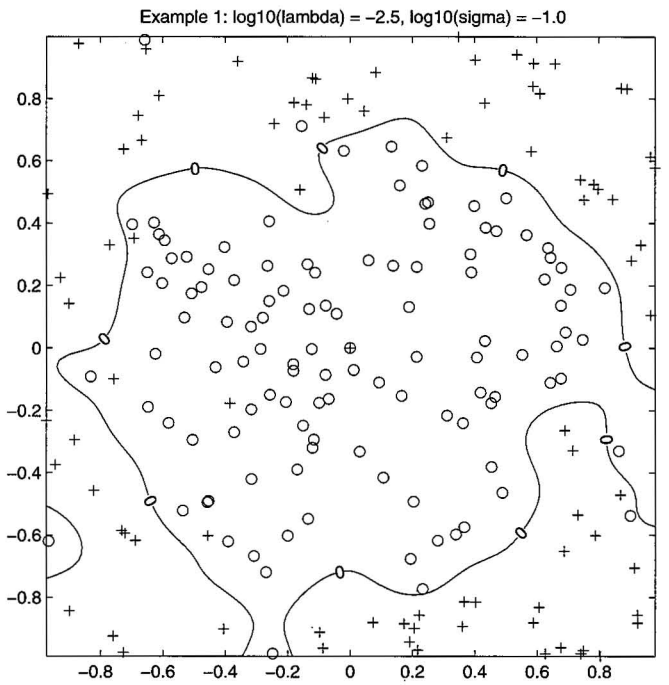

Figure 16.5 Level curve for $f_\lambda = 0$.

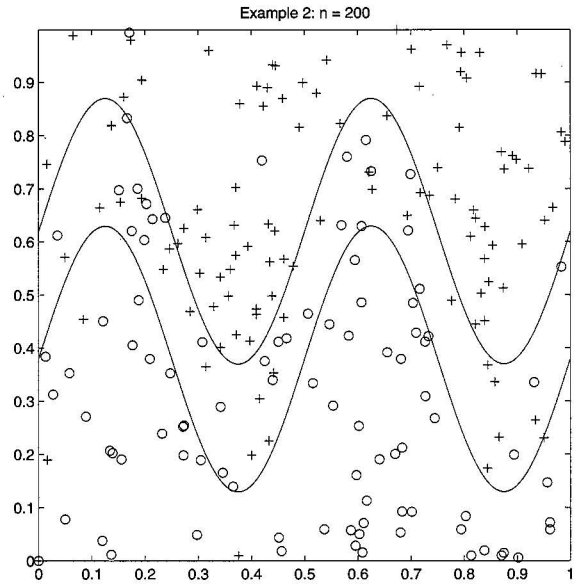

Figure 16.6 Data for Example 2, and Regions of Constant (Generating) Probability.

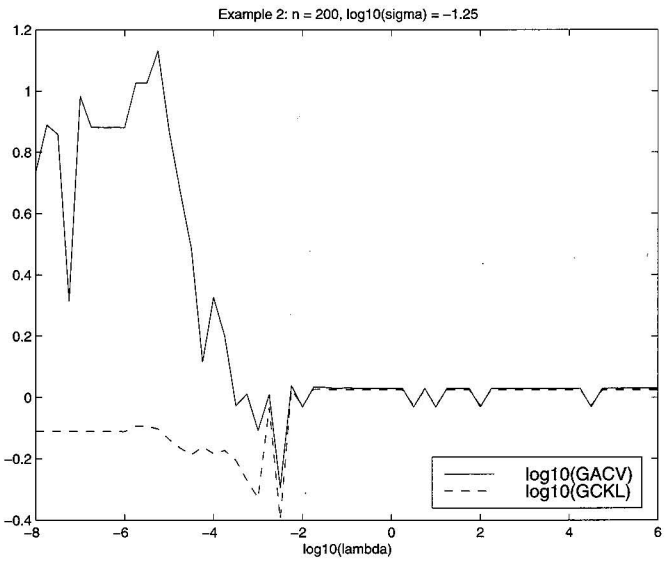

Figure 16.7 Plot of $\log_{10} GACV$ and $\log_{10} GCKL$ as a function of $\log_{10} \lambda$ for $\log_{10} \sigma = -1.25$.

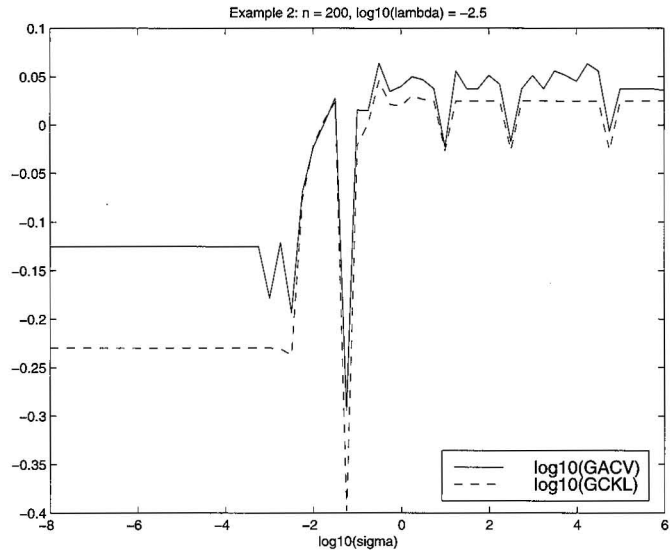

Figure 16.8 Plot of $\log_{10} GACV$ and $\log_{10} GCKL$ as a function of $\log_{10} \sigma$ for $\log_{10} \lambda = -2.5$.

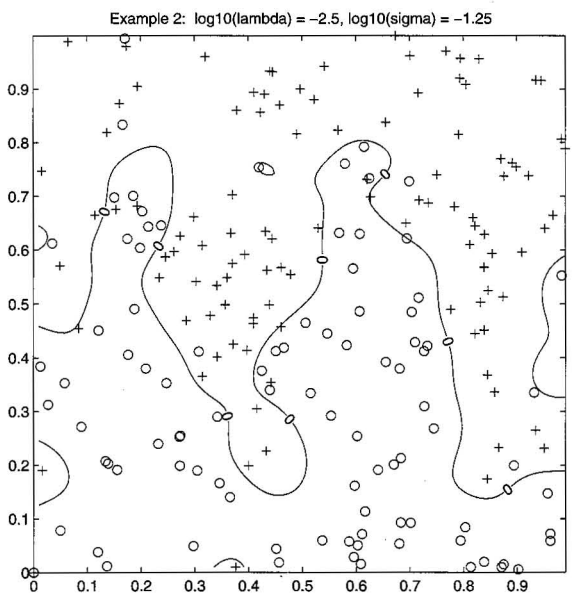

Figure 16.9 Level curve for $f_\lambda = 0$.