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

TECHNICAL REPORT NO. 1006
April 6, 1999

GENERALIZED APPROXIMATE CROSS VALIDATION FOR SUPPORT VECTOR MACHINES, OR, ANOTHER WAY TO LOOK AT MARGIN-LIKE QUANTITIES

Grace Wahba, Yi Lin and Hao Zhang

wahba,yilin,hzhang@stat.wisc.edu

http://stat.wisc.edu/~wahba, ~yilin, ~hzhang

[1]This paper was the basis of a talk at the NIPS*98 Workshop on Large Margin Classifiers, Breckenridge CO, December 5, 1998 and is submitted for 'Advances in Large Margin Classifiers', A. Smola, P. Bartlett, B. Scholkopf and D. Schurmans, Eds. This replaces an earlier version of TR1006 which was posted Feb 25, 1999. We have fixed some typos and added Section 6 to the original version. Partly supported by NSF Grant DMS9704758 and NIH Grant EY09946.

Generalized Approximate Cross Validation for Support Vector Machines, or, Another Way to Look at Margin-Like Quantities [2]

Grace Wahba, Yi Lin and Hao Zhang

wahba,yilin,hzhang@stat.wisc.edu

http://stat.wisc.edu/~wahba, ~yilin, ~hzhang

April 6, 1999

1 Introduction

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 & Wahba (1971), Wahba (1990), Girosi (1997), Poggio & Girosi (1998), the papers and references in Schoelkopf, Burges & Smola (1999), and elsewhere. In this note, which is a sequel to Wahba (1999), we look at the SVM paradigm from the point of view of a regularization problem, which allows a comparison with penalized 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.

We first review the steps connecting the SVM paradigm in RKHS and its connection to 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 as a proxy for the GCKL proposed in Wahba (1999) 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 (1999), 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.

2 The SVM variational problem

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 K. See Wahba (1990), Wahba (1999), Lin, Wahba, Xiang, Gao, Klein & Klein (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

[2]Corresponding author address: Prof. Grace Wahba, Department of Statistics, University of Wisconsin, 1210 W. Dayton St., Madison WI 53706. Research supported in part by NIH Grant EY09946 and NSF Grant DMS9704758.

http://svm.first.gmd.de/. Recently Poggio & Girosi (1998) interestingly observed how different scales may be accomodated 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

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

here $f_i = f(t_i)$, and $(\tau)_+ = \tau, \tau > 0; = 0$ otherwise. 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 (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 $e = (1, \cdots, 1)', c = (c_1, \cdots, c_n)', (f(t_1), \cdots f(t_n))' = (f_1, \cdots, f_n)'$, 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

$$f = Kc + ed \tag{2}$$

and the variational problem (1) becomes: find (c, d) to minimize

$$\frac{1}{n}\sum_{i=1}^{n}(1 - y_i f_i)_+ + \lambda c' K c. \tag{3}$$

3 The Dual Problem

The primal problem (3) is equivalent to the following quadratic programming problem by introducing a new vector $z = (z_1, \cdots, z_n)'$

$$min_{z,c,d} e'z + n\lambda c' K c, \ subject\ to \begin{cases} 0 \le z \\ e - YKc - Yed \le z \end{cases}$$

where Y is the $n \times n$ diagonal matrix with y_i in the iith position. The original problem is sometimes more ill-conditioned than the dual problem. The dual problem is the one typically solved in the SVM literature, and our arguments involve the dual form. We now obtain the dual form of our problem. Introducing two new vectors $\alpha = (\alpha_1, \cdots, \alpha_n)'$ and $r = (r_1, \cdots, r_n)'$, we have

$$max_{c,d,z,\alpha,r} L(c, d, z, \alpha, r) = e'z + n\lambda c' K c - \sum_{i=1}^{n} r_i z_i + \sum_{i=1}^{n} \alpha_i (1 - y_i f_i - z_i)$$

$$\text{subject to} \begin{cases} \frac{\partial L}{\partial c} = 0 \\ \frac{\partial L}{\partial d} = 0 \\ \frac{\partial L}{\partial z} = 0 \\ 0 \le \alpha \\ 0 \le r \end{cases}$$

Letting $y = (y_1, \cdots, y_n)'$, we get the matrix form of L as follows:

$$L = e'z + n\lambda c'Kc - r'z + e'\alpha - \alpha'Y(Kc + ed) - \alpha'z$$

By differentiation, we have the following equations:

$$\frac{\partial L}{\partial c} = 2n\lambda Kc - KY\alpha = 0, \tag{4}$$

which gives

$$c = \frac{1}{2n\lambda}K^{-1}KY\alpha = \frac{1}{2n\lambda}Y\alpha \tag{5}$$

$$\frac{\partial L}{\partial d} = -e'Y\alpha = 0, \tag{6}$$

and

$$\frac{\partial L}{\partial z} = e - \alpha - r = 0, \tag{7}$$

Finally, letting $H = \frac{1}{2n\lambda}YKY$, we have

$$maxL = -\frac{1}{2}\alpha' H\alpha + e'\alpha \tag{8}$$

$$\text{subject to } \begin{cases} 0 \leq \alpha \leq 1 \\ e'Y\alpha = y'\alpha = 0 \end{cases}$$

this being the usual form in which the SVM is computed.

MINOS or other optimization routine can be used to find α, and then (5) gives c. The support vectors are those $K(\cdot, t_i)$ for which $\alpha_i \neq 0$. d can be found from any of the support vectors for which $0 < \alpha_i < 1$. As we know, the Kuhn-Tucker conditions are satisfied by the solutions:

$$(1 - \alpha_i)z_i = 0 \tag{9}$$

$$\alpha_i(1 - y_i f_i - z_i) = 0 \tag{10}$$

where $f(t_i) \equiv f_i = \sum_{j=1}^n c_j K(t_i, t_j) + d$. Thus $z_i = 0$ from (9) as long as $0 < \alpha_i < 1$ for some i. By (10) $1 - y_i f_i = 0$ implies that $d = [1 - y_i(\sum_{j=1}^n c_j K(t_i, t_j))]/y_i$ which implies that $d = 1/y_i - \sum_{j=1}^n c_j K(t_i, t_j)$.

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

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

See Cortes & Vapnik (1995), Bartlett & 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 $\gamma = \frac{1}{\|h\|_{\mathcal{H}_\mathcal{K}}} = (c'Kc)^{-1/2}$. This equality 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:

$$maxL = -\frac{1}{2}\alpha' H\alpha + e'\alpha$$

subject to $\alpha_i \geq 0$ and $y'\alpha = 0$.

Introducing the Lagrangian multipliers $\xi = (\xi_1, \cdots, \xi_n)'$ 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

$$\begin{aligned} c'Kc &= \frac{1}{2n\lambda}\alpha'_\lambda H \alpha_\lambda \\ &= \frac{1}{2n\lambda}\left[\alpha'_\lambda e - \beta \alpha'_\lambda y - \alpha'_\lambda \xi\right] \\ &= \frac{1}{2n\lambda}\left[\alpha'_\lambda e\right] \end{aligned}$$

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

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

4 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

$$GCKL(p_{true}, f_\lambda) \doteq GCKL(\lambda) = E_{true}\frac{1}{n}\sum_{j=1}^{n} g(y_j f_{\lambda j}). \tag{11}$$

Here f_λ is considered fixed and the expectation is taken over future, unobserved y_j. If $g(\tau) = ln(1 + e^{-\tau})$, then $GCKL(\lambda)$ reduces to the usual CKL for Bernoulli data [3] averaged over the

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

attribute vectors of the training set. More details may be found in Wahba (1999). If $g(\tau) = [-\tau]_*$, then

$$\begin{aligned} E_{true}[-y_j f_{\lambda j}]_* &= p_{[true]j}[-f_{\lambda j}]_* + (1 - p_{[true]j})[f_{\lambda j}]_* && (12) \\ &= p_{[true]j}, \quad f_{\lambda j} < 0 && (13) \\ &= (1 - p_{[true]j}), \quad f_{\lambda j} > 0, && (14) \end{aligned}$$

where $p_{[true]j} = p_{[true]}(t_j)$, 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

$$\begin{aligned} E_{true}(1 - y_j f_{\lambda j})_+ &= p_{[true]j}(1 - f_{\lambda j}), \quad f_{\lambda j} < -1 && (15) \\ &= 1 + (1 - 2p_{[true]j}) f_{\lambda j}, \quad -1 \le f_{\lambda j} \le 1 && (16) \\ &= (1 - p_{[true]j})(1 + f_{\lambda j}), \quad f_{\lambda j} > 1. && (17) \end{aligned}$$

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 1.

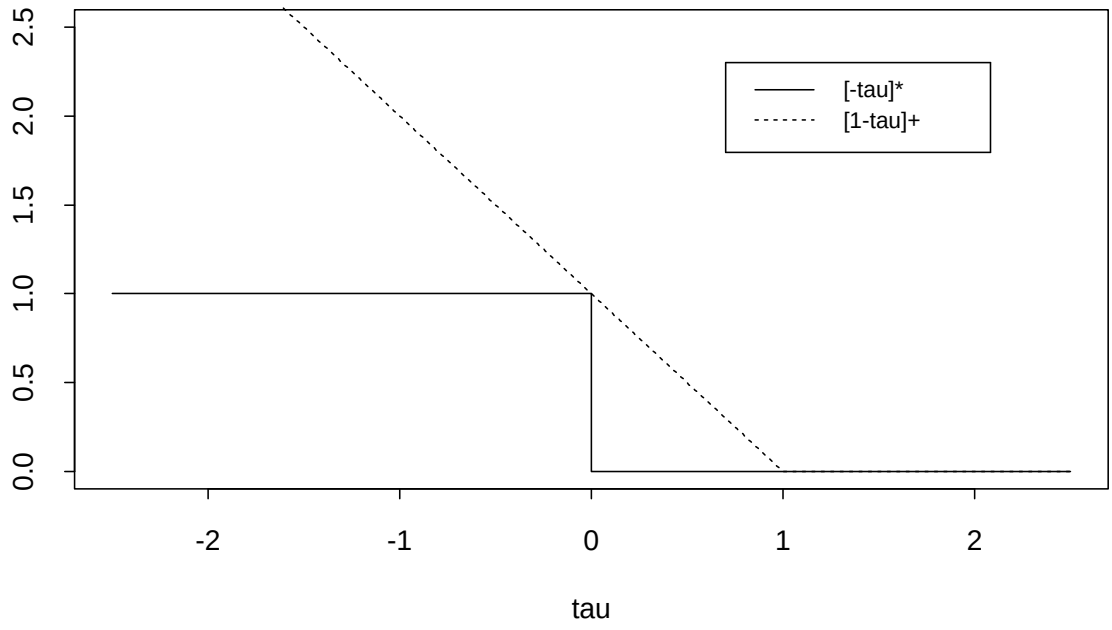

Figure 1: $g(\tau) = (1 - \tau)_+$ and $g(\tau) = [-\tau]_*$ compared.

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, Campbell

& Shawe-Taylor (1998), Kearns, Ng, Mansour & Ron (to appear), surely not a complete list. Important references in the statistics literature that are related include Efron & Tibshirani (1997), Ye & Wong (1997). X. 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.

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

$$\frac{1}{n}\sum_{\substack{j=1\\ j\neq i}}^{n} g(y_j f_j) + \lambda \|h\|^2_{\mathcal{H}_K} \tag{18}$$

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

$$V_0(\lambda) = \frac{1}{n}\sum_{i=1}^{n} g(y_i f_{\lambda i}^{[-i]}). \tag{19}$$

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

$$V_0(\lambda) = OBS(\lambda) + D(\lambda), \tag{20}$$

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

$$OBS(\lambda) = \frac{1}{n}\sum_{i=1}^{n} g(y_i f_{\lambda i}) \tag{21}$$

and

$$D(\lambda) = \frac{1}{n}\sum_{i=1}^{n} [g(y_i f_{\lambda i}^{[-i]}) - g(y_i f_{\lambda i})]. \tag{22}$$

Using a first order Taylor series expansion gives

$$D(\lambda) \approx -\frac{1}{n}\sum_{i=1}^{n} \frac{\partial g}{\partial f_{\lambda i}}(f_{\lambda i} - f_{\lambda i}^{[-i]}). \tag{23}$$

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

$$\mu_i = \mu(f_i) = \sum_{y\in\{+1,-1\}} \frac{\partial}{\partial f_i} g(y_i f_i). \tag{24}$$

When $g(\tau) = ln(1 + e^{-\tau})$ then $\mu(f) = 2p - 1 = E\{y|p\}$. For $g(\tau) = (1-\tau)_+, \mu(f) = -1, f < -1; \mu(f) = 0, -1 \leq f \leq 1$ and $\mu(f) = 1$ for $f > 1$.

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),

$$D(\lambda) \approx -\frac{1}{n}\sum_{i=1}^{n} \frac{\partial g}{\partial f_{\lambda i}} \frac{(f_{\lambda i} - f_{\lambda i}^{[-i]})}{(y_i - \mu_{\lambda i}^{[-i]})}(y_i - \mu_{\lambda i}^{[-i]}) \tag{25}$$

This is equation (1.40) in Wahba (1999). We now provide somewhat different arguments than in Wahba (1999) 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 1 [4] 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 (1999) 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 aproximation

$$\frac{f_{\lambda i}-f_{\lambda i}^{[-i]}}{y_i-\mu_{\lambda i}^{[-i]}}\equiv\frac{f_{\lambda i}[y_i]-f_{\lambda i}[x]}{y_i-x}\approx\frac{\partial f_{\lambda i}}{\partial y_i}. \tag{26}$$

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 (25) by

$$D(\lambda)\approx-\frac{1}{n}\sum_{i=1}^{n}\frac{\partial g}{\partial f_{\lambda i}}\frac{\partial f_{\lambda i}}{\partial y_i}(y_i-\mu_{\lambda i}). \tag{27}$$

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

$$D(\lambda)\approx\frac{1}{n}\sum_{y_i f_{\lambda i}<-1}2\frac{\partial f_{\lambda i}}{\partial y_i}+\frac{1}{n}\sum_{y_i f_{\lambda i}\in[-1,1]}\frac{\partial f_{\lambda i}}{\partial y_i}. \tag{28}$$

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

$$\frac{\partial f_{\lambda i}}{\partial y_i}=K(t_i,t_i)\frac{\alpha_{\lambda i}}{2n\lambda}\equiv\|K(\cdot,t_i)\|^2_{\mathcal{H}_K}\frac{\alpha_{\lambda i}}{2n\lambda}. \tag{29}$$

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

$$GACV(\lambda)=\frac{1}{n}\sum_{i=1}^{n}(1-y_i f_{\lambda i})_+ +\hat{D}(\lambda), \tag{30}$$

where

$$\hat{D}(\lambda)=\frac{1}{n}\left[2\sum_{y_i f_{\lambda i}<-1}\frac{\alpha_{\lambda i}}{2n\lambda}\cdot\|K(\cdot,t_i)\|^2_{\mathcal{H}_K}+\sum_{y_i f_{\lambda i}\in[-1,1]}\frac{\alpha_{\lambda i}}{2n\lambda}\cdot\|K(\cdot,t_i)\|^2_{\mathcal{H}_K}\right]. \tag{31}$$

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 (31) does not appear and $\hat{D}(\lambda)=K(0,0)/n\gamma^2$.

We note that Opper & Winther (1999) have obtained a different approximation for $f_{\lambda i}-f_{\lambda i}^{[-i]}$.

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

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 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 3 gives a plot of $log(GACV)$ of (30) and $log(GCKL)$ of (11) as a function of $log\lambda$, for $log\sigma = -1$. Figure 4 gives the corresponding plot as a function of $log\sigma$ for $log\lambda = -2.5$, which was the minimizer of $log10(GACV)$ in Figure 3. Figure 5 shows the level curves for $f_\lambda = 0$ for $log\lambda = -2.5$ and $log\sigma = -1.0$, which was the minimizer of $log(GACV)$ over the two plots. This can be compared to the theoretically optimal classifier, which according to the Neyman-Pearson Lemma would be any curve between the inner and outer circles, where the theoretical log-odds ratio is 0. For the second example, Figure 6 corresponds to Figure 2, with $p_{[true]} = .95, .5$ and $.05$ respectively in the three regions, starting from the top. Figure 7 gives a plot of $log(GACV)$ and $log(GCKL)$ as a function of $log\lambda$ for $log\sigma = -1.25$. and Figure 8 gives $log(GACV)$ and $log(GCKL)$ as a function of $log\sigma$ for $log\lambda = -2.5$, which was the minimizer of Figure 7. Figure 9 gives the level curves for f_λ at 0 for $log\lambda = -2.5$, $log\sigma = -1.25$, which was the minimizer of $log(GACV)$ over Figures 7 and 8. This can also be compared to the theoretically optimal classifier, which would be any curve falling between the two sine waves of Figure 7.

It can be seen that $log_{10}GACV$ tracks $log_{10}GCKL$ very well in Figures 3, 4, 7and 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 first in $log\lambda$ for a trial value of $log\sigma$ and then in $log\sigma$ for the minimizing value of $log\lambda$ (in the GACV curve), and so forth, to get to the plots shown. A more serious effort to obtain the global minimizers over of $log(GACV)$ over $log\lambda$ and $log\sigma$ 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 Figure 7 of Wahba (1982). 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.

7 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.

References

Bartlett, P. & Shawe-Taylor, J. (1999), Generalization performance of support vector machines and other pattern classifiers, *in* B. Schoelkopf, C. Burges & A. Smola, eds, 'Advances in Kernel Methods-Support Vector Learning', MIT Press, pp. 43–54.

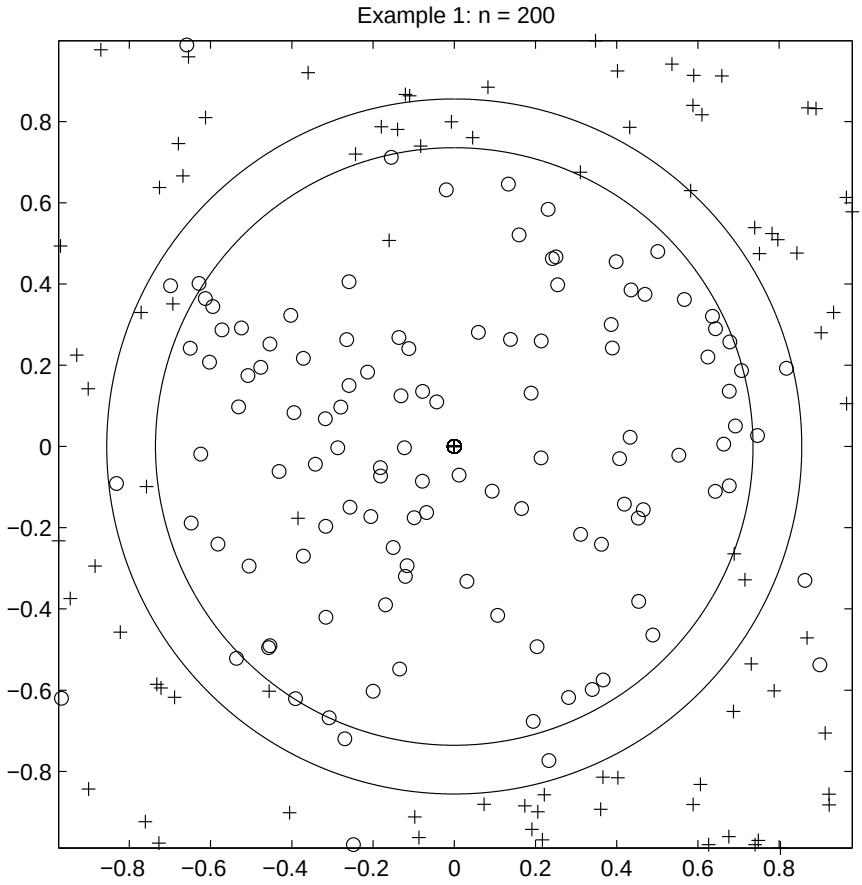

Figure 2: Data for Example 1, With Regions of Constant (Generating) Probability.

Burges, C. (1998), 'A tutorial on support vector machines for pattern recognition', *Data Mining and Knowledge Discovery* **2**, 121–167.

Cortes, C. & Vapnik, V. (1995), 'Support vector networks', *Machine Learning* **20**, 1–25.

Cristianini, N., Campbell, C. & Shawe-Taylor, J. (1998), Dynamically adapting kernels in support vector machines, Technical Report NC2-TR-1998-017, Royal Holloway University of London, Surrey England.

Efron, B. & Tibshirani, R. (1997), 'Improvements on cross-validation: the .632+ bootstrap method', *J. Amer. Statist. Assoc* **92**, 548–560.

Girosi, F. (1997), An equivalence between sparse approximation and support vector machines, Technical Report A. I. 1606, MIT artificial Intelligence Laboratory, Boston MA.

Kearns, M., Ng, A., Mansour, Y. & Ron, D. (to appear), 'An experimantal and theoretical comparison of model selection methods', *Machine Learning* **xx**, xx.

Kimeldorf, G. & Wahba, G. (1971), 'Some results on Tchebycheffian spline functions', *J. Math. Anal. Applic.* **33**, 82–95.

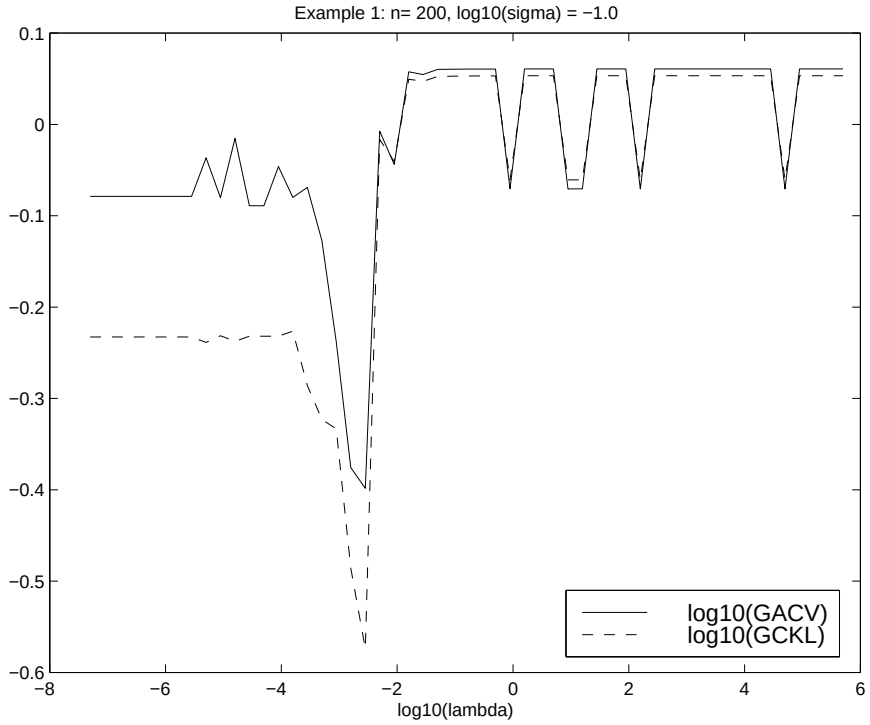

Figure 3: Plot of $log_{10}GACV$ and $log_{10}GCKL$ as a function of $log_{10}\lambda$ for $log_{10}\sigma = -1.0$.

Lin, X., Wahba, G., Xiang, D., Gao, F., Klein, R. & Klein, B. (1998), Smoothing spline ANOVA models for large data sets with Bernoulli observations and the randomized GACV, Technical Report 998, Department of Statistics, University of Wisconsin, Madison WI.

Opper, M. & Winther, O. (1999), Gaussian process classification and SVM: Mean field results and leave-out-one estimator, *in* 'Advances in Large Margin Classifiers', MIT Press.

Poggio, T. & Girosi, F. (1998), 'A sparse representation for function approximation', *Neural Computation* **10**, 1445–1454.

Schoelkopf, B., Burges, C. & Smola, A. (1999), *Advances in Kernel Methods-Support Vector Learning*, MIT Press.

Wahba, G. (1982), 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.

Wahba, G. (1990), *Spline Models for Observational Data*, SIAM. CBMS-NSF Regional Conference Series in Applied Mathematics, v. 59.

Wahba, G. (1999), Support vector machines, reproducing kernel Hilbert spaces and the randomized GACV, *in* B. Schoelkopf, C. Burges & A. Smola, eds, 'Advances in Kernel Methods-Support Vector Learning', MIT Press, pp. 69–88.

Ye, J. & Wong, W. (1997), 'Evaluation of highly complex modeling procedures with Binomial and Poisson data', manuscript, University of Chicago School of Business.

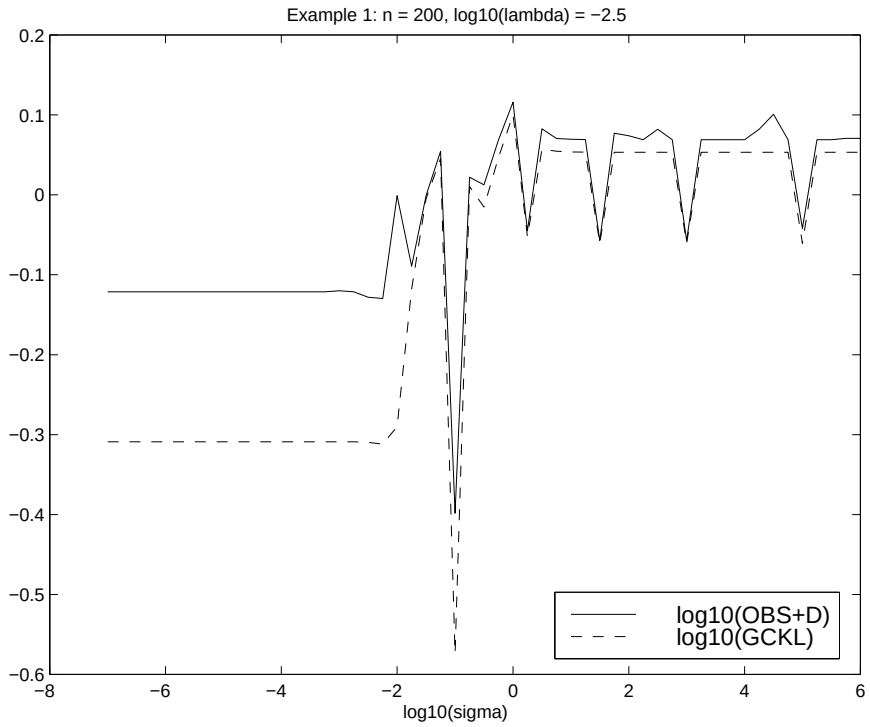

Figure 4: Plot of $log_{10}GACV$ and $log_{10}GCKL$ as a function of $log_{10}\sigma$ for $log_{10}\lambda = -2.5$.

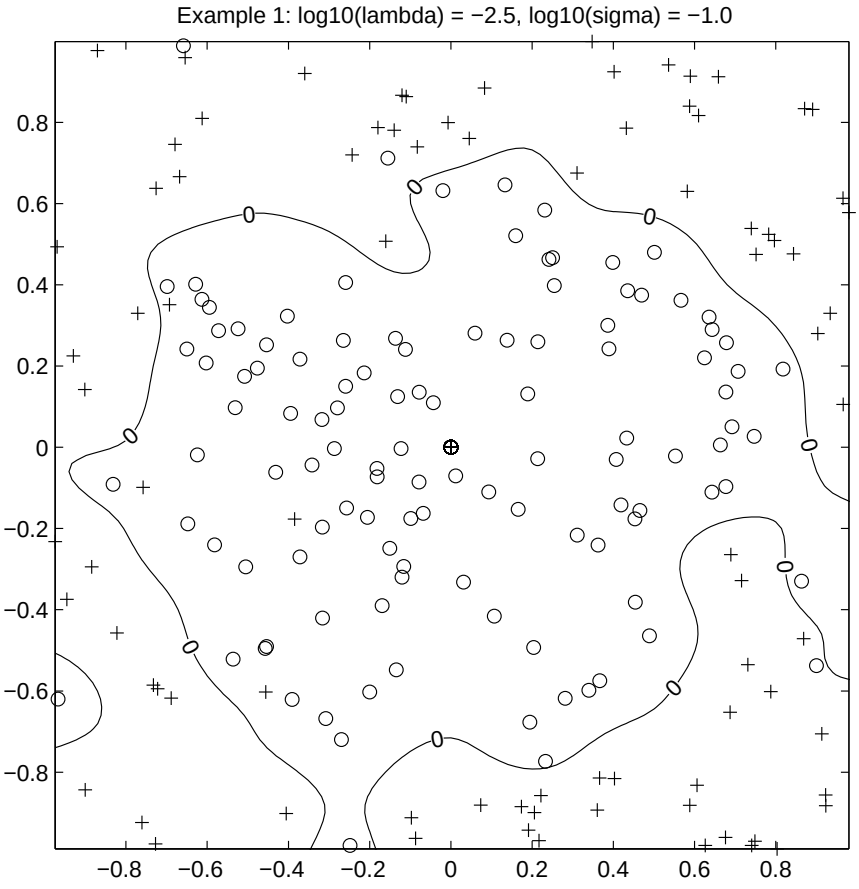

Figure 5: Level curve for $f_\lambda = 0$.

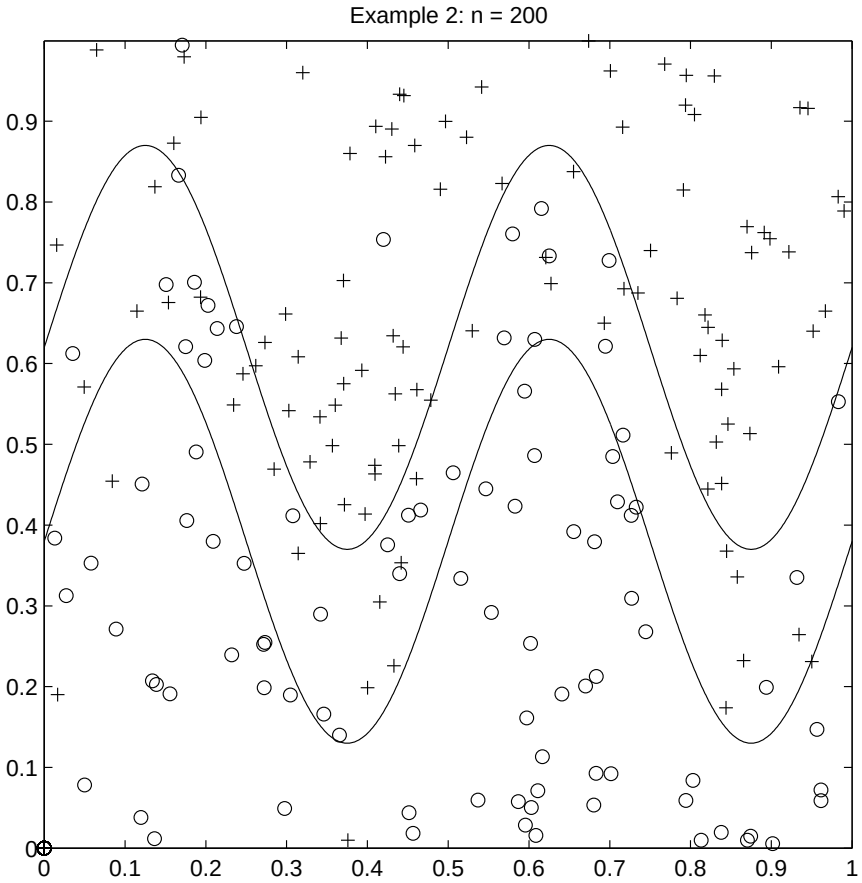

Figure 6: Data for Example 2, and Regions of Constant (Generating) Probability.

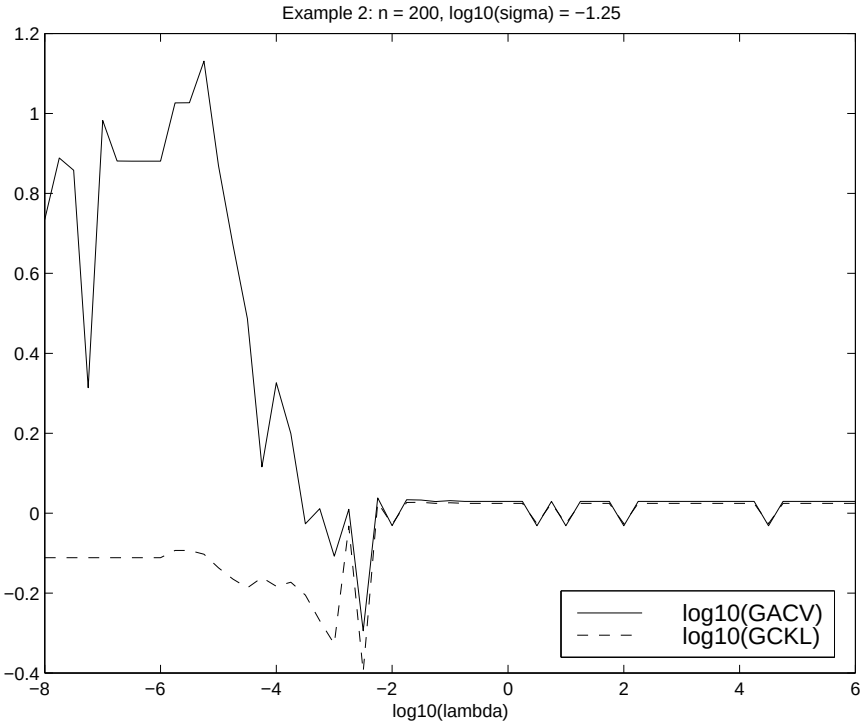

Figure 7: Plot of $log_{10}GACV$ and $log_{10}GCKL$ as a function of $log_{10}\lambda$ for $log_{10}\sigma = -1.25$.

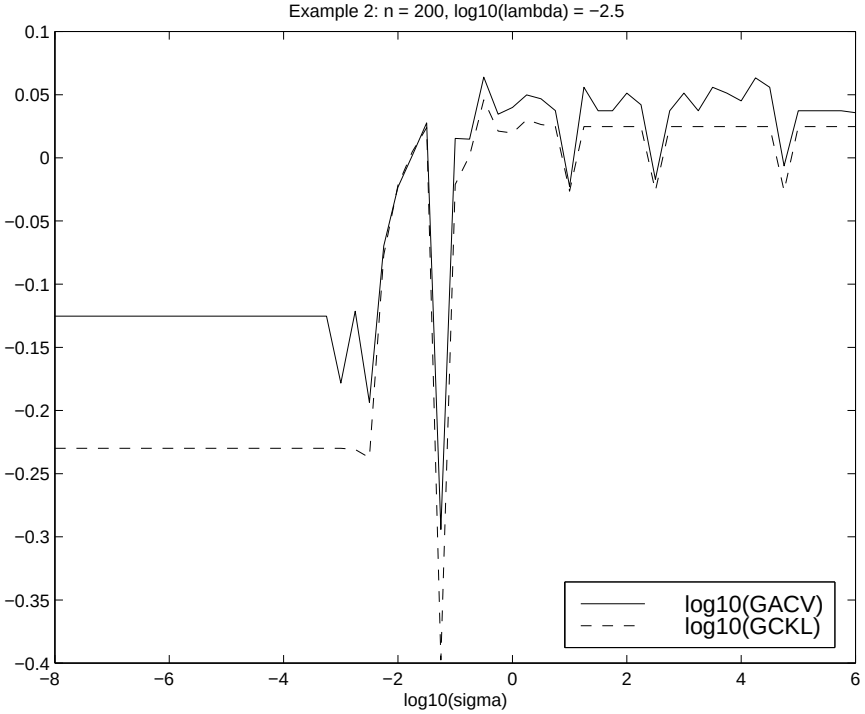

Figure 8: Plot of $log_{10}GACV$ and $log_{10}GCKL$ as a function of $log_{10}\sigma$ for $log_{10}\lambda = -2.5$.

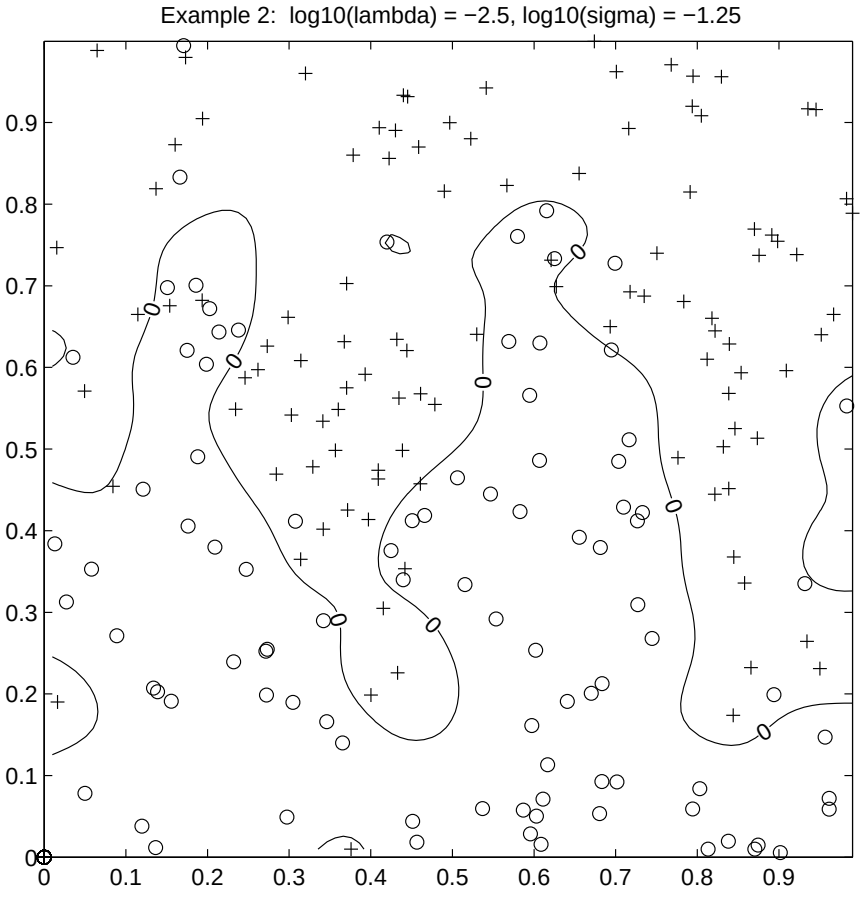

Figure 9: Level curve for $f_\lambda = 0$.