♣♣ Tuning the SVM (part 1).

Recall that the penalized log likelihood estimate was tuned by a criteria which chose λ to minimize a proxy for

$$R(\lambda) = E\frac{1}{n}\sum_{i=1}^{n} -y_{new\cdot i} f_\lambda(x_i) + \log(1 + e^{f(x_i)}).$$

$R(\lambda)$ is the expected 'distance' or negative log likelihood for a new observation with the same x_i. For the SVM classifier we will say that it is optimally tuned if we have a criteria which chooses λ to minimize a proxy for

$$R(\lambda) = E\frac{1}{n}\sum_{i=1}^{n} (1 - y_{new\cdot i} f_\lambda(x_i))_+.$$

That is, it is choosing λ (and possibly other parameters in K) to minimize a proxy for an upper bound on the misclassification rate.

♣♣ Yi Lin's Lemma [Lin99]:

The minimizer of $E(1 - y_{new} f(x))_+$ is sign $f(x)$

$= \text{sign}\ (p(x) - \frac{1}{2})$

where $f(x) = \log p(x)/(1 - p(x))$.

AS A CONSEQUENCE: Find $f_\lambda = d + h$ which minimizes

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

where λ is chosen to minimize (a proxy for) $R(\lambda)$,

is estimating sign $f(x)$ – EXACTLY WHAT YOU NEED

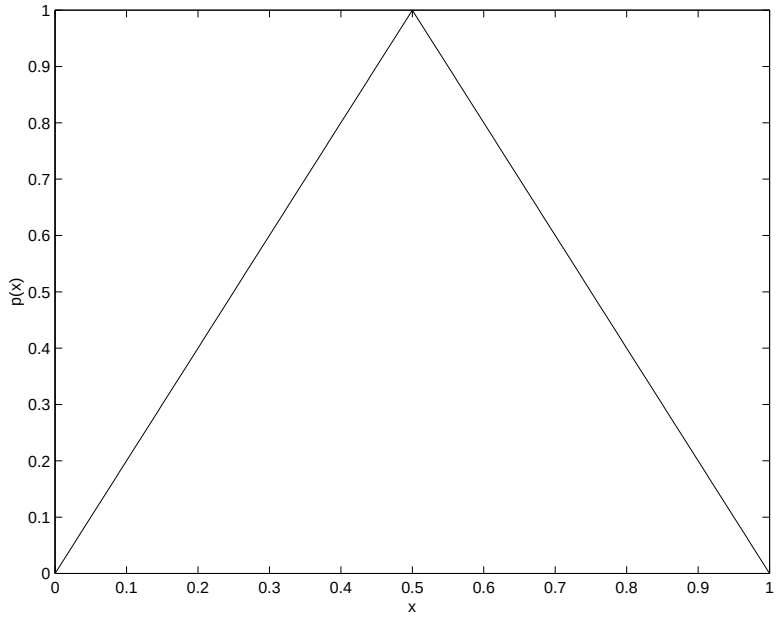
to minimize the misclassification rate!

$$R(\lambda) = E\frac{1}{n}\sum_{i=1}^{n}(1 - y_{new \cdot i} f_\lambda(x_i))_+.$$

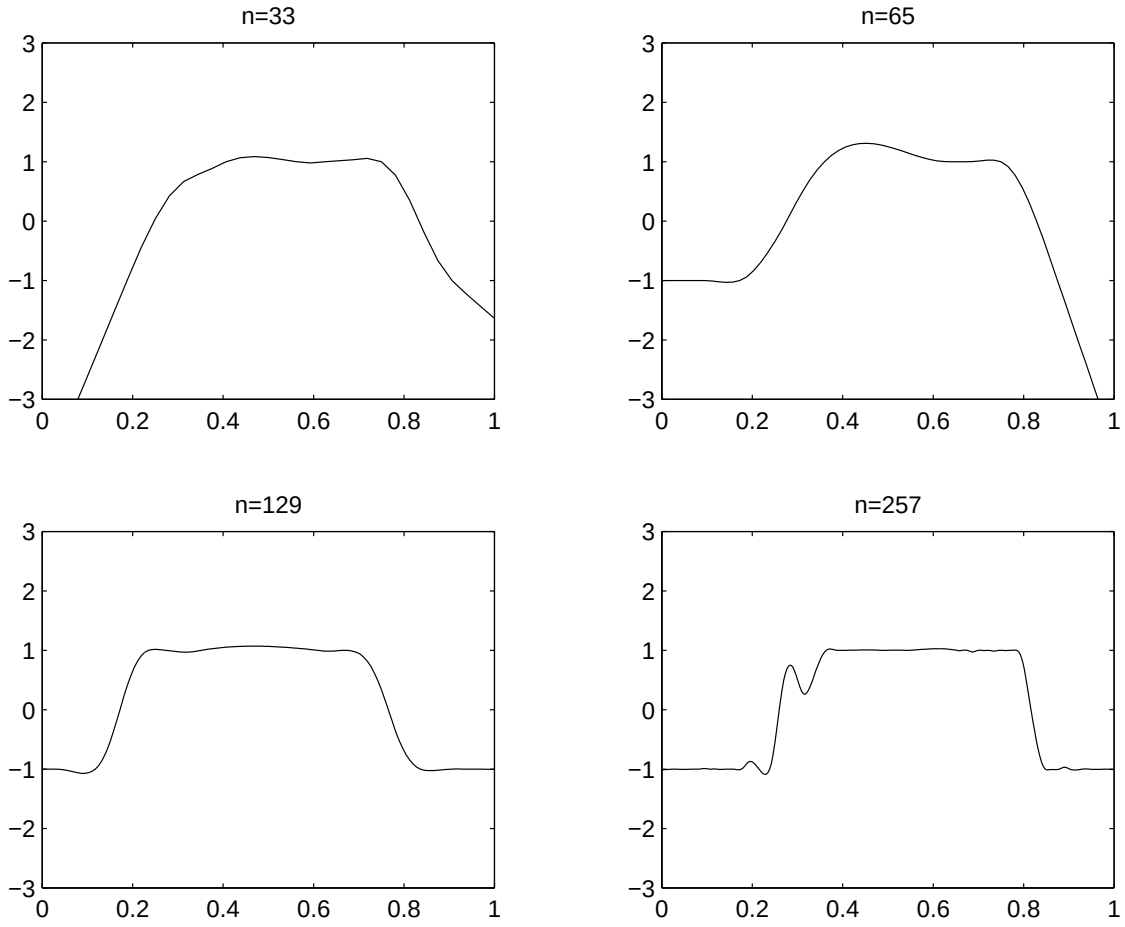
$$E(1 - y_{new} f_\lambda) =$$
$$\left\{ \begin{array}{lll} p(1 - f_\lambda), & & f_\lambda < -1 \\ p(1 - f_\lambda) + (1 - p)(1 + f_\lambda), & -1 < & f_\lambda < +1 \\ (1 - p)(1 + f_\lambda), & & f_\lambda > +1. \end{array} \right\}$$

$R(\lambda)$ is also known as the $GCKL(\lambda)$, the Generalized Kullback-Leibler Distance). Since the true p is only known in a simulation experiment, $GCKL$ is also only known in experiments.

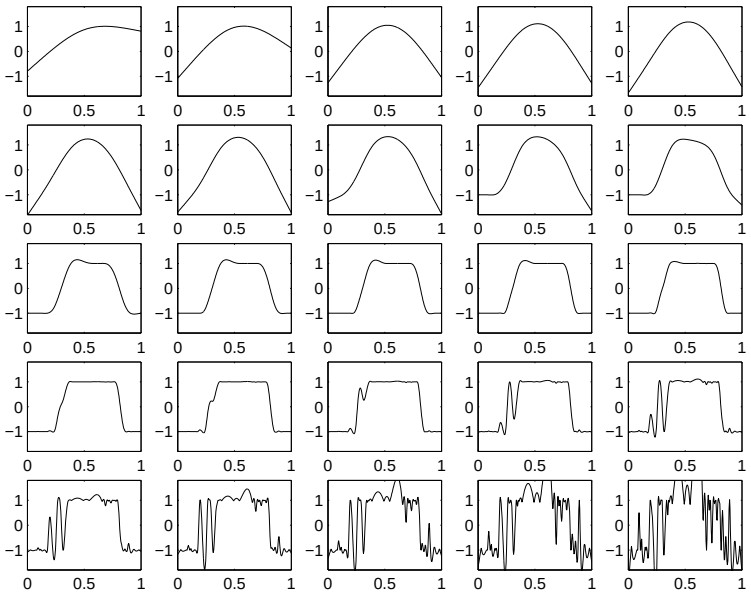
Experiments to follow are courtesy Yi Lin, reprinted from [Lin99].



From [Lin99]. The underlying conditional probability function $p(x) = Prob\{y = 1|x\}$ in our simulation. The function sign $[p(x) - 1/2]$ is 1, for $0.25 < x < 0.75$; -1 otherwise.



From [Lin99]. SVM estimates, Sobolev Hilbert space kernel (spline kernel), for samples of size 33, 65, 129, 257. The training set is generated using p from the preceeding slide and the x_i equally spaced on $[0, 1]$. The tuning parameter λ is chosen to minimize the $GCKL$ in each case. Note that as the sample size becomes larger, the curve becomes more like the step function sign $(p(x) - 1/2)$.



From [Lin99] For the same $n = 257$ sample as in the preceeding figure- the solutions to the SVM regularization $n\lambda = 2^{-1}, 2^{-2}, \ldots, 2^{-25}$, left to right starting with the top row. . We see that solution is close to *sign*$[p(x) - 1/2]$ when $n\lambda$ is in the neighborhood of 2^{-18}. 2^{-18} was the minimizer of the $GCKL$, suggesting that it is necessary to tune the SVM to estimate sign $(p - 1/2)$ well.

♣♣. Expected Misclassification Rate:

Let $\eta(x)$ be any classification function-

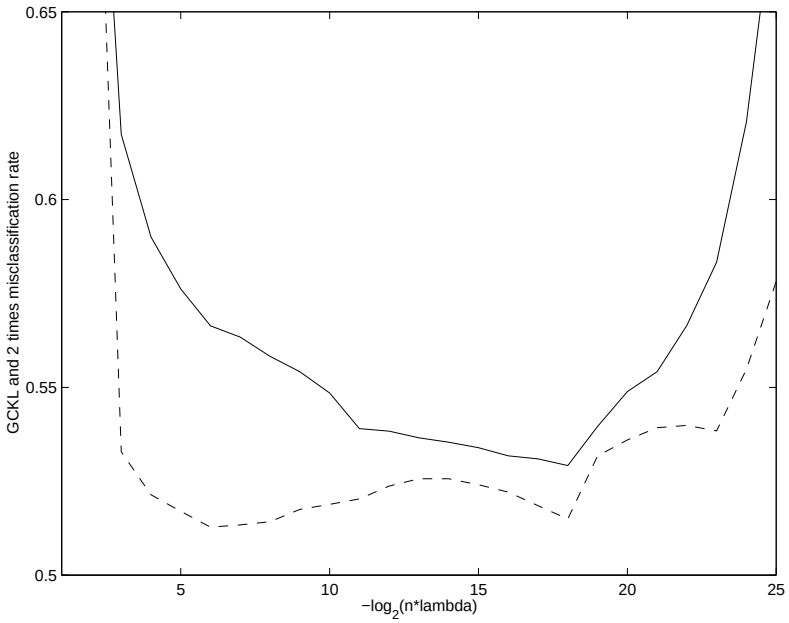
$$\eta(x) \rightarrow \begin{matrix} +1 \\ -1 \end{matrix}.$$

Then $E(1 - y\eta(x))_+$

$= \quad 2 \times$ expected misclass. rate

$= \quad p(x)(1 - \eta(x)) + (1 - p(x))(1 + \eta(x))$

$= \quad [Pr\ y = 1|\eta(x) = -1] + [Pr\ y = -1|\eta(x) = +1].$



From [L in99] GCKL (solid line) and $2 \times$ misclass. rate (dashed line) as a function of λ for the same sample with $n = 257$ as before. Larger values of λ correspond to the points on the left. The minimizer of the $GACV$ is at about 2^{-18}.

♣♣ The GACV for choosing λ (and other parameters in K):

Want a proxy for the (unobservable)

$$R(\lambda) = E\frac{1}{n}\sum_{i=1}^{n}(1 - y_{new.i}\ f_\lambda(x_i))_+.$$

Start with leaving-out-one. Let $f_\lambda^{[-k]}$ be the minimizer of the form $f = d + h$ with $h \in \mathcal{H}_K$ to min

$$\frac{1}{n}\sum_{\substack{i=1\\ i\neq k}}(1 - f(x_i))_+ + \lambda\|h\|_K^2.$$

Let

$$V_0(\lambda) = \frac{1}{n}\sum_{k=1}^{n}(1 - y_k f_\lambda^{[-k]}(x_k))_+.$$

♣♣ The GACV for choosing λ (and other parameters in K). (continued)

Let

$$V_0(\lambda) \equiv \frac{1}{n}\sum_{k=1}^{n}(1 - y_k f_\lambda^{[-k]}(x_k))_+$$

$$\equiv \mathrm{OBS}(\lambda) + D(\lambda),$$

where

$$\mathrm{OBS}(\lambda) = \frac{1}{n}\sum_{k=1}^{n}(1 - y_k f_\lambda(x_k))_+.$$

[WLZ00] showed that

$$D(\lambda) \approx \hat{D}(\lambda)$$

where $\hat{D}(\lambda)$ is given by

$$\frac{1}{n}\left[\sum_{y_i f_\lambda(x_i)<1} 2\frac{\partial f_\lambda(x_i)}{\partial y_i} + \sum_{y_i f_\lambda(x_i)\in[-1,1]} \frac{\partial f_\lambda(x_i)}{\partial y_i}\right].$$

♣♣ The GACV for choosing λ (and other parameters in K). (continued)

$$V_0(\lambda) \approx \mathrm{OBS}(\lambda) + \hat{D}(\lambda),$$

where $\hat{D}(\lambda)$ is given by

$$\frac{1}{n}\left[\sum_{y_i f_\lambda(x_i)<1} 2\frac{\partial f_\lambda(x_i)}{\partial y_i} + \sum_{y_i f_\lambda(x_i)\in[-1,1]} \frac{\partial f_\lambda(x_i)}{\partial y_i}\right].$$

(y_i is treated as a continuous variable for this.)

$\hat{D}(\lambda)$ may be compared to trace $A(\lambda)$ in GCV and unbiased risk estimates.

♣♣ How to interpret $\frac{\partial f_\lambda(x_i)}{\partial y_i}$??

Let $K_{n\times n} = \{K(x_i, x_j)\}, D_y =$Diag $\begin{pmatrix} y_1 & & \\ & \ddots & \\ & & y_n \end{pmatrix}$

$$= \begin{pmatrix} f(x_1) \\ \vdots \\ f(x_n) \end{pmatrix} = Kc + ed \ , \ \ e = \begin{pmatrix} 1 \\ 1 \\ \cdots \\ 1 \\ 1 \end{pmatrix}.$$

Find (c, d) to min

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

The dual problem is: Find $\alpha = \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{pmatrix}$ to

$$\max -\frac{1}{2}\alpha' \left(\frac{1}{2n\lambda} D_y K D_y\right) \alpha + e'\alpha$$

subject to $\begin{pmatrix} 0 \\ 0 \\ \cdots \\ 0 \\ 0 \end{pmatrix} \leq \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{pmatrix} \leq \begin{pmatrix} 1 \\ 1 \\ \cdots \\ 1 \\ 1 \end{pmatrix}$

where $y'\alpha = 0$,

$$c = \frac{1}{2n\lambda} D_y \alpha$$

$$\begin{pmatrix} f(x_1) \\ \vdots \\ f(x_n) \end{pmatrix} = \frac{1}{2n\lambda} K D_y \alpha + ed,$$

$$\frac{\partial f(x_i)}{\partial y_i} = \frac{1}{2n\lambda} K(x_i, x_i)\alpha_i.$$

$$\hat{D}(\lambda) = \frac{1}{n}\left[2 \sum_{y_i f_\lambda(x_i) < -1} \frac{\alpha_i}{2n\lambda} K(x_i, x_i) \right.$$

$$\left. + \sum_{y_i f_\lambda(x_i) \in [-1,1]} \frac{\alpha_i}{2n\lambda} K(x_i, x_i)\right]$$

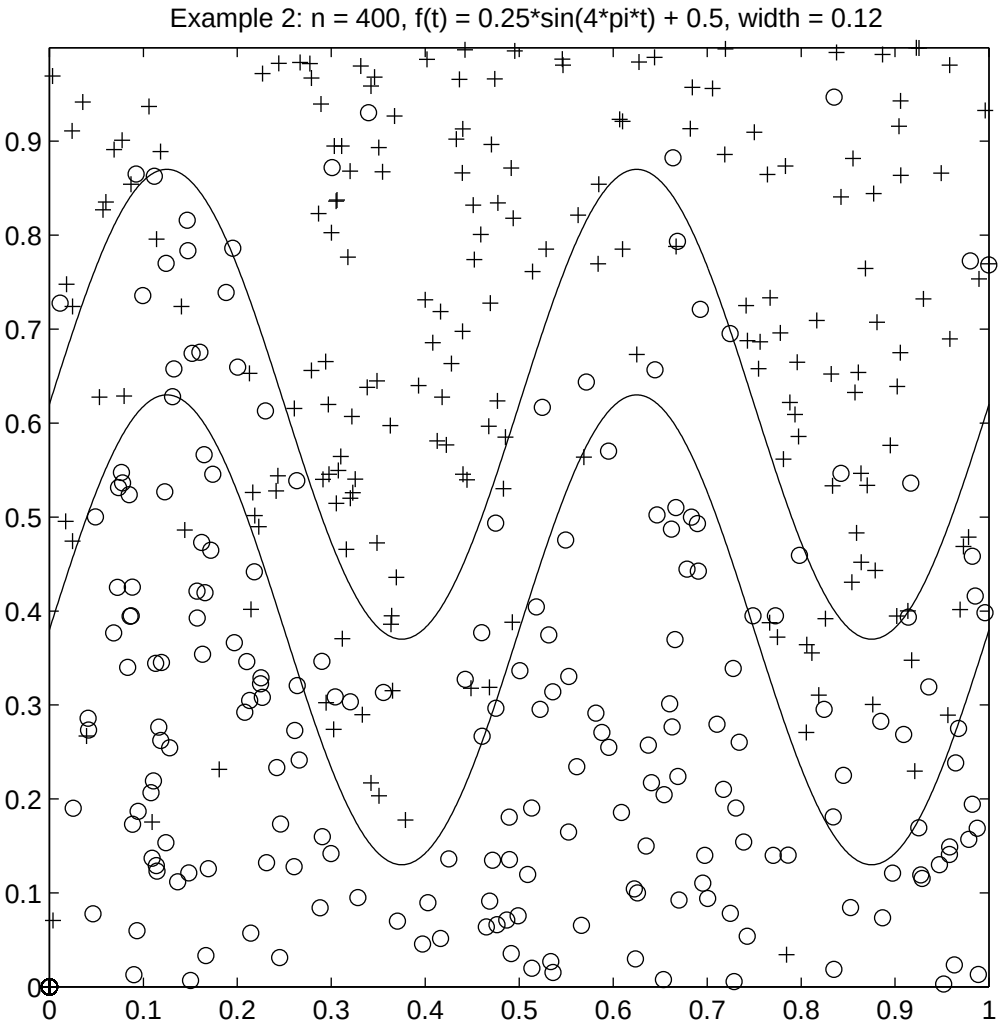
$$\boxed{\text{GACV}(\lambda) = \text{OBS}(\lambda) + \hat{D}(\lambda)}$$

(Remark: If the training set can be separated exactly then the margin γ is given by

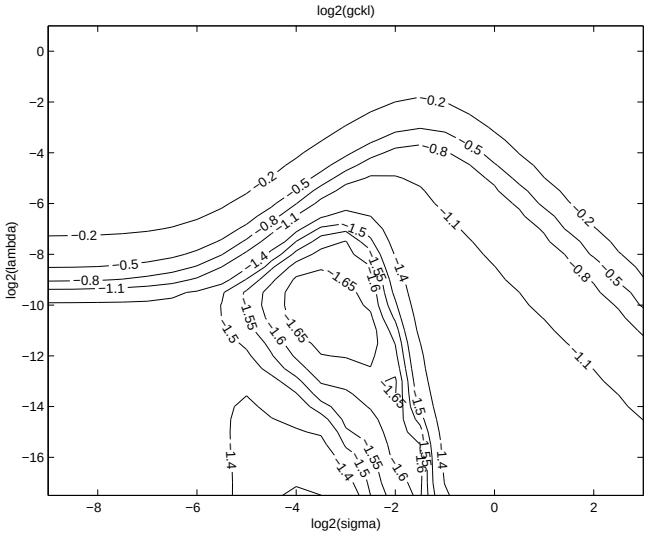
$$\gamma^2 = \left[\frac{\sum_{y_i f_\lambda(x_i) \leq 1} \alpha_i}{2n\lambda}\right]^{-1}).$$

GACV(λ) = Generalized Approximate Cross Validation, a computable proxy for $R(\lambda)$.

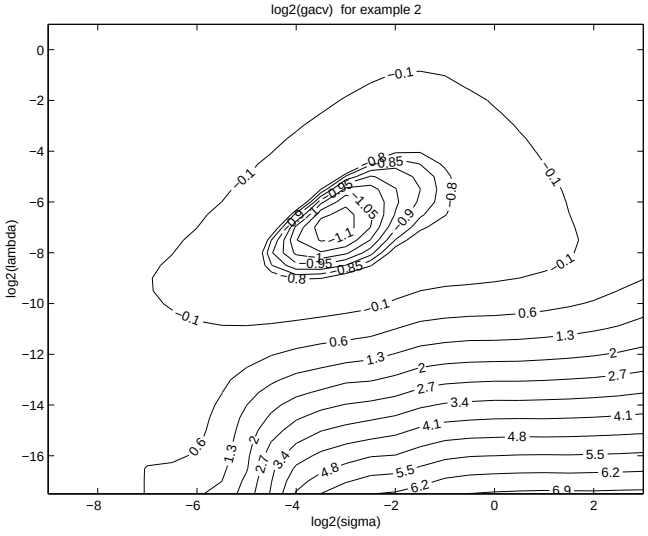
Next: A simulation study to examine GACV(λ)



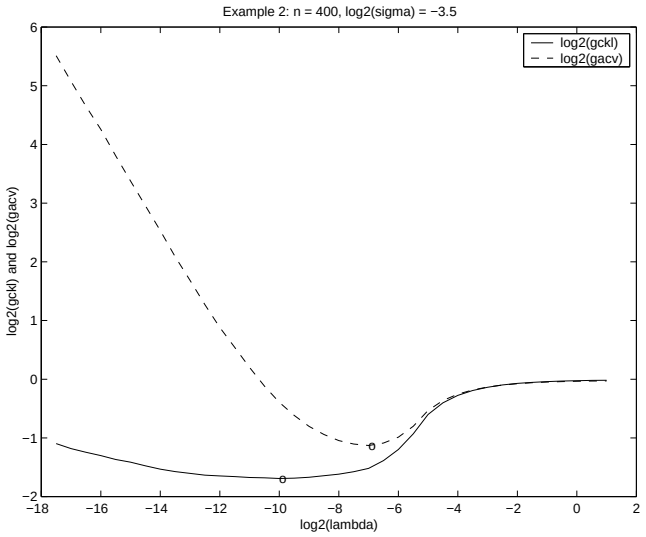
Regions of constant $p(x)$ and training data for the experiment. $p(x) = .95, .5$, and $.05$. $n = 400$. Figures courtesy Hao Zhang.



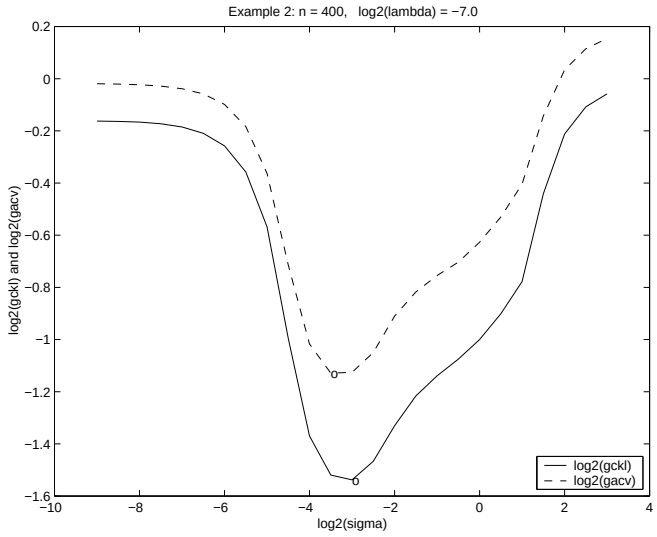
Plot of $\log_2$GCKL as a function of $\log_2 \lambda$ and $\log_2 \sigma$.
($K(s,t) = exp - \frac{1}{\sigma^2}\|s-t\|^2$.)



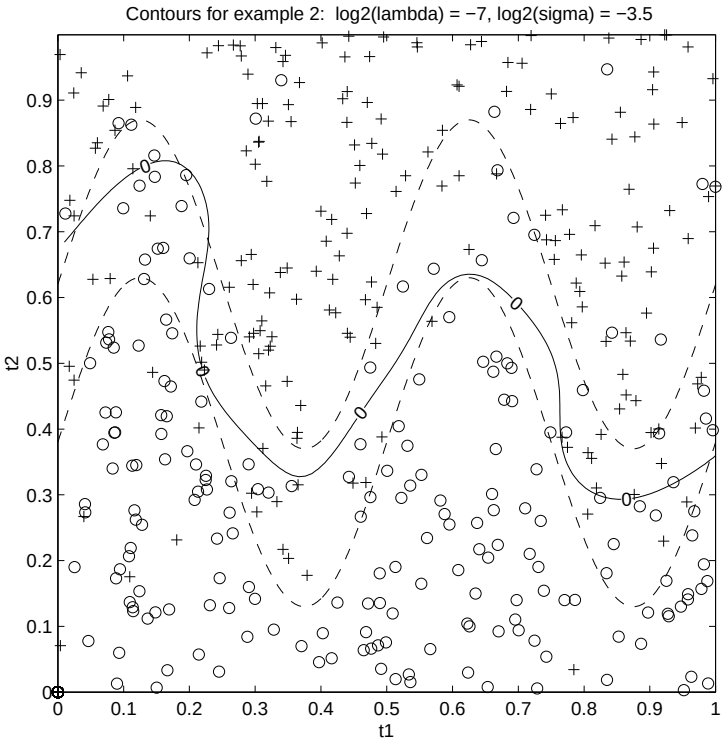
Plot of $\log_2$GACV as a function of $\log_2 \lambda$ and $\log_2 \sigma$.



Tuning λ for fixed $\widehat{\sigma} = 2^{-3.5}$



Tuning σ for fixed $\widehat{\lambda} = 2^{-7}$



Decision surface given by GACV, along with true boundaries.

♣♣ The Nonstandard Situation

$\pi_{\mathcal{A}} =$ prob. an observation in the population is an $\mathcal{A}$

$\pi_{\mathcal{B}} = 1 - \pi_{\mathcal{A}} =$ prob. an observation in the population is a $\mathcal{B}$ (as before)

$\pi^s_{\mathcal{A}} =$ fraction of training set that are $\mathcal{A}$'s

$\pi^s_{\mathcal{B}} = 1 - \pi^s_{\mathcal{A}} =$ fraction of training set that are $\mathcal{B}$'s

Let

$$p_s(x) = \frac{\pi^s_{\mathcal{A}} h_{\mathcal{A}}(x)}{\pi^s_{\mathcal{A}} h_{\mathcal{A}}(x) + \pi^s_{\mathcal{B}} h_{\mathcal{B}}(x)}$$

$$= \text{Prob.}\{y^s = \mathcal{A} | x\}$$

y^s=element of training set

♣♣ The Nonstandard Situation (continued)

Since p_s is more directly accessible we re-express the Bayes classification rule to minimize the expected cost for a random sample from the population: to get

$$\phi_{\text{OPT}}(x) = \left\{ \begin{array}{ll} \mathcal{A} & \text{if} \frac{p_s(x)}{1-p_s(x)} > \frac{c_{\mathcal{A}}}{c_{\mathcal{B}}} \frac{\pi^s_{\mathcal{A}}}{\pi^s_{\mathcal{B}}} \frac{\pi_{\mathcal{B}}}{\pi_{\mathcal{A}}} \\ \mathcal{B} & \text{otherwise} \end{array} \right\}$$

Letting $\begin{array}{rcl} L(\mathcal{B}) & = & c_{\mathcal{A}} \pi^s_{\mathcal{A}} \pi_{\mathcal{B}} \\ L(\mathcal{A}) & = & c_{\mathcal{B}} \pi^s_{\mathcal{B}} \pi_{\mathcal{A}} \end{array}$

gives

$$\begin{array}{rcl} \phi_{\text{OPT}}(x) & = & \mathcal{A} \quad \text{if } p_s(x) - \frac{L(-1)}{L(-1)+L(1)} > 0 \\ & = & \mathcal{B} \quad \text{if } p_s(x) - \frac{L(-1)}{L(-1)+L(1)} < 0 \end{array}$$

♣♣ The Nonstandard Situation (continued).

Find $f(x) = d + h(x)$ with $h \in \mathcal{H}_K$ to min

$$\frac{1}{n}\sum_{i=1}^{n} L(y_i)(1 - y_i f(x_i))_+ + \lambda \|h\|_K^2$$

(only the ratio $L(\mathcal{A})/L(\mathcal{B})$ counts if a constant is absorbed in λ).

Lemma [LLW00]

The minimizer of

$$E\ L(y^s_{new})(1 - y^s_{new} f(x))_+ \text{ is}$$

$$\text{sign}\quad \left(p_s(x) - \frac{L(-1)}{L(-1) + L(1)}\right)$$

"s" – training set

↗ replaces $\frac{1}{2}$

♣♣ The Nonstandard Situation (continued)

Define

$$R(\lambda) = E\frac{1}{n}\sum_{i=1}^{n} L(y^s_{new\cdot i})(1 - y^s_{new\cdot i} f_\lambda(x_i))_+$$

(a.k.a GCKL (λ)) (Its an upper bound for the expected cost in the general population). Define

$$\boxed{\text{GACV}(\lambda) = \text{OBS}(\lambda) + \hat{D}(\lambda)}$$

where

$$\text{OBS}(\lambda) = \frac{1}{n}\sum_{i=1}^{n} L(y_i)(1 - y_i f_\lambda(x_i))_+$$

$$\hat{D}(\lambda) = \frac{1}{n}\left[2\sum_{y_i f_\lambda(x_i)<1} L(y_i)\frac{\alpha_i}{2n\lambda}K(x_i,x_i)\right.$$

$$\left. + \sum_{y_i f_\lambda(x_i)\in[-1,1]} L(y_i)\frac{\alpha_i}{2n\lambda}K(x_i,x_i)\right].$$

♣♣ The Nonstandard Situation (continued)

$$f_\lambda(x) \rightarrow \text{ sign } \left[p_s - \frac{L(-1)}{L(-1)+L(1)}\right]$$

Then $R(\lambda) \rightarrow$ expected cost. Choose λ to min GACV(λ) (a proxy for $R(\lambda)$).

Experiment: Population:

$$\mathcal{A}: \quad \mathcal{N}\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & 0\\0 & 1\end{pmatrix}$$

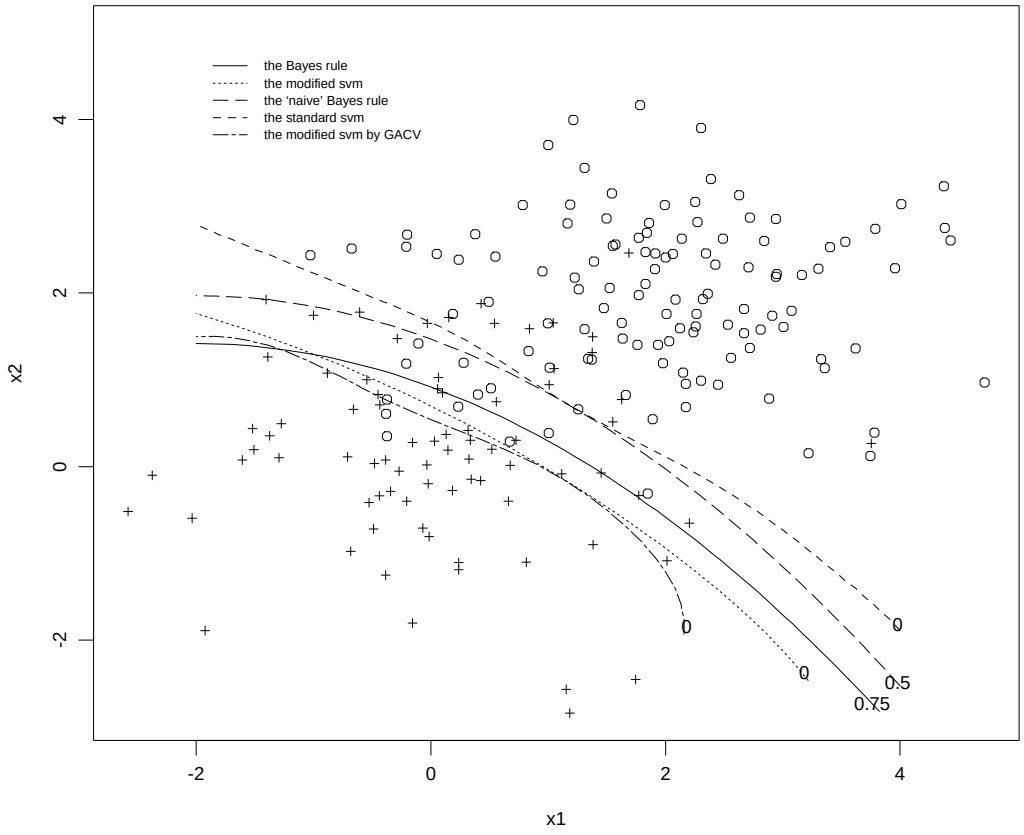
$$\mathcal{B}: \quad \mathcal{N}\begin{pmatrix}2\\2\end{pmatrix}, \begin{pmatrix}2 & 0\\0 & 1\end{pmatrix}$$

$\pi_{\mathcal{A}} = .1 \quad \pi_{\mathcal{B}} = .9 \quad c_{\mathcal{A}} = 2c_{\mathcal{B}}$

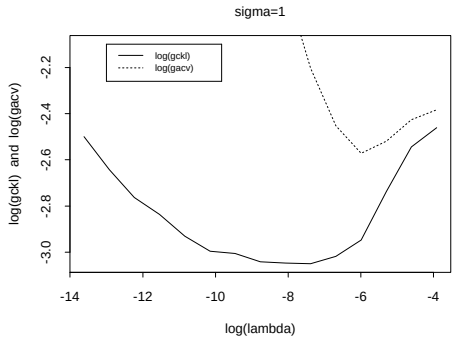
Sample:

$$\pi^s_{\mathcal{A}} = .4, \pi^s_{\mathcal{B}} = .6, L(-1) = .36, L(1) = .12.$$

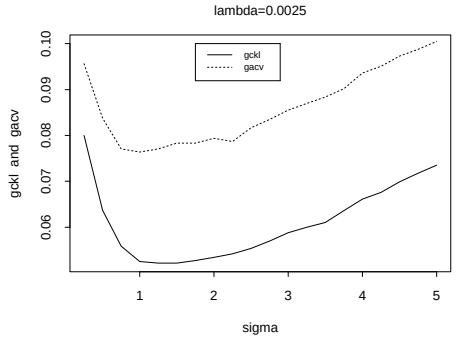
$n = 200 \qquad n_{tune} = 200$ generated to compare GACV to the use of a tuning set.



From [LLW00] Decision surfaces given by the modified and standard support vector machines, the Bayes rule, and the 'naive' (standard) Bayes rule. The Bayes rules are computed using knowledge of the underlying populations. The standard SVM is tuned using a second set of observations generated and used for tuning. The modified support vector machine is implemented both with the tuning set and with the GACV.



From [LLW00] GCKL and GACV plot as a function of λ when σ is fixed at 1. We can see the minimizer of GACV is a decent estimate of the minimizer of GCKL.



From [LLW00] GCKL and GACV plot as a function of σ when λ is fixed at 0.0025. Again we can see the minimizer of GACV is a decent estimate of the minimizer of GCKL.