Nonstandard, Multicategory Support Vector Machines for Classification

Grace Wahba

`http://www.stat.wisc.edu/~wahba`

→ *TRLIST*

Based on the thesis of Yoonkyung Lee

`http://www.stat.wisc.edu/~yklee`

some recent work of Yi Lin

`http://www.stat.wisc.edu/~yilin`

and some joint Lee, Lin and Wahba work.

Harvard University Statistics Department

April 1, 2002

Yi is at Madison

Yoon will be joining the Statistics Department at Ohio State this fall.

...

Abstract

Support Vector Machines (SVM's) are a modern and very popular method for building (supervised) classifiers. We define (two-category) SVMS and compare them in their variational formulation with penalized log likelihood estimates for the log odds ratio for two classes, $\mathcal{A}$ and $\mathcal{B}$, given a labeled representative training set of attribute vectors. SVM's return a yes-no response for a new attribute vector x to be in class $\mathcal{A}$ or class $\mathcal{B}$, while penalized likelihood estimates return an estimate of the probability $p(x)$ of membership in class $\mathcal{A}$ for a new attribute vector x. We describe a version of the generalized approximate cross validation (GACV) for tuning or controlling the bias-variance tradeoff a. k. a. goodness of fit/complexity tradeoff for the SVM case.

A result of Yi Lin (UW-Madison Statistics Dept TR 1014, 1999) that the (tuned) SVM in the standard case (representative sample, equal costs of misclassification), is asymptotically estimating sign$[p(x) - 1/2]$ is noted. In that case it can be shown that the GACV is asymptotically tuning SVM's against the misclassification rate.

We generalize the (usual) SVM to the unbalanced case where the fraction of members of class $\mathcal{A}$ in the training set is different than that in the general population, and the costs of misclassification for the two kinds of errors are different.

Then we generalize the SVM to the nonstandard multicategory case, where the sample is not necessarily representative and the costs of different kinds of misclassification are not necessarily equal.

References: Lin and Wahba papers available via their home pages.

Y. Lin, G. Wahba, H. Zhang, and Y. Lee. Statistical properties and adaptive tuning of support vector machines. *Machine Learning*, 48:115–136, 2002.

Y. Lin, Y. Lee, and G. Wahba. Support vector machines for classification in nonstandard situations. *Machine Learning*, 46:191–202, 2002.

Y. Lee, Y. Lin, and G. Wahba. Multicategory support vector machines. Technical Report 1043, Department of Statistics, University of Wisconsin, Madison WI, 2001. To appear, The Interface Foundation, Computing Science and Statistics, v 33.

G. Wahba, Y. Lin, Y. Lee, and H. Zhang. Optimal properties and adaptive tuning of standard and non-standard support vector machines. Technical Report 1045, Statistics Department University of Wisconsin, Madison WI, 2001. Proceedings of the MSRI Berkeley Workshop on Nonlinear Estimation and Classification.

Y. Lin. Some Asymptotic Results On the Support Vector Machine. Technical Report 1029, Department of Statistics, University of Wisconsin, Madison WI, 2002.

Y. Lin. Support Vector Machines and the Bayes Rule in Classification, To appear *Data Mining and Knowledge Discovery*, also University of Wisconsin-Madison Statistics Dept TR 1014, 1999
.

G. Wahba, Y. Lin, and H. Zhang. Generalized approximate cross validation for support vector machines. In

A. Smola, P. Bartlett, B. Scholkopf, and D. Schuurmans, editors, *Advances in Large Margin Classifiers*, pages 297–311. MIT Press, 2000.

O. Chapelle, V. Vapnik, O. Bousquet, and S. Mukherjee. Choosing multiple parameters for support vector machines. *Machine Learning*, 46:131–159, 2002.

K. Duan, S. Keerthi, and A. Poo. Evaluation of simple performance measures for tuning svm hyperparameters. Technical Report CD-01-11, Dept. of Mechanical Engineering, National University of Singapore, Singapore, 2001.

T. Joachims. Estimating the generalization performance of an SVM efficiently. In *Proceedings of the International Conference on Machine Learning*, San Francisco, 2000. Morgan Kaufman.

The website `http://www.kernel-machines.org` has links to papers, tutorials, books, researchers connected with SVM's, and other methods which use (reproducing) kernels (!)

OUTLINE

1. Review of Optimal Classification.

2. Comparison of penalized likelihood and SVM classifiers.

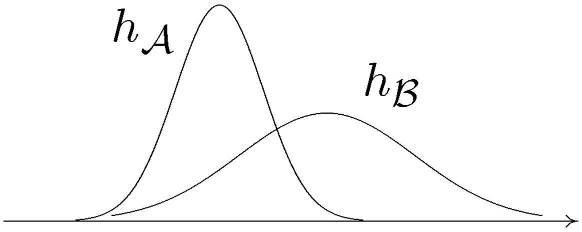
3. The standard case SVM – equal cost of misclassification and representative training set. GACV tuning for the standard case.

4. Yi Lin's theorem: The (tuned) SVM is estimating the sign of the log-odds ratio and minimizing the expected misclassification rate.

5. Extension to the non-standard case:
Non-representative training set, unequal costs.

6. Extension to the non-standard multicategory case.

♣♣ Optimal Classification and the Neyman-Pearson Lemma:



$h_{\mathcal{A}}(\cdot), h_{\mathcal{B}}(\cdot)$ densities of x

for class $\mathcal{A}$ and class $\mathcal{B}$.

NOTATION:

$\pi_{\mathcal{A}} =$ prob. next observation (Y) is an $\mathcal{A}$

$\pi_{\mathcal{B}} = 1 - \pi_{\mathcal{A}} =$ prob. next observation is a $\mathcal{B}$

$$\begin{aligned} p(x) &= prob\{Y = \mathcal{A}|x\} \\ &= \frac{\pi_{\mathcal{A}} h_{\mathcal{A}}(x)}{\pi_{\mathcal{A}} h_{\mathcal{A}}(x) + \pi_{\mathcal{B}} h_{\mathcal{B}}(x)} \end{aligned}$$

Let $c_{\mathcal{A}} =$ cost to falsely call a $\mathcal{B}$ an $\mathcal{A}$

$c_{\mathcal{B}} =$ cost to falsely call an $\mathcal{A}$ a $\mathcal{B}$

Bayes classification rule: Let

$\phi(x)$: $\quad x \rightarrow \{{}^{\mathcal{A}}_{\mathcal{B}}\}$

Expected cost:

$E\{c_{\mathcal{A}}[1 - p(x)]\ I(\phi(x) = \mathcal{A})\}$

get a $\mathcal{B}$ and call it an $\mathcal{A}$

$+E\{c_{\mathcal{B}}[p(x)]\ \ I(\phi(x) = \mathcal{B})\}$

get an $\mathcal{A}$ and call it $\mathcal{B}$

Optimum (Bayes) classifier:

$$\phi_{\mathsf{OPT}}(x) = \begin{cases} \mathcal{A} & \text{if } \frac{p(x)}{1-p(x)} > \frac{c_{\mathcal{A}}}{c_{\mathcal{B}}}, \\ \mathcal{B} & \text{otherwise.} \end{cases}$$

To estimate $p(x)$, alternatively let $f(x) = \log p(x)/(1-p(x))$, the log odds ratio a.k.a. the logit. “Standard” case: Training set

$$\{y_i, x_i\} \qquad \begin{array}{l} y_i \in \{\mathcal{A}, \mathcal{B}\} \\ x_i \in \mathcal{T}, \text{ some index set} \end{array}.$$

Relative frequency of $\mathcal{A}$’s in the training set is about the same as in the general population.

Penalized log likelihood estimation:

Estimate f by penalized likelihood. If $c_{\mathcal{A}}/c_{\mathcal{B}} = 1$, then the optimal classifier is

$$f(x) > 0 \text{ (equivalently, } p(x) - \tfrac{1}{2} > 0) \rightarrow \mathcal{A}$$
$$f(x) < 0 \text{ (equivalently, } p(x) - \tfrac{1}{2} < 0) \rightarrow \mathcal{B}$$

♣♣ Penalized log likelihood estimation of the logit $f = \log[p/(1-p)]$.

$$y = \begin{matrix} 1 & = \mathcal{A} \\ 0 & = \mathcal{B} \end{matrix} \text{ (important)}$$

The probability distribution function (likelihood) for $y \mid p$ is: $\mathcal{L} = p^y(1-p)^{1-y} = \begin{cases} p & \text{if } y = 1 \\ (1-p) & \text{if } y = 0 \end{cases}$
and the negative log likelihood is

$$-\log \mathcal{L} = -\log[p^y(1-p)^{1-y}]$$
$$= -y \log p - (1-y)\log(1-p).$$

Using $p = e^f/(1+e^f)$ gives
$-\log \mathcal{L} = -yf + \log(1+e^f)$

♣♣ Penalized log likelihood estimation of f (continued) (special case).

$$\{y_i, x_i\}, \ \ y_i = \begin{matrix} 1 \\ 0 \end{matrix} \ \ , x_i \in \mathcal{T}$$

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

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

where $\mathcal{H}_K$ is the reproducing kernel Hilbert space (RKHS) with reproducing kernel

$$K(s, t), \ \ s, t, \in \mathcal{T}.$$

Theorem: [KW71]

$$f_\lambda(x) = d + \sum_{i=1}^{n} c_i K(x, x_i).$$

♣♣ Penalized log likelihood estimation of f (continued)

$$f_\lambda(x) = d + \sum_{i=1}^{n} c_i K(x, x_i)$$

Find $d, c = (c_1, \ldots, c_n) = c_\lambda$ to minimize

$$\frac{1}{n}\sum_{i=1}^{n}\left[-y_i f(x_i) + \log(1 + e^{f(x_i)})\right] + \lambda \|h\|^2_{\mathcal{H}_K}.$$

Here

$$\|h\|^2_{\mathcal{H}_K} \equiv \sum_{i,j=1}^{n} c_i c_j \; K(x_i, x_j).$$

Given λ, this is a nice strictly convex optimization problem. Choose λ by GACV (Xiang and Wahba, Statistica Sinica 1996 et seq). Target for GACV is to minimize the Comparative Kullback-Liebler (CKL) distance of the estimate from the true distribution:

$$R(\lambda) = E_{f_{true}} \sum_{i=1}^{n} -y_{new.i} f_\lambda(x_i) + \log(1 + e^{f_\lambda(x_i)}).$$

♣♣ Support Vector Machines

$$y = \begin{matrix} +1 = \mathcal{A} \\ -1 = \mathcal{B} \end{matrix} \text{ (note different coding)}$$

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

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

where $(\tau)_+ = \tau, \tau > 0, = 0$ otherwise.

Then

$$f_\lambda(x) = d + \sum_{i=1}^{n} c_i K(x, x_i). \qquad (*)$$

Substitute (*) into (**), choose λ, given λ, find c and d. The classifier is

$$f_\lambda(x) > 0 \rightarrow \mathcal{A}$$

$$f_\lambda(x) < 0 \rightarrow \mathcal{B}$$

♣♣ Comparison of the penalized log likelihood estimate f_λ of the log odds ratio $\log p/(1-p)$ and f_λ, the SVM classifier:

Suspicion: They are related...

Let us relabel y in the likelihood –

$$\tilde{y} = \begin{cases} +1 & \text{if } \mathcal{A}, \\ -1 & \text{if } \mathcal{B}. \end{cases}$$

Then

$$-yf + \log(1 + e^f) \rightarrow \log(1 + e^{-\tilde{y}f})$$

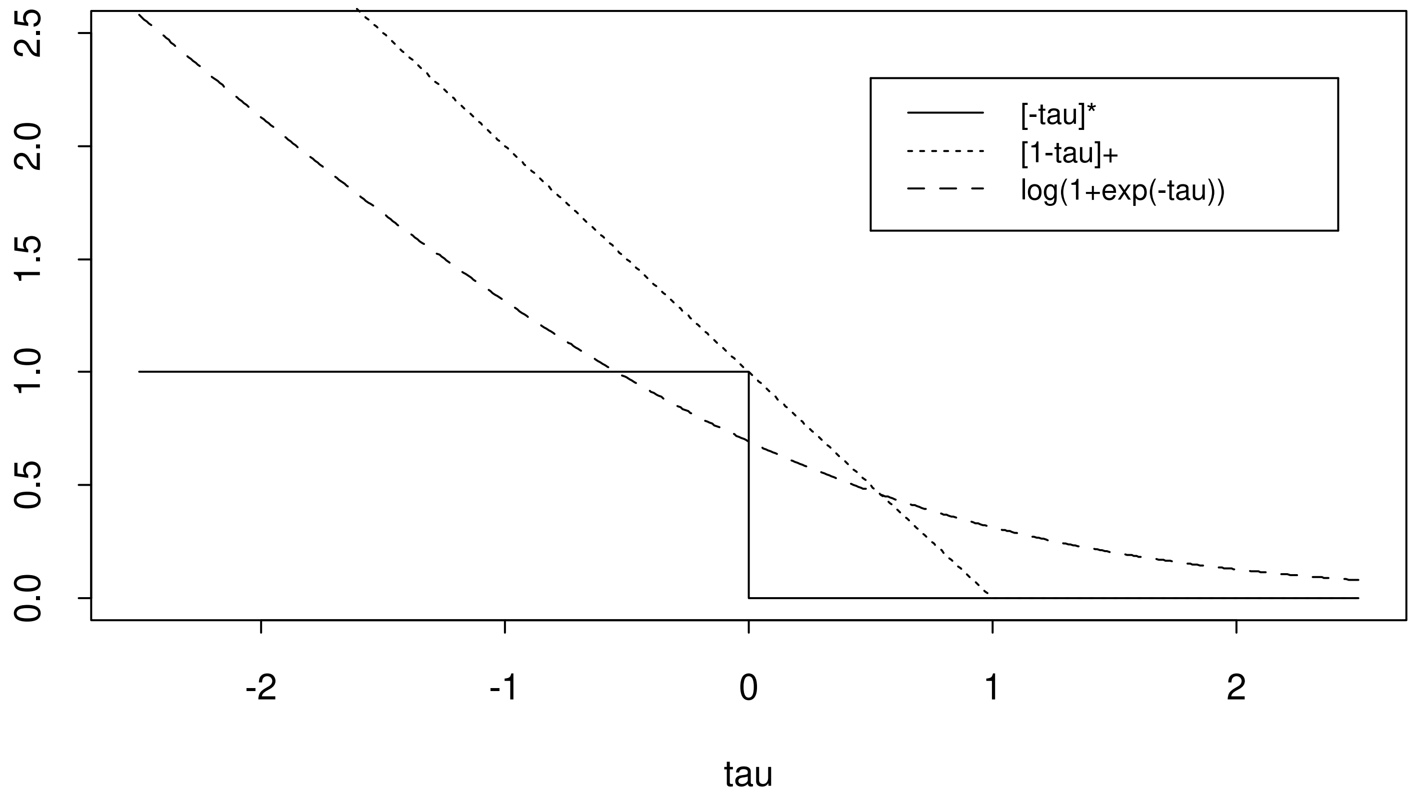
Figure 1 compares

$$\log(1 + e^{-yf}), \quad (1 - yf)_+ \text{ and } (-yf)_*$$

where

$$(\tau)_* = \begin{cases} 1 & \text{if } \tau > 0, \\ 0 & \text{otherwise.} \end{cases}$$

($(-yf)_*$ is the misclassification counter).



Comparison of $(-\tau)_*$, $(1-\tau)_+$ and $log_e(1+e^{-\tau})$. Any strictly convex non-negative function that goes through 1 at $\tau = 0$ will be an upper bound on the missclassification function and will be a looser bound than some SVM function.

♣♣ 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]:

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

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

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) $R(\lambda)$,

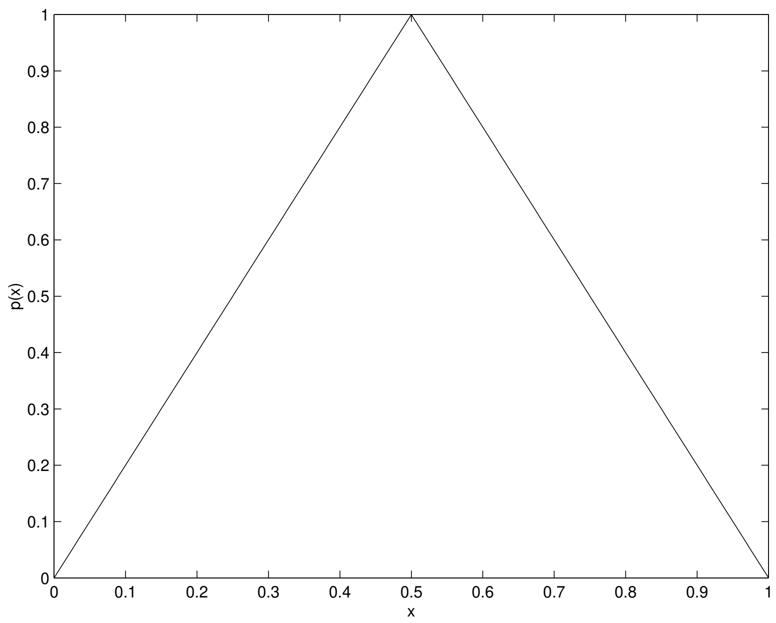
is estimating sign $f(x)$ – EXACTLY WHAT YOU NEED to minimize the misclassification rate!

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

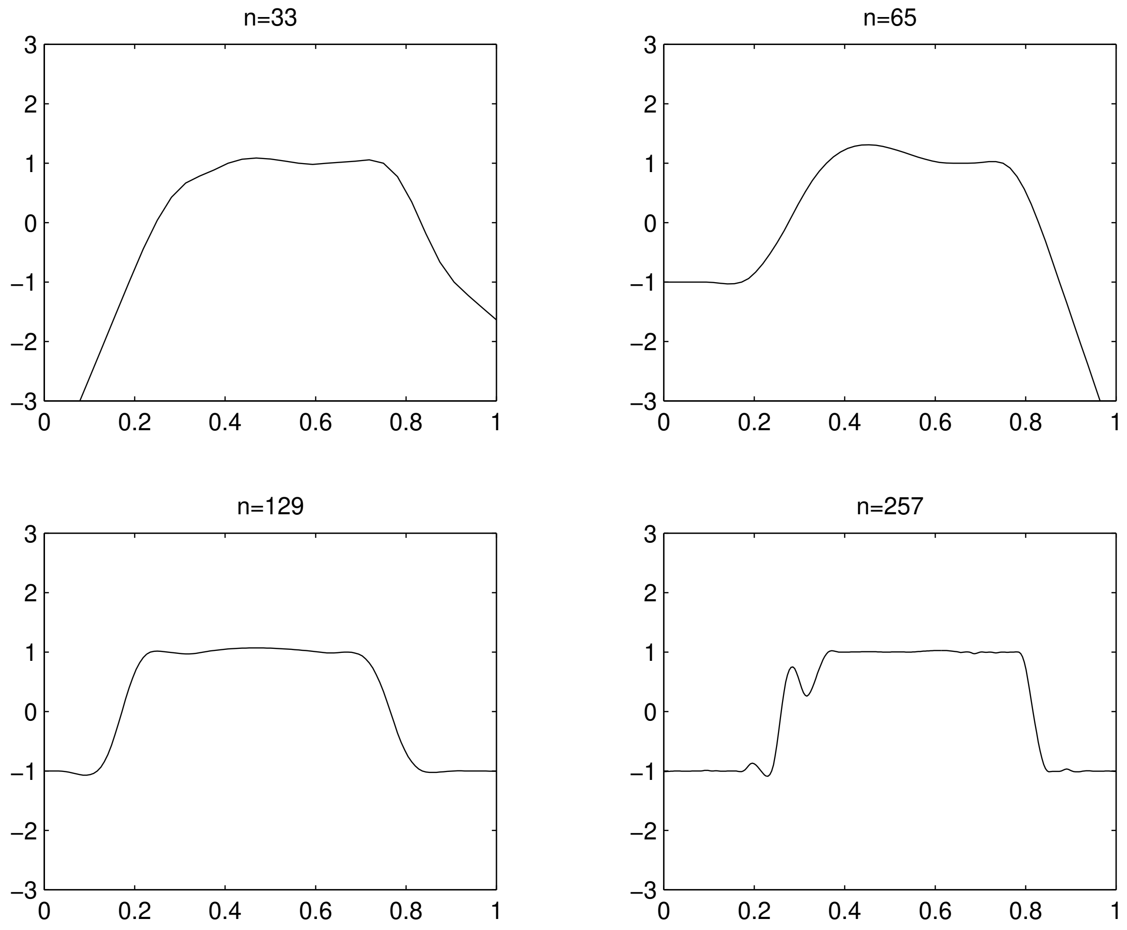
$$E(1 - y_{new} f_\lambda)_+ = \left\{ \begin{array}{lrl} p(1 - f_\lambda), & f_\lambda & < -1 \\ p(1 - f_\lambda) + (1 - p)(1 + f_\lambda), \quad -1 < & f_\lambda & < +1 \\ (1 - p)(1 + f_\lambda), & f_\lambda & > +1. \end{array} \right\}$$

$R(\lambda)$ 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.

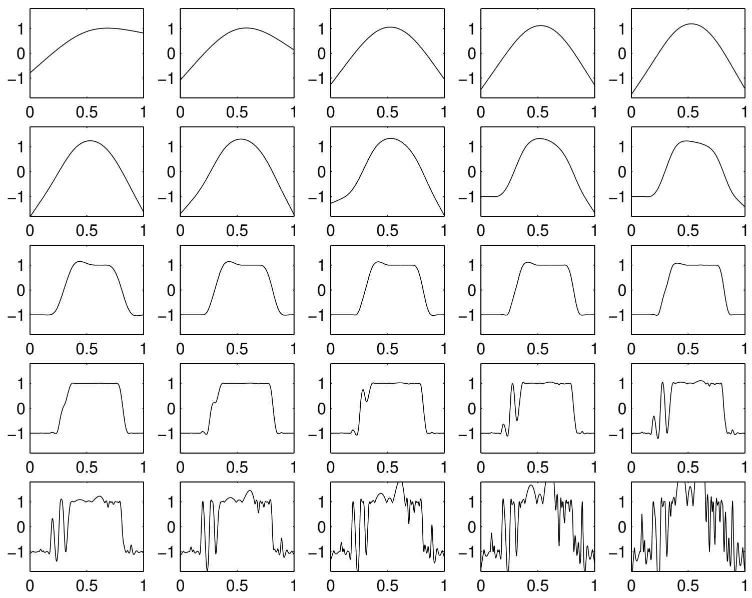
Experiments to follow are courtesy Yi Lin, reprinted from [Lin99].



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.

♣♣ 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

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

Wahba, Lin, Zhang, 2000 showed that

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

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

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

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

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

$$V(\lambda) = \mathrm{OBS}(\lambda) + \widehat{D}(\lambda),$$

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

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

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

$$\boxed{\mathrm{GACV}(\lambda) = \mathrm{OBS}(\lambda) + \widehat{D}(\lambda)}$$

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

Loosely speaking, if f_λ were differentiable with respect to y_i, one could make the association that $\widehat{D}(\lambda)$ is approximated by

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

Then $\boxed{\widehat{D}(\lambda) \text{ may be compared to trace } A(\lambda) \text{ in}}$ $\boxed{\text{GCV and unbiased risk estimates.}}$
See also Joachims, Keerthi et al and references in Chapelle et al for closely related estimates.

♣♣ 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

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

Letting
$$\begin{aligned} L(-1) &= c_{\mathcal{A}} \pi_{\mathcal{B}} / \pi^s_{\mathcal{B}} \\ L(+1) &= c_{\mathcal{B}} \pi_{\mathcal{A}} / \pi^s_{\mathcal{A}} \end{aligned}$$

gives

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

♣♣ 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 $L(+1)/L(-1)$ counts if a constant is absorbed in λ).

Lemma (Lin, Lee, Wahba 2000)

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

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

where

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

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

♣♣ The Nonstandard Situation (continued)

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

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

Experiment: Population:

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

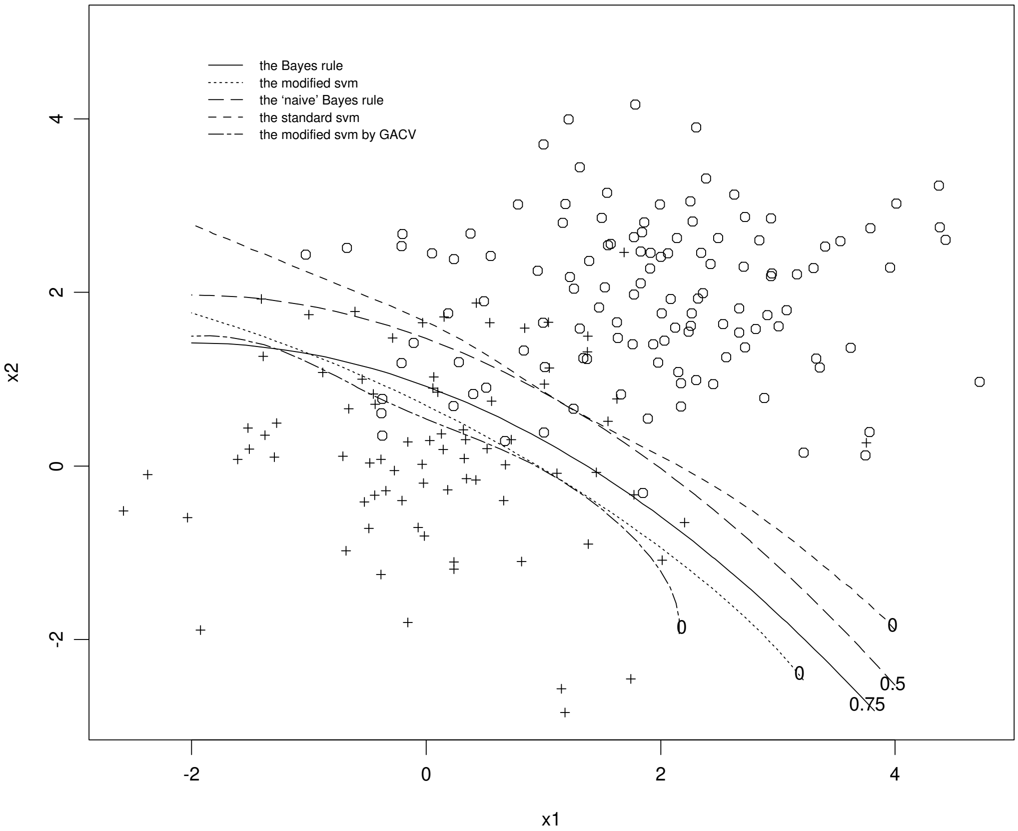
$$\mathcal{B}: \quad \mathcal{N} \begin{pmatrix} 2 \\ 2 \end{pmatrix}, \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix}$$

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

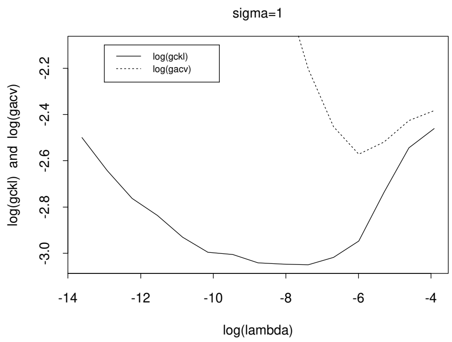
Sample:

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

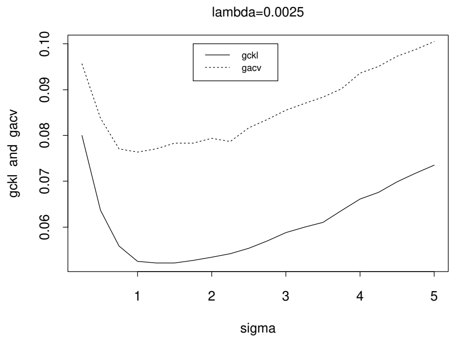
$n = 200 \qquad n_{tune} = 200$ generated to compare GACV to the use of a tuning set. $K(u,v) = e^{-\|u-v\|^2/\sigma^2}$.



From Lin, Lee, Wahba, 2000. Decision surfaces given by the nonstandard 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 nonstandard SVM is implemented both with the tuning set and with the GACV.



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.



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.

♣♣ The Multicategory Support Vector Machine

If the cost of misclassification is the same for each category, then the optimum classifier would choose category k if $p_k(x) = max_r p_r(x)$

Standard case: cost of misclassification the same for each k, training sample representative.

Many classification problems involve more than two categories and most of the SVM authors use one-versus-many or multiple pairwise comparisons.

We code the class label y_i as a k dimensional vector with 1 in the jth position if example i is in category j and $-\frac{1}{k-1}$ otherwise, thus $y_i = (1, -\frac{1}{k-1}, \cdots, -\frac{1}{k-1})$ indicates that the ith example is in category 1.

Define a k-tuple of separating functions: $f(x) = (f^1(x), \cdots f^k(x))$, with each $f^j = d^j + h^j$ with $h^j \in \mathcal{H}_K$, and required to satisfy a sum-to-zero constraint, $\sum_{j=1}^k f^j(x) = 0$, for all x in $\mathcal{T}$. Let y_{ir} be the rth component of y_i, $L_{jr}(y)$ be the cost if y is a category j and it is classified as a category r. In the standard case, $L_{jr} = 1, j \neq r$ and 0 if $j = r$. The multicategory SVM (with a representative sample) is defined as the vector of functions f_λ, satisfying the sum-to-zero constraint, which minimizes

$$\frac{1}{n} \sum_{i=1}^{n} \sum_{r=1}^{k} L_{jr}(y_i)(f^r(x_i) - y_{ir})_+ + \lambda \sum_{j=1}^{k} \|h^j\|^2_{\mathcal{H}_K}.$$

It can be shown that the $k = 2$ case reduces to the usual 2-category SVM just discussed. Lee, Lin and Wahba 2001 show that the target for the multicategory SVM is $f(x) = (f^1(x), \cdots, f^k(x))$ with $f^\ell(x) = 1$ if $p_\ell(x) = max_r p_r(x)$ and $f^r(x) = -\frac{1}{k-1}, r \neq \ell$.

♣♣ The Nonstandard Multicategory Support Vector Machine

Allow the proportion $\pi_j^s, j = 1, \cdots, k$ of examples in the training set be not representative of the proportion π_j in the population as a whole, as well as allowing costs of misclassification to be different for different mistakes. Let $c_{\ell j}$ be the cost of classifying an ℓ subject as a j. Then the Bayes rule (which minimizes expected cost) is to choose the j for which $\sum_{\ell=1}^{k} c_{\ell j} p_\ell(x)$ is minimized, where $p_\ell(x)$ is the probability that a subject (in the population as a whole) with attribute vector x is in category ℓ. Given the π_j, π_j^s and $p_j(x)$, then $p_j^s(x)$, the (theoretical) probability that a subject with attribute vector x is in category j, is determined. (See Lee, Lin Wahba 2001).

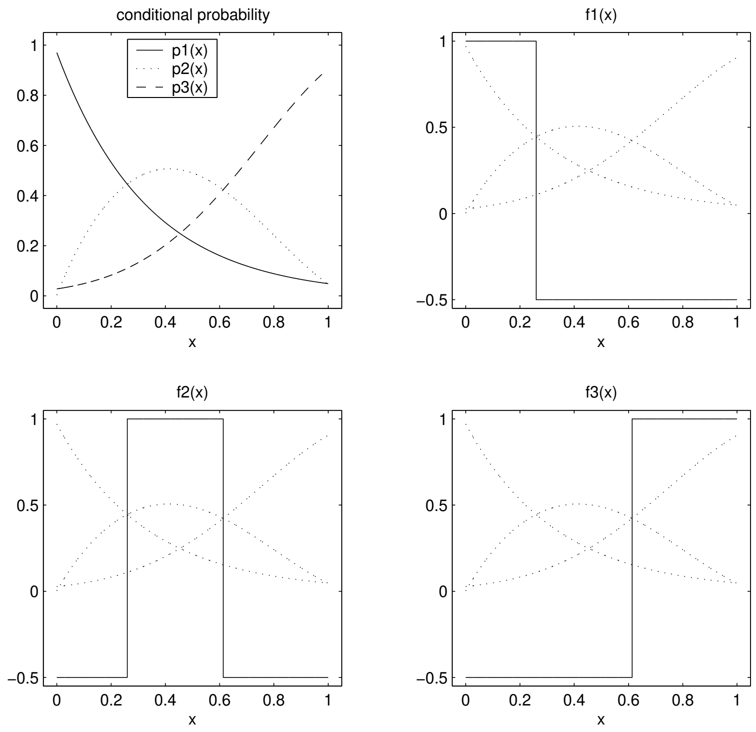
Let now

$$L_{jr} = c_{jr}\pi_j/\pi_j^s \quad (*).$$

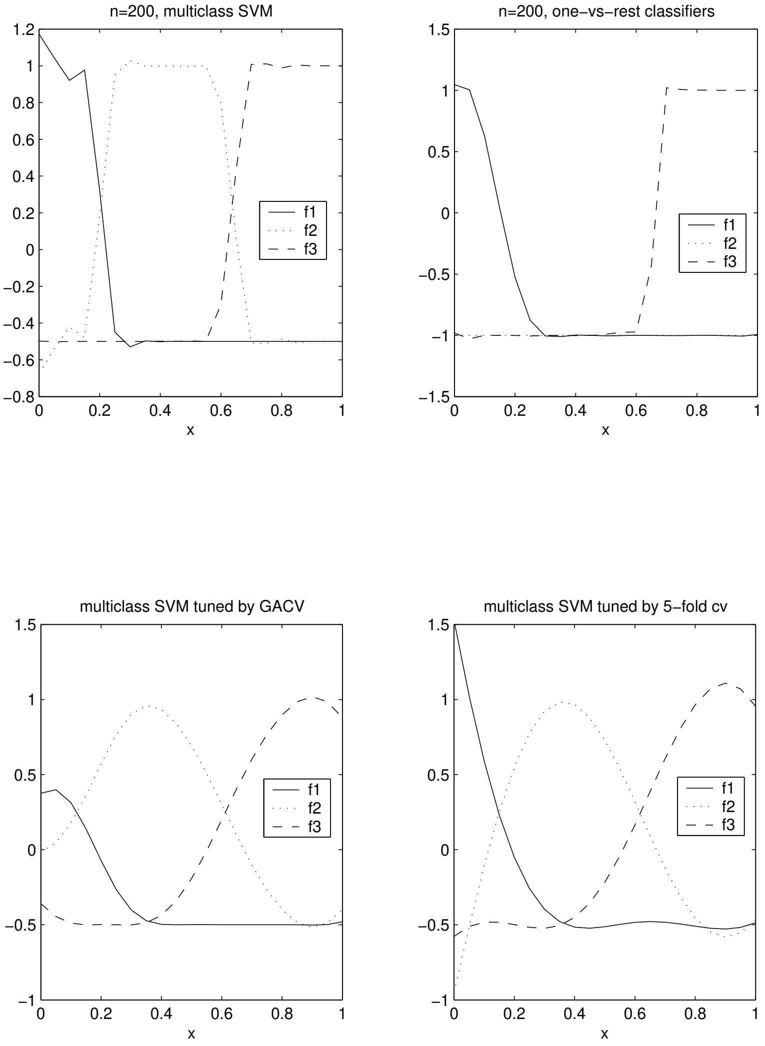
One can then show that the optimum classifier chooses category j if the minimizer over $\ell = 1, 2, \cdots k$ of $\sum_{r=1}^{k} L_{r\ell} p_r^s(x)$ is j. The nonstandard multicategory SVM is defined as the vector of functions f_λ satisfying the sum-to-zero constraint, which minimizes

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{r=1}^{k} L_{jr}(y_i)(f^r(x_i) - y_{ir})_+ + \lambda \sum_{j=1}^{k} \|h^j\|^2_{\mathcal{H}_K}.$$

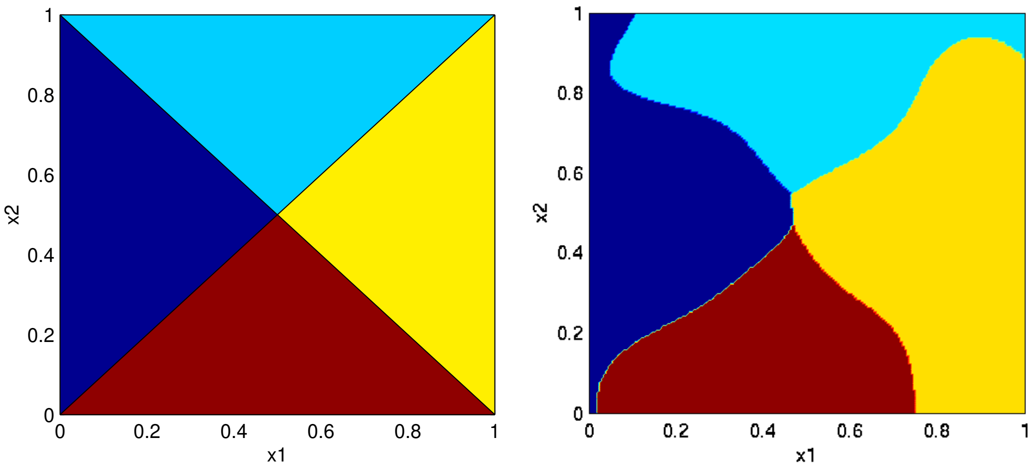
where $L_{jr}(y_i)$ is given by $(*)$ when y_i is from class j. It is shown in Lee, Lin and Wahba 2001 that the target of this SVM is: $f^j(x) = 1$ if j is the minimizer over ℓ of $\sum_{r=1}^{k} c_{r\ell} p_r(x)$ and $-\frac{1}{k-1}$ otherwise.



Conditional probabilities and multicategory SVM target functions for a three class example.



Top: comparison between the multicategory SVM (l.) and the one-vs-rest method (r). Tuning parameters λ and σ chosen by GCKL (i. e. ideally tuned). Botton λ, and σ are chosen by GACV (l.) and 5-fold CV (r.).



Classification boundaries determined by the Bayes rule (left) and estimated classification boundaries by the multicategory SVM (right). $n = 300$. The $p_j(x)$ are Gaussian densities, and a Gaussian kernel was used (ideally tuned).

♣♣ Summary and Conclusions

1. SVM's are wildly popular. An advanced google search yielded 'about 5770' entries.

2. Yi Lin has provided an explanation, from the point of view of a statistician, why.

3. We have generalized the SVM to the nonstandard case, where the sample is not representative and misclassification costs are not equal.

4. We have provided an in-sample method (GACV) for tuning the SVM. Several others have (earlier!) obtained an in-sample tuning procedure via essentially the same approximate leaving-out-one

argument for the standard case. In particular some thing very close is implemented in Joachims' SVM-light, see `www.kernel-machines.org` for links to that and many other codes for the two category case.

5. We have generalized the SVM to the nonstandard, multicategory case, along with the GACV.