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

TECHNICAL REPORT NO. 1039

June 15, 2001

On the Relation Between the GACV and Joachims' $\xi\alpha$ Method for Tuning Support Vector Machines, With Extensions to the Non-Standard Case

Grace Wahba, Yi Lin, Yoonkyung Lee and Hao Zhang

wahba,yilin,yklee,hzhang@stat.wisc.edu

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

Research partly supported by NSF Grant DMS0072292 and NIH Grant EY09946.

On the Relation Between the GACV and Joachims' $\xi\alpha$ Method for Tuning Support Vector Machines, With Extensions to the Non-Standard Case

Grace Wahba, Yi Lin, Yoonkyung Lee and Hao Zhang
Department of Statistics
University of Wisconsin
1210 W. Dayton St, Madison WI 53706
wahba, yilin, yklee, hzhang@stat.wisc.edu

Abstract

We rederive a form of Joachims' $\xi\alpha$ method for tuning Support Vector Machines by the same approach as was used to derive the GACV, and show how the two methods are related. We generalize the $\xi\alpha$ method to the nonstandard case of nonrepresentative training set and unequal misclassification costs and compare the result to the GACV estimate for the standard and nonstandard cases.

1 Introduction

Support Vector Machines (SVM's) are an important and increasingly popular tool for classification, and in their nonlinear forms can handle a wide variety of classification problems. The Generalized Approximate Cross Validation (GACV) method for choosing the tuning parameter(s) in Support Vector Machines (SVM's) was proposed in [7]. The $\xi\alpha$ (XA) method for choosing the tuning parameter(s) in SVM's was proposed in [1] and it is included in recent versions of the code SVM^{light}, `http://ais.gmd.de/~thorsten/svm_light/`. We have found an interesting relationship between the XA and the GACV, which can be computed at essentially no cost alongside the XA. It is the first purpose of this report to note the aforementioned relationship, and to provide an alternative derivation of (one form of) the XA, which reaches the particular form studied as an approximation to a leaving out one misclassification rate estimate, rather than an upper bound as given by Joachims. We compare the two estimates in a context quite different than the text classification problems considered in detail in [1]. A modest simulation example shown here shows that the GACV may tune (very marginally) closer to the minimizer of the misclassification rate, but the estimates are very close and both do very well. These and Joachims' results, along with the rapidly growing body of other numerical results referenced, e. g. in `http://www.kernel-machines.org`, as well as the theoretical results of [2] which show that tuned SVM's are implementing the Bayes rule, serve to further buttress the claim that tuned SVM's are the method of choice for a truly vast array of classification problems. Background on

SVM's may be found in the 'Tutorial' section of the kernel-machines.org website. [3], [4] proposed the non-standard SVM for classification problems where the relative proportion of the two classes in the training set is not equal to that in the general population, and where the costs of the two kinds of misclassification are not equal. The non-standard SVM proposed there is shown to implement the Bayes rule (minimize Bayes Risk (BR)) in the nonstandard situation when tuned, and the GACV estimate for tuning it was proposed. The second purpose of this paper is to use the alternative derivation of the XA noted above to obtain a generalization of the XA estimate to the non-standard SVM, (to be called BRXA for the Bayes Risk $\xi\alpha$). Again a modest example suggests that both the (nonstandard) GACV and the BRXA provide tuning parameters which can be used to optimize for Bayes risk in the nonstandard case. The alternative derivation argument here is believed to shed light on more general tuning processes, where the tuning is with respect to a different target than that embodied in the variational problem being solved.

2 The Support Vector Machine

The training set for an SVM consists of pairs (y_i, x_i) from one of two classes, $\mathcal{A}$, or $\mathcal{B}$. $y_i = 1$ if the ith sample is from $\mathcal{A}$, and $y_i = -1$ if the ith sample is from $\mathcal{B}$. x_i is the attribute vector for the ith sample, where x_i is in some index set $\mathcal{X}$. The SVM with reproducing kernel $K(\cdot,\cdot)$ finds the minimizer of

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

over all functions of the form $f(x) = h(x) + b$, and $h \in \mathcal{H}_K$, where $\mathcal{H}_K$ is the Reproducing Kernel Hilbert Space with Reproducing Kernel K. Here $(\tau)_+ = \tau$ if $\tau > 0$ and 0 otherwise. Once the minimizer f_λ is found, the SVM classification rule is $f_\lambda(x) > 0 \longrightarrow \mathcal{A}$, $f_\lambda(x) < 0 \longrightarrow \mathcal{B}$. f_λ will depend on λ and may depend as well on tuning parameters in K, with some abuse of notation we will let λ stand for the collective set of tuning parameters. The minimizer of (1) has the form

$$f(\cdot) = \sum_{i=1}^{n} c_i K(\cdot, x_i) + b. \tag{2}$$

Letting $e = (1, ..., 1)'$, $y = (y_1, y_2, ..., y_n)'$, $c = (c_1, c_2, ..., c_n)'$, and with some abuse of notation, letting $f = (f(x_1), f(x_2), ..., f(x_n))' = (f_1, f_2, \cdots, f_n)'$ and K now be the $n \times n$ matrix with ijth entry $K(x_i, x_j)$, we have $f = Kc + eb$, and the regularization problem (1) becomes: find (c, b) to minimize $\frac{1}{n}\sum_{i=1}^{n}(1 - y_i f_i)_+ + \lambda c'Kc$, or equivalently, find (c, b) to minimize $\frac{1}{n}\sum_{i=1}^{n}\xi_i + \lambda c'Kc$, subject to constraints: $y_i f_i \geq 1 - \xi_i, \forall i$, $\xi_i \geq 0, \forall i$. A standard way of solving this problem is to consider its dual problem. Let Y be the $n \times n$ diagonal matrix with y_i in the iith position, and let $H = \frac{1}{2n\lambda}YKY$. The dual problem has the form $\max W = -\frac{1}{2}\alpha' H\alpha + e'\alpha$ subject to $0 \leq \alpha_i \leq 1, i = 1, 2, ..., n$ and $y'\alpha = 0$. Here $\alpha = (\alpha_1, \alpha_2, ..., \alpha_n)'$. Once we get the α's, we get c's by $c = \frac{1}{2n\lambda}Y\alpha$. The issue at hand is the problem of estimating a good value of λ, and, possibly, other parameters in K, from the training set.

3 GCKL, MISCLASS, GACV, and XA

The GCKL (Generalized Kullback-Liebler Distance, see [5]) for SVM's is defined as

$$GCKL(\lambda) = E_{true}\frac{1}{n}\sum_{i=1}^{n}(1 - y_i f_{\lambda i})_+ \equiv \frac{1}{n}\sum_{i=1}^{n}\{p_i(1 - f_{\lambda i})_+ + (1 - p_i)(1 + f_{\lambda i})_+\} \tag{3}$$

where f_λ is the minimizer of (1), $f_{\lambda i} = f_\lambda(x_i)$, $p_i = p(x_i)$ is the conditional probability that $y_i = 1$ given x_i, and where the expectation is taken over new y_i's at the observed x_i's. GCKL is an upper bound for the misclassification rate (MISCLASS) (over a new set of observations with the same attribute vectors). The GACV is a computable proxy for the GCKL, that is, choosing λ (and any other tunable parameters) to minimize the GACV is supposed to come close to minimizing the GCKL. The GACV is defined as

$$GACV(\lambda) = \frac{1}{n}\left[\sum_{i=1}^{n}\xi_i + 2\sum_{y_i f_{\lambda i} < -1}\frac{\alpha_i}{2n\lambda}K_{ii} + \sum_{y_i f_{\lambda i}\in[-1,1]}\frac{\alpha_i}{2n\lambda}K_{ii}\right]. \tag{4}$$

where $\xi_i = (1 - y_i f_{\lambda i})_+$, and $K_{ij} = K(x_i, x_j)$. It was derived and studied in [7], see also [3], [4]. A more direct target is the misclassification rate, defined (conditional on the observed set of attribute variables) as

$$MISCLASS(\lambda) = E_{true}\frac{1}{n}\sum_{i=1}^{n}[-y_i f_{\lambda i}]_* \equiv \frac{1}{n}\sum_{i=1}^{n}\{p_i[-f_{\lambda i}]_* + (1-p_i)[f_{\lambda i}]_*\} \tag{5}$$

where $[\tau]_* = 1$ if $\tau \geq 0$, $= 0$ otherwise, and E_{true} has the same meaning. Joachims [1], Equation (7) proposed the $\xi\alpha$ (to be called XA_J here) proxy for MISCLASS as: $XA_J(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left[\xi_i + \rho\frac{\alpha_i}{2n\lambda}K - 1\right]_*$ where $\rho = 2$ and here (with some abuse of notation) K is an upper bound on $K_{ii} - K_{ij}$. Letting $\theta_i = \rho\frac{\alpha_i}{2n\lambda}K$, it can be shown that the sum in $XA_J(\lambda)$ counts all of the samples for which $y_i f_{\lambda i} \leq \theta_i$. Note that $y_i f_{\lambda i} > 1 \Rightarrow \alpha_i = 0$, so that XA_J may also be written

$$XA_J(\lambda) = \frac{1}{n}\left[\sum_{i=1}^{n}[-y_i f_{\lambda i}]_* + \sum_{y_i f_{\lambda i}\leq 1} I_{[\frac{\rho\alpha_i}{2n\lambda}K]}(y_i f_{\lambda i})\right], \tag{6}$$

where $I_{[\theta]}(\tau) = 1$ if $\tau \in (0, \theta]$ and 0 otherwise. Equivalently the sum in XA_J counts the misclassified cases in the training set plus all of the cases where $y_i f_{\lambda i} \in (0, \rho\frac{\alpha_i}{2n\lambda}K]$ (adopting the convention that if $f_{\lambda i}$ is exactly 0 then the example is considered misclassified). In some of his experiments Joachims (empirically) set $\rho = 1$ because it achieved a better estimate of the misclassification rate than did the XA with $\rho = 2$.

4 A Parallel Derivation for XA

The relations between XA_J and GACV that are apparent in (4) and (6) are not a coincidence. We will carry out the same argument that resulted in GACV, whose target is $E_{true}(1 - y_i f_{\lambda i})_+$ to obtain the XA (with $\rho = 1$ and K replaced by K_{ii}), whose target is $E_{true}[-y_i f_{\lambda i}]_*$. The purpose of this argument is to provide insight as to how estimates of the difference between a target and its leaving out one version may be used to construct estimates when the 'fit' is not the same as the target - here the 'fit' is $(1 - y_i f_{\lambda i})_+$, while the 'target' is $[-y_i f_{\lambda i}]_*$. We believe that this may prove to be useful in other 'tuning' problems where the target is different than the fit. We will also use the argument in a straightforward way to generalize the XA to the nonstandard case in the same way that the GACV is generalized to its nonstandard version.

Let $f_\lambda^{[-i]}$ be the minimizer of (1) with the ith data point left out, and let $f_{\lambda i}^{[-i]} = f_\lambda^{[-i]}(x_i)$. Suppose we have the approximation $y_i f_{\lambda i} \approx y_i f_{\lambda i}^{[-i]} + \theta_i$, with $\theta_i \geq 0$. A leaving out one estimate of the misclassification rate is given by $V_0 = \frac{1}{n}\sum_{i=1}^{n}[-y_i f_{\lambda i}^{[-i]}]_*$. Now $V_0 = \frac{1}{n}\sum_{i=1}^{n}[-y_i f_{\lambda i}]_* + D(\lambda)$ where $D(\lambda) =$

$\frac{1}{n}\sum_{i=1}^{n}\{[-y_i f_{\lambda i}^{[-i]}]_* - [-y_i f_{\lambda i}]_*\}$. Now, the ith term in $D(\lambda) = 0$ unless $y_i f_{\lambda i}^{[-i]}$ and $y_i f_{\lambda i}$ have different signs. For $\theta_i > 0$ this can only happen if $y_i f_{\lambda i} \in (0, \theta_i]$. Returning to the derivation of the GACV in [7], Section 16.5 (this paper is also available on the website of the first author, see Equations (26) and (29)), the approximation $\frac{f_{\lambda i} - f_{\lambda i}^{[-i]}}{y_i - \mu_\lambda^{[-i]}(f_{\lambda i})} \approx \theta_i \equiv \frac{\alpha_i}{2n\lambda} K_{ii}$, where $\mu_\lambda^{[-i]}(f_{\lambda i})$ is a function which is 0 for $f_{\lambda i} \in [-1, 1]$. This is the only case we are interested in, since if $y_i f_{\lambda i}$ is negative and θ_i is positive $y_i f_{\lambda i}$ and $y_i f_{\lambda i}^{[-i]}$ cannot have different signs, and if $y_i f_{\lambda i} > 1$, then the basis function corresponding to x_i is not a support vector and so leaving it out has no effect (that is, $\alpha_i = 0$). The conclusion is, that, to the extent that these approximations are valid, for $y_i f_{\lambda i} \in (0, 1]$, $y_i f_{\lambda i} \approx y_i f_{\lambda i}^{[-i]} + \frac{\alpha_i}{2n\lambda} K_{ii}$. This tells us that $\frac{1}{n}\sum_{y_i f_{\lambda i} \le 1} I_{[\frac{\alpha_i}{2n\lambda} K_{ii}]}(y_i f_{\lambda i})$, can be taken as an approximation to $D(\lambda)$, resulting in (our version of the)

$$XA(\lambda) = \frac{1}{n}\left[\sum_{i=1}^{n}[-y_i f_{\lambda i}]_* + \sum_{y_i f_{\lambda i} \le 1} I_{[\frac{\alpha_i}{2n\lambda} K_{ii}]}(y_i f_{\lambda i})\right], \tag{7}$$

providing an alternate derivation as well as an alternative interpretation of XA with $\rho = 1$, K replaced by K_{ii}. It can be interpreted as an approximation to a leaving out one estimate, whereas the original XA_J was derived by Joachims as an upper bound to a leaving out one estimate.

5 The Nonstandard SVM and the Nonstandard GACV

We now review the nonstandard case, from [4]. Let π_s^+ and π_s^- be the relative frequencies of the + and - classes in the training (sample) set, and let π^+ and π^- be the relative frequencies of the two classes in the target population. Let C^+ and C^- be the costs of a false positive and a false negative respectively. Let $g^+(x)$ and $g^-(x)$ be the densities of x in the + and $-$ classes respectively. Then the probability that a subject from the target population with attribute x belongs to the + class is $p(x) = \frac{\pi^+ g^+(x)}{\pi^+ g^+(x) + \pi^- g^-(x)}$, and the probability that a subject with attribute x chosen with from a population with the same distribution as the training set, belongs to the + class, is $p_s(x) = \frac{\pi_s^+ g^+(x)}{\pi_s^+ g^+(x) + \pi_s^- g^-(x)}$. Letting $\phi(x)$ be the decision rule, that is, a map from $x \in \mathcal{X}$ to $\{-1, 1\}$, the expected cost, using $\phi(x)$ is $E_{x_{true}}\{C^- p(x)[-\phi(x)]_* + C^+(1 - p(x))[\phi(x)]_*\}$, where the expectation is taken over the distribution of x in the target population. The Bayes rule, which minimizes the expected cost is $\phi(x) = +1$ if $\frac{p(x)}{1-p(x)} > \frac{C^+}{C^-}$ and -1 otherwise. Since we don't observe a sample from the true distribution but only from the sampling distribution, we need to express the Bayes rule in terms of the sampling distribution p_s. It is shown in [4] that the Bayes rule can be written in terms of p_s as $\phi(x) = +1$ if $\frac{p_s(x)}{1-p_s(x)} > \frac{C^+}{C^-}\frac{\pi_s^+}{\pi_s^-}\frac{\pi^-}{\pi^+}$ and -1 otherwise. Let $L(-1) = C^+\pi^-/\pi_s^-$ and $L(1) = C^-\pi^+/\pi_s^+$. Then the Bayes rule can be expressed as $\phi(x) = sign\left[p_s(x) - \frac{L(-1)}{L(-1)+L(1)}\right]$. [4] proposed the nonstandard SVM to handle this nonstandard case as:

$$\min \frac{1}{n}\sum_{i=1}^{n} L(y_i)[(1 - y_i f(x_i))_+] + \lambda\|h\|_{H_K}^2 \tag{8}$$

over all the functions of the form $f(x) = h(x) + b$, with $h \in H_K$. This definition is justified there by showing that, if the RKHS is rich enough and λ is chosen suitably, the minimizer of (8) tends to $f(x) = sign\left[p_s(x) - \frac{L(-1)}{L(-1)+L(1)}\right]$. The minimizer of (8) has same form as in (2). [3] show that the dual problem becomes $\max W = -\frac{1}{2}\alpha' H\alpha + e'\alpha$ subject to $0 \leq \alpha_i \leq L(y_i)$, $i = 1, 2, ..., n$, and $y'\alpha = 0$, where, once the α's are obtained, the c's are found by $c = \frac{1}{2n\lambda} Y\alpha$. The GACV for nonstandard problems was proposed there, in an argument generalizing the usual case, as:

$$GACV(\lambda) = \frac{1}{n}\left[\sum_{i=1}^{n} L(y_i)\xi_i + 2\sum_{y_i f_{\lambda i} < -1} L(y_i)\frac{\alpha_i}{2n\lambda}K_{ii} + \sum_{y_i f_{\lambda i} \in [-1,1]} L(y_i)\frac{\alpha_i}{2n\lambda}K_{ii}\right]. \tag{9}$$

It was shown to be a proxy for the GCKL given by

$$GCKL(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\{L(1)p_s(x_i)(1 - f_{\lambda i})_+ + L(-1)(1 - p_s(x_i))(1 + f_{\lambda i})_+\}. \tag{10}$$

where the expectation is with respect to new observations obtained according to the sampling distribution and the observed x_i. We now propose a generalization, BRXA, of the XA as a computable proxy for the Bayes risk in the nonstandard case. Putting together the arguments which resulted in the the GACV of (4), the XA in the form that it appears in (7) and the nonstandard GACV of (9), we obtain the BRXA:

$$BRXA(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left[L(y_i)[-y_i f_{\lambda i}]_* + \sum_{y_i f_{\lambda i} \leq 1} L(y_i) I_{[\frac{\alpha_i}{2n\lambda}K_{ii}]}(y_i f_{\lambda i}),\right]. \tag{11}$$

The BRXA is a proxy for BRMISCLASS, given by

$$BRMISCLASS(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\{L(1)p_s(x_i)[-f_{\lambda i}]_* + L(-1)(1 - p_s(x_i))[f_{\lambda i}]_*\}. \tag{12}$$

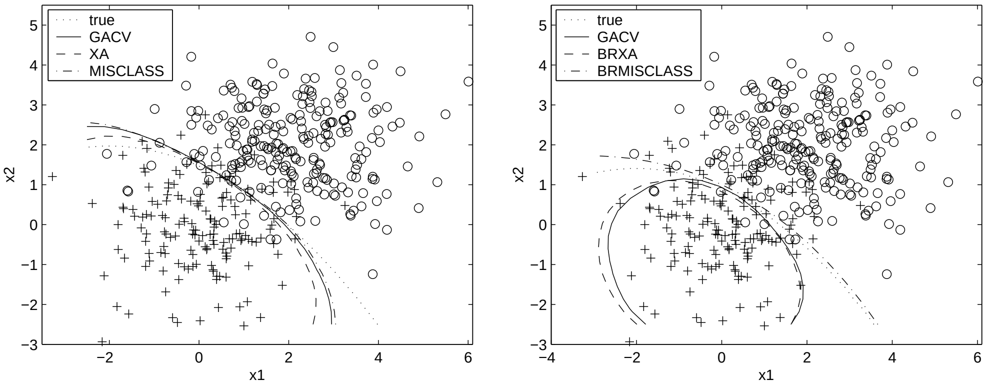

Figure 1: Observations, and true, GACV, XA and MISCLASS Decision Curves for the Standard Case (Left) and true, GACV, BRXA and BRMISCLASS Decision Curves for the Nonstandard Case (Right).

6 Simulation Results and Conclusions

Both panels of Figure 1 show the same simulated training set. The sample proportions of the + and o (-) classes are .4 and .6 respectively. The conditional distribution of x given that the sample is from the + class is bivariate Normal with mean (0,0) and covariance matrix diag (1,1). The distribution for x from the negative class is bivariate Normal with mean (2,2) and covariance diag (2,1). The left panel in Figure 1 is for the standard case, assuming that misclassification costs are the same for both kinds of misclassification, and the target population has the same proportions of the + and o as the sample. For the right panel, we assume that the costs of the two types of errors are different, and that the target population has different relative frequencies than the training set. We took $C^+ = 1$ $C^- = 2$, $\pi^+ = 0.1$, $\pi^- = 0.9$. As before, $\pi_s^+ = 0.4$, and $\pi_s^- = 0.6$, yielding $L(-1) = C^+\pi^-/\pi_s^- = 1.5$, and $L(1) = C^-\pi^+/\pi_s^+ = 0.5$. Since the distributions generating the data and the distributions of the target populations are known and involve Gaussians, the theoretical best decision rules (for an infinite future population) are known, and are given by the curves marked 'true' in both panels. The Gaussian kernel $K(x, x') = \exp\{-\|x - x'\|^2/2\sigma^2\}$ was used, where $x = (x_1, x_2)$, and σ is to be tuned along with λ. The curves selected by the GACV of (4) and the XA of (7) in the standard case are shown in the left panel, along with MISCLASS of (5), which is only known in a simulation experiment. The right panel gives the curves chosen by the nonstandard GACV of (9), the BRXA of (11) and the BRMISCLASS of (12). The optimal (λ, σ) pair in each case for the tuned curves was chosen by a global search. It can be seen from both panels in Figure 1 that the MISCLASS curve, which is based on the (finite) observed sample is quite close to the theoretical true curve (based on an infinite future population), we make this observation because it will be easier to compare the GACV and the XA against MISCLASS than against the true, similarly for the BRMISCLASS curve. In both panels it can be seen that the decision curves determined by the GACV and the XA(BRXA) are very close. We have computed the inefficiency of these estimates with respect to MISCLASS(BRMISCLASS), by inefficiency is meant the ratio of MISCLASS(BRMISCLASS) at the estimated (λ, σ) pair to its minimum value, a value of 1 means that the estimated pair is as accurate as possible, with respect to the (uncomputable) minimizer of MISCLASS(BRMISCLASS). The results for the standard case were: $GACV : 1.0064$, $XA : 1.0062 - 1.0094$ (due to multiple neighboring minima in the grid search, the 1.0062 case is in Figure 1); and for the nonstandard case: $GACV : 1.151, BRXA : 1.166$. Figure 2 gives contour plots for GCKL, GACV, BRMISCLASS and BRXA as a function of λ and σ in the nonstandard case. It can be seen that the GACV and BRXA curves have nearly the same minima. The GCKL and BRMISCLASS curves both have long, shallow, tilted cigar-shaped minima, and the GACV and BRXA minima are near the lower right end. For the standard case (not shown) the minima are somewhat more pronounced and the GACV and XA minima are closer to the MISCLASS minimum, and this is reflected in inefficiencies nearer to 1. (BR)MISCLASS curves in other simulation studies we have done show this same behavior. We have observed (as did Joachims) that the value of XA in the standard case is a good estimate of the value of MISCLASS at its minimizer, only slightly pessimistic, one-half value of GACV (which should be divided by 2) is somewhat more pessimistic. We note that once one obtains the solution to the problem the computation of both GACV and (BR)XA are equally trivial. The GACV in (quadratically) penalized likelihood cases generally hits the minimizer of its target (analogous to GCKL)(see [6]) but here, both the GACV and the BRXA (along with the standard case) appear to be biased towards larger λ. The (BR)MISCLASS surfaces are so flat in λ in our examples this does not seem to be a serious problem (less so in the standard case). In the absence of a possible second order correction to these estimates, we believe that these two estimates will prove to be extremely useful as internal tuning methods.

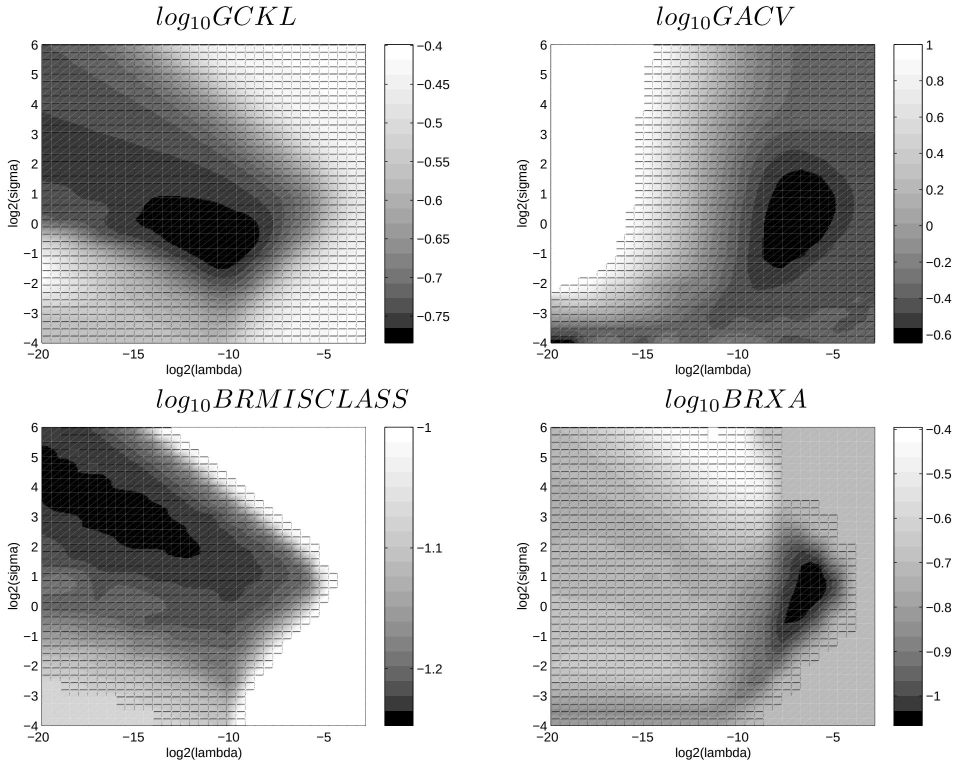

Figure 2: GCKL, GACV, BRMISCLASS, BRXA as functions of λ and σ^2, for the nonstandard example. Note different logarithmic scales in λ and σ.

References

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

[2] Y. Lin. Support vector machines and the Bayes rule in classification. Technical Report 1014, Department of Statistics, University of Wisconsin, Madison WI, to appear, *Data Mining and Knowledge Discovery*, 1999.

[3] Y. Lin, Y. Lee, and G. Wahba. Support vector machines for classification in nonstandard situations. Technical Report 1016, Department of Statistics, University of Wisconsin, Madison WI, 2000. To appear, *Machine Learning*.

[4] Y. Lin, G. Wahba, H. Zhang, and Y. Lee. Statistical properties and adaptive tuning of support vector machines. Technical Report 1022, Department of Statistics, University of Wisconsin, Madison WI, 2000. To appear, *Machine Learning*.

[5] G. Wahba. Support vector machines, reproducing kernel Hilbert spaces and the randomized GACV. In B. Scholkopf, C. Burges, and A. Smola, editors, *Advances in Kernel Methods-Support Vector Learning*, pages 69–88. MIT Press, 1999.

[6] G. Wahba, X. Lin, F. Gao, D. Xiang, R. Klein, and B. Klein. The bias-variance tradeoff and the randomized GACV. In M. Kearns, S. Solla, and D. Cohn, editors, *Advances in Information Processing Systems 11*, pages 620–626. MIT Press, 1999.

[7] 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.