SELECTION OF TUNING PARAMETERS FOR SUPPORT VECTOR MACHINES

Victor Solo

Department of Electrical Engineering and Computer Science
University of Michigan, Ann Arbor
Ann Arbor, MI, 48109
email: vsolo@umich.edu

ABSTRACT

Support Vector machines have become important in classification, biometrics, machine learning and pattern recognition. But successful application requires selection of various tuning parameters such as kernel parameters and penalty or margin parameters. We apply a new technique for this problem which provides very simple structure for the automatic selector.

1. INTRODUCTION

In the last decade or so significant new methods have been developed for pattern recognition or classification especially within the machine learning literature. Certainly one of the main new techniques is support vector machines (SVMs) [1],[2],[3].

The pattern recognition problem of interest is a so-called supervised learning problem where training data pairs $(y_i, x_i), i = 1, \cdots, N$ are observed. Here $y_i = 0, 1$ are class labels (below we also use $\overline{y}_i = 2y_i - 1 = -1, 1$) and x_i are variously called predictors,covariates,features. The aim is to construct a regression function $f(x)$ whose associated classifier, say $sign(f(x))$ minimizes a criterion such as the misclassification error (when $f(x) \neq \overline{y}$) on future cases.

This function estimation problem based on binary dependent data is an ill-condiitoned inverse problem [4] and its solution requires that the class of functions being fitted be constrained in some way. This is typically managed by a penalty term which measures the size of the function in some way and whose weighting is controlled by a penalty parameter to be chosen by the user.

2. SUPPORT VECTOR MACHINES

It is thus satisfying to discover as shown by [5] (see also [2]) that the SVM problem [3] can be formulated as a penalised optimization problem. We represent the function by a basis expansion $f(x) = \beta_0 + \psi(x)^T\beta$ and choose β, β_0 to solve the problem

$$\begin{aligned} \hat{\beta}, \hat{\beta}_0 &= \min_{\beta,\beta_0} \overline{J}(\beta_0, \beta) \qquad (2.1) \\ \overline{J}(\beta_0, \beta) &= \Sigma_1^N |1 - \overline{y}_i f_i|_+ + \frac{\lambda}{2} \parallel \beta \parallel^2 \\ f_i &= f(x_i) \\ |u|_+ &= u, u \geq 0; = 0, \text{ otherwise} \end{aligned}$$

Here λ is a penalty parameter which controls the magnitude of the basis coefficients. The maginitude of λ has a profound effect on the results and successful implementation requires its choice. To continue we must be more specific about the basis expansion. The case we pursue here is where the basis arises as an eigen-expansion of a positive definite kernel

$$K(x, x^{'}) = \Sigma_1^\infty \phi_k(x)\phi_k(x^{'})\lambda_k$$

and where $\psi(x) = \sqrt{\lambda_k}\phi_k(x)$. Popular kernels are [2] the Radial basis kernel:

$$K_\sigma(x, x^{'}) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{\| x - x^{'} \|^2}{2\sigma^2}}$$

and the polynomial kernel:

$$\begin{aligned} K_d(x, x^{'}) &= (1 + < x, x^{'} >)^d \\ < x, x^{'} > &= \frac{1}{4} \| x + x^{'} \|^2 + \| x - x^{'} \|^2 \end{aligned}$$

We see that each of these kernels depends on an additional tuning parameter σ or d.

Thus to successfully solve the SVM problem we must select values for two tuning parameters λ and σ ot λ and d.

Continuing, the penalised problem can be reformulated [5],[2] as a problem of estimating N+1 Lagrange multipliers α_i, α_0 by solving

$$\begin{aligned} \hat{\alpha}, \hat{\alpha}_0 &= \min_{\alpha,\alpha_0} J(\alpha_0, \alpha) \qquad (2.2) \\ J(\alpha_0, \alpha) &= \Sigma_1^N |1 - \overline{y}_i f_i|_+ + \frac{\lambda}{2} \alpha^T K \alpha \\ f_i &= f(x_i) \end{aligned}$$

$$
\begin{aligned}
f(x) &= \alpha_0 + \Sigma_1^N \alpha_i K(x, x_i) \\
K &= [K(x_i, x_j)]
\end{aligned}
$$

3. CHOICE OF TUNING PARAMETERS

Methods for choosing regularizing parameters in ill-conditioned inverse problems are reviewed in [6] where the approach to be applied here was developed. Methods such as AIC [7] or MDL [8] are not obviously applicable because they require the tuning parameter to be a model dimension.A method such as cross-validation [7] would be computationally demanding because it requires that the SVM problem be solved over and over as data points are left out one at a time. Sometimes Taylor series expansion can be used to finesse this problem and such an approach is pursued by [5]. The criterion developed there is not computationally demanding but is completely different from that developed below. The problem has also been discussed recently by [9]. The emphasis there is on development of a steepest descent procedure for optimizing a selection criterion with respect to relevant tuning parameters. But the criterion being optimized is itself computationally demanding. Bayesian method could also be applied but in general require a huge developmental and computational investment.

The advantage of the approach pursued here, is its easy applicability to non-linear problems and that the selection criteria are often computationally very simple. The technique has been successfully applied to a number of problems: threshold selection for wavelets in coloured noise [10];penalty parameter selection for total variation denoising in white noise [11] and coloured noise [12];estimation of neighbourhood size in optical flow [13] and robust optical flow [14];selection of stopping iteration for anisotropic diffusion [15],[16].

To apply our approach in the current setting we first need to put the problem into a statistical setting. The traditional statistical approach to binary function estimation is by logistic regression [2]. The model is that

$$
p_i = P(Y_i = 1|x_i) = \frac{1}{1 + e^{-f_i}}
$$

While logistic regression provides a binary function estimator that performs very well in practice an important difference from SVM is that the SVM fit often uses only a fraction of the data. Of interest to us here is the fitting procedure for logistic regression. This is an iterively reweighted least squares technique which, as with many non-linear regression problems, can be interpreted as as a weighted least squares fit to 'pseudo' Gaussian data [2]. Given a current estimate of the coefficients α_0, α, the pseudo data is

$$
z_i = \alpha_0 + k_i^T \alpha + \frac{y_i - p_i}{p_i(1 - p_i)}, i = 1, \cdots, N \qquad (3.1)
$$

where here, k_i is the ith column of K. And the next iterate is

$$
(\alpha_0, \alpha)^T = (X^T W X)^{-1} X^T W z
$$

where $X = K$,$z = (z_1, \cdots, z_N)^T$. and $W = diag(w_i), w_i = p_i(1 - p_i)$. Our point of departure now is (3.1) wherein we will henceforth treat z as being Gaussian.

We emphasize that this reference to the logistic framework is only for theoretical development. We do not fit the logistic model at any stage. We rewrite the pseudo Gaussian model as

$$
\begin{aligned}
z &= \mu + n \\
\mu &= 1\alpha_0 + K\alpha, 1 \text{ is a vector of } 1s \\
n &\sim N(0, \Omega) \\
\Omega &= W^{-1} = diag(\frac{1}{p_i(1 - p_i)})
\end{aligned}
$$

This is justified since y given x is a Bernoulli random variable with mean p and variance $p(1 - p)$. We use a simple quadratic measure of reconstruction quality.

$$
R_{\lambda,\sigma} = E[(\hat{\mu} - \mu)\hat{\Omega}^{-1}(\hat{\mu} - \mu)]
$$

where $\hat{\mu} = 1\hat{\alpha}_0 + K\hat{\alpha}$ and we have used variance equalizing weighting due to (3.1). Also quantities with a ˆ refer to (2.2). The criterion partly measures the quality of the regression function but the weighting modifies this substantially to account for the importance of classification probability.

Ideally we would choose (λ, σ) to minimize $R_{\lambda,\sigma}$. However $R_{\lambda,\sigma}$ cannot be computed if only because $f(x)$ is unknown.

We try then to find an empirically computable, statistically unbiassed estimator of $R_{\lambda,\sigma}$, call it $\hat{R}_{\lambda,\sigma}$, and minimize that instead. The unbiassedness is important since that ensures that on average the minimizer of $\hat{R}_{\lambda,\sigma}$ should be close to the minimizer of $R_{\lambda,\sigma}$. Using only the Gaussian assumption ,a simple integration by parts argument and some differentiations involving the Euler equation for (2.1) it can be shown from results in [6] that an empirically computable (nearly) unbiassed estimator of $R_{\lambda,\sigma}$ is

$$
\begin{aligned}
\hat{R}_{\lambda,\sigma} &= \chi^2_{\lambda,\sigma} + mc_{\lambda,\sigma} && (3.2) \\
\chi^2_{\lambda,\sigma} &= \Sigma_1^N \frac{(y_i - \hat{p}_i)^2}{\hat{p}_i(1 - \hat{p}_i)} && (3.3) \\
mc_{\lambda,\sigma} &= \frac{4}{\lambda}\Sigma_1^N(\hat{w}_i s_i K(x_i, x_i)) && (3.4) \\
&= \frac{4}{\sqrt{2\pi}\sigma\lambda}\Sigma_1^N(\hat{w}_i s_i), \text{ for RBF} \\
s_i &= 1, \text{ if } \overline{y}_i f_i < 1; \text{ else } = 0 \\
\hat{f}_i &= \hat{\alpha}_0 + k_i^T \hat{\alpha} \\
\hat{p}_i &= \frac{1}{1 + e^{-\hat{f}_i}} \\
\hat{w}_i &= \hat{p}_i(1 - \hat{p}_i)
\end{aligned}
$$

We interpret $mc_{\lambda,\sigma}$ as a model complexity term. The idea then is to plot $\hat{R}_{\lambda,\sigma}$ for a minimum in λ, σ. We note the extreme simplicity of the criterion. The probabilities are obtained from the fitted SVM model while the model complexity term is a simple well conditioned sum. The additional computation required once the SVM is fitted is miniscule. The criterion (3.2,3.3,3.4) applies to any kernel and a similar expression applies when a finite basis expansion is used (whose tuning parameter is then the number of basis coefficients).

We call this selector SURE (Stein's unbiassed risk estimator for the originator, in a different context, of the integration by parts argument: see [6]).

4. RESULTS

We now develop a small simulation study to illustrate the new method. Our simulation example is a simple example of those typically used. The two class density functions are each a mixture of Gaussians,

$$\begin{aligned} p_1(x) &= \frac{1}{2}N(\mu_1, \gamma^2 I) + \frac{1}{2}N(-\mu_1, \gamma^2 I) \\ p_0(x) &= \frac{1}{2}N(\mu_2, \gamma^2 I) + \frac{1}{2}N(-\mu_2, \gamma^2 I) \\ \mu_1 &= (2,2)^T, \mu_0 = (-2,2)^T, \gamma^2 = 1; \end{aligned}$$

We simulated a sample of 100 independent samples from each class density and used these $n = 200$ pairs as training data. We solved the SVM problem with a simple steepest descent algorithm. As long as one uses small steps this works well.

Fig.1 shows plots of $\hat{R}_{\lambda,\sigma}$ for a range of λ, σ values. We see a joint minimum with respect to λ, σ near $\lambda, \sigma = 1.1, .7$ or $1.1, .6$. In practice one should investigate fits in the vicinity of the minimum. For this reason a procedure that only supplies a point minimum will be inadequate in general. Also sampling fluctuations can sometimes deliver local minima, as is evident in Fig 2. Also Fig 2 shows the individual components of the SURE criterion as well as the number of retained training pairs as a function of λ. We see that about $\frac{3}{4}$ of the data pairs are zeroed out near the minimum.

Finally Fig 3. shows details of iterations near the minumum (λ, σ) combination. We see convergence is rapid.

5. SUMMARY

In this work we have presented a new criterion for choosing tuning parameters (specifically a kernel parameter and a penalty parameter) in support vector machine classification. The criterion requires minimal additonal computation once a SVM model has been fitted. We have illustrated the method by applying it to a simulated example to choose

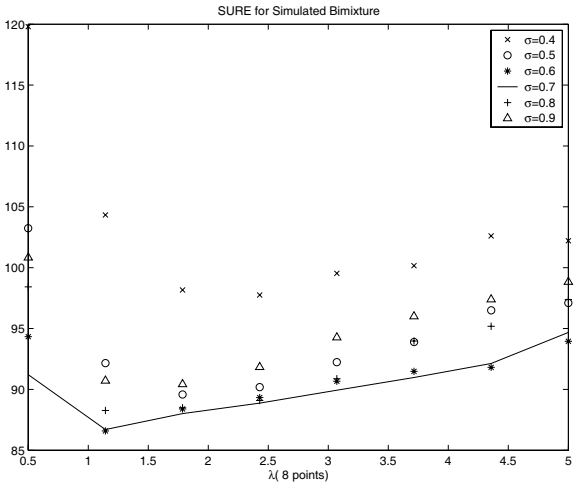

Fig. 1. $\hat{R}_{\lambda,\sigma}$ versus λ, σ.

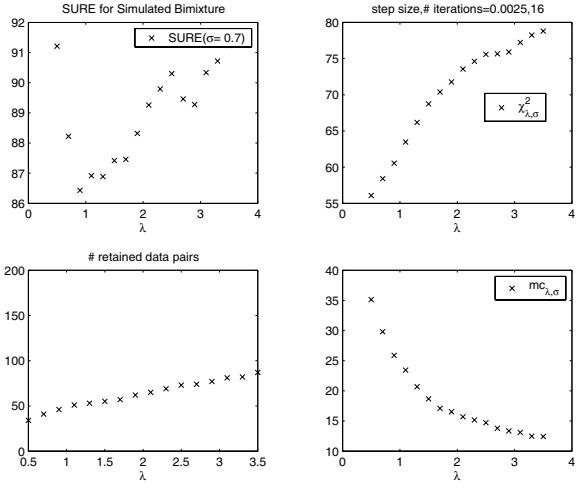

Fig. 2. Iterates of $\hat{R}_{\lambda,\sigma}$ and related quantities.

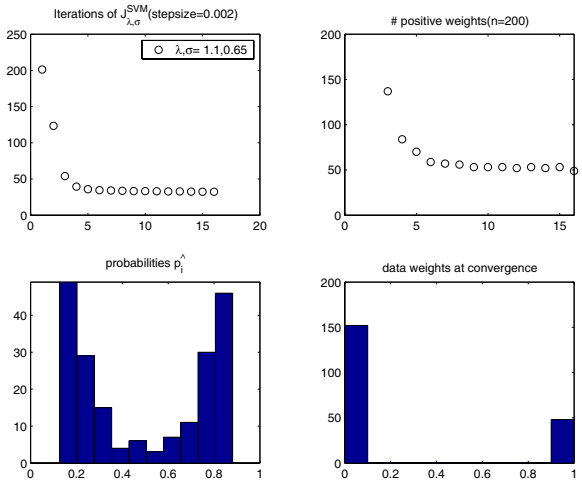

Fig. 3. Convergence of Steepest Descent Iteration for solving SVM problem.

both the kernel width for a radial basis function kernel and a penalty parameter related to the classifier margin.

6. REFERENCES

[1] B Scholkopf and A Smola, *Learning with Kernels*, MIT Press, Cambridge MA, 2000.

[2] T Hastie, R Tibshirani, and J Friedman, *The Elements of Statistical Learning*, Springer, New York, 2001.

[3] V N Vapnik, *The Nature of Statistical Learning Theory*, Springer, New York, 1995.

[4] V N Vapnik, *Estimation of Dependences Based on Empirical Data*, Springer-Verlag, Berlin, 1982.

[5] G Wahba, Y Lin, and H Zhang, "Gacv for support vector machines," in *Advances in Large Margin Classifiers*. MIT Press, 2000, pp. 297–311.

[6] V. Solo, "A SURE-fired way to choose smoothing parameters in ill-conditioned inverse problems," in *Proc. IEEE ICIP96*. IEEE, 1996, IEEE Press.

[7] H. Linhart and W. Zucchini, *Model selection*, J. Wiley, New York, 1986.

[8] J Rissanen, *Stochastic Complexity in Statistical Enquiry*, World Scientific, Singapore, 1989.

[9] O Chapelle, V Vapnik, O Bousquet, and S Mukherjee, "Choosing multiple tuning parameters for Support Vector Machines," *Mach Learn*, vol. 46, pp. 131–159, 2002.

[10] V. Solo, "Wavelet signal estimation in coloured noise with extension to transfer function estimation," in *Proc 37th IEEE CDC*, Tampa, FL, 1998, IEEE.

[11] V Solo, "Selection of regularization parameters for total variation denoising," in *Proc. ICASSP99*. IEEE, 1999.

[12] V. Solo, "Total variation denoising in coloured noise," in *Proc ICASSP2000*, Istanbul, Turkey, 2000, IEEE.

[13] L Ng and V Solo, "Errors-in-variables modelling in optical flow estimation," *IEEE Trans. Im.Proc.*, vol. 10, pp. 1528–1540, 2001.

[14] M Shi and V Solo, "Empirical choice of smoothing parameters in optical flow with correlated errors," submitted to IEEE ICASSP 2003, April 2003, Hong Kong, 2003.

[15] V. Solo, "Automatic stopping criterion for anisotropic diffusion," in *Proc ICASSP2001*, Salt Lake City, Utah, 2001, IEEE.

[16] V Solo, "A fast automatic stopping criterion for anisotropic diffusion," in *Proc ICASSP02 Orlando FL.* IEEE, 2002.