

A New Implementation for SVM Regression Based on Mean Field Analysis

J.B. Gao, S.R. Gunn, and C.J. Harris
Image, Speech and Intelligent System Research Group
Department of Electronics and Computer Science
University of Southampton, Southampton SO17 1BJ, UK.
Email: {jg,srg,cjh}@ecs.soton.ac.uk

Abstract

This paper deals with two subjects. First, we will show how support vector machine (SVM) regression problem can be solved as the maximum a posteriori (MAP) prediction in the Bayesian framework. The second part describes an approximation technique that is useful in performing calculations for SVMs based on the mean field algorithm which was originally proposed in Statistical Physics of disordered systems. An advantage is that the approach makes a better approximation to the mean and variance of SVM posterior, with respect to previous approaches, which are analytically intractable.

1 Introduction

Recently, there has been a great deal of interest in non-parametric Bayesian approaches to regression and classification problems which are based on the concept of Gaussian processes [9]. It is well-known that Support Vector Machines (SVM) can be interpreted as the maximum a posteriori (MAP) prediction with a Gaussian prior, under the Bayesian framework so that some statistical quantities such as error bars can be determined, [3, 8]. Instead of defining prior distributions over parameters of a learning machine, one directly defines a Gaussian prior over the function space on which the machine computes. These ideas provide a probabilistic interpretation for regression and classification problems. The standard SVM solution can be obtained by a quadratic optimization algorithm that maximizes the posterior distribution through a dual optimisation problem. This is closely related to other kernel-based methods [7].

Bayesian methods have a number of virtues, particularly their uniform treatment of uncertainty at all levels of the modelling process. Another virtue of Bayesian framework is that it gives prediction statistics so that one can easily obtain error bars. The Bayesian method has been successfully applied to the L_2 network regularisation and some classification cases with a Gaussian prior [9]. The main difficulty in adopting a probabilistic framework for SVMs is due to the SVM likelihood (loss) function, i.e., the non-normalized likelihood in SVM classification and the likelihood defined by Vapnik's ϵ -insensitive loss function which results in an intractable high-dimension integral. In the L_2 network, the likelihood on the training dataset is a Gaussian, so that the posterior distribution given data is also a Gaussian when a Gaussian prior distribution is given. However, for some statistical models like the ones used for classification or SVMs, the high dimensional integrals which occur in performing *a posteriori* averages can only be treated by approximative methods. An approximation to these integrations can be based on Markov Chain Monte Carlo (MCMC) sampling which, for large data sets, may be time consuming. There are

other possible approaches which can be used to approximate the posterior distribution or its statistical average, such as variational algorithm [5] or Laplace's methods. The Laplace approximation has been used for both classification problems [9], SVR problems [3] and SVC problems [8]. In an SVM the MAP solution has to be determined by a Quadratic Programming (QP) problem which is very time consuming when dealing with a large training dataset and then the posterior distribution is simply approximated by a Gaussian distribution centered at the MAP solution based on the first order expansion. We should notice here that such an approximation is only needed when we try to put the SVM in a Bayesian framework.

2 SVR with Gaussian Prior

Consider the supervised learning problem: A training set, $\mathcal{D} = \{(\mathbf{x}_i, t_i) | i = 1, 2, \dots, N\}$ of input vectors \mathbf{x}_i and associated targets t_i is given and the goal is to infer the output t for a new input \mathbf{x} . Here we consider the special case of the SVR problem with Vapnik's ϵ -insensitive loss function defined as

$$L_\epsilon(t - y(\mathbf{x})) = L(t, y(\mathbf{x})) = \begin{cases} 0 & |t - y(\mathbf{x})| \leq \epsilon, \\ |t - y(\mathbf{x})| - \epsilon & |t - y(\mathbf{x})| > \epsilon \end{cases} \quad (2.1)$$

where $\epsilon \geq 0$ is a prespecified constant controlling the noise tolerances. There is no penalty at \mathbf{x}_i when $|t_i - y(\mathbf{x}_i)| \leq \epsilon$.

In order to construct a Bayesian framework under Vapnik's ϵ -insensitive loss function L_ϵ , we employ the probabilistic model in which the probability of the output t , the likelihood $P[t|y(\mathbf{x})]$ is assumed by the following relationship

$$P[t|y(\mathbf{x})] = \frac{C}{2(\epsilon C + 1)} \exp\{-CL_\epsilon(t - y(\mathbf{x}))\}. \quad (2.2)$$

Thus equation (2.2) can be interpreted as an additive noise model of the target t . It has been shown that the standard SVR framework is a special case of regularisation network [2] with this particular noise model.

The probabilistic interpretation of SVRs can be regarded as the following likelihood defined by the ϵ -insensitive loss function,

$$P[\mathcal{D}|\mathbf{y}(\mathbf{X})] = \left[\frac{1}{2} \frac{C}{\epsilon C + 1} \right]^N \exp \left\{ -C \sum_{i=1}^N L_\epsilon(t_i - y(\mathbf{x}_i)) \right\} \quad (2.3)$$

where $\mathbf{y}(\mathbf{X}) = [y(\mathbf{x}_1), y(\mathbf{x}_2), \dots, y(\mathbf{x}_N)]$. We take the prior probability distribution $P[y(\mathbf{x})]$ as a functional Gaussian process. For any finite point set $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N]$ the prior probability distribution $P[\mathbf{y}(\mathbf{X})]$ is specified as a Gaussian process with a zero mean and a covariance function $K(\mathbf{x}, \mathbf{x}')$,

$$P[\mathbf{y}(\mathbf{X})] = \frac{1}{\sqrt{\det 2\pi K_N}} \exp \left\{ -\frac{1}{2} \mathbf{y}(\mathbf{X})^T K_N^{-1} \mathbf{y}(\mathbf{X}) \right\}. \quad (2.4)$$

where $K_N = [K(\mathbf{x}_i, \mathbf{x}_j)]$ is the covariance matrix at the points \mathbf{X} .

Combining (2.3) and (2.4) and applying Bayes' rule results in the following posterior about the process function $\mathbf{y}(\mathbf{X})$,

$$P[\mathbf{y}(\mathbf{X})|\mathcal{D}] \propto \exp \left\{ -C \sum_{i=1}^N L_\epsilon(t_i - y(\mathbf{x}_i)) - \frac{1}{2} \mathbf{y}(\mathbf{X})^T K_N^{-1} \mathbf{y}(\mathbf{X}) \right\} \quad (2.5)$$

where $K(C, \epsilon) = \frac{C}{2(\epsilon C + 1)}$ and $P[\mathcal{D}]$ is the normalization constant. It is obvious that the MAP estimate of the posterior distribution $P[\mathbf{y}(\mathbf{X})|\mathcal{D}]$ is the minimizer of

$$\min_{\mathbf{y}(\mathbf{X})} C \sum_{i=1}^N L_\epsilon(t_i - y(\mathbf{x}_i)) + \frac{1}{2} \mathbf{y}(\mathbf{X})^T K_N^{-1} \mathbf{y}(\mathbf{X}) \quad (2.6)$$

This is the original SVM setting, and (2.6) can be converted into a quadratic program (QP) [7, 4] by introducing some slack variables and dual variables. Due to the size of the optimization problems arising from the SVM, one has to pay special attention as to how these problems can be solved efficiently. Several algorithms can be used to solve the quadratic programming problem arising in SVR. Most of them can be shown to share some common strategy that can be understood well in the view of duality theory. These algorithms include the interior point algorithm, the subset selection algorithms, the Sequential Minimal Optimization (SMO) see survey [7]. However, all of these algorithms are limited by the scale of the problem and also require *a priori* control factor C and precision size ϵ .

3 Mean Field Theory of SVR

The calculation of the posterior average needed to derive Bayes algorithm is typically intractable and approximation techniques are required. Recently Opper and Winther [6] have introduced an advanced mean field theory approach based on ideas of statistical mechanics to cope with the Gaussian classification problem. This approach is equivalent to the so-called TAP mean field theory.

In this section we will follow the discussion in [6]. From the posterior distribution defined in (2.5) the prediction on a new test input \mathbf{x} is given by

$$\langle y(\mathbf{x}) \rangle = \frac{K(C, \epsilon)^N}{\sqrt{\det(2\pi K_{N+1}^{-1})}} \int y(\mathbf{x}) \frac{\exp \left\{ -C \sum_{i=1}^N L_\epsilon(t_i - y(\mathbf{x}_i)) - \frac{1}{2} \mathbf{Y}^T K_{N+1}^{-1} \mathbf{Y} \right\}}{P[\mathcal{D}]} d\mathbf{Y} \quad (3.1)$$

where $\mathbf{Y} = [y(\mathbf{x}_1), \dots, y(\mathbf{x}_N), y(\mathbf{x})]^T$ and K_{N+1} is obtained from K_N with an additional row and column of $k_N(\mathbf{X}) = [K(\mathbf{x}_1, \mathbf{x}), K(\mathbf{x}_2, \mathbf{x}), \dots, K(\mathbf{x}_N, \mathbf{x})]$. Note that

$$y(\mathbf{x}) \exp \left\{ -\frac{1}{2} \mathbf{Y}^T K_{N+1}^{-1} \mathbf{Y} \right\} = \sum_{i=1}^{N+1} K(\mathbf{x}, \mathbf{x}_i) \frac{\partial}{\partial y(\mathbf{x}_i)} \exp \left\{ -\frac{1}{2} \mathbf{Y}^T K_{N+1}^{-1} \mathbf{Y} \right\}$$

where \mathbf{x} denotes \mathbf{x}_{N+1} , then by substituting the above relation into (3.1) and applying integration by parts,

$$\langle y(\mathbf{x}) \rangle = \sum_{i=1}^N K(\mathbf{x}, \mathbf{x}_i) w_i, \quad (3.2)$$

where w_i s are constant defined as

$$w_i = \frac{K(C, \epsilon)^N}{P[\mathcal{D}]} \int N(\mathbf{y}(\mathbf{X}) | \mathbf{0}, K_N) \frac{\partial}{\partial y(\mathbf{x}_i)} \exp \left\{ -C \sum_{j=1}^N L_\epsilon(t_j - y(\mathbf{x}_j)) \right\} d\mathbf{y}(\mathbf{X}). \quad (3.3)$$

Let us define a new distribution for each i as follows

$$P[y(\mathbf{x}_i) | \overline{\mathcal{D}}_i] = \frac{\int N(\mathbf{y}(\mathbf{X}) | \mathbf{0}, K_N) \exp \left\{ -C \sum_{j \neq i} L_\epsilon(t_j - y(\mathbf{x}_j)) \right\} d\mathbf{y}(\overline{\mathbf{X}}_i)}{\int N(\mathbf{y}(\mathbf{X}) | \mathbf{0}, K_N) \exp \left\{ -C \sum_{j \neq i} L_\epsilon(t_j - y(\mathbf{x}_j)) \right\} d\mathbf{y}(\mathbf{X})} \quad (3.4)$$

where $\overline{\mathcal{D}}_i$ and $\overline{\mathbf{X}}_i$ are obtained by removing the data pattern (\mathbf{x}_i, t_i) from \mathcal{D} . In fact $P[y(\mathbf{x}_i) | \overline{\mathcal{D}}_i]$ is the predictive distribution at the ‘‘test’’ point \mathbf{x}_i given the dataset $\overline{\mathcal{D}}_i$. Denoting an average with respect to this predictive distribution by $\langle \dots \rangle_i$ we can rewrite the coefficient in (3.3) as,

$$w_i = \frac{\left\langle K(C, \epsilon) \frac{\partial}{\partial y(\mathbf{x}_i)} \exp \{ -C L_\epsilon(t_i - y(\mathbf{x}_i)) \} \right\rangle_i}{\left\langle K(C, \epsilon) \exp \{ -C L_\epsilon(t_i - y(\mathbf{x}_i)) \} \right\rangle_i} \quad (3.5)$$

The magnitude of w_i can be interpreted as the normalised variant rate of the likelihood. Thus the weight coefficients in the SVM solution (3.2) can be determined by the likelihood variant rates with respect to the local predictive distribution $P[y(\mathbf{x}_i) | \overline{\mathcal{D}}_i]$. In order to calculate such weights, a simple and direct method is to apply some Gaussian approximation to the local predictive distribution as follows

$$P[y(\mathbf{x}_i) | \overline{\mathcal{D}}_i] \approx \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp \left\{ -\frac{(y(\mathbf{x}_i) - \langle y(\mathbf{x}_i) \rangle_i)^2}{2\sigma_i^2} \right\} \quad (3.6)$$

with the variance defined as $\sigma_i^2 = \langle y(\mathbf{x}_i)^2 \rangle_i - \langle y(\mathbf{x}_i) \rangle_i^2$. Inserting (3.6) into (3.5) we derive the following expression

$$w_i \approx \frac{F(\langle y(\mathbf{x}_i) \rangle_i, \sigma_i^2)}{G(\langle y(\mathbf{x}_i) \rangle_i, \sigma_i^2)} \quad (3.7)$$

where F and G are computed, respectively, by the following explicit formula,

$$F(\langle y(\mathbf{x}_i) \rangle_i, \sigma_i^2) = \frac{C}{2} \beta_+^i \exp \left\{ \frac{C}{2} \alpha_+^i \right\} - \frac{C}{2} \beta_-^i \exp \left\{ \frac{C}{2} \alpha_-^i \right\} \quad (3.8)$$

$$G(\langle y(\mathbf{x}_i) \rangle_i, \sigma_i^2) = \gamma_+^i - \gamma_-^i + \frac{\beta_+^i}{2} \exp \left\{ \frac{C}{2} \alpha_+^i \right\} + \frac{\beta_-^i}{2} \exp \left\{ \frac{C}{2} \alpha_-^i \right\} \quad (3.9)$$

where $\alpha_{\pm}^i = 2\epsilon + C\sigma_i^2 \pm (2\langle y(\mathbf{x}_i) \rangle_i - 2t_i)$, $\beta_{\pm}^i = 1 - \operatorname{erf}\left[\frac{\epsilon + C\sigma_i^2 \pm (\langle y(\mathbf{x}_i) \rangle_i - t_i)}{\sqrt{2\sigma_i^2}}\right]$ and $\gamma_{\pm}^i = \frac{1}{2}\operatorname{erf}\left[\frac{t_i - \langle y(\mathbf{x}_i) \rangle_i \pm \epsilon}{\sqrt{2\sigma_i^2}}\right]$. Equations (3.7), (3.8) and (3.9) are called the mean field equations corresponding to the weight parameters. In order to work out the weight coefficients, one has to determine the local predictive average $\langle y(\mathbf{x}_i) \rangle_i$ and variance σ_i^2 in the approximated Gaussian (3.6). Recently Opper and Winther [6] have derived an effective mean field equation for both $\langle y(\mathbf{x}_i) \rangle_i$ and σ_i^2 by the TAP linear respondent approach. Denote by $\langle y(\mathbf{x}_i) \rangle$ the posterior average at \mathbf{x}_i which is given by (3.2) as $\langle y(\mathbf{x}_i) \rangle = \sum_{j=1}^N K(\mathbf{x}_i, \mathbf{x}_j)w_j$, then the formulas for $\langle y(\mathbf{x}_i) \rangle_i$ and σ_i^2 can be explicitly represented as,

$$\langle y(\mathbf{x}_i) \rangle_i \approx \langle y(\mathbf{x}_i) \rangle - \sigma_i^2 w_i, \quad \sigma_i^2 \approx \frac{1}{[(\Sigma + K)^{-1}]_{ii}} - \Sigma_i, \quad (3.10)$$

with $\Sigma = \operatorname{diag}(\Sigma_1, \Sigma_2, \dots, \Sigma_N)$ and $\Sigma_i = -\sigma_i^2 - \left(\frac{\partial w_i}{\partial \langle y(\mathbf{x}_i) \rangle_i}\right)^{-1}$. An explicit expression for $\frac{\partial w_i}{\partial \langle y(\mathbf{x}_i) \rangle_i}$ can be obtained from (3.7) as

$$\frac{\partial w_i}{\partial \langle y(\mathbf{x}_i) \rangle_i} \approx C^2 - w_i^2 - \frac{w_i \langle y(\mathbf{x}_i) \rangle_i + \sigma_i^2 C^2 \int_{t_i - \epsilon}^{t_i + \epsilon} P[y(\mathbf{x}_i) | \overline{\mathcal{D}}_i] dy(\mathbf{x}_i)}{\sigma_i^2 G(\langle y(\mathbf{x}_i) \rangle_i, \sigma_i^2)} \quad (3.11)$$

Since $P[y(\mathbf{x}_i) | \overline{\mathcal{D}}_i]$ is assumed to be Gaussian, the integral in (3.11) can easily be computed using the error function.

4 Algorithm for the SVR Mean Field Equations

The required information needed for prediction at a test input \mathbf{x} in (3.2) is w_i and the local predictive posterior average $\langle y(\mathbf{x}_i) \rangle_i$ and the local predictive posterior variance σ_i^2 . These variables satisfy the non-linear mean field equations (3.7) and (3.10) etc. The mean field equation can be solved by an iteration method,

1. Initialization: set the learning rate η , e.g. $\eta = 0.05$, and draw w_i from the prior;
2. Calculate the Kernel matrix K and let $\sigma_i^2 = K_{ii}$;
3. Iterate steps 4 to 6 until the change in w_i is below a given tolerance;
4. For $i = 1, \dots, N$ do

$$\langle y(\mathbf{x}_i) \rangle := \sum_{j=1}^N K(\mathbf{x}_i, \mathbf{x}_j)w_j \text{ and } \langle y(\mathbf{x}_i) \rangle_i := \langle y(\mathbf{x}_i) \rangle - \sigma_i^2 w_i$$

Compute F_i and G_i by (3.8) and (3.9)

5. Update w_i by $w_i := w_i + \eta(F_i/G_i - w_i)$
6. For every M iterations of w_i , update

$$Dw_i := C^2 - w_i^2 - \frac{w_i \langle y(\mathbf{x}_i) \rangle_i + \sigma_i^2 C^2 (\gamma_+^i - \gamma_-^i)}{\sigma_i^2 G(\langle y(\mathbf{x}_i) \rangle_i, \sigma_i^2)}$$

$$\Sigma_i := -\sigma_i^2 - 1/Dw_i; \quad \sigma_i^2 := \frac{1}{[(\Sigma + K)^{-1}]_{ii}} - \Sigma_i$$

In the above iteration the steps 4 and 5 are called the inner iteration and step 6 the outer iteration. It is obvious that the most expensive step in the above mean field algorithm is the inversion of the matrix $K + \Sigma$ in the outer iteration cycle. To save computing time we choose to make less iterations for the outer iteration than for the inner iteration. For example, after $M = 30$ inner iterations update Σ_i and σ_i^2 in the outer iteration.

5 Error Bar Estimation

In the mean field method the posterior distribution was characterized by the first two moments of the posterior distribution, the mean $\langle y(\mathbf{x}) \rangle$ and the variance $\sigma_{\mathbf{x}}^2 = \langle y(\mathbf{x})^2 \rangle - \langle y(\mathbf{x}) \rangle^2$. The implicit assumption of this approach is that the posterior distribution $P[y(\mathbf{x})|\mathcal{D}]$ can be approximated by a Gaussian distribution with the mean $\langle y(\mathbf{x}) \rangle$ and variance $\sigma_{\mathbf{x}}^2$. The posterior mean $\langle y(\mathbf{x}) \rangle$ has been calculated by the mean field algorithm. An approximation formula for the variance of the posterior distribution, $\sigma_{\mathbf{x}}^2$, has been derived in [6] by applying a linear response argument, such that

$$\sigma_{\mathbf{x}}^2 \approx K(\mathbf{x}, \mathbf{x}) - \mathbf{K}(\mathbf{x}, \mathbf{X})^T [K + \Sigma]^{-1} \mathbf{K}(\mathbf{x}, \mathbf{X}) \quad (5.1)$$

where $\mathbf{K}(\mathbf{x}, \mathbf{X}) = [K(\mathbf{x}, \mathbf{x}_1), K(\mathbf{x}, \mathbf{x}_2), \dots, K(\mathbf{x}, \mathbf{x}_N)]^T$ and $\Sigma = \text{diag}(\Sigma_1, \Sigma_2, \dots, \Sigma_N)$.

Thus in the mean field method the posterior distribution is approximated by $P[y(\mathbf{x})|\mathcal{D}] \approx N(y(\mathbf{x})|\langle y(\mathbf{x}) \rangle, \sigma_{\mathbf{x}}^2)$. Consider the prediction for a new test data point \mathbf{x} . The prediction distribution for the target t given the data set \mathcal{D} can be generated from the ϵ -insensitive noise model (2.2) and the approximated posterior $N(y(\mathbf{x})|\langle y(\mathbf{x}) \rangle, \sigma_{\mathbf{x}}^2)$ as follows

$$P[t|\mathcal{D}] = \int K(C, \epsilon) \exp\{-CL_{\epsilon}(t - y(\mathbf{x}))\} N(y(\mathbf{x})|\langle y(\mathbf{x}) \rangle, \sigma_{\mathbf{x}}^2) dy(\mathbf{x}) \quad (5.2)$$

Let us first note that the likelihood of SVM case is described by $K(C, \epsilon) \exp\{-CL_{\epsilon}(t - y(\mathbf{x}))\}$. This distribution function with respect to t can be represented as a superposition of Gaussian processes. By using the technique in [3] the error bar of the SVR based on the mean field algorithm, is then given by

$$\sigma_t^2 = \sigma_{\mathbf{x}}^2 + \frac{2}{C^2} + \frac{\epsilon^3 + 3\epsilon^2}{2C^2(\epsilon + 1)}. \quad (5.3)$$

This error bar has two components, see equation (5.3). The first $\sigma_{\mathbf{x}}^2$ is an estimate of the width of the posterior over the hidden function $y(\mathbf{x})$, i.e., the function uncertainty. The second term can be viewed as the measure for the uncertainty induced in the target noise determined by the control factor C and ϵ .

A similar error bar formula for the SVR problem has been given in our previous paper [3] based on the Laplacian approximation to the posterior distribution at the SVM solution. In this paper the posterior distribution has been also approximated by a Gaussian distribution but with the moments determined by the mean field.

6 Conclusions

We have shown that the mean field approach can be used in the SVR problem as the same as in the classification problem. Due to limited space we have not yet include the experimental results in this paper. Based on the mean field equation for a Gaussian process an efficient iterative implementation algorithm has been derived in this paper. Another point to note is that the mean field SVR method is moderately easy to implement and use. However the control factor C and precision parameter ϵ should be prespecified in the current form of algorithm. The standard SVM algorithm gives a deterministic solution but does not provide any statistical information, thus no confidence estimate, such as the error bars are available. Although we can interpret the SVM regression method under the probabilistic framework [3], the error bar estimation is calculated from the whole Gaussian approximation at the MAP solution based on the support vectors. However, under the mean field framework here a Gaussian distribution with covariance $K + \Sigma$ is used to approximate the posterior distribution with much more possible accuracy.

References

- [1] D. Goldberg, *Genetic Algorithms in Search, Optimisation and Machine Learning*, Addison Wesley, Massachusetts, USA, 1989.
- [2] Evgeniou, T., M. Pontil, and T. Poggio, A unified framework for regularization networks and support vector machines, A.I. Memo 1654, AI Lab, MIT, Massachusetts, 1999.
- [3] Gao, J., S. Gunn, C. Harris, and M. Brown, A probabilistic framework for SVM regression and error bar estimation, *Machine Learning* in press 2001.
- [4] S.R. Gunn, Support vector machines for classification and regression, Technical report, ISIS, Department of Electronics and Computer Science, University of Southampton, 1998.
- [5] Jaakkola, T. and M. Jordan, Bayesian parameter estimation through variational methods, *Statistics and Computing*, to appear 2000.
- [6] Opper, M. and O. Winther, Gaussian processes for classification: Mean Field Algorithms, *Neural Computation*, to appear 2001.
- [7] Smola, A. *Learning with Kernels*, Ph. D. thesis, Technischen Universität Berlin, Berlin, Germany, 1998.
- [8] Sollich, P. Probabilistic interpretations and Bayesian methods for support vector machines. Technical report, King's College London, London, UK, 1998
- [9] Williams, C. . Prediction with gaussian processes: from linear regression to linear prediction and beyond. In M. Jordan (Ed.), *Learning in Graphical Models*, pp. 599–621. Cambridge, Massachusetts: MIT Press, 1998.