MASSACHUSETTS INSTITUTE OF TECHNOLOGY
ARTIFICIAL INTELLIGENCE LABORATORY
and
CENTER FOR BIOLOGICAL AND COMPUTATIONAL LEARNING
DEPARTMENT OF BRAIN AND COGNITIVE SCIENCES

A.I. Memo No. 1651 October 1998
C.B.C.L Paper No. 168

On the Noise Model of Support Vector Machine Regression

Massimiliano Pontil, Sayan Mukherjee, Federico Girosi

This publication can be retrieved by anonymous ftp to publications.ai.mit.edu.
The pathname for this publication is: ai-publications/1500-1999/AIM-1651.ps

Abstract

Support Vector Machines Regression (SVMR) is a regression technique which has been recently introduced by V. Vapnik and his collaborators (Vapnik, 1995; Vapnik, Golowich and Smola, 1996). In SVMR the goodness of fit is measured not by the usual quadratic loss function (the mean square error), but by a different loss function called *Vapnik's ϵ-insensitive loss function*, which is similar to the "robust" loss functions introduced by Huber (Huber, 1981). The quadratic loss function is well justified under the assumption of Gaussian additive noise. However, the noise model underlying the choice of Vapnik's loss function is less clear. In this paper the use of Vapnik's loss function is shown to be equivalent to a model of additive and Gaussian noise, where the variance and mean of the Gaussian are random variables. The probability distributions for the variance and mean will be stated explicitly. While this work is presented in the framework of SVMR, it can be extended to justify non-quadratic loss functions in any Maximum Likelihood or Maximum A Posteriori approach. It applies not only to Vapnik's loss function, but to a broader class of loss functions.

This report describers research done at the Center for Biological & Computational Learning and the Artificial Intelligence Laboratory of the Massachusetts Institute of Technology. This research was sponsored by the Office of Naval Research under contract No. N00014-93-1-0385 and contract No. N00014-95-1-0600. Partial support was also provided by Daimler-Benz AG, Eastman Kodak, Siemens Corporate Research, Inc., and AT&T.

1 Introduction

Recently, there has been a growing interest in a novel function approximation technique called Support Vector Machines Regression (SVMR) [7, 9]. This technique was motivated by the framework of statistical learning theory, and also has a close relationship with the classical regularization theory approach to function approximation. One feature of SVMR is that it measures the interpolation error by *Vapnik's ϵ-Insensitive Loss Function* (ILF), a loss function similar to the "robust" loss functions introduced by Huber [4]. While the quadratic loss function commonly used in regularization theory is well justified under the assumption of Gaussian, additive noise, it is less clear what precise noise model underlies the choice of Vapnik's ILF. Understanding the nature of this noise is important for at least two reasons: 1) it can help us decide under which conditions it is appropriate to use SVMR rather than regularization theory; and 2) it may help to better understand the role of the parameter ϵ, which appears in the definition of Vapnik's ILF, and is one of the two free parameters in SVMR.
In this paper we demonstrate the use of Vapnik's ILF is justified under the assumption that the noise affecting the data is additive and Gaussian, but *its variance and mean are random variables whose probability distributions can be explicitly computed.* The result is derived by using the same Bayesian framework which can be used to derive the regularization theory approach, and it is an extension of existing work on noise models and "robust" loss functions [1].
The plan of the paper is as follows: in section 2 we briefly review SVMR and Vapnik's ILF; in section 3 we introduce the Bayesian framework necessary to prove our main result, which is shown in section 4; in section 5 we show some additional results which relate to the topic of robust statistics, while section 6 summarizes the paper.

2 Vapnik's ϵ-insensitive loss function

Consider the following problem: we are given a data set $g = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$, obtained by sampling, with noise, some unknown function $f(\mathbf{x})$ and we are asked to recover the function f, or an approximation of it, from the data g. A common strategy consists of choosing as a solution the minimum of a functional of the following form:

$$H[f] = \sum_{i=1}^{l} V(y_i - f(\mathbf{x}_i)) + \alpha\Phi[f] \tag{1}$$

where $V(x)$ is some loss function used to measure the interpolation error, α is a positive number, and $\Phi[f]$ is a smoothness functional. SVMR correspond to a particular choice for V, that is Vapnik's ILF, plotted below in figure (1):

$$V(x) \equiv |x|_\epsilon \equiv \begin{cases} 0 & \text{if } |x| < \epsilon \\ |x| - \epsilon & \text{otherwise.} \end{cases} \tag{2}$$

(Details about minimizing the functional (1) and the specific form of the smoothness functional (1) can be found in [7] and [2]). Vapnik's ILF is similar to some of the functions used in robust statistics (Huber, 1981), which are known to provide robustness against outliers and was motivated by it (see Vapnik, 1998)i [8]. However the function (2) is not only a robust cost function, because of its linear behavior outside the interval $[-\epsilon, \epsilon]$, but also assigns zero cost to errors smaller then ϵ. In other words, for the cost function V_ϵ any function closer than ϵ to the

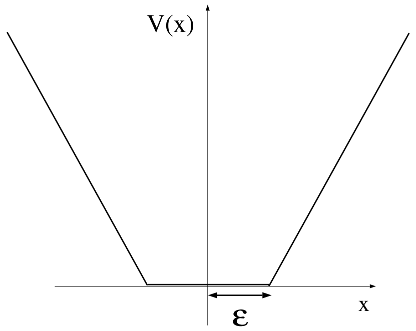

Figure 1: Vapnik's ILF $V_\epsilon(x)$.

data points is a perfect interpolant.
It is important to notice that if we choose $V(x) = x^2$, then the functional (1) is the usual regularization theory functional [10, 3], and its minimization leads to models which include Radial Basis Functions or multivariate splines. Vapnik's ILF represents therefore a crucial difference between SVMR and more classical models such as splines and Radial Basis Functions. What is the rationale for using Vapnik's ILF rather than a quadratic loss function like in regularization theory? In the next section we will introduce a Bayesian framework that will allow us to answer this question.

3 Bayes approach to SVMR

In this section, a simple Bayesian framework is used to interpret the variational approach of the type eq. (1). Rigorous work on this topic was originally done by Kimeldorf and Wahba, and we refer to [5, 10, 3, 6] for details.
Suppose that the set $g = \{(\mathbf{x}_i, y_i) \in R^n \times R\}_{i=1}^N$ of data has been obtained by randomly sampling a function f, defined on R^n, in the presence of additive noise, that is

$$f(\mathbf{x}_i) = y_i + \delta_i, \quad i = 1, \ldots, N \tag{3}$$

where δ_i are random independent variables with a given distribution. We want to recover the function f, or an estimate of it, from the set of data g. We take a probabilistic approach, and regard the function f as the realization of a random field with a known prior probability distribution. We are interested in maximizing the a posteriori probability of f given the data g, which can be written, using Bayes' theorem, as following:

$$\mathcal{P}[f|g] \propto \mathcal{P}[g|f]\ \mathcal{P}[f], \tag{4}$$

where $\mathcal{P}[g|f]$ is the conditional probability of the data g given the function f and $\mathcal{P}[f]$ is the *a priori* probability of the random field f, which is often written as $\mathcal{P}[f] \propto e^{-\alpha\Phi[f]}$, where $\Phi[f]$ is usually a smoothness functional. The probability $\mathcal{P}[g|f]$ is essentially a model of the noise, and if the noise is additive, as in eq. (3), and i.i.d. with probability distribution $P(\delta)$, it can be written as:

$$\mathcal{P}[g|f] = \prod_{i=1}^N P(\delta_i) \tag{5}$$

Substituting eq. (5) in eq. (4), it is easy to see that the function that maximizes the posterior probability of f given the data g is the one that minimizes the following functional:

$$H[f] = -\sum_{i=1}^{N} \log P(f(\mathbf{x}_i) - y_i) + \alpha\Phi[f] \ . \tag{6}$$

This functional is of the same form as equation (1), once we identify the loss function $V(x)$ as the log-likelihood of the noise. If we assume that the noise in eq. (3) is Gaussian, with zero mean and variance σ, then the functional above takes the form:

$$H[f] = \frac{1}{2\sigma^2} \sum_{i=1}^{N} (y_i - f(\mathbf{x}_i))^2 + \alpha\Phi[f] \ .$$

which corresponds to the classical regularization theory approach [10, 3]. In order to obtain SVMR in this approach one would have to assume that the probability distribution of the noise is $P(\delta) = e^{-|\delta|_\epsilon}$. Unlike an assumption of Gaussian noise, it is not clear what motivates in this Bayesian framework such a choice. The next section will address this question.

4 Main Result

In this section we build on the probabilistic approach described in the previous section and on work done by Girosi [1], and derive a novel class of noise models and loss functions.

4.1 A novel noise model

We start by modifying eq. (5), and drop the assumption that noise variables have all identical probability distributions. Different data points may have been collected at different times, under different conditions, so it is more realistic to assume that the noise variables δ_i have probability distributions P_i which are not necessarily identical. Therefore we write:

$$\mathcal{P}[g|f] = \prod_{i=1}^{N} P_i(\delta_i). \tag{7}$$

Now we assume that the noise distributions P_i are actually Gaussians, *but do not have necessarily zero mean*, and define P_i as:

$$P_i(\delta_i) \propto e^{-\beta_i(\delta_i - t_i)^2} \tag{8}$$

While this model is realistic, and takes into account the fact that the noise could be biased, it is not practical because it is unlikely that we know the set of parameters $\boldsymbol{\beta} \equiv \{\beta_i\}_{i=1}^N$ and $\mathbf{t} = \{t_i\}_{i=1}^N$. However, we may have some information about $\boldsymbol{\beta}$ and $\mathbf{t}$, for example a range for their values, or the knowledge that most of the time they assume certain values. It is therefore natural to model the uncertainty on $\boldsymbol{\beta}$ and $\mathbf{t}$ by considering them as i.i.d. random variables, with probability distributions $\mathcal{P}(\boldsymbol{\beta}, \mathbf{t}) = \prod_{i=1}^N P(\beta_i, t_i)$. Under this assumption, eq. (8) can be interpreted as $P(\delta_i|\beta_i, t_i)$, the conditional probability of δ_i *given* β_i and t_i. Taking this in account, we can rewrite eq. (4) as:

$$\mathcal{P}[f|g, \boldsymbol{\beta}, \mathbf{t}] \propto \prod_{i=1}^{N} P(\delta_i|\beta_i, t_i)\mathcal{P}[f]. \tag{9}$$

Since we are interested in computing the conditional probability of f given g, independently of $\boldsymbol{\beta}$ and $\mathbf{t}$, we compute the marginal of the distribution above, integrating over $\boldsymbol{\beta}$ and $\mathbf{t}$:

$$\mathcal{P}^*[f|g] \propto \int d\boldsymbol{\beta} \int d\mathbf{t} \prod_{i=1}^{N} P(\delta_i|\beta_i, t_i)\mathcal{P}[f]\mathcal{P}(\boldsymbol{\beta}, \mathbf{t}) \tag{10}$$

Using the assumption that $\boldsymbol{\beta}$ and $\mathbf{t}$ are i.i.d., so that $\mathcal{P}(\boldsymbol{\beta}, \mathbf{t}) = \prod_{i=1}^{N} P(\beta_i, t_i)$, we can easily see that the function that maximizes the a posteriori probability $\mathcal{P}^*[f|g]$ is the one that minimizes the following functional:

$$H[f] = \sum_{i=1}^{N} V(f(\mathbf{x}_i) - y_i) + \alpha\Phi[f] \tag{11}$$

where V is given by:

$$V(x) = -\log \int_0^\infty d\beta \int_{-\infty}^{\infty} dt \sqrt{\beta} e^{-\beta(x-t)^2} P(\beta, t) \tag{12}$$

where the factor $\sqrt{\beta}$ appears because of the normalization of the Gaussian (other constant factors have been disregarded). Equation (11) defines a novel class of loss functions, and provides a probabilistic interpretation for them: using a loss function V with an integral representation of the form (12) is equivalent to assuming that the noise is Gaussian, but the mean and the variance of the noise are random variables with probability distribution $P(\beta, t)$.
The class of loss functions defined by equation (12) is an extension of the model discussed by Girosi [1] where the case of unbiased noise distributions is considered:

$$V(x) = -\log \int_0^\infty d\beta \sqrt{\beta} e^{-\beta x^2} P(\beta) \tag{13}$$

Equation (13) can be obtained from equation (12) by setting $P(\beta, t) = P(\beta)\delta(t)$. The class of loss function of type (13) can be completely identified. To this purpose observe that given V the probability function $P(\beta)$ is essentially the inverse Laplace transform of $e^{V(\sqrt{x})}$. So $V(x)$ verify equation (13) whenever the inverse Laplace transform of $\exp -V(\sqrt{x})$ is nonnegative and integrable. In practice this is difficult if not impossible to check. However alternative conditions that guarantee V to be in the class (13) can be employed (see [1] and references there). A simple example of loss function of type (13) is $V(x) = |x|^a,\ a \in (0, 2]$. When $a = 2$ this corresponds to the classical quadratic loss function which solves equation (13) with $P(\beta) = \delta(t)$. The case $a = 1$ corresponds to the robust L_1 error measure and equation (13) is solved by: $P(\beta) = \beta^{-2} exp(-\frac{1}{4\beta})$ as can be seen computing the inverse Laplace transform of $e^{\sqrt{x}}$.

4.2 The noise model for Vapnik's ILF

In order to provide a probabilistic interpretation for Vapnik's ILF we need to find a probability distribution $P(\beta, t)$ such that eq. (12) is verified when we set $V(x) = |x|_\epsilon$. This is a difficult problem, which requires the solution of a linear integral equation. Here we state a solution under the assumption that β and t are independent variable, that is:

$$P(t, \beta) = \mu(\beta)\lambda(t) \tag{14}$$

Plugging equation (14) in equation (12) and computing the integral with respect to β we obtain:

$$e^{-V(x)} = \int_{-\infty}^{+\infty} dt\lambda(t)G(x-t) \tag{15}$$

where we have defined:

$$G(t) = \int_0^{\infty} d\beta\mu(\beta)\sqrt{\beta}e^{-\beta t^2} \tag{16}$$

Observe that the function G is a density distribution, because both the function in the r.h.s. of equation (16) are densities.
In order to compute G we observe that for $\epsilon = 0$ the function $e^{-|x|_\epsilon}$ becomes the Laplace distribution which we have seen above be in the class (13). Then $\lambda_{\epsilon=0}(t) = \delta(t)$ and from equation (15):

$$G(t) = e^{-|t|}. \tag{17}$$

Observe also that in view of the example discussed at the end of section (4.1) and of equation (17) the function $\nu(\beta)$ in equation (16) is:

$$P(\beta) = \beta^{-2}e^{-\frac{1}{4\beta}}. \tag{18}$$

It remains to obtain the expression of $\lambda(t)$ for $\epsilon > 0$. To this purpose we write equation (15) in Fourier space:

$$\tilde{F}[e^{-|x|_\epsilon}] = \tilde{G}(\omega)\tilde{\lambda}_\epsilon(\omega) \tag{19}$$

with:

$$\tilde{F}[e^{-|x|_\epsilon}] = \frac{\sin(\epsilon\omega) + w\cos(\epsilon\omega)}{\omega(1+\omega^2)}. \tag{20}$$

and:

$$\tilde{G}(\omega) = \frac{1}{1+\omega^2}. \tag{21}$$

Plugging equations (20) and (21) in equation (19) we obtain:

$$\tilde{\lambda}_\epsilon(\omega) = \frac{sin\epsilon\omega}{\omega} + cos\epsilon\omega. \tag{22}$$

Finally taking the inverse Fourier Transform and normalizing:

$$\lambda_\epsilon(t) = \frac{1}{2(\epsilon+1)}\left(\chi_{[-\epsilon,\epsilon]}(t) + \delta(t-\epsilon) + \delta(t+\epsilon)\right) \tag{23}$$

where $\chi_{[-\epsilon,\epsilon]}$ is the characteristic function of the interval $[-\epsilon,\epsilon]$

$$P_\epsilon(x) = \int_{-\infty}^{+\infty} dt\lambda_\epsilon(t)e^{-|t-x|} \tag{24}$$

The shape of the functions in eq. (18) and (23) are shown in figure (2). The above model has a simple interpretation: using Vapnik's ILF is equivalent to assuming that the noise affecting the data is Gaussian. However, the variance and the mean of the Gaussian noise are random

(a) (b)

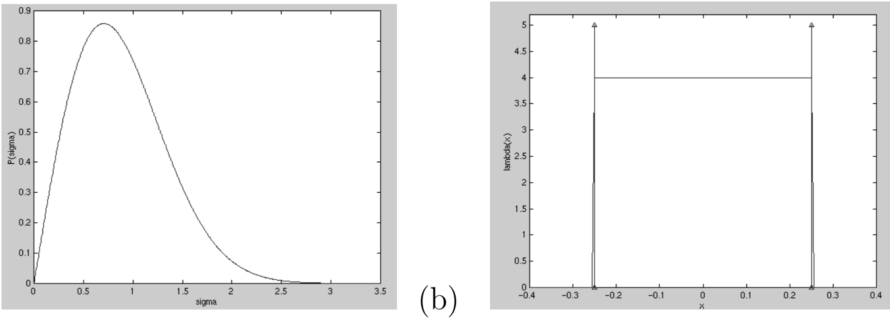

Figure 2: a) The probability distribution $\bar{P}(\sigma)$, where $\sigma^2 = \frac{1}{2\beta}$ and $\bar{P}(\beta)$ is given by eq. 18 ; b) The probability distribution $\lambda_\epsilon(x)$ for $\epsilon = .25$ (see eq.23).

variables: the variance ($\sigma^2 = \frac{1}{2\beta}$) has a unimodal distribution that does not depend on ϵ, and the mean has a distribution which is uniform in the interval $[-\epsilon, \epsilon]$, (except for two delta functions at $\mp\epsilon$, which ensures that the mean has not zero probability to be equal to $\mp\epsilon$). The distribution of the mean is consistent with the current understanding of Vapnik's ILF: errors smaller than ϵ do not count because they may be due entirely to the bias of the Gaussian noise.
Finally observe that equation (15) establishes an indirect representation of the density e^{-V} as a superposition of translates of the density G, where the coefficients are given by the distribution of the mean $\lambda(t)$, the length of translation being in the interval $[-\epsilon, \epsilon]$.

5 Additional Results

While it is difficult to give a simple characterization of the class of loss functions with an integral representation of the type (12), it is possible to extend the results of the previous section to a particular sub-class of loss functions, ones of the form:

$$V_\epsilon(x) = \begin{cases} h(x) & \text{if } |x| < \epsilon \\ \\ |x| & \text{otherwise,} \end{cases} \tag{25}$$

where $h(x)$ is some symmetric function with some restriction that will become clear later. A well known example is one of Huber's robust loss functions [4], for which $h(x) = \frac{x^2}{2\epsilon} + \frac{\epsilon}{2}$ (see figure (3.a)). For loss functions of the form (25), it can be shown that a *function* $P(\beta, t)$ that solves eq. (12) always exists, and it has a form which is very similar to the one for Vapnik's ILF. More precisely, we have that $P(\beta, t) = \bar{P}(\beta)\lambda_\epsilon(t)$, where $\bar{P}(\beta)$ is given by eq. (18), and $\lambda_\epsilon(t)$ is the following compact-support distribution:

$$\lambda_\epsilon(t) = \begin{cases} P(t) - P''(t) & \text{if } \ |t| < \epsilon \\ 0 & \text{otherwise,} \end{cases} \tag{26}$$

where we have defined $P(x) = e^{-V_\epsilon(x)}$. This result does not guarantee, however, that λ_ϵ is a measure, because $P(t) - P''(t)$ may not be positive on the whole interval $[-\epsilon, \epsilon]$, depending on h. The positivity constraint defines the class of "admissible" functions h. A precise characterization

(a) (b)

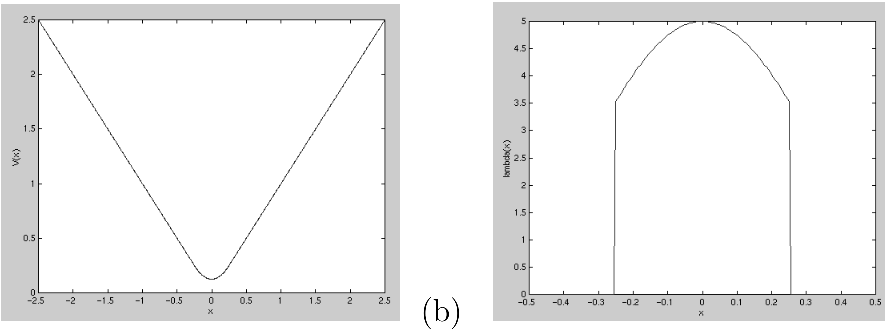

Figure 3: a) The Huber loss function; b) the corresponding $\lambda_\epsilon(x)$, $\epsilon = .25$. Notice the difference between this distribution and the one that corresponds to Vapnik's ILF: while for this one the mean of the noise is zero most of the times, in Vapnik's ILF all the values of the mean are equally likely.

of the class of admissible h, and therefore the class of "shapes" of the functions which can be derived in this model is currently under study. It is easy to verify that the Huber's loss function described above is admissible, and corresponds to a probability distribution for which the mean is equal to $\lambda_\epsilon(t) = (1 + \frac{1}{\epsilon} - (\frac{t}{\epsilon})^2)e^{-\frac{t^2}{2\epsilon}}$ over the interval $[-\epsilon, \epsilon]$ (see figure (3.b)).
Finally note that for the class of loss functions (25) the noise distribution can be written as the convolution between the distribution of the mean λ_ϵ and the Laplace distribution:

$$P_\epsilon(x) = \int_{-\epsilon}^{+\epsilon} dt \lambda_\epsilon(t) e^{-|t-x|} \tag{27}$$

Equation (27) establishes a representation of the noise distribution P as a continuos superposition of translate Laplace functions in the interval $[-\epsilon, \epsilon]$.

6 Conclusions and future work

In this work an interpretation of Vapnik's ILF was presented in a simple Bayesian framework. This will hopefully lead to a better understanding of the assumptions that are implicitly made when using SVMR. We believe this work is important for at least two reasons: 1) it makes more clear under which conditions it is appropriate to use SVMR rather than regularization theory; and 2) it may help to better understand the role of the parameter ϵ, which appears in the definition of Vapnik's loss function, and is one of the two free parameters in SVMR. We demonstrated that the use of Vapnik's ILF is justified under the assumption that the noise affecting the data is additive and Gaussian, but not necessarily zero mean, and that its variance and mean are random variables with given probability distributions. Similar results can be derived for some other loss functions of the "robust" type. A clear characterization of the class of loss functions which can be derived in this framework is still missing, and it is subject of current work . While we present this work in the framework of SVMR, similar reasoning can be applied to justify non-quadratic loss functions in any Maximum Likelihood or Maximum A Posteriori approach.

Acknowledgments F. Girosi wishes to thank J. Lemm for useful discussions and suggestions.

References

[1] F. Girosi. Models of noise and robust estimates. A.I. Memo 1287, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 1991.

[2] F. Girosi. An equivalence between sparse approximation and Support Vector Machines. *Neural Computation*, 10(6):1455–1480, 1998.

[3] F. Girosi, M. Jones, and T. Poggio. Regularization theory and neural networks architectures. *Neural Computation*, 7:219–269, 1995.

[4] P.J. Huber. *Robust Statistics.* John Wiley and Sons, New York, 1981.

[5] G.S. Kimeldorf and G. Wahba. A correspondence between Bayesan estimation on stochastic processes and smoothing by splines. *Ann. Math. Statist.*, 41(2):495–502, 1971.

[6] T. Poggio and F. Girosi. A Sparse Representation for Function Approximation. *Neural Computation*, 10(6), 1998.

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

[8] V. Vapnik. *Statistical Learning Theory.* John Wiley and sons, New York, 1998.

[9] V. Vapnik, S.E. Golowich, and A. Smola. Support vector method for function approximation, regression estimation, and signal processing. In M. Mozer, M. Jordan, and T. Petsche, editors, *Advances in Neural Information Processings Systems 9*, pages 281–287, Cambridge, MA, 1997. The MIT Press.

[10] G. Wahba. *Splines Models for Observational Data.* Series in Applied Mathematics, Vol. 59, SIAM, Philadelphia, 1990.