Linear Dependency Between ϵ and The Input Noise in ϵ-Support Vector Regression

James T. Kwok

Department of Computer Science,
Hong Kong University of Science and Technology, Hong Kong,
jamesk@cs.ust.hk

Abstract. In using the ϵ-support vector regression (ϵ-SVR) algorithm, one has to decide on a suitable value of the insensitivity parameter ϵ. Smola *et al.* [6] determined its "optimal" choice based on maximizing the statistical efficiency of a location parameter estimator. While they successfully predicted a linear scaling between the optimal ϵ and the noise in the data, the value of the theoretically optimal ϵ does not have a close match with its experimentally observed counterpart. In this paper, we attempt to better explain the experimental results there, by analyzing a toy problem with a closer setting to the ϵ-SVR. Our resultant predicted choice of ϵ is much closer to the experimentally observed value, while still demonstrating a linear trend with the data noise.

1 Introduction

In recent years, uses of the support vector machines (SVMs) on various classification and regression problems have been increasingly popular. SVMs are motivated by results from the statistical learning theory and, unlike other machine learning methods, their generalization performances do not depend on the dimensionality of the feature space. In this paper, we focus on the ϵ-support vector regression (ϵ-SVR) algorithm [7], which has shown promising results in a number of different applications (e.g. [4]).

One issue about ϵ-SVR is on how to set the insensitivity parameter ϵ. Data-resampling techniques such as cross-validation can be used [4], but they are usually very expensive in terms of computation and/or data. A more efficient approach is to use another variant called the ν-SVR [5], which determines ϵ by using another parameter ν. It is claimed that ν may be easier to specify than ϵ. Another approach by Smola *et al.* [6] is to find the "optimal" choice of ϵ based on maximizing the statistical efficiency of a location parameter estimator. They showed that the asymptotically optimal ϵ should scale linearly with the noise in the data, and this was verified experimentally. However, their predicted value of the optimal ϵ does not have a close match with their experimental results.

In this paper, we attempt to better explain the experimental results in [6] by analyzing a toy problem with a closer setting to the ϵ-SVR. The rest of this paper is organized as follows. Brief introduction to the ϵ-SVR is given in Section 2. The specification and analysis of the toy problem are given in Section 3, while the last section gives some concluding remarks.

2 ϵ-Support Vector Regression

Let the training set D be $\{(\mathbf{x}_i, y_i)\}_{i=1}^N$, with input $\mathbf{x}_i$ and output $y_i \in \Re$. In ϵ-SVR, $\mathbf{x}$ is first mapped to $\mathbf{z} = \phi(\mathbf{x})$ in feature space $\mathcal{F}$, then a linear function $f(\mathbf{x}, \mathbf{w}) = \mathbf{w}^T\mathbf{z} + b$ is constructed in $\mathcal{F}$ such that it deviates least from the training data according to Vapnik's ϵ-insensitive loss function

$$|y - f(\mathbf{x})|_\epsilon = \begin{cases} 0 & \text{if } |y - f(\mathbf{x})| \leq \epsilon, \\ |y - f(\mathbf{x})| - \epsilon & \text{otherwise,} \end{cases}$$

while at the same time is as "flat" as possible. Mathematically, this means:

$$\min \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^N (\xi_i + \xi_i^*), \text{ subject to } \begin{cases} y_i - f_i \leq \epsilon + \xi_i, \\ f_i - y_i \leq \epsilon + \xi_i^*, \ i = 1, \ldots, N, \\ \xi_i, \xi_i^* \geq 0, \end{cases} \quad (1)$$

where $f_i = f(\mathbf{x}_i, \mathbf{w})$ and C is a user-defined constant. It is now well-known that (1) can be transformed to the following quadratic programming (QP) problem:

$$\max \ -\frac{1}{2}\sum_{i,j=1}^N (\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)K(\mathbf{x}_i, \mathbf{x}_j) - \epsilon\sum_{i=1}^N(\alpha_i + \alpha_i^*) + \sum_{i=1}^N y_i(\alpha_i - \alpha_i^*), \quad (2)$$

subject to $\sum_{i=1}^N (\alpha_i - \alpha_i^*) = 0$, and $\alpha_i, \alpha_i^* \in [0, C]$. Here, K is a *kernel* function such that $K(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i)^T\phi(\mathbf{x}_j)$.

3 Toy Problem

As in [6], we study the problem of estimating the location parameter $\tilde{w}$ of a set D of data points

$$y_i = \tilde{w} + \eta_i, \quad i = 1, \ldots, N,$$

where η_i's are i.i.d. noise belonging to some distribution $\phi(\cdot)$. The corresponding density function of y will be denoted in the sequel by $\tilde{p}(y)$[1]. Studies of this simple problem have often gained insights into various regularization schemes [1, 2]. In the following, we assume that $\phi(\eta)$ is symmetrical around zero, which usually holds in practice. Moreover, we employ the ϵ-insensitive cost function together with the same regularization term[2] as in ϵ-SVR. Consequently, while the approach in [6] is based on the maximum likelihood estimator, ϵ-SVR as well as our analysis of this toy problem in Section 3.2 are based on the *maximum a posteriori* (MAP) estimator.

In Section 3.1, we first show that this toy problem also leads to a QP problem similar to that for ϵ-SVR. In Section 3.2, we discuss a probabilistic model, and show how our "optimal" estimate for ϵ can be obtained. We then discuss some common noise models in Section 3.4.

[1] Obviously, $\tilde{p}(y) = \phi(y - \tilde{w})$.

[2] While location parameters typically should not be regularized, Goutte and Hansen [1] argued that the use of regularization in this toy problem can be used to demonstrate more clearly the effect of the prior.

3.1 The Corresponding Quadratic Programming Problem

Similar to ϵ-SVR, we try to find an estimator w such that it deviates least from the training data $\{y_i\}_{i=1}^N$ according to the ϵ-insensitive loss function, while at the same time is as small as possible. Again, we formulate this as a constrained optimization problem by introducing slack variables ξ_i, ξ_i^*'s, and obtain

$$\min \frac{1}{2}w^2 + C\sum_{i=1}^{N}(\xi_i + \xi_i^*), \text{ subject to } \begin{cases} y_i - w \le \epsilon + \xi_i, \\ w - y_i \le \epsilon + \xi_i^*, \; i = 1, \ldots, N, \\ \xi_i, \xi_i^* \ge 0. \end{cases} \quad (3)$$

Using the standard technique of transforming this to its dual problem, we have

$$\max \; -\frac{1}{2}\sum_{i,j=1}^{N}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*) - \epsilon\sum_{i=1}^{N}(\alpha_i + \alpha_i^*) + \sum_{i=1}^{N} y_i(\alpha_i - \alpha_i^*),$$

subject to $\alpha_i, \alpha_i^* \in [0, C]$. Putting $K(\mathbf{x}_i, \mathbf{x}_j) = 1$ for all $\mathbf{x}_i, \mathbf{x}_j$, this is thus the same as the QP problem for ϵ-SVR in (2), with the exception that the bias b is not used in the approximating function $f(\mathbf{x}, \mathbf{w})$.

3.2 Estimating the Optimal Values of β and ϵ

The ϵ-insensitive cost function corresponds to the following noise model [7]:

$$p(y_i|\mathbf{x}_i, \mathbf{w}, \beta, \epsilon) = \frac{\beta}{2(1+\epsilon\beta)}\exp(-\beta|y_i - f_i|_\epsilon).$$

Notice that [7] does not have the factor β, but is introduced here to play the important role of controlling the noise variance. In ϵ-SVR, by using the Gaussian prior $p(\mathbf{w}|\alpha) \propto \exp(-\frac{\alpha}{2}\|\mathbf{w}\|^2)$, optimizing (1) can be interpreted as finding the MAP estimate of $\mathbf{w}$ by setting $C = \beta/\alpha$.

For the toy problem we address here, we adopt a similar prior $p(w) = \sqrt{\frac{\alpha}{2\pi}}\exp(-\frac{\alpha}{2}w^2)$, and the likelihood model

$$p(y_i|w, \beta, \epsilon) = \frac{\beta}{2(1+\epsilon\beta)}\exp(-\beta|y_i - w|_\epsilon). \quad (4)$$

Here, β and ϵ are regarded as hyperparameters, whereas α is treated as a constant. Using the Bayes rule, $p(w|D, \beta, \epsilon) \propto p(D|w, \beta, \epsilon)p(w)$, we obtain

$$\log p(w|D, \beta, \epsilon) = -\frac{\alpha}{2}w^2 - \beta\sum_{i=1}^{N}|y_i - w|_\epsilon + N\log\frac{\beta}{2(1+\epsilon\beta)} + \text{const}.$$

As for ϵ-SVR, by setting $C = \beta/\alpha$, optimizing (3) can be interpreted as finding the MAP estimate $\hat{w}$ of w for given β, ϵ. The value of $\hat{w}$ depends on the particular training set and a closed-form solution is difficult to obtain. To simplify the analysis, we replace $\sum_{i=1}^{N}|y_i - w|_\epsilon$ by its expected value and obtain

$$M(w, \beta, \epsilon) = -\frac{\alpha}{2}w^2 - \beta N E(|y - w|_\epsilon) + N\log\frac{\beta}{2(1+\epsilon\beta)} + \text{const}.$$

Setting its partial derivative w.r.t. w to zero, it can be shown that $\hat{w}$ satisfies

$$\alpha\hat{w} + \beta N \left(\int_{-\infty}^{\hat{w}-\epsilon} \tilde{p}(y)dy - \int_{\hat{w}+\epsilon}^{\infty} \tilde{p}(y)dy \right) = 0.$$

To find suitable values of β and ϵ, we use the method of type II maximum likelihood prior[3] (ML-II) in Bayesian statistics. The ML-II solutions of β and ϵ are those that maximize $p(D|\beta,\epsilon)$, which in turn can be obtained by integrating out w, as $p(D|\beta,\epsilon) = \int p(w,D|\beta,\epsilon)dw = \int p(D|w,\beta,\epsilon)p(w)dw$. Assume $p(D|w,\beta,\epsilon)p(w)$ to be sharply peaked at its maximum $\hat{w}$, maximizing $p(D|\beta,\epsilon)$ w.r.t. β and ϵ thus becomes maximizing $p(D|\hat{w},\beta,\epsilon)p(\hat{w})$. This is the same as maximizing $\log p(\hat{w}|D,\beta,\epsilon)$, which equals $M(\hat{w},\beta,\epsilon)$ approximately. Taking the derivative of M w.r.t. ϵ and β to zero, we obtain

$$\frac{1}{1+\epsilon\beta} = \int_{-\infty}^{\hat{w}-\epsilon} \tilde{p}(y)dy + \int_{\hat{w}+\epsilon}^{\infty} \tilde{p}(y)dy = P(y \notin [\hat{w}-\epsilon, \hat{w}+\epsilon]), \tag{5}$$

and

$$\frac{1}{\beta(1+\epsilon\beta)} = E(|y-\hat{w}|_\epsilon). \tag{6}$$

Here, we have assumed that both $\frac{\partial\hat{w}}{\partial\epsilon}$ and $\frac{\partial\hat{w}}{\partial\beta}$ are negligible, which can be shown to be valid as $\tilde{p}(\cdot)$ is symmetrical around $\tilde{w}$. Finally, (5) and (6) can then be used to solve for β and ϵ.

3.3 Linear Dependency Between ϵ and the Input Noise

From (5), (6), we obtain

$$\ell \equiv \frac{1}{\beta} = \frac{E(|y-\hat{w}|_\epsilon)}{P(y \notin [\hat{w}-\epsilon, \hat{w}+\epsilon])}, \tag{7}$$

and

$$\frac{\epsilon}{\ell} = \frac{P(y \in [\hat{w}-\epsilon, \hat{w}+\epsilon])}{P(y \notin [\hat{w}-\epsilon, \hat{w}+\epsilon])}. \tag{8}$$

ℓ can be interpreted as the distance from the ϵ-tube averaged over those points that are outside. The denominator acts as a normalizing factor as only that portion of points contribute towards $E(|y-\hat{w}|_\epsilon)$. Moreover, recall our probability model in (4), for $y \in [\hat{w}-\epsilon, \hat{w}+\epsilon]$,

$$p(y|\hat{w},\beta,\epsilon) = \frac{\beta}{2(1+\epsilon\beta)} = \frac{1}{2} \cdot \frac{P(y \notin [\hat{w}-\epsilon, \hat{w}+\epsilon])}{\ell},$$

using (7) and (8). As the total probability mass outside (both sides of) the ϵ-tube is $P(y \notin [\hat{w}-\epsilon, \hat{w}+\epsilon])$, ℓ can be interpreted as the "spread" of the density function outside the tube, by extending the density at the boundary of the ϵ-tube outside (Figure 1). This interpretation can also be seen from (8).

[3] Notice that MacKay's evidence framework [3], which has been popularly used in the neural networks community, is computationally equivalent to the ML-II method.

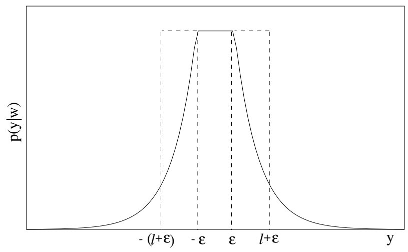

Fig. 1. The role of ℓ as the effective spread of the density function $p(y|\hat{w}, \beta, \epsilon)$ outside the ϵ-tube. Here, for simplicity, $\hat{w}$ is shifted to zero.

It was observed in [6] that there exists a linear dependency between ϵ and the scale parameter of the input noise model. This can be seen as a sufficient condition for ensuring the relationship in (8). When the input noise is changed by a factor λ, $\tilde{p}(\cdot)$ becomes stretched/compressed around $\tilde{w}$ by the same factor. By having a linear relationship between ϵ and the input noise, ϵ becomes $\epsilon' = \lambda\epsilon$. If we assume that $\hat{w}$ is close to $\tilde{w}$ and remains relatively unchanged, then $P(y \in [\hat{w} - \epsilon', \hat{w} + \epsilon'])$ will be unchanged, while $E(|y - \hat{w}|_{\epsilon'}) = \lambda E(|y - \hat{w}|_{\epsilon})$. Hence, by (7), $\ell' = \lambda\ell$, and (8) still holds for the new values of ϵ' and ℓ'.

3.4 Applications to the Gaussian and Laplacian Noise Models

For the Gaussian noise model, $\phi(\eta) = \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{\eta^2}{2\sigma^2})$. Define $\delta = \hat{w} - \tilde{w}, b_1 = \frac{1}{\sqrt{2}}(\frac{\epsilon}{\sigma} - \frac{\delta}{\sigma}), b_2 = \frac{1}{\sqrt{2}}(\frac{\epsilon}{\sigma} + \frac{\delta}{\sigma})$. Using (5) and (6), it can be shown that

$$\frac{1}{1 + \epsilon\beta} = \frac{1}{2}(\mathrm{erfc}(b_1) + \mathrm{erfc}(b_2)), \tag{9}$$

and

$$\frac{1}{\beta(1 + \epsilon\beta)} = -\frac{\sigma b_1}{2}\mathrm{erfc}(b_1) - \frac{\sigma b_2}{2}\mathrm{erfc}(b_2) + \frac{\sigma}{\sqrt{2\pi}} \exp(-b_1^2) + \frac{\sigma}{\sqrt{2\pi}} \exp(-b_2^2), \tag{10}$$

where $\mathrm{erfc}(x) = \frac{2}{\sqrt{\pi}} \int_x^{\infty} \exp(-t^2) dt$ is the complementary error function. Substituting (9) into (10), we see that ϵ always appears as the factor ϵ/σ in the solution, and thus there is always a linear relationship between ϵ and σ. If $|\delta|$ is small compared to ϵ relative to σ, $b_1 \simeq b_2 \simeq \epsilon/\sqrt{2}\sigma$, and we obtain $\mathrm{erfc}(\frac{\epsilon}{\sqrt{2}\sigma})(1 + \sqrt{\frac{\pi}{2}} \frac{\epsilon}{\sigma} \exp(\frac{\epsilon^2}{2\sigma^2})) = 1$. Solving, we have

$$\epsilon = 0.8485\sigma.$$

While the theoretical results in [6] suggest the "optimal" choice of $\epsilon = 0.6166\sigma$, our predicted choice is much closer to the experimentally observed optimal value of $\epsilon = 0.9\sigma$ in [6].

We next consider the Laplacian noise model, $\phi(\eta) = \frac{1}{2\sigma} \exp(-\frac{|\eta|}{\sigma})$. Using (5), (6) and assume that $|\hat{w} - \tilde{w}| \leq \epsilon$, we obtain $\frac{1}{1+\epsilon\beta} = \frac{1}{2}(\exp(-\frac{\epsilon}{\sigma} + \frac{\hat{w}-\tilde{w}}{\sigma}) +$

$\exp(-\frac{\epsilon}{\sigma} - \frac{\hat{w}-\tilde{w}}{\sigma}))$, and $\frac{1}{\beta(1+\epsilon\beta)} = \frac{\sigma}{2}(\exp(-\frac{\epsilon}{\sigma} + \frac{\hat{w}-\tilde{w}}{\sigma}) + \exp(-\frac{\epsilon}{\sigma} - \frac{\hat{w}-\tilde{w}}{\sigma}))$. Solving, then ϵ satisfies the equation

$$\frac{1}{1+\frac{\epsilon}{\sigma}} = \frac{1}{2}\left(\exp(-\frac{\epsilon}{\sigma} + \frac{\hat{w}-\tilde{w}}{\sigma}) + \exp(-\frac{\epsilon}{\sigma} - \frac{\hat{w}-\tilde{w}}{\sigma})\right).$$

Hence, again, there is a linear dependency between ϵ and σ. In the special case when $\hat{w} = \tilde{w}$, the above equations reduce to $\frac{1}{1+\epsilon\beta} = \exp(-\frac{\epsilon}{\sigma})$, and $\frac{1}{\beta(1+\epsilon\beta)} = \sigma\exp(-\frac{\epsilon}{\sigma})$. Solving, we obtain $\epsilon = 0$, which was also obtained in [6].

4 Conclusion

In this paper, we analyze the location estimation problem with a similar setting as for ϵ-SVR, and then derive its optimal choice of ϵ. In comparison to [6], their results are based on maximum likelihood estimation, while our analysis employs the same regularization term as in ϵ-SVR and is based on MAP estimation. Our predicted optimal choice of ϵ under Gaussian noise is also much closer to its experimentally observed optimal value. This prediction will be most accurate when the estimator $\hat{w}$ is close to the true value $\tilde{w}$. Besides, we accord with [6] in that ϵ should scale linearly with the input noise of the training data.

A pitfall of the results here is that their practical use is limited to cases when the noise model is known or at least can be estimated. Nevertheless, they can still be of tremendous help especially in performing simulation experiments. Typically, a researcher/practitioner may be experimenting with new techniques on ϵ-SVR, while not addressing the issue of finding the optimal value of ϵ. The results here can then be used to ascertain that a suitable value/range of ϵ has been chosen in the simulation.

References

1. C. Goutte and L.K. Hansen. Regularization with a pruning prior. *Neural Networks*, 10(6):1053–1059, 1997.
2. L.K. Hansen and C.E. Rasmussen. Pruning from adaptive regularization. *Neural Computation*, 6:1223–1232, 1994.
3. D.J.C. MacKay. Bayesian interpolation. *Neural Computation*, 4(3):415–447, May 1992.
4. K.R. Müller, A.J. Smola, G. Rätsch, B. Schölkopf, J. Kohlmorgen, and V. Vapnik. Predicting time series with support vector machines. In *Proceedings of the International Conference on Artificial Neural Networks*, 1997.
5. B. Schölkopf and A.J. Smola. New support vector algorithms. NeuroCOLT2 Technical Report NC2-TR-1998-031, GMD FIRST, 1998.
6. A.J. Smola, N. Murata, B. Schölkopf, and K.-R. Müller. Asymptotically optimal choice of ϵ-loss for support vector machines. In *Proceedings of the International Conference on Artificial Neural Networks*, 1998.
7. A.J. Smola and B. Schölkopf. A tutorial on support vector regression. NeuroCOLT2 Technical Report NC2-TR-1998-030, Royal Holloway College, 1998.