SUPPORT VECTOR REGRESSION WITH A GENERALIZED QUADRATIC LOSS

Filippo Portera and Alessandro Sperduti
Dipartimento di Matematica Pura ed Applicata
Universitą di Padova, Padova, Italy
{portera,sperduti}@math.unipd.it

Abstract The standard SVR formulation for real-valued function approximation on multi-dimensional spaces is based on the ϵ-insensitive loss function, where errors are considered not correlated. Due to this, local information in the feature space which can be useful to improve the prediction model is disregarded. In this paper we address this problem by defining a generalized quadratic loss where the co-occurrence of errors is weighted according to a kernel similarity measure in the feature space. We show that the resulting dual problem can be expressed as a hard margin SVR in a different feature space when the co-occurrence error matrix is invertible. We compare our approach against a standard SVR on two regression tasks. Experimental results seem to show an improvement in the performance.

Keywords: Regression, Support Vector Machines, Loss Functions, Kernel Methods.

1. Introduction

Statistical Learning Theory [Vapnik, 1998] provides a very effective framework for classification and regression tasks involving numerical features. Support Vectors Machines are directly derived from this framework and they work by solving a constrained quadratic problem where the convex objective function to minimize is given by the combination of a loss function with a regularization term (the norm of the weights). While the regularization term is directly linked, through a theorem, to the VC-dimension of the hypothesis space, and thus fully justified, the loss function is usually (heuristically) chosen on the basis of the task at hand. For example, when considering binary classification tasks, the ideal loss would be the 0-1 loss, which however cannot directly be plugged into the objective function because it is not convex. Thus, convex upper bounds to the 0-1 loss are used, e.g., the Hinge loss or the quadratic loss. In general, however, the used loss does not exploit the correlation that the input patterns may exhibit. A first attempt to exploit this type of information for

classification tasks has been presented in [Portera and Sperduti, 2004], where a family of generalized quadratic loss is defined. The basic idea is to first of all take into consideration the correlation between input patterns (eventually corrected by the targets of the involved examples), which can be coded as cross-coefficients of pairs of errors in a fully quadratic form, and then to modulate the strength of these cross-coefficients through a new hyperparameter. The "right" value of this new hyperparameter is then chosen by a search in the hyperparameters space (eventually involving a validation set) of the machine so to optimize the final performance [Zhang and Oles, 2001]. The experimental results presented in [Portera and Sperduti, 2004] seem to indicate a systematic improvement in the performance.

In this paper, we show that the same idea and advantages can be extended to real-valued function regression. Specifically, we suggest to use a loss function that weights every error associated to two patterns proportionally to the pattern similarity. This can be done by modifying the primal objective function of the SVR model with a loss that is a quadratic expression of the slack variables, weighting couples of errors by a pattern similarity measure based on a kernel function. In addition, signed slack variables are used so that given two distinct patterns, the modified SVR solution will penalize couple of errors (of similar patterns) that are both due to an overestimate (or underestimate) of the target values versus couple of errors (of similar patterns) that are due to an overestimate of one of the target values and an underestimate of the other target value. This method should bias the learning towards solutions where the local concentration of errors of the same type (either underestimate or overestimate) is discouraged.

We show that using this generalized quadratic loss function in a Support Vector Regression method, the resulting dual problem can be expressed as a hard margin SVR in a new feature space which is related to the original feature space via the inverse of the similarity matrix and the target information. Thus, in order to get a well-formed dual formulation we need to work with a similarity matrix which is invertible.

We compare our approach against a standard SVR with ϵ-insensitive loss on a couple of regression tasks. The experimental results seem to show an improvement in the performance.

2. SVR definition for a generalized quadratic loss

Suppose that l inputs $(\vec{x}_1, y_1), \ldots, (\vec{x}_l, y_l)$ are given, where $\mathbf{x_i} \in \mathbb{R}^{\mathbf{d}}$ are the input patterns, and $y_i \in \mathcal{R}$ are the related target values of our supervised regression problem. The standard SVR model for 2-norm ϵ-insensitive loss

function [Cristianini and Shawe-Taylor, 2000], that we denote QSVR, is:

$$\begin{array}{ll} & \min_{\vec{w},b,\vec{\xi},\vec{\xi^*}} ||\vec{w}||^2 + c(\vec{\xi}'\vec{\xi} + \vec{\xi^*}'\vec{\xi^*}) \\ \text{s.t.:} & \\ & \vec{w}\cdot\vec{x}_i + b - y_i \leq \epsilon + \xi_i, \qquad i = 1,\ldots,l \\ & y_i - \vec{w}\cdot\vec{x}_i + b \leq \epsilon + \xi_i^*, \qquad i = 1,\ldots,l \end{array} \tag{1}$$

where $\vec{w}$ and b are the parameters of the linear regressor $\vec{w}\vec{x} + b$, ξ_i is the slack variable associated to an over-estimate of the linear regressor over input $\vec{x}_i$ and ξ_i^* is the slack variable associated to an under-estimate on the same pattern; ϵ determines the size of the approximation tube and c is the constant that controls the tradeoff between the empirical error as measured by the loss function and the regularization term. Note that non negativity constraints over $\vec{\xi}$ and $\vec{\xi^*}$ components are redundant. The solution of (1) can be expressed in general using a kernel function $K(\vec{x},\vec{y})$ with $f(\vec{x}) = \frac{1}{2}\sum_{i=1}^{l}(\alpha_i^{*+} - \alpha_i{}^{+})K(\vec{x}_i,\vec{x}) + b^+$ where $\vec{\alpha^{*+}}$, $\vec{\alpha^+}$ is the dual optimal solution and an optimal bias value b^+ can be derived from the KKT conditions.

To weight the co-occurrence of errors corresponding to close patterns we adopted the following formulation :

$$\begin{array}{ll} & \min_{\vec{w},b,\vec{\xi},\vec{\xi^*}} ||\vec{w}||^2 + c(\vec{\xi} - \vec{\xi^*})'S(\vec{\xi} - \vec{\xi^*})) \\ \text{s.t.:} & \\ & \vec{w}\cdot\vec{x}_i + b - y_i \leq \epsilon + \xi_i, \qquad i = 1,\ldots,l \\ & y_i - \vec{w}\cdot\vec{x}_i - b \leq \epsilon + \xi_i^*, \qquad i = 1,\ldots,l \end{array} \tag{2}$$

where S is a positive definite matrix. Defining $\delta_i = \xi_i - \xi_i^*$ we obtain:

$$\begin{array}{ll} & \min_{\vec{w},b,\vec{\delta}} ||\vec{w}||^2 + c\vec{\delta}'S\vec{\delta} \\ \text{s.t.:} & \\ & \vec{w}\cdot\vec{x}_i + b - y_i \leq \epsilon + \delta_i + \xi_i^*, \quad i = 1,\ldots,l \\ & y_i - \vec{w}\cdot\vec{x}_i + b \leq \epsilon - \delta_i + \xi_i, \quad i = 1,\ldots,l \end{array} \tag{3}$$

and since when one of the first constraints is active, the related ξ_i^* is 0, and viceversa, when one the second constraints is active, the related ξ_i is 0, we can write:

$$\begin{array}{ll} & \min_{\vec{w},b,\vec{\delta}} ||\vec{w}||^2 + c\vec{\delta}'S\vec{\delta} \\ \text{s.t.:} & \\ & \vec{w}\cdot\vec{x}_i + b - y_i \leq \epsilon + \delta_i, \quad i = 1,\ldots,l \\ & y_i - \vec{w}\cdot\vec{x}_i + b \leq \epsilon - \delta_i, \quad i = 1,\ldots,l \end{array} \tag{4}$$

Finally we obtain:

$$\begin{array}{ll} & \min_{\vec{w},b,\vec{\delta}} ||\vec{w}||^2 + c\vec{\delta}'S\vec{\delta} \\ \text{s.t.:} & \\ & -\epsilon \leq \vec{w}\cdot\vec{x}_i + b - y_i - \delta_i \leq \epsilon, \quad i = 1,\ldots,l \end{array} \tag{5}$$

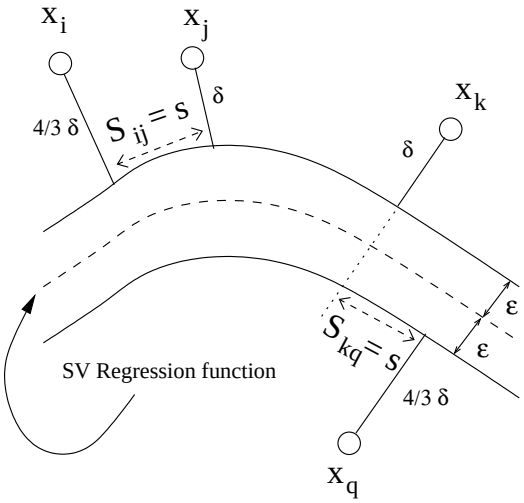

Figure 1. In our generalized quadratic loss, the error configuration generated by patterns $\vec{x}_i$ and $\vec{x}_j$ is more expensive than the error configuration generated by patterns $\vec{x}_k$ and $\vec{x}_q$. Here we assume that $S_{ij} = S_{kq} = s$.

A solution of this problem is a function with the best tradeoff between its smoothness and a uniform error on the training set. In addition, since we are considering signed slack variables ($\vec{\delta}$), we penalize errors on close patterns of the same sign, preferring errors with opposite signs. In Figure 1 we give a graphical exemplification about which type of error co-occurrence we prefer to penalize. Let X be the $l \times d$ matrix of input patterns. Given problem (5) the corresponding Lagrangian objective function is:

$$L = ||\vec{w}||^2 + c\vec{\delta}' S\vec{\delta} + \vec{\alpha}'(X\vec{w} + b\vec{1} - \vec{y} - \vec{\delta} - \epsilon\vec{1}) + \vec{\alpha^*}'(\vec{\delta} - X\vec{w} - b\vec{1} + \vec{y} - \epsilon\vec{1}) \quad (6)$$

where $\alpha_i \geq 0$, $\alpha_i^* \geq 0$ for $i = 1, \ldots, l$.

The Kuhn Tucker conditions for optimality are:

$$\begin{array}{l} \frac{\partial L}{\partial \vec{w}} = 2\vec{w} + X'(\vec{\alpha} - \vec{\alpha^*}) = 0 \Rightarrow \vec{w} = \frac{1}{2}X'(\vec{\alpha^*} - \vec{\alpha}) \\ \frac{\partial L}{\partial b} = (\vec{\alpha} - \vec{\alpha^*})'\vec{1} = 0 \Rightarrow (\vec{\alpha^*} - \vec{\alpha})'\vec{1} = 0 \\ \frac{\partial L}{\partial \vec{\delta}} = 2cS\vec{\delta} - (\vec{\alpha} - \vec{\alpha^*}) = 0 \Rightarrow \vec{\delta} = \frac{S^{-1}(\vec{\alpha} - \vec{\alpha^*})}{2c} \end{array} \quad (7)$$

if S is invertible. Supposing that S^{-1} exists, substituting (7) in (6) gives:

$$\begin{array}{c} \max_{\vec{\alpha}, \vec{\alpha^*}} (\vec{\alpha^*} - \vec{\alpha})\vec{y} - \epsilon(\vec{\alpha^*} + \vec{\alpha})'\vec{1} - \frac{1}{2}(\vec{\alpha^*} - \vec{\alpha})'\frac{1}{2}(K + \frac{S^{-1}}{c})(\vec{\alpha^*} - \vec{\alpha}) \\ \text{s.t.: } (\vec{\alpha^*} - \vec{\alpha})'\vec{1} = 0, \;\; \alpha_i \geq 0, \alpha_i^* \geq 0 \;\; i = 1, \ldots, l \end{array} \quad (8)$$

Notice that when S^{-1} exists, problem (8) is equivalent to a hard margin SVR problem with a kernel matrix equal to $\frac{1}{2}(K + \frac{S^{-1}}{c})$, while the regression

function is defined over the feature space induced by kernel K. Actually, in this case it is also possible to explicitly build a feature map. Let consider the following mapping $\phi : \mathbb{R}^d \rightarrow \mathbb{R}^{d+l}$ that, for all $i \in [1, \ldots, l]$, maps $\vec{x}_i \mapsto \phi(\vec{x}_i)$: $\phi(\vec{x}_i) = [\vec{x}'_i, (\sqrt{\frac{S^{-1}}{c}}\vec{e}_i)']'$ where $\vec{e}_i$ is the i-th vector of the canonical base of $\mathbb{R}^l$. It is not difficult to see that the kernel matrix obtained with this transformation is equal to $K + \frac{S^{-1}}{c}$. In the following we denote the overall method with QLSVR.

3. Definition of the similarity matrix

The dual solution of problem (5) is based on the inversion of S. Note that when all patterns are distinct points and S is generated by a Gaussian RBF kernel then S is invertible ([Micchelli, 1998]). Under some experimental conditions, however, a similarity matrix defined in this way may be ill-conditioned and inversion can be problematic.

For this reason we also considered an exponential kernel $e^{\nu K}$, defined by $e^{\nu K} = \sum_{i=0}^{+\infty} \frac{\nu^i}{i!} K^i$. A kernel matrix obtained by this formula is always invertible and its inverse is $(e^{\nu K})^{-1} = e^{-\nu K}$. Experimentally we never had problems in computing the inverse of the exponential matrix.

A similarity matrix generated by an RBF kernel can be understood as a way to take into account local similarity between patterns, where the amount of locality is regulated by the width of the RBF function. The exponential kernel, besides to guarantee the invertibility of the S matrix, has been proposed in the context of discrete domains [Kondor and Lafferty, 2002], and it appears to be particularly suited when the instance space is composed of structured objects, such as sequences or trees.

4. Experiments

To measure the performance of the regression methods we used the average absolute error ($AAE = \frac{1}{l}\sum_{i=1}^{m} |y_i - f(\vec{x}_i)|$) and the average squared error ($ASE = \frac{1}{l}\sum_{i=1}^{m}(y_i - f(\vec{x}_i))^2$). Since the reported performances are averaged across different shuffles, we also report their standard deviation computed as $\sigma = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(E_i - \mu_E)^2}$, where n is the number of data shuffles, E_i is the AAE (or ASE) error on the i-th shuffle and μ_E is the mean AAE (or ASE) error on the shuffles set.

We tested the two regression methods on two datasets: the Abalone dataset from the UCI repository and a QSPR problem involving alkanes, i.e. chemical compounds represented as trees. For both datasets we report the results obtained by SVR and QLSVR. We employed a modified version of SVMLight 5.0 [Joachims, 1998] enabled to work with a kernel matrix generated by Scilab 2.7 ©INRIA-ENPC.

The Abalone dataset comprises 3000 training patterns and 1177 test patterns and the input patterns are normalized to zero mean and unit variance coordinate-wise. We considered 10 independent shuffles of the Abalone dataset and we calibrated the hyperparameters using a split of each original training set. The calibration procedure is based on the first 2000 patterns for training and on the last 1000 patterns for validation.

For the SVR algorithm we adopted a RBF kernel $K(\vec{x}, \vec{y}) = e^{-\gamma||\vec{x}-\vec{y}||^2}$ for the input feature space. We applied on each shuffle of the dataset a calibration process that involved a 5×5 mesh of powers of 10 starting from $10, 0.1$ for c and γ, while the ϵ parameter was increased by steps of size 0.3 starting from 0 up to 1.2. For each shuffle we selected the hyperparameters set that gave the best performance in terms of ASE, we trained the SVR on the original training set, and finally the obtained regressor was evaluated on the original test problem.

For QLSVR we considered the same setting as the SVR and a similarity matrix S generated by an RBF kernel with parameter γ_S. During the calibration phase γ_S was varied from 4 to 24 by steps of size 5. Hyperparameters selection and final evaluation were performed using the same procedure as adopted for SVR.

We also considered a QSPR problem consisting in the prediction of the boiling point for a group of acyclic hydrocarbons (alkanes). The dataset comprises 150 alkanes with up to 10 carbon atoms, each represented as a tree (for more details, see [Bianucci et al., 2000; Bianucci et al., 2003]). The target values are in the range [-164 , 174] in Celsius degrees.

In order to deal with trees as input instances, we have chosen the most popular and used Tree Kernel proposed in [Collins and Duffy, 2002]. It is based on counting matching subtrees between two input trees.

For the calibration of SVR hyperparameters, we shuffled the 150 compounds and we created 30 splits of 5 patterns each. The calibration involved a set of 3 parameters: the SVR training error weight constant c, the Tree Kernel downweighting factor λ and the SVR regression tube width ϵ. On the last 3 splits we applied a 3-fold cross validation that involved a 5×5 mesh of powers of 10 starting from $10, 0.1$ for c and $\sqrt{\lambda}$, while the ϵ parameter is increased by steps of size 0.01 starting from 0 up to 0.04. We selected the parameter vector that gave the median of the best AAE on the three splits and then we used these parameters on 10 different splits of the original dataset to obtain the final test results.

For QLSVR we considered the same setting as the SVR and a similarity matrix generated by an exponential kernel ($S = e^{\nu TK}$), since the exponential kernel has been proposed in the context of discrete domains [Kondor and Lafferty, 2002], such as set of trees. During the calibration phase ν was varied

Table 1. Results for the Abalone dataset. We report also the unbiased standard deviation measured on the 10 different shuffles of the dataset. SVR $_{\text{Chu}}$ refers to [Chu et al., 2004].

Method	*AAE tr*	*ASE tr*	*AAE ts*	*ASE ts*
SVR_{Chu}	-	-	0.454±0.009	0.441±0.021
SVR	0.432±0.008	0.397±0.017	0.456±0.010	0.435±0.020
QLSVR	0.006±2.2E-4	3.4E-5±2.9E-6	0.461±0.009	0.424±0.019

from 0.5 to 0.65 by steps of size 0.015. Hyperparameters selection and final evaluation were performed using the same procedure adopted for SVR.

The results for the Abalone dataset, both for the training set (tr) and the test set (ts), are shown in Table 1 where we report also the results obtained for SVR in [Chu et al., 2004]. From the experimental results it can be concluded that the proposed approach and the SVR method give a similar result in terms of the absolute mean error, while the quadratic loss produces an improved mean squared error with a reduced standard deviation.

Table 2 reports the results obtained for the Alkanes dataset, including the values for the hyperparameters, as returned by the calibration process described above. Also in this case we got a similar result in terms of the absolute mean error, while the quadratic loss produces an slightly improved mean squared error, but with an increased standard deviation.

These results, however, should be considered very preliminary for the QLSVR method, since the presence of an additional hyperparameter for the generation of the similarity matrix S, as well as the possibility to use different methods for its generation, require a more intensive set of experiments in order to get a better coverage for S.

Table 2. Results for the alkanes dataset. We report also the unbiased standard deviation measured on the 10 different shuffles of the dataset.

Method	*Parameters*	*AAE tr*	*ASE tr*	*AAE ts*	*ASE ts*
SVR	$c = 1E5$ $\lambda = 0.25$ $\epsilon = 0.02$	1.68±0.03	3.19±0.08	3.82±0.97	30.27±32.08
QLSVR	$c = 1E4$ $\lambda = 0.25$ $\epsilon = 0.02$ $\nu = 0.8$	1.67±0.02	3.16±0.06	3.82±1.09	30.00± 32.63

5. Conclusions

In this paper we proposed a generalized quadratic loss for regression problems that exploits the similarity of the input patterns. In fact, the proposed generalized quadratic loss weights co-occurrence of errors on the basis of the similarity of the corresponding input patterns. Moreover errors of similar patterns of the same sign are discouraged. We derived a SVR formulation for the proposed loss showing that if the similarity matrix is invertible the problem is equivalent to a hard margin SVR problem with a kernel matrix which depends also on the inverse of the similarity loss matrix. Experimental results on two regression tasks seem to show an improvement in the performance.

A problem with this approach is the need to invert the similarity matrix and how to define it in a meaningful way. Thus further study will be devoted to these issues and to the extension of the framework to multiclass and ranking problems. Finally, the robustness of the approach should be studied, both theoretically and empirically.

References

Bianucci, A.M., Micheli, A., Sperduti, A., and Starita, A. (2000). Application of cascade correlation networks for structures to chemistry. *Journal of Applied Intelligence (Kluwer Academic Publishers)*, 12:117–146.

Bianucci, A.M., Micheli, A., Sperduti, A., and Starita, A. (2003). A novel approach to QSPR/QSAR based on neural networks for structures. In Sztandera, L.M. and Cartwright, H.M., editors, *Soft Computing Approaches in Chemistry*. Springer-Verlag.

Chu, W., Keerthi, S. S., and Ong, C. J. (2004). Bayesian support vector regression using a unified loss function. *IEEE Trans. on Neural Networks*, 15(1):29–44.

Collins, M. and Duffy, N. (2002). Convolution kernels for natural language. In *NIPS 14*, Cambridge, MA. MIT Press.

Cristianini, N. and Shawe-Taylor, J. (2000). *An Introduction to Support Vector Machines and other Kernel-based Learning Methods*. Cambridge University Press, Cambridge.

Joachims, T. (1998). Text categorization with support vector machines: learning with many relevant features. In *Proceedings of ECML-98, 10th European Conference on Machine Learning*, pages 137–142.

Kondor, R. and Lafferty, J. (2002). Diffusion kernels on graphs and other discrete input spaces. In *Proceedings of the Int. Conf. on Machine Learning, 2002*.

Micchelli, C.A. (1998). Algebraic aspects of interpolation. In *Proceedings of Symposia in Applied Mathematics*, pages 36:81–102.

Portera, Filippo and Sperduti, Alessandro (2004). A generalized quadratic loss for support vector machines. In *Proceedings of* 16^{th} *European Conference on Artificial Intelligence*, pages 628–632.

Vapnik, V.N. (1998). *Statistical Learning Theory*. Wiley, New York.

Zhang, T. and Oles, F. J. (2001). Text categorization based on regularized linear classification methods. *Information Retrieval*, 4:5–31.