A GENERAL FORMULATION FOR SUPPORT VECTOR MACHINES

Wei Chu, S. Sathiya Keerthi, Chong Jin Ong

Control Division, Department of Mechanical Engineering, National University of Singapore
10 Kent Ridge Crescent, Singapore, 119260
engp9354@nus.edu.sg, mpessk@nus.edu.sg, mpeongcj@nus.edu.sg

ABSTRACT

In this paper, we derive a general formulation of support vector machines for classification and regression respectively. L_e loss function is proposed as a patch of L_1 and L_2 soft margin loss functions for classifier, while soft insensitive loss function is introduced as the generalization of popular loss functions for regression. The introduction of the two loss functions results in a general formulation for support vector machines.

1. INTRODUCTION

As computationally powerful tools for supervised learning, support vector machines (SVMs) are widely used in classification and regression problems [10]. Let us suppose that a data set $\mathcal{D} = \{(x_i, y_i)|i = 1, \ldots, n\}$ is given for training, where the input vector $x_i \in \mathbb{R}^d$ and y_i is the target value. SVMs take the idea to map these input vectors into a high dimensional reproducing kernel Hilbert space (RKHS), where a linear machine is constructed by minimizing a regularized functional. The linear machine takes the form of $f(x) = \langle \mathbf{w} \cdot \phi(x) \rangle + b$, where $\phi(\cdot)$ is the mapping function, b is known as the bias, and the dot product $\langle \phi(x) \cdot \phi(x') \rangle$ is also the reproducing kernel $K(x, x')$ in the RKHS. The regularized functional is usually defined as

$$\mathcal{R}(\mathbf{w}, b) = C \cdot \sum_{i=1}^{n} \ell\big(y_i, f(x_i)\big) + \frac{1}{2}\|\mathbf{w}\|^2 \qquad (1)$$

where the regularization parameter $C > 0$, the norm of $\mathbf{w}$ in the RKHS is the stabilizer and $\sum_{i=1}^{n} \ell\big(y_i, f(x_i)\big)$ is empirical loss term. In standard SVMs, the regularized functional (1) can be minimized by solving a convex quadratic programming optimization problem that guarantees a unique global minimum solution.

Various loss functions can be used in SVMs that results in quadratic programming. In SVMs for classification [1], hard margin, L_1 soft margin and L_2 soft margin loss functions are widely used. For regression, a lot of common loss functions have been discussed in [9], such as Laplacian, Huber's, ϵ-insensitive and Gaussian etc.

In this paper, we generalize these popular loss functions and put forward new loss functions for classification and regression respectively. The two new loss functions are C^1 smooth. By introducing these loss functions in the regularized functional of classical SVMs in place of popular loss functions, we derive a general formulation for SVMs.

2. SUPPORT VECTOR CLASSIFIER

In classical SVMs for binary classification (SVC) in which the target values $y_i \in \{-1, +1\}$, the hard margin loss function is defined as

$$\ell_h(y_x \cdot f_x) = \begin{cases} 0 & if \quad y_x \cdot f_x \geq +1; \\ +\infty & otherwise. \end{cases} \qquad (2)$$

The hard margin loss function is suitable for noise-free data sets. For other general cases, a soft margin loss function is popularly used in classical SVC, which is defined as

$$\ell_\rho(y_x \cdot f_x) = \begin{cases} 0 & if \quad y_x \cdot f_x \geq +1; \\ \dfrac{1}{\rho}\,(1 - y_x \cdot f_x)^\rho & otherwise, \end{cases} \qquad (3)$$

where ρ is a positive integer. The minimization of the regularized functional (1) with the soft margin (3) as loss function leads to a convex programming problem for any positive integer ρ; for L_1 ($\rho = 1$) or L_2 ($\rho = 2$) soft margin, it is also a convex quadratic programming problem.

2.1. L_e Loss Function

We generalize the L_1 and L_2 soft margin loss functions as the L_e soft margin loss function, which is defined as

$$\ell_e(y_x \cdot f_x) = \begin{cases} 0 & if \quad y_x \cdot f_x > 1; \\ \dfrac{(1 - y_x \cdot f_x)^2}{4\epsilon} & if \quad 1 \geq y_x \cdot f_x \geq 1 - 2\epsilon; \\ (1 - y_x \cdot f_x) - \epsilon & otherwise, \end{cases} \qquad (4)$$

where the parameter $\epsilon > 0$.

Wei Chu gratefully acknowledges the financial support provided by the National University of Singapore through Research Scholarship.

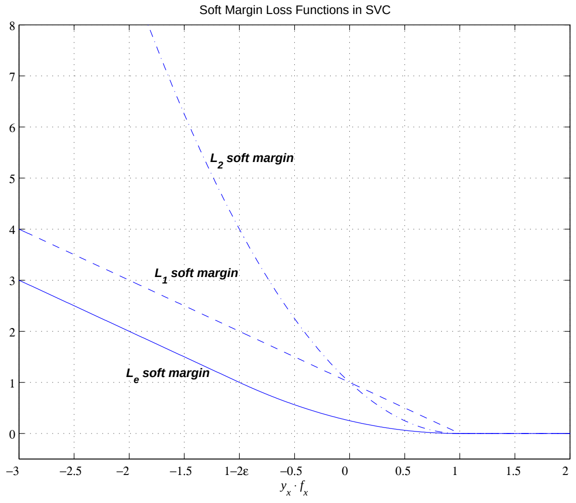

Figure 1: Graphs of soft margin loss functions, where ϵ is set at 1.

From their definitions and Figure 1, we find that the L_e soft margin approaches to L_1 soft margin as the parameter $\epsilon \rightarrow 0$. Let ϵ be fixed at some large value, the L_e soft margin approaches the L_2 soft margin for all practical purposes.

2.2. Optimization Problem

The minimization problem in SVC (1) with the L_e soft margin loss function can be rewritten as the following equivalent optimization problem by introducing slack variables $\xi_i \geq 1 - y_i \cdot \big(\langle \mathbf{w} \cdot \phi(x_i) \rangle + b\big)\ \forall i$, which we refer to as the *primal* problem

$$\min_{\mathbf{w},b,\boldsymbol{\xi}} \mathcal{R}(\mathbf{w},b,\boldsymbol{\xi}) = C \cdot \sum_{i=1}^{n} \psi_e\big(\xi_i\big) + \frac{1}{2}\|\mathbf{w}\|^2 \tag{5}$$

subject to

$$\begin{cases} y_i \cdot \big(\langle \mathbf{w} \cdot \phi(x_i) \rangle + b\big) \geq 1 - \xi_i, & \forall i \\ \xi_i \geq 0, \quad \forall i \end{cases} \tag{6}$$

where

$$\psi_e(\xi) = \begin{cases} \dfrac{\xi^2}{4\epsilon} & if \quad \xi \in [0, 2\epsilon]; \\ \xi - \epsilon & if \quad \xi \in (2\epsilon, +\infty). \end{cases} \tag{7}$$

Standard Lagrangian techniques [4] are used to derive the *dual* problem. Let $\alpha_i \geq 0$ and $\gamma_i \geq 0$ be the corresponding Lagrange multipliers for the inequalities in the *primal* problem (6), and then the Lagrangian for the *primal* problem would be:

$$\begin{aligned} L(\mathbf{w},b,\boldsymbol{\xi}) \quad = \quad & C \cdot \textstyle\sum_{i=1}^{n} \psi_e\big(\xi_i\big) + \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{n} \gamma_i \cdot \xi_i \\ & - \sum_{i=1}^{n} \alpha_i \cdot \big(y_i \cdot (\langle \mathbf{w} \cdot \phi(x_i) \rangle + b) - 1 + \xi_i\big) \end{aligned} \tag{8}$$

The KKT conditions for the *primal* problem (5) require

$$\mathbf{w} = \sum_{i=1}^{n} y_i \cdot \alpha_i \cdot \phi(x_i) \tag{9}$$

$$\sum_{i=1}^{n} y_i \cdot \alpha_i = 0 \tag{10}$$

$$C \cdot \frac{\partial \psi_e(\xi_i)}{\partial \xi_i} = \alpha_i + \gamma_i \quad \forall i \tag{11}$$

Based on the definition of $\psi_e(\cdot)$ given in (7) and the constraint condition (11), an equality constraint on Lagrange multipliers can be explicitly written as

$$\begin{array}{l} C \cdot \frac{\xi_i}{2\epsilon} = \alpha_i + \gamma_i \quad if \quad 0 \leq \xi_i \leq 2\epsilon \\ C = \alpha_i + \gamma_i \quad if \quad \xi_i > 2\epsilon \quad \forall i \end{array} \tag{12}$$

If we collect all terms involving ξ_i in the Lagrangian (8), we get $T_i = C\psi_e(\xi_i) - (\alpha_i + \gamma_i)\xi_i$. Using (7) and (12) we can rewrite T_i as

$$T_i = \begin{cases} -\dfrac{\epsilon}{C}(\alpha_i + \gamma_i)^2 & if \quad \xi \in [0, 2\epsilon]; \\ -C\epsilon & if \quad \xi \in (2\epsilon, +\infty). \end{cases} \tag{13}$$

Thus the ξ_i can be eliminated if we set $T_i = -\dfrac{\epsilon}{C}(\alpha_i + \gamma_i)^2$ and introduce the additional constraints $0 \leq \alpha_i + \gamma_i \leq C$. Then the *dual* problem can be stated as a maximization problem in terms of the positive dual variables α_i and γ_i:

$$\begin{aligned} \max_{\boldsymbol{\alpha},\boldsymbol{\gamma}} \mathcal{R}(\boldsymbol{\alpha},\boldsymbol{\gamma}) = & \sum_{i=1}^{n} \alpha_i - \frac{\epsilon}{C} \sum_{i=1}^{n} (\alpha_i + \gamma_i)^2 \\ & - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \cdot y_i \cdot \alpha_j \cdot y_j \cdot \langle \phi(x_i) \cdot \phi(x_j) \rangle \end{aligned} \tag{14}$$

subject to

$$\alpha_i \geq 0, \gamma_i \geq 0, 0 \leq \alpha_i + \gamma_i \leq C, \forall i \text{ and } \sum_{i=1}^{n} \alpha_i \cdot y_i = 0. \tag{15}$$

It is noted that $\mathcal{R}(\boldsymbol{\alpha},\boldsymbol{\gamma}) \leq \mathcal{R}(\boldsymbol{\alpha},0)$ for any $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$ satisfying (15). Hence the maximization of (14) over $(\boldsymbol{\alpha},\boldsymbol{\gamma})$ can be found as maximizing $\mathcal{R}(\boldsymbol{\alpha},0)$ over $0 \leq \alpha_i \leq C$ and $\sum_{i=1}^{n} \alpha_i \cdot y_i = 0$. Therefore, the *dual* problem can be finally simplified as

$$\min_{\boldsymbol{\alpha}} \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \cdot y_i \cdot \alpha_j \cdot y_j \cdot K(x_i, x_j) - \sum_{i=1}^{n} \alpha_i + \frac{\epsilon}{C} \sum_{i=1}^{n} \alpha_i^2 \tag{16}$$

subject to $0 \leq \alpha_i \leq C, \forall i$ and $\sum_{i=1}^{n} \alpha_i \cdot y_i = 0$. With the equality (9), the linear classifier can be obtained from the solution of (16) as $f(x) = \sum_{i=1}^{n} \alpha_i \cdot y_i \cdot K(x_i, x) + b$ where b can be easily obtained as a byproduct in the solution. In most of the cases, only some of the Lagrange multipliers, α_i, differ from zero at the optimal solution. They define the support vectors (SVs) of the problem. More exactly, the training samples (x_i, y_i) associated with α_i satisfying $0 < \alpha_i < C$ are called off-bound SVs, the samples with $\alpha_i = C$ are called on-bound SVs, and the samples with $\alpha_i = 0$ are called non-SVs.

2.3. Discussion

The above formulation (16) is a general framework for classical SVC. There are three special cases of the formulation:

1. L_1 soft margin: the formulation is just the SVC using L_1 soft margin if we set $\epsilon = 0$.
2. Hard margin: when we set $\epsilon = 0$ and keep C large enough to prevent any α_i from reaching the upper bound C, the solution of this formulation is identical to the standard SVC with hard margin loss function.
3. L_2 soft margin: when we set $\dfrac{C}{2\epsilon}$ equal to the regularization parameter in SVC with L_2 soft margin, and keep C large enough to prevent any α_i from reaching the upper bound at the optimal solution, the solution will be same as that of the standard SVC with L_2 soft margin loss function.

In practice, such as on unbalanced data sets, we would like to use different regularization parameters C_+ and C_- for the samples with positive label and negative label separately.[1] As an extremely general case, we can use different regularization parameter C_i for every sample (x_i, y_i). It is straightforward to obtain the general *dual* problem as

$$\min_{\boldsymbol{\alpha}} \frac{1}{2} \sum_{i=1}^{n} \sum_{i=1}^{n} \alpha_i \cdot y_i \cdot \alpha_j \cdot y_j \cdot K(x_i, x_j) - \sum_{i=1}^{n} \alpha_i + \frac{\epsilon}{C_i} \sum_{i=1}^{n} \alpha_i^2 \tag{17}$$

subject to $0 \leq \alpha_i \leq C_i, \forall i$ and $\sum_{i=1}^{n} \alpha_i \cdot y_i = 0$. Obviously, the *dual* problems (17) is a constrained convex quadratic programming problem. Denoting $\hat{\boldsymbol{\alpha}} = [y_1\alpha_1, y_2\alpha_2, \ldots, y_n\alpha_n]$, $\boldsymbol{P} = [-y_1, -y_2, \ldots, -y_n]^T$ and $\boldsymbol{Q} = \boldsymbol{K} + \boldsymbol{\Lambda}$ where $\boldsymbol{\Lambda}$ is a $n \times n$ diagonal matrix with $\boldsymbol{\Lambda}_{ii} = \frac{2\epsilon}{C_i}$ and $\boldsymbol{K}$ is the kernel matrix with $\boldsymbol{K}_{ij} = K(x_i, x_j)$, (17) can be written in a general form as

$$\min_{\hat{\boldsymbol{\alpha}}} \frac{1}{2} \hat{\boldsymbol{\alpha}}^T \boldsymbol{Q} \hat{\boldsymbol{\alpha}} + \boldsymbol{P}^T \hat{\boldsymbol{\alpha}} \tag{18}$$

subject to $l_i \leq \hat{\alpha}_i \leq u_i$, $\forall i$ and $\sum_{i=1}^{n} \hat{\alpha}_i = 0$ where $l_i = 0$, $u_i = C_i$ when $y_i = +1$ and $l_i = -C_i$, $u_i = 0$ when $y_i = -1$. As to the algorithm design for the solution, matrix-based quadratic programming techniques that use the "chunking" idea can be employed here. Popular SMO algorithms [7, 5] could be easily adapted for the solution. For a program design and source code, refer to [2].

3. SUPPORT VECTOR REGRESSION

A lot of loss functions for support vector regression (SVR) have been discusses in [9] that leads to a general optimization problem. There are four popular loss functions widely used for regression problems. They are

1. Laplacian loss function: $\ell_l(\delta) = |\delta|$.
2. Huber's loss function:
$$\ell_h(\delta) = \begin{cases} \frac{\delta^2}{4\epsilon} & if\ |\delta| \leq 2\epsilon \\ |\delta| - \epsilon & otherwise. \end{cases}$$
3. ϵ-insensitive loss function:
$$\ell_\epsilon(\delta) = \begin{cases} 0 & if\ |\delta| \leq \epsilon \\ |\delta| - \epsilon & otherwise. \end{cases}$$
4. Gaussian loss function: $\ell_g(\delta) = \frac{1}{2}\delta^2$.

We introduce another loss function as the generalization of these popular loss functions.

3.1. Soft Insensitive Loss Function

The loss function known as soft insensitive loss function (SILF) is defined as:

$$\ell_{\epsilon,\beta}(\delta) = \begin{cases} |\delta| - \epsilon & if \quad |\delta| > (1+\beta)\epsilon \\ \frac{(|\delta| - (1-\beta)\epsilon)^2}{4\beta\epsilon} & if \quad (1+\beta)\epsilon \geq |\delta| \geq (1-\beta)\epsilon \\ 0 & if \quad |\delta| < (1-\beta)\epsilon \end{cases} \tag{19}$$

where $0 < \beta \leq 1$ and $\epsilon > 0$. There is a profile of SILF as shown in Figure 2. The properties of SILF are entirely controlled by two parameters, β and ϵ. For a fixed ϵ, SILF approaches the ϵ-ILF as $\beta \to 0$; on the other hand, as $\beta \to 1$, it approaches the Huber's loss function. In addition, SILF becomes the Laplacian loss function as $\epsilon \to 0$. Hold ϵ fixed at some large value and let $\beta \to 1$, the SILF approach the quadratic loss function for all practical purposes. The application of SILF in Bayesian SVR has been discussed in [3].

3.2. Optimization Problem

We introduce SILF into the regularized functional (1) that will leads to a quadratic programming problem that could work as a general framework. As usual, two slack variables

[1] The ratio between C_+ and C_- is usually fixed at $\frac{C_+}{C_-} = \frac{N_-}{N_+}$, where N_+ is the number of samples with positive label and N_- is the number of samples with negative label.

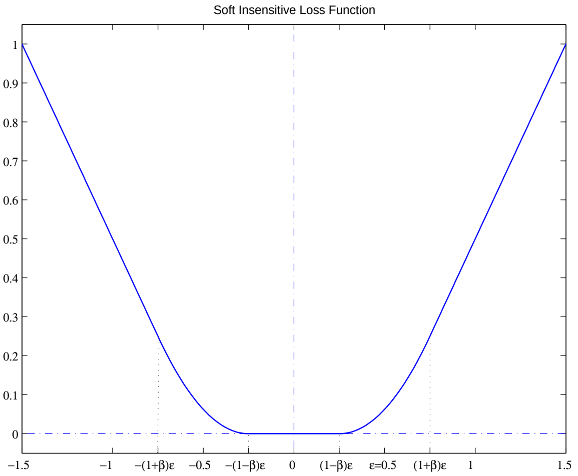

Figure 2: Graphs of soft insensitive loss functions, where $\epsilon = 0.5$ and $\beta = 0.5$.

ξ_i and ξ_i^* are introduced as $\xi_i \geq y_i - \langle \mathbf{w} \cdot \phi(x_i) \rangle - b - (1-\beta)\epsilon$ and $\xi_i^* \geq \langle \mathbf{w} \cdot \phi(x_i) \rangle + b - y_i - (1-\beta)\epsilon \; \forall i$. The minimization of the regularized functional (1) with SILF as loss function could be rewritten as the following equivalent optimization problem, which is usually called *primal* problem:

$$\min_{\mathbf{w}, b, \boldsymbol{\xi}, \boldsymbol{\xi}^*} \mathcal{R}(\mathbf{w}, b, \boldsymbol{\xi}, \boldsymbol{\xi}^*) = C \sum_{i=1}^{n} \big(\psi_s(\xi_i) + \psi_s(\xi_i^*) \big) + \frac{1}{2} \|\mathbf{w}\|^2 \tag{20}$$

subject to

$$\begin{cases} y_i - \langle \mathbf{w} \cdot \phi(x_i) \rangle - b \leq (1-\beta)\epsilon + \xi_i; \\ \langle \mathbf{w} \cdot \phi(x_i) \rangle + b - y_i \leq (1-\beta)\epsilon + \xi_i^*; \\ \xi_i \geq 0; \quad \xi_i^* \geq 0 \quad \forall i \end{cases} \tag{21}$$

where

$$\psi_s(\xi) = \begin{cases} \dfrac{\xi^2}{4\beta\epsilon} & if \quad \xi \in [0, 2\beta\epsilon]; \\ \xi - \beta\epsilon & if \quad \xi \in (2\beta\epsilon, +\infty). \end{cases} \tag{22}$$

Let $\alpha_i \geq 0$, $\alpha_i^* \geq 0$, $\gamma_i \geq 0$ and $\gamma_i^* \geq 0 \; \forall i$ be the corresponding Lagrangian multipliers for the inequalities in (21). The KKT conditions for the *primal* problem require

$$\mathbf{w} = \sum_{i=1}^{n} (\alpha_i - \alpha_i^*) \cdot \phi(x_i) \tag{23}$$

$$\sum_{i=1}^{n} (\alpha_i - \alpha_i^*) = 0 \tag{24}$$

$$C \cdot \frac{\partial \psi_e(\xi_i)}{\partial \xi_i} = \alpha_i + \gamma_i \quad \forall i \tag{25}$$

$$C \cdot \frac{\partial \psi_e(\xi_i^*)}{\partial \xi_i^*} = \alpha_i^* + \gamma_i^* \quad \forall i \tag{26}$$

Based on the definition of ψ_s given by (22), the constraint condition (26) could be explicitly written as equality constraints:

$$\begin{aligned} \alpha_i + \gamma_i &= C \cdot \frac{\xi_i}{2\beta\epsilon} && if \quad 0 \leq \xi_i \leq 2\beta\epsilon \\ \alpha_i + \gamma_i &= C && if \quad \xi_i > 2\beta\epsilon \end{aligned} \tag{27}$$

$$\begin{aligned} \alpha_i^* + \gamma_i^* &= C \cdot \frac{\xi_i^*}{2\beta\epsilon} && if \quad 0 \leq \xi_i^* \leq 2\beta\epsilon \\ \alpha_i^* + \gamma_i^* &= C && if \quad \xi_i^* > 2\beta\epsilon \end{aligned} \tag{28}$$

Following the analogous arguments as SVC did in (13), we can also eliminate ξ_i and ξ_i^* here. That yields a maximization problem in terms of the positive *dual* variables $\mathbf{w}, b, \boldsymbol{\xi}, \boldsymbol{\xi}^*$:

$$\begin{aligned} \max_{\boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\alpha}^*, \boldsymbol{\gamma}^*} \quad & -\frac{1}{2} \sum_{i=1}^{n} \sum_{i=1}^{n} (\alpha_i - \alpha_i^*)(\alpha_j - \alpha_i^*) \langle \phi(x_i) \cdot \phi(x_j) \rangle \\ & + \sum_{i=1}^{n} (\alpha_i - \alpha_i^*) y_i - \sum_{i=1}^{n} (\alpha_i + \alpha_i^*)(1-\beta)\epsilon \\ & - \frac{\beta\epsilon}{C} \sum_{i=1}^{n} \left((\alpha_i + \gamma_i)^2 + (\alpha_i^* + \gamma_i^*)^2 \right) \end{aligned} \tag{29}$$

subject to $\alpha_i \geq 0$, $\alpha_i^* \geq 0$, $\gamma_i \geq 0$, $\gamma_i^* \geq 0$, $0 \leq \alpha_i + \gamma_i \leq C, \forall i$, $0 \leq \alpha_i^* + \gamma_i^* \leq C, \forall i$ and $\sum_{i=1}^{n} (\alpha_i - \alpha_i^*) = 0$. As γ_i and γ_i^* only appear in the last term in (29), the functional (29) is maximal when $\gamma_i = 0$ and $\gamma_i^* = 0 \; \forall i$. Thus, the *dual* problem can be simplified as

$$\begin{aligned} \min_{\boldsymbol{\alpha}, \boldsymbol{\alpha}^*} \quad \mathcal{R}(\boldsymbol{\alpha}, \boldsymbol{\alpha}^*) = & \frac{1}{2} \sum_{i=1}^{n} \sum_{i=1}^{n} (\alpha_i - \alpha_i^*)(\alpha_j - \alpha_i^*) K(x_i, x_j) \\ & - \sum_{i=1}^{n} (\alpha_i - \alpha_i^*) y_i + \sum_{i=1}^{n} (\alpha_i + \alpha_i^*)(1-\beta)\epsilon \\ & + \frac{\beta\epsilon}{C} \sum_{i=1}^{n} (\alpha_i^2 + \alpha_i^{*2}) \end{aligned} \tag{30}$$

subject to $0 \leq \alpha_i \leq C, \forall i$, $0 \leq \alpha_i^* \leq C, \forall i$ and $\sum_{i=1}^{n} (\alpha_i - \alpha_i^*) = 0$. With the equality (23), the dual form of the regression function can be written as $f(x) = \sum_{i=1}^{n} (\alpha_i - \alpha_i^*) K(x_i, x_j) + b$.

3.3. Discussion

Like SILF, the *dual* problem (30) is a generalization of the several SVR formulations. More exactly,

1. when we set $\beta = 0$, (30) becomes the classical SVR using ϵ-insensitive loss function;

2. When $\beta = 1$, (30) becomes that when the Huber's loss function is used.

3. When $\beta = 0$ and $\epsilon = 0$, (30) becomes that for the case of the Laplacian loss function.

4. As for the optimization problem (1) using Gaussian loss function with the variance σ^2, that is equivalent to the general SVR (30) with $\beta = 1$ and $\dfrac{2\epsilon}{C} = \sigma^2$ provided that we keep upper bound C large enough to prevent any α_i and α_i^* from reaching the upper bound at the optimal solution. That is also the case of least-square support vector machines. As for a SMO algorithm design, refer to [6].

The *dual* problem (30) is also a constrained convex quadratic programming problem. Let us denote

$$\begin{aligned}
&\hat{\boldsymbol{\alpha}} = [\alpha_1, \ldots, \alpha_n, -\alpha_1^*, \ldots, -\alpha_n^*]^T, \\
&\boldsymbol{P} = [-y_1 + (1-\beta)\epsilon, \ldots, -y_n + (1-\beta)\epsilon, \\
&\qquad -y_1 - (1-\beta)\epsilon, \ldots, -y_n - (1-\beta)\epsilon]^T \\
&\boldsymbol{Q} = \begin{bmatrix} \boldsymbol{K} + \frac{2\beta\epsilon}{C}\boldsymbol{I} & \boldsymbol{K} \\ \boldsymbol{K} & \boldsymbol{K} + \frac{2\beta\epsilon}{C}\boldsymbol{I} \end{bmatrix}
\end{aligned}$$

where $\boldsymbol{K}$ is the kernel matrix with ij-entry $K(x_i, x_j)$, (30) can be rewritten as

$$\min_{\hat{\boldsymbol{\alpha}}} \frac{1}{2}\hat{\boldsymbol{\alpha}}^T \boldsymbol{Q} \hat{\boldsymbol{\alpha}} + \boldsymbol{P}^T \hat{\boldsymbol{\alpha}} \qquad (31)$$

subject to $l_i \leq \hat{\alpha}_i \leq u_i$, $\forall i$ and $\sum_{i=1}^{2n} \hat{\alpha}_i = 0$ where $l_i = 0$, $u_i = C$ for $1 \leq i \leq n$ and $l_i = -C$, $u_i = 0$ for $n+1 \leq i \leq 2n$.

4. CONCLUSION

In this paper, we introduce two loss functions as the generalization of popular loss functions for SVC and SVR respectively. The dual problem of SVMs arising from the introduction of the two loss functions could be used as a general formulation for SVMs. The optimization problems in SVMs with popular loss functions could be obtained as the special cases in the general formulation we derived. As a byproduct, the general formulation also provides a framework that facilitates the algorithm implementation.

5. REFERENCES

[1] Burges, C. J. C. A tutorial on support vector machines for pattern recogintion, *Data Mining and Knowledge Discovery*, 2(2) pp. 121-167, 1998.

[2] Chu, W and Keerthi, S. S. and Ong, C. J., Extended SVMs - Theory and Implementation, *Technical Report CD-01-14*, Mechanical Engineering, National University of Singapore, 2001. http://guppy.mpe.nus.edu.sg/∼chuwei/svm/esvm.pdf

[3] Chu, W and Keerthi, S. S. and Ong, C. J., A unified loss function in Bayesian framework for support vector regression, *Proceeding of the 18th International Conference on Machine Learning*, 2001. http://guppy.mpe.nus.edu.sg/∼mpessk/svm/icml.pdf

[4] Fletcher, R., *Practical methods of optimization*, John Wiley, 1987.

[5] Keerthi, S. S. and Shevade, S. K. and Bhattacharyya, C. and Murthy, K. R. K., Improvements to Platt's SMO algorithm for SVM classifier design, *Neural Computation*, 13 pp. 637-649, 2001.

[6] Keerthi, S. S. and Shevade, SMO algorithm for least squares SVM formulations, *Technical Report CD-02-08*, Dept. of Mechanical Engineering, National University of Singapore, 2002. http://guppy.mpe.nus.edu.sg/∼mpessk/papers/lssvm_smo.ps.gz

[7] Platt, J. C., Fast training of support vector machines using sequential minimal optimization, *Advances in Kernel Methods - Support vector learning*, pp. 185-208, 1998.

[8] Shevade, S. K. and Keerthi, S. S. and Bhattacharyya, C. and Murthy, K. R. K., Improvements to the SMO algorithm for SVM regression, *IEEE Transactions on Neural Networks*, 11 pp. 1188-1194, 2000.

[9] Smola, A. J. and Schölkopf, B., A tutorial on support vector regression, *Technical Report NC2-TR-1998-030*, GMD First, October, 1998.

[10] Vapnik, V. N. *The Nature of Statistical Learning Theory*, New York: Springer-Verlag, 1995.