EE6550 Machine Learning, Spring 2016

A Supplement of SMO Algorithm for HW#5 Programming Problem

Professor Chung-Chin Lu
Department of Electrical Engineering, National Tsing Hua University
June 20, 2016

This article is a detailed description of the sequential minimal optimization (SMO) algorithm to help students implement this efficient algorithm in HW# 5 programming problem for a kernel-based support vector regression (SVR).

The Kuhn-Tucker conditions :
On page 79 of Lecture 10: Regression, any feasible point $(\mathbf{w}, b, \eta, \eta')$, i.e.,

$$\begin{aligned}
(\mathbf{w}\cdot\Phi(\omega_i)+b)-c(\omega_i) &\le \epsilon+\eta_i,\ i\in[1,m], & (1)\\
c(\omega_i)-(\mathbf{w}\cdot\Phi(\omega_i)+b) &\le \epsilon+\eta_i',\ i\in[1,m], & (2)\\
\eta_i &\ge 0,\ i\in[1,m], & (3)\\
\eta_i' &\ge 0,\ i\in[1,m], & (4)
\end{aligned}$$

which satisfies the Kuhn-Tucker necessary conditions

$$\begin{aligned}
\mathbf{w} &= \sum_{i=1}^{m}(\lambda_i'-\lambda_i)\Phi(\omega_i), & (5)\\
0 &= \sum_{i=1}^{m}(\lambda_i'-\lambda_i), & (6)\\
C &= \lambda_i+\mu_i,\ i\in[1,m], & (7)\\
C &= \lambda_i'+\mu_i',\ i\in[1,m]; & (8)\\
\lambda_i((\mathbf{w}\cdot\Phi(\omega_i)+b)-c(\omega_i)-\epsilon-\eta_i) &= 0,\ i\in[1,m], & (9)\\
\lambda_i'(c(\omega_i)-(\mathbf{w}\cdot\Phi(\omega_i)+b)-\epsilon-\eta_i') &= 0,\ i\in[1,m], & (10)\\
\mu_i\eta_i &= 0,\ i\in[1,m], & (11)\\
\mu_i'\eta_i' &= 0,\ i\in[1,m]; & (12)\\
\lambda_i &\ge 0,\ i\in[1,m], & (13)\\
\lambda_i' &\ge 0,\ i\in[1,m], & (14)\\
\mu_i &\ge 0,\ i\in[1,m], & (15)\\
\mu_i' &\ge 0,\ i\in[1,m]. & (16)
\end{aligned}$$

is a global minimum solution of the primal problem of SVR.

Consequences of the Kuhn-Tucker conditions (1)-(16) : As discussed on page 80 of Lecture 10, for each $i\in[1,m]$, at most one of λ_i' and λ_i in any SVR solution is nonzero, otherwise a contradiction will occur by the complementary slackness conditions in (9) and (10). From (7), (8) and (13)-(16), there are five cases to consider:

Case 1: $\lambda_i = \lambda'_i = 0$.

Then by (7), (8), we have $\mu_i = C - \lambda_i = C = C - \lambda'_i = \mu'_i$ and then $\eta_i = \eta'_i = 0$ by (11), (12) so that

$$-\epsilon \leq (\mathbf{w} \cdot \Phi(\omega_i) + b) - c(\omega_i) \leq \epsilon$$

by (1) and (2).

Case 2: $0 < \lambda_i < C$ and $\lambda'_i = 0$.

Then by (9), we have

$$(\mathbf{w} \cdot \Phi(\omega_i) + b) - c(\omega_i) - \epsilon - \eta_i = 0.$$

Since $\mu_i = C - \lambda_i > 0$ by (7), we have $\eta_i = 0$ by (11) and then

$$(\mathbf{w} \cdot \Phi(\omega_i) + b) - c(\omega_i) = \epsilon.$$

Case 3: $\lambda_i = 0$ and $0 < \lambda'_i < C$.

Then by (10), we have

$$c(\omega_i) - (\mathbf{w} \cdot \Phi(\omega_i) + b) - \epsilon - \eta'_i = 0.$$

Since $\mu'_i = C - \lambda'_i > 0$ by (7), we have $\eta'_i = 0$ by (11) and then

$$(\mathbf{w} \cdot \Phi(\omega_i) + b) - c(\omega_i) = -\epsilon.$$

Case 4: $\lambda_i = C$ and$\lambda'_i = 0$.

Then by (9), we have

$$(\mathbf{w} \cdot \Phi(\omega_i) + b) - c(\omega_i) - \epsilon - \eta_i = 0.$$

Since $\mu_i = C - \lambda_i = 0$ by (7), we have $\eta_i \geq 0$ by (3) and then

$$(\mathbf{w} \cdot \Phi(\omega_i) + b) - c(\omega_i) \geq \epsilon.$$

Case 5: $\lambda_i = 0$ and $\lambda'_i = C$.

Then by (10), we have

$$c(\omega_i) - (\mathbf{w} \cdot \Phi(\omega_i) + b) - \epsilon - \eta'_i = 0.$$

Since $\mu'_i = C - \lambda'_i = 0$ by (8), we have $\eta'_i \geq 0$ by (4) and then

$$(\mathbf{w} \cdot \Phi(\omega_i) + b) - c(\omega_i) \leq -\epsilon.$$

The prediction error $(\mathbf{w} \cdot \Phi(\omega_i) + b) - c(\omega_i)$ will be denoted as E_i:

$$\begin{aligned} E_i &= (\mathbf{w} \cdot \Phi(\omega_i) + b) - c(\omega_i) = \sum_{k=1}^{m} (\lambda'_k - \lambda_k) \Phi(\omega_k) \cdot \Phi(\omega_i) + b - c(\omega_i) \\ &= \sum_{k=1}^{m} (\lambda'_k - \lambda_k) K(\omega_k, \omega_i) + b - c(\omega_i) \qquad (17) \end{aligned}$$

by (5), where K is the PDS kernel associated with the feature mapping Φ.

To summarize, the Kuhn-Tucker conditions imply:

$$\begin{aligned}
\lambda_i = \lambda_i' = 0 &\Rightarrow -\epsilon \le E_i \le \epsilon, \\
0 < \lambda_i < C \text{ and } \lambda_i' = 0 &\Rightarrow E_i = \epsilon, \\
\lambda_i = 0 \text{ and } 0 < \lambda_i' < C &\Rightarrow E_i = -\epsilon, \\
\lambda_i = C \text{ and } \lambda_i' = 0 &\Rightarrow E_i \ge \epsilon, \\
\lambda_i = 0 \text{ and } \lambda_i' = C &\Rightarrow E_i \le -\epsilon.
\end{aligned}$$

Thus in the following three cases, the Kuhn-Tucker conditions are violated:

1. Either $(\lambda_i < C, \lambda_i' = 0)$ or $(\lambda_i = 0, \lambda_i' < C)$ and $|E_i| > \epsilon$,
2. $(\lambda_i > 0, \lambda_i' = 0)$ and $E_i < \epsilon$,
3. $(\lambda_i = 0, \lambda_i' > 0)$ and $E_i > -\epsilon$.

Checking Kuhn-Tucker conditions (1)-(16) without using the offset b : As the Lagrangian dual problem of SVR does not solve for the offset b directly, the improvement proposed by Keerthi et al.[1] avoids the use of the offset b in checking the Kuhn-Tucker conditions for SVM classification. Here we extend the idea to SVR.

The prediction error in (17) can also be written as

$$E_i = \sum_{k=1}^{m} (\lambda_k' - \lambda_k) K(\omega_k, \omega_i) + b - c(\omega_i) = F_i + b$$

where

$$F_i = \sum_{k=1}^{m} (\lambda_k' - \lambda_k) K(\omega_k, \omega_i) - c(\omega_i). \tag{18}$$

so that

$$E_i - E_j = F_i - F_j. \tag{19}$$

Now consequences of the Kuhn-Tucker conditions can be rewritten as

$$\begin{aligned}
\lambda_i' - \lambda_i = 0 &\Rightarrow -\epsilon - b \le F_i \le \epsilon - b, && (20)\\
-C < \lambda_i' - \lambda_i < 0 &\Rightarrow F_i = \epsilon - b, && (21)\\
0 < \lambda_i' - \lambda_i < C &\Rightarrow F_i = -\epsilon - b, && (22)\\
\lambda_i' - \lambda_i = -C &\Rightarrow F_i \ge \epsilon - b, && (23)\\
\lambda_i' - \lambda_i = C &\Rightarrow F_i \le -\epsilon - b. && (24)
\end{aligned}$$

Let

$$\begin{aligned}
I_0 &= \{i \in [1, m] \mid \lambda_i' - \lambda_i = 0\}, && (25)\\
I_{\epsilon,+} &= \{i \in [1, m] \mid -C < \lambda_i' - \lambda_i < 0\}, && (26)\\
I_{\epsilon,-} &= \{i \in [1, m] \mid 0 < \lambda_i' - \lambda_i < C\}, && (27)\\
I_{+} &= \{i \in [1, m] \mid \lambda_i' - \lambda_i = -C\}, && (28)\\
I_{-} &= \{i \in [1, m] \mid \lambda_i' - \lambda_i = C\}. && (29)
\end{aligned}$$

[1] S.S. Keerthi, S.K. Shevade, C. Bhattacharyya, and K.R.K. Murthy, “Improvements to platt’s SMO algorithm for SVM classifier design,” *Neural Computation*, 13(3):637–649, March 2001.

Define

$$\begin{aligned} B_{up} &= \max\{F_j : j \in I_0\}, \\ B_{low} &= \min\{F_j : j \in I_0\}. \end{aligned}$$

Then the Kuhn-Tucker conditions imply

$$\begin{aligned} B_{up} - B_{low} &\leq 2\epsilon, & (30) \\ F_i &\geq B_{up},\ \forall\, i \in I_{\epsilon,+} \cup I_+, & (31) \\ F_k &\leq B_{low},\ \forall\, k \in I_{\epsilon,-} \cup I_-. & (32) \end{aligned}$$

These comparisons do not use the offset b.

The SMO algorithm : The sequential minimal optimization (SMO) algorithm proposed by Platt[2] is an efficient iterative algorithm commonly used to solve the Lagrangian dual problem for kernel-based SVM for classification. Here we will extend the SMO algorithm to solve the Lagrangian dual problem for kernel-based SVR as stated on page 94 of Lecture 10:

$$\begin{array}{ll} \text{Maximize} & \theta(\lambda, \lambda') = -\epsilon \sum_{i=1}^m (\lambda'_i + \lambda_i) + \sum_{i=1}^m (\lambda'_i - \lambda_i) c(\omega_i) \\ & -\frac{1}{2} \sum_{i,j=1}^m (\lambda'_i - \lambda_i) K(\omega_i, \omega_j)(\lambda'_j - \lambda_j), \\ \text{Subject to} & \lambda_i, \lambda'_i \geq 0, i \in [1, m] \\ & C - \lambda_i \geq 0, i \in [1, m] \\ & C - \lambda'_i \geq 0, i \in [1, m] \\ & \sum_{i=1}^m (\lambda'_i - \lambda_i) = 0 \\ & \lambda, \lambda' \in \mathbb{R}^m. \end{array} \qquad (33)$$

Since $\lambda_i \lambda'_i = 0$ for all $i \in [1, m]$, we will follow the idea of Flake and Lawrence[3] to transform the Lagrangian dual problem in (33) to an equivalent problem by letting $\beta_i \triangleq \lambda'_i - \lambda_i, i \in [1, m]$:

$$\begin{array}{ll} \text{Maximize} & \Theta(\beta) = -\epsilon \sum_{i=1}^m |\beta_i| + \sum_{i=1}^m \beta_i c(\omega_i) \\ & -\frac{1}{2} \sum_{i,j=1}^m \beta_i \beta_j K(\omega_i, \omega_j), \\ \text{Subject to} & C - \beta_i \geq 0, i \in [1, m] \\ & \beta_i + C \geq 0, i \in [1, m] \\ & \sum_{i=1}^m \beta_i = 0 \\ & \beta \in \mathbb{R}^m. \end{array} \qquad (34)$$

Note that the objective function $\Theta(\beta)$ is neither quadratic nor differentiable any more. However, it is still convex and therefore the equivalent problem has the global maximum. In each iteration, the SMO algorithm solves the equivalent problem which involves only two variables β_i and β_j, by fixing the values of other variables $\beta_k, k \neq i, j$ to their most recently updated values, to update the values of the two variables β_i and β_j. We next describe the update rules of the SMO algorithm.

[2] J.C. Pratt, "Sequential minimal optimization: A fast algorithm for training support vector machines," Microsoft Research Technical Report MSR-TR-98-14, April 21, 1998.

[3] G.W. Flake and S. Lawrence, "Efficient SVM regression training with SMO," *Machine Learning*, vol. 46, pp. 271V290, 2002.

The update rules of the SMO algorithm : Let $\beta_1^*, \ldots, \beta_m^*$ be the most recently updated values of the variables $\beta_1, \ldots, \beta_m$. By fixing $\beta_k, k \neq i, j$, to their most recently updated values $\beta_k^*, k \neq i, j$, the equivalent problem in (34) becomes

$$\begin{array}{ll} \text{Maximize} & \Psi_1(\beta_i, \beta_j) = -\epsilon(|\beta_i| + |\beta_j|) + \beta_i c(\omega_i) + \beta_j c(\omega_j) \\ & -\frac{1}{2}\beta_i^2 K_{ii} - \frac{1}{2}\beta_j^2 K_{jj} - \beta_i \beta_j K_{ij} \\ & -\beta_i v_i^* - \beta_j v_j^* \\ \text{Subject to} & -C \leq \beta_i, \beta_j \leq C \ \text{ and } \ \beta_i + \beta_j = \gamma_{ij}^*, \end{array} \tag{35}$$

where $K_{ln} = K(\omega_l, \omega_n)\ \forall\ l, n \in [1, m]$, $v_i^* = \sum_{k \neq i,j} \beta_k^* K_{ik}$, $v_j^* = \sum_{k \neq i,j} \beta_k^* K_{jk}$ and $\gamma_{ij}^* = -\sum_{k \neq i,j} \beta_k^* = \beta_i^* + \beta_j^*$. When $\gamma_{ij}^* = 2C$, i.e., $\beta_i^* = \beta_j^* = C$, the solution for the optimization problem in (35) is exactly $\beta_i^{new} = \beta_j^{new} = C$ and the SMO does not change anything. Similar situation occurs when $\gamma_{ij}^* = -2C$. Thus we will assume that $-2C < \gamma_{ij}^* < 2C$.

By substituting $\beta_i = \gamma_{ij}^* - \beta_j$, the equivalent problem in (35) becomes

$$\begin{array}{ll} \text{Maximize} & \Psi_2(\beta_j) = -\epsilon(|\gamma_{ij}^* - \beta_j| + |\beta_j|) + (\gamma_{ij}^* - \beta_j)c(\omega_i) + \beta_j c(\omega_j) \\ & -\frac{1}{2}(\gamma_{ij}^* - \beta_j)^2 K_{ii} - \frac{1}{2}\beta_j^2 K_{jj} - (\gamma_{ij}^* - \beta_j)\beta_j K_{ij} \\ & -(\gamma_{ij}^* - \beta_j)v_i^* - \beta_j v_j^* \\ \text{Subject to} & -C \leq \beta_j \leq C \ \text{ and } \ -C \leq \gamma_{ij}^* - \beta_j \leq C. \end{array} \tag{36}$$

Although the object function $\Psi_2(\beta_j)$ in the equivalent problem in (36) is not differentiable, it is still convenient to let

$$\frac{d|\gamma_{ij}^* - \beta_j|}{\beta_j} \triangleq -\mathrm{sgn}(\gamma_{ij}^* - \beta_j) \ \text{ and } \ \frac{d|\beta_j|}{d\beta_j} \triangleq \mathrm{sgn}(\beta_j),$$

where $\mathrm{sgn}(x)$ is equal to $+1$ if $x > 0$ and -1 if $x < 0$, and then obtain

$$\begin{aligned} \frac{d\Psi_2(\beta_j)}{d\beta_j} &= -\epsilon(\mathrm{sgn}(\beta_j) - \mathrm{sgn}(\gamma_{ij}^* - \beta_j)) - c(\omega_i) + c(\omega_j) + (\gamma_{ij}^* - \beta_j)K_{ii} - \beta_j K_{jj} \\ &\quad -(\gamma_{ij}^* - 2\beta_j)K_{ij} + v_i^* - v_j^* \\ &= \epsilon(\mathrm{sgn}(\gamma_{ij}^* - \beta_j) - \mathrm{sgn}(\beta_j)) + (v_i^* - v_j^*) - (c(\omega_i) - c(\omega_j)) \\ &\quad +\gamma_{ij}^*(K_{ii} - K_{ij}) - \beta_j(K_{ii} + K_{jj} - 2K_{ij}). \end{aligned} \tag{37}$$

Let

$$h^*(\omega) = \sum_{k=1}^{m} \beta_k^* K(\omega_k, \omega) + b^*$$

be the hypothesis based on the most recently updated values of the variables $\beta_1, \ldots, \beta_m$ and the offset b. It can be seen that

$$\begin{aligned} h^*(\omega_i) &= \beta_i^* K_{ii} + \beta_j^* K_{ji} + \sum_{k \neq i,j} \beta_k^* K_{ki} + b^* = \beta_i^* K_{ii} + \beta_j^* K_{ji} + v_i^* + b^* \\ h^*(\omega_j) &= \beta_i^* K_{ij} + \beta_j^* K_{jj} + \sum_{k \neq i,j} \beta_k^* K_{kj} + b^* = \beta_i^* K_{ij} + \beta_j^* K_{jj} + v_j^* + b^* \end{aligned}$$

so that

$$\begin{aligned} v_i^* - v_j^* &= h^*(\omega_i) - h^*(\omega_j) + \beta_i^*(K_{ij} - K_{ii}) + \beta_j^*(K_{jj} - K_{ij}) \\ &= h^*(\omega_i) - h^*(\omega_j) + (\gamma_{ij}^* - \beta_j^*)(K_{ij} - K_{ii}) + \beta_j^*(K_{jj} - K_{ij}) \\ &= h^*(\omega_i) - h^*(\omega_j) + \gamma_{ij}^*(K_{ij} - K_{ii}) + \beta_j^*(K_{ii} + K_{jj} - 2K_{ij}). \end{aligned}$$

Table 1: The values of $\Lambda_{ij} \triangleq \text{sgn}(\gamma_{ij}^* - \beta_j) - \text{sgn}(\beta_j)$ and the computation of $\beta_j^{raw}(\Lambda_{ij})$ in various conditions of γ_{ij}^* and β_j.

		Λ_{ij}	$\beta_j^{raw}(\Lambda_{ij})$
$\gamma_{ij}^* = 0$	$\beta_j < 0$	2	$\beta_j^* + (F_i^* - F_j^* + 2\epsilon)/\eta_{ij}$
	$0 < \beta_j$	-2	$\beta_j^* + (F_i^* - F_j^* - 2\epsilon)/\eta_{ij}$
$\gamma_{ij}^* > 0$	$\beta_j < 0$	2	$\beta_j^* + (F_i^* - F_j^* + 2\epsilon)/\eta_{ij}$
	$0 < \beta_j < \gamma_{ij}^*$	0	$\beta_j^* + (F_i^* - F_j^*)/\eta_{ij}$
	$\gamma_{ij}^* < \beta_j$	-2	$\beta_j^* + (F_i^* - F_j^* - 2\epsilon)/\eta_{ij}$
$\gamma_{ij}^* < 0$	$\beta_j < \gamma_{ij}^*$	2	$\beta_j^* + (F_i^* - F_j^* + 2\epsilon)/\eta_{ij}$
	$\gamma_{ij}^* < \beta_j < 0$	0	$\beta_j^* + (F_i^* - F_j^*)/\eta_{ij}$
	$\beta_j > 0$	-2	$\beta_j^* + (F_i^* - F_j^* - 2\epsilon)/\eta_{ij}$

Now the derivative in (37) becomes

$$\begin{aligned}\frac{d\Psi_2(\beta_j)}{d\beta_j} &= \epsilon(\text{sgn}(\gamma_{ij}^* - \beta_j) - \text{sgn}(\beta_j)) + (E_i^* - E_j^*) - (\beta_j - \beta_j^*)\eta_{ij} \\ &= \epsilon(\text{sgn}(\gamma_{ij}^* - \beta_j) - \text{sgn}(\beta_j)) + (F_i^* - F_j^*) - (\beta_j - \beta_j^*)\eta_{ij}, \qquad (38)\end{aligned}$$

where $E_i^* = h^*(\omega_i) - c(\omega_i)$ and $E_j^* = h^*(\omega_j) - c(\omega_j)$ are prediction errors by the hypothesis $h^*(\omega)$ and

$$\begin{aligned} F_i^* &= \sum_{k=1}^m \beta_k^* K(\omega_k, \omega_i) - c(\omega_i), & (39) \\ F_j^* &= \sum_{k=1}^m \beta_k^* K(\omega_k, \omega_j) - c(\omega_j), & (40) \\ \eta_{ij} &= K_{ii} + K_{jj} - 2K_{ij}. & (41) \end{aligned}$$

Since K is a PDS kernel, we have

$$\eta_{ij} = [1 \ \ -1] \begin{bmatrix} K_{ii} & K_{ij} \\ K_{ji} & K_{jj} \end{bmatrix} [1 \ \ -1]^T \geq 0$$

and we have assumed that $\eta_{ij} > 0$ here. Table 1 lists the value of $\Lambda_{ij} \triangleq \text{sgn}(\gamma_{ij}^* - \beta_j) - \text{sgn}(\beta_j)$ in various conditions of γ_{ij}^* and β_j and shows that the derivative $\frac{d\Psi_2(\beta_j)}{d\beta_j}$ in (38) is a strictly decreasing function of β_j and has two jumps of size 2ϵ at $\beta_j = 0, \gamma_{ij}^*$ when $\gamma_{ij}^* \neq 0$ and one jump of size 2ϵ at $\beta_j = 0$ when $\gamma_{ij}^* = 0$.

By setting $\frac{d\Psi_2(\beta_j)}{d\beta_j} = 0$, the β_j^{raw} which maximizes $\Psi_2(\beta_j)$ without the inequality constraints in (36) is

$$\beta_j^{raw}(\Lambda_{ij}) = \beta_j^* + \frac{F_i^* - F_j^* + \Lambda_{ij}\epsilon}{\eta_{ij}}, \qquad (42)$$

which is a function of $\Lambda_{ij} = \text{sgn}(\gamma_{ij}^* - \beta_j) - \text{sgn}(\beta_j)$. Λ_{ij} takes on one of the values $2, 0, -2$, depending on γ_{ij}^* and β_j as shown in Table 1.

The inequality constraints in the optimization problem in (36) are

$$\begin{aligned} & -C \le \beta_j \le C \quad \text{and} \quad \gamma^*_{ij} - C \le \beta_j \le \gamma^*_{ij} + C \\ \Leftrightarrow \quad & L = \max(-C, \gamma^*_{ij} - C) \le \beta_j \le \min(C, \gamma^*_{ij} + C) = H. \end{aligned} \tag{43}$$

The updated value β_j^{new} of β_j in various conditions of γ^*_{ij} and $\Lambda_{ij} \triangleq \text{sgn}(\gamma^*_{ij} - \beta_j) - \text{sgn}(\beta_j)$ are given in Table 2, where ‡ indicates that the derivative $\frac{\Psi_2(\beta_j)}{d\beta_j}$ in (38) has a jump of size 2ϵ there.

When $\eta_{ij} = K_{ii} + K_{jj} - 2K_{ij} = 0$, the optimization problem in (36) becomes

$$\begin{array}{ll} \text{Maximize} & \Psi_3(\beta_j) = -\epsilon(|\gamma^*_{ij} - \beta_j| + |\beta_i|) + (F^*_i - F^*_j)\beta_j \\ \text{Subject to} & L = \max(-C, \gamma^*_{ij} - C) \le \beta_j \le \min(C, \gamma^*_{ij} + C) = H. \end{array} \tag{44}$$

Table 3 lists the object function $\Psi_3(\beta_j)$ in various conditions of γ^*_{ij} and β_j when $\eta_{ij} = 0$. It shows that $\Psi_3(\beta_j)$ is a piecewise linear continuous function of β_j in $[L, H]$. The updated value β_j^{new} of β_j after solving the optimal problem in (44) is given in Table 4 and dependent on the slopes of linear pieces of the object function $\Psi_3(\beta_j)$.

Here is a pseudocode to determine the updated value β_j^{new} based on Tables 2 and 4:

```
$L \leftarrow \max(-C, \gamma^*_{ij} - C)$
$H \leftarrow \min(C, \gamma^*_{ij} + C)$
$\Delta F^*_{ij} \leftarrow F^*_i - F^*_j$
$\eta_{ij} \leftarrow K_{ii} + K_{jj} - 2K_{ij}$
if $\eta_{ij} > 0$, then
    $\beta^{raw}_{j,2} \leftarrow \beta^*_j + \frac{\Delta F^*_{ij} + 2\epsilon}{\eta_{ij}}$
    $\beta^{raw}_{j,0} \leftarrow \beta^*_j + \frac{\Delta F^*_{ij}}{\eta_{ij}}$
    $\beta^{raw}_{j,-2} \leftarrow \beta^*_j + \frac{\Delta F^*_{ij} - 2\epsilon}{\eta_{ij}}$
    if $\gamma^*_{ij} = 0$, then
        if $\beta^{raw}_{j,2} \le L$, then $\beta^{new}_j \leftarrow L$
        else if $L < \beta^{raw}_{j,2} < 0$, then $\beta^{new}_j \leftarrow \beta^{raw}_{j,2}$
        else if $\beta^{raw}_{j,-2} \ge H$, then $\beta^{new}_j \leftarrow H$
        else if $0 < \beta^{raw}_{j,-2} < H$, then $\beta^{new}_j \leftarrow \beta^{raw}_{j,-2}$
        else $\beta^{new}_j \leftarrow 0$
    else if $0 < \gamma^*_{ij} < C$, then
        if $\beta^{raw}_{j,2} \le L$, then $\beta^{new}_j \leftarrow L$
        else if $L < \beta^{raw}_{j,2} < 0$, then $\beta^{new}_j \leftarrow \beta^{raw}_{j,2}$
        else if $\beta^{raw}_{j,0} \le 0$, then $\beta^{new}_j \leftarrow 0$
        else if $0 < \beta^{raw}_{j,0} < \gamma^*_{ij}$, then $\beta^{new}_j \leftarrow \beta^{raw}_{j,0}$
        else if $\beta^{raw}_{j,-2} \ge H$, then $\beta^{new}_j \leftarrow H$
        else if $\gamma^*_{ij} < \beta^{raw}_{j,-2} < H$, then $\beta^{new}_j \leftarrow \beta^{raw}_{j,-2}$
```

Table 2: The updated value β_j^{new} of β_j in various conditions of γ_{ij}^* and $\Lambda_{ij} \triangleq \text{sgn}(\gamma_{ij}^* - \beta_j) - \text{sgn}(\beta_j)$, where ‡ indicates that the derivative $\frac{\Psi_2(\beta_j)}{d\beta_j}$ in (38) has a jump of size 2ϵ.

γ_{ij}^*	L	H		Λ_{ij}		β_j^{new}
$\gamma_{ij}^* = 0$	$-C$	C	$\beta_j < 0$	2	$\beta_j^{raw}(2) \leq L$	L
					$L < \beta_j^{raw}(2) < 0$	$\beta_j^{raw}(2)$
			$\ddagger\beta_j = 0$		$\beta_j^{raw}(-2) \leq 0 \leq \beta_j^{raw}(2)$	0
			$0 < \beta_j$	-2	$0 < \beta_j^{raw}(-2) < H$	$\beta_j^{raw}(-2)$
					$\beta_j^{raw}(-2) \geq H$	H
$0 < \gamma_{ij}^* < C$	$\gamma_{ij}^* - C$	C	$\beta_j < 0$	2	$\beta_j^{raw}(2) \leq L$	L
					$L < \beta_j^{raw}(2) < 0$	$\beta_j^{raw}(2)$
			$\ddagger\beta_j = 0$		$\beta_j^{raw}(0) \leq 0 \leq \beta_j^{raw}(2)$	0
			$0 < \beta_j < \gamma_{ij}^*$	0	$0 < \beta_j^{raw}(0) < \gamma_{ij}^*$	$\beta_j^{raw}(0)$
			$\ddagger\beta_j = \gamma_{ij}^*$		$\beta_j^{raw}(-2) \leq \gamma_{ij}^* \leq \beta_j^{raw}(0)$	γ_{ij}^*
			$\gamma_{ij}^* < \beta_j$	-2	$\gamma_{ij}^* < \beta_j^{raw}(-2) < H$	$\beta_j^{raw}(-2)$
					$\beta_j^{raw}(-2) \geq H$	H
$\gamma_{ij}^* = C$	0	C	$\ddagger\beta_j = 0$		$\beta_j^{raw}(0) \leq 0 = L$	L
			$0 < \beta_j < \gamma_{ij}^*$	0	$L < \beta_j^{raw}(0) < H$	$\beta_j^{raw}(0)$
			$\ddagger\beta_j = \gamma_{ij}^*$		$\beta_j^{raw}(0) \geq \gamma_{ij}^* = H$	H
$C < \gamma_{ij}^* < 2C$	$\gamma_{ij}^* - C$	C	$0 < \beta_j < \gamma_{ij}^*$	0	$\beta_j^{raw}(0) < L$	L
					$L \leq \beta_j^{raw}(0) \leq H$	$\beta_j^{raw}(0)$
					$\beta_j^{raw}(0) > H$	H
$-C < \gamma_{ij}^* < 0$	$-C$	$\gamma_{ij}^* + C$	$\beta_j < \gamma_{ij}^*$	2	$\beta_j^{raw}(2) \leq L$	L
					$L < \beta_j^{raw}(2) < \gamma_{ij}^*$	$\beta_j^{raw}(2)$
			$\ddagger\beta_j = \gamma_{ij}^*$		$\beta_j^{raw}(0) \leq \gamma_{ij}^* \leq \beta_j^{raw}(2)$	γ_{ij}^*
			$\gamma_{ij}^* < \beta_j < 0$	0	$\gamma_{ij}^* < \beta_j^{raw}(0) < 0$	$\beta_j^{raw}(0)$
			$\ddagger\beta_j = 0$		$\beta_j^{raw}(-2) \leq 0 \leq \beta_j^{raw}(0)$	0
			$\beta_j > 0$	-2	$0 < \beta_j^{raw}(-2) < H$	$\beta_j^{raw}(-2)$
					$\beta_j^{raw}(-2) \geq H$	H
$\gamma_{ij}^* = -C$	$-C$	0	$\ddagger\beta_j = \gamma_{ij}^*$		$\beta_j^{raw}(0) \leq \gamma_{ij}^* = L$	L
			$\gamma_{ij}^* < \beta_j < 0$	0	$L < \beta_j^{raw}(0) < H$	$\beta_j^{raw}(0)$
			$\ddagger\beta_j = 0$		$\beta_j^{raw}(0) \geq 0 = H$	H
$-2C < \gamma_{ij}^* < -C$	$-C$	$\gamma_{ij}^* + C$	$\gamma_{ij}^* < \beta_j < 0$	0	$\beta_j^{raw}(0) < L$	L
					$L \leq \beta_j^{raw}(0) \leq H$	$\beta_j^{raw}(0)$
					$\beta_j^{raw}(0) > H$	H

Table 3: The object function $\Psi_3(\beta_j)$ in various conditions of γ_{ij}^* and β_j when $\eta_{ij} = 0$.

		$\lvert\gamma_{ij} - \beta_j\rvert$	$\lvert\beta_j\rvert$	$\Psi_3(\beta_j)$
$\gamma_{ij}^* = 0$	$\beta_j \leq 0$	$-\beta_j$	$-\beta_j$	$(2\epsilon + F_i^* - F_j^*)\beta_j$
	$0 \leq \beta_j$	β_j	β_j	$(-2\epsilon + F_i^* - F_j^*)\beta_j$
$\gamma_{ij}^* > 0$	$\beta_j \leq 0$	$\gamma_{ij}^* - \beta_j$	$-\beta_j$	$-\epsilon\gamma_{ij}^* + (2\epsilon + F_i^* - F_j^*)\beta_j$
	$0 \leq \beta_j \leq \gamma_{ij}^*$	$\gamma_{ij}^* - \beta_j$	β_j	$-\epsilon\gamma_{ij}^* + (F_i^* - F_j^*)\beta_j$
	$\gamma_{ij}^* \leq \beta_j$	$\beta_j - \gamma_{ij}^*$	β_j	$\epsilon\gamma_{ij}^* + (-2\epsilon + F_i^* - F_j^*)\beta_j$
$\gamma_{ij}^* < 0$	$\beta_j \leq \gamma_{ij}^*$	$\gamma_{ij}^* - \beta_j$	$-\beta_j$	$-\epsilon\gamma_{ij}^* + (2\epsilon + F_i^* - F_j^*)\beta_j$
	$\gamma_{ij}^* \leq \beta_j \leq 0$	$\beta_j - \gamma_{ij}^*$	$-\beta_j$	$\epsilon\gamma_{ij}^* + (F_i^* - F_j^*)\beta_j$
	$0 \leq \beta_j$	$\beta_j - \gamma_{ij}^*$	β_j	$\epsilon\gamma_{ij}^* + (-2\epsilon + F_i^* - F_j^*)\beta_j$

```
            else $\beta_j^{new} \leftarrow \gamma_{ij}^*$
        else if $\gamma_{ij}^* = C$, then
            if $\beta_{j,0}^{raw} \leq L$, then $\beta_j^{new} \leftarrow L$
            else if $L < \beta_{j,0}^{raw} < H$, then $\beta_j^{new} \leftarrow \beta_{j,0}^{raw}$
            else $\beta_{j,0}^{raw} \leftarrow H$
        else if $\gamma_{ij}^* > C$, then
            if $\beta_{j,0}^{raw} < L$, then $\beta_j^{new} \leftarrow L$
            else if $L \leq \beta_{j,0}^{raw} \leq H$, then $\beta_j^{new} \leftarrow \beta_{j,0}^{raw}$
            else $\beta_{j,0}^{raw} \leftarrow H$
        else if $-C < \gamma_{ij}^* < 0$, then
            if $\beta_{j,2}^{raw} \leq L$, then $\beta_j^{new} \leftarrow L$
            else if $L < \beta_{j,2}^{raw} < \gamma_{ij}^*$, then $\beta_j^{new} \leftarrow \beta_{j,2}^{raw}$
            else if $\beta_{j,0}^{raw} \leq \gamma_{ij}^*$, then $\beta_j^{new} \leftarrow \gamma_{ij}^*$
            else if $\gamma_{ij}^* < \beta_{j,0}^{raw} < 0$, then $\beta_j^{new} \leftarrow \beta_{j,0}^{raw}$
            else if $\beta_{j,-2}^{raw} \geq H$, then $\beta_j^{new} \leftarrow H$
            else if $0 < \beta_{j,-2}^{raw} < H$, then $\beta_j^{new} \leftarrow \beta_{j,-2}^{raw}$
            else $\beta_j^{new} \leftarrow 0$
        else if $\gamma_{ij}^* = -C$, then
            if $\beta_{j,0}^{raw} \leq L$, then $\beta_j^{new} \leftarrow L$
            else if $L < \beta_{j,0}^{raw} < H$, then $\beta_j^{new} \leftarrow \beta_{j,0}^{raw}$
            else $\beta_{j,0}^{raw} \leftarrow H$
        else if $\gamma_{ij}^* < -C$, then
            if $\beta_{j,0}^{raw} < L$, then $\beta_j^{new} \leftarrow L$
            else if $L \leq \beta_{j,0}^{raw} \leq H$, then $\beta_j^{new} \leftarrow \beta_{j,0}^{raw}$
            else $\beta_{j,0}^{raw} \leftarrow H$
    else                $\triangleright$ $\eta = 0$
```

Table 4: The updated value β_i^{new} of β_j when $\eta_{ij} = 0$.

γ_{ij}^*	L	H		β_j^{new}
$\gamma_{ij}^* = 0$	$-C$	C	$F_i^* - F_j^* < -2\epsilon$	L
			$-2\epsilon \leq F_i^* - F_j^* \leq 2\epsilon$	0
			$2\epsilon < F_i^* - F_j^*$	H
$0 < \gamma_{ij}^* < C$	$\gamma_{ij}^* - C$	C	$F_i^* - F_j^* < -2\epsilon$	L
			$-2\epsilon \leq F_i^* - F_j^* \leq 0$	0
			$0 < F_i^* - F_j^* < 2\epsilon$	γ_{ij}^*
			$2\epsilon \leq F_i^* - F_j^*$	H
$\gamma_{ij}^* = C$	0	C	$F_i^* - F_j^* \leq 0$	L
			$0 < F_i^* - F_j^*$	H
$C < \gamma_{ij}^* < 2C$	$\gamma_{ij}^* - C$	C	$F_i^* - F_j^* < 0$	L
			$0 \leq F_i^* - F_j^*$	H
$-C < \gamma_{ij}^* < 0$	$-C$	$\gamma_{ij}^* + C$	$F_i^* - F_j^* \leq -2\epsilon$	L
			$-2\epsilon < F_i^* - F_j^* < 0$	γ_{ij}^*
			$0 \leq F_i^* - F_j^* \leq 2\epsilon$	0
			$2\epsilon < F_i^* - F_j^*$	H
$\gamma_{ij}^* = -C$	$-C$	0	$F_i^* - F_j^* < 0$	L
			$0 \leq F_i^* - F_j^*$	H
$-2C < \gamma_{ij}^* < -C$	$-C$	$\gamma_{ij}^* + C$	$F_i^* - F_j^* \leq 0$	L
			$0 < F_i^* - F_j^*$	H

```
    if $\gamma_{ij}^* = 0$, then
        if $\Delta F_{ij}^* < -2\epsilon$, then $\beta_j^{new} \leftarrow L$
        else if $\Delta F_{ij}^* > 2\epsilon$, then $\beta_j^{new} \leftarrow H$
        else $\beta_j^{new} \leftarrow 0$
    else if $0 < \gamma_{ij}^* < C$, then
        if $\Delta F_{ij}^* < -2\epsilon$, then $\beta_j^{new} \leftarrow L$
        else if $\Delta F_{ij}^* \geq 2\epsilon$, then $\beta_j^{new} \leftarrow H$
        else if $0 < \Delta F_{ij}^* < 2\epsilon$, then $\beta_j^{new} \leftarrow \gamma_{ij}^*$
        else $\beta_j^{new} \leftarrow 0$
    if $\gamma_{ij}^* = C$, then
        if $\Delta F_{ij}^* \leq 0$, then $\beta_j^{new} \leftarrow L$
        else $\beta_j^{new} \leftarrow H$
    if $\gamma_{ij}^* > C$, then
        if $\Delta F_{ij}^* < 0$, then $\beta_j^{new} \leftarrow L$
        else $\beta_j^{new} \leftarrow H$
    else if $-C < \gamma_{ij}^* < 0$, then
        if $\Delta F_{ij}^* < -2\epsilon$, then $\beta_j^{new} \leftarrow L$
        else if $\Delta F_{ij}^* > 2\epsilon$, then $\beta_j^{new} \leftarrow H$
        else if $-2\epsilon < \Delta F_{ij}^* < 0$, then $\beta_j^{new} \leftarrow \gamma_{ij}^*$
        else $\beta_j^{new} \leftarrow 0$
    if $\gamma_{ij}^* = -C$, then
        if $\Delta F_{ij}^* < 0$, then $\beta_j^{new} \leftarrow L$
        else $\beta_j^{new} \leftarrow H$
    if $\gamma_{ij}^* < -C$, then
        if $\Delta F_{ij}^* \leq 0$, then $\beta_j^{new} \leftarrow L$
        else $\beta_j^{new} \leftarrow H$
```

Now the newly updated value β_i^{new} of β_i together with the newly updated value β_j^{new} of β_j is the solution of the optimization problem with inequality and equality constraints in (35) and must satisfy the equality

$$\beta_i^{new} + \beta_j^{new} = \gamma_{ij}^* = \beta_i^* + \beta_j^*.$$

Thus the newly updated value of β_i is

$$\beta_i^{new} = \beta_i^* - (\beta_j^{new} - \beta_j^*). \tag{45}$$

Updating $F_k, 1 \leq k \leq m$, and weight vector w after a successful optimization iteration :

$$F_k^{new} = F_k^* + \Delta\beta_i K(\omega_i, \omega_k) + \Delta\beta_j K(\omega_j, \omega_k) = F_k^* + \Delta\beta_j(K(\omega_j, \omega_k) - K(\omega_i, \omega_k)) \tag{46}$$

where

$$\begin{aligned} \Delta\beta_j &= \beta_j^{new} - \beta_j^*, \\ \Delta\beta_i &= \beta_i^{new} - \beta_i^* = -\Delta\beta_j. \end{aligned}$$

And if the feature space is $\mathbb{R}^N$, we have

$$\mathbf{w}^{new} = \mathbf{w}^* + \Delta\beta_j(\Phi(\omega_j) - \Phi(\omega_i)).$$

Heuristics for picking two items i and j for joint optimization : In order to speed up convergence, heuristics are suggested to choose which two variables β_i and β_j to jointly optimize as follows:

1. If (30) is violated, i.e., $B_{up} - B_{low} > 2(\epsilon + \tau)$, , where τ is a preset tolerance, select a β_i such that $F_i^* = B_{up}$ and a β_j such that $F_j^* = B_{low}$. Continue until (30) is not violated.
2. If (31) is violated, i.e., $\min_{k \in I_+} F_k^* < B_{up} - \tau$, select a β_i such that $F_i^* = \min_{k \in I_+} F_k^*$ and a β_j such that $F_j^* = B_{up}$. Go to 1.
3. If (32) is violated, i.e., $B_{low} + \tau < \max_{k \in I_-} F_k^*$, select a β_i such that $F_i^* = \max_{k \in I_-} F_k^*$ and a β_j such that $F_j^* = B_{low}$. Go to 1.
4. The algorithm terminates when none of (30)-(32) is violated.

Initialization : We will set

- $\beta_i^{ini} = 0$ for all $i \in [1, m]$;
- $F_i^{ini} = -c(\omega_i)$ for all $i \in [1, m]$.

Then we have

$$\begin{aligned} I_0^{ini} &= [1, m], \\ I_+^{ini} &= \emptyset, \\ I_-^{ini} &= \emptyset \end{aligned}$$

and

$$\begin{aligned} B_{up}^{ini} &= \max\{F_j : j \in I_0\} = -\min\{c(\omega_j) : j \in [1, m]\}, \\ B_{low}^{ini} &= \min\{F_j : j \in I_0\} = -\max\{c(\omega_j) : j \in [1, m]\}. \end{aligned}$$

Termination : When the entire training set obeys the three consequences (30)-(32) of the Kuhn-Tucker conditions within the tolerance τ, the SMO algorithm terminates.

Output : The returned hypothesis h_S^{SVR} from the SMO-SVR algorithm with respect to a sample $S = (\omega_1, \ldots, \omega_m)$ of size m is

$$h_S^{SVR}(\omega) = \sum_{k=1}^{m} \beta_k^{SVR} K(\omega_k, \omega) + b^{SVR},$$

where

$$b^{SVR} = \epsilon - F_j^{SVR}$$

for any $j \in I_{\epsilon,+}$ by (26) and (21) or

$$b^{SVR} = -\epsilon - F_j^{SVR}$$

for any $j \in I_{\epsilon,-}$ by (27) and (22). Since we have a preset tolerance τ, a better value for b^{SVR} is

$$\begin{aligned} b^{SVR} &= \frac{1}{|I_{\epsilon,+}| + |I_{\epsilon,-}|} \left(\sum_{j \in I_{\epsilon,+}} (\epsilon - F_j^{SVR}) + \sum_{j \in I_{\epsilon,-}} (-\epsilon - F_j^{SVR}) \right) \\ &= \epsilon \frac{|I_{\epsilon,+}| - |I_{\epsilon,-}|}{|I_{\epsilon,+}| + |I_{\epsilon,-}|} - \frac{\sum_{j \in I_{\epsilon,+} \cup I_{\epsilon,-}} F_j^{SVR}}{|I_{\epsilon,+}| + |I_{\epsilon,-}|}. \end{aligned}$$