The Entire Regularization Path for the Interval Support Vector Machine

May 2, 2012

1 Problem formulation

Suppose n data points are given: (x_i, y_i) for $i = 1, ..., n$. Let $\mathcal{I}_+ = \{i : y_i = +1\}$ and $\mathcal{I}_- = \{i : y_i = -1\}$ be the set of positive class and negative class respectively. Also let $n_+ = |\mathcal{I}_+|$ and $n_- = |\mathcal{I}_-|$ be the number of data points in these two classes. In this section, we will derive an efficient algorithm to compute the entire regularization path for the following interval svm problem:

$$\min_{\beta} \sum_{i=1}^{n} (1 - y_i(\beta_0 + \beta^T x_i) + \rho\sigma_i^T |\beta|)_+ + \frac{\lambda}{2} ||\beta||_2^2. \tag{1}$$

(**??**) (**??**) Introducing slack variables to get rid of the hinge loss function and absolute values, the problem is equivalent to:

$$\begin{aligned} \min \sum_{i=1}^{n} \xi_i + \lambda \frac{\beta^T \beta}{2}, \qquad\qquad & \qquad (2)\\ \text{subject to } \xi_i \geq 1 - y_i(\beta_0 + \beta^T x_i) + \rho\sigma_i^T t, &\\ \xi_i \geq 0, \text{ for } i = 1, 2, ..., n; &\\ -t_j \leq \beta_j \leq t_j, &\\ t_i \geq 0, \text{ for } j = 1, 2, ..., p. & \end{aligned}$$

Let the above problem be the primal problem. The Lagragian for this problem is

$$\begin{aligned} L(\xi, \beta_0, \beta, t, \alpha, \gamma, \nu, \mu, c) = & \sum_{i=1}^{n} \xi_i + \lambda \tfrac{\beta^T \beta}{2} + \sum_{i=1}^{n} \alpha_i (1 - y_i(\beta_0 + \beta^T x_i) + \rho\sigma_i^T t - \xi_i) \\ & - \sum_{j=1}^{p} \mu_j (t_j - \beta_j) - \sum_{j=1}^{p} \nu_j (t_j + \beta_j) - \sum_{i=1}^{n} \gamma_i \xi_i - \sum_{i=1}^{n} c_j t_j. \\ = & \sum_{i=1}^{n} (1 - \alpha_i - \gamma_i) \xi_i + \lambda \tfrac{\beta^T \beta}{2} - (\sum_{i=1}^{n} \alpha_i y_i x_i - (\mu - \nu))^T \beta \\ & - \sum_{i=1}^{n} \alpha_i y_i \beta_0 + \sum_{j=1}^{p} (\rho \alpha_i \sigma_{ji} - \mu_j - \nu_j - c_j) t_j. \end{aligned}$$

Minimizing with respect to the primal variables (ξ, β_0, β, t) we derive the dual problem as follows:

$$\begin{aligned} \max_{\alpha,\mu,\nu} \sum_{i=1}^{n} \alpha_i - \frac{1}{2\lambda} || \sum_{i=1}^{n} \alpha_i y_i x_i - (\mu - \nu) ||_2^2, \qquad & \qquad (3)\\ \text{subject to } \sum_{i=1}^{n} \alpha_i y_i = 0, \ \rho \sum_{i=1}^{n} \alpha_i \sigma_i \geq \mu + \nu, &\\ \alpha \in [0, 1], \ \mu \geq 0, \ \nu \geq 0. & \end{aligned}$$

Remarks: if the precision matrix is zero, i.e. $\Sigma = 0$, then second dual constraint implies $\mu + \nu \leq 0$. Furthermore by the nonnegativity constraints we have $\mu = 0$ and $\nu = 0$. So the dual problem becomes

$$\begin{array}{rl} \max_\alpha & \sum_{i=1}^n \alpha_i - \frac{1}{2\lambda}||\sum_{i=1}^n \alpha_i y_i x_i||_2^2, \\ \text{subject to} & \sum_{i=1}^n \alpha_i y_i = 0,\ \alpha \in [0,1], \end{array}$$

which is exactly the dual to the standard SVM (see, for example, (12.13) of EML).

One can easily determine the KKT conditions of the Interval SVM problem:
(1) Primal stationarity:

$$\frac{\partial}{\partial \beta} : \beta = \frac{1}{\lambda}(\sum_{i=1}^n \alpha_i y_i x_i + \nu - \mu) \tag{4}$$

$$\frac{\partial}{\partial \beta_0} : \sum_{i=1}^n \alpha_i y_i = 0 \tag{5}$$

$$\frac{\partial}{\partial \xi} : 1 - \alpha - \gamma = 0. \tag{6}$$

(2) Complementary slackness:

$$\alpha_i(1 - y_i(\beta_0 + \beta^T x_i) + \rho\sigma_i^T t - \xi_i) = 0 \tag{7}$$

$$\gamma_i \xi_i = 0 \tag{8}$$

$$\mu_j(t_j - \beta_j) = 0 \tag{9}$$

$$\nu_j(t_j + \beta_j) = 0 \tag{10}$$

$$c_j t_j = 0 \tag{11}$$

$$\rho \sum_{i=1}^n \alpha_i \sigma_{ji} - (\mu_j + \nu_j + c_j) = 0 \tag{12}$$

$$\tag{13}$$

(3) Primal feasibility and; (4) Dual feasibility.

From the optimality conditions, it is easy to observe the following statements hold:
(i) If $\xi_i > 0$, then $\gamma_i = 0$ and so $\alpha_i = 1$. Therefore by equation (7) we have $y_i(\beta_0 + \beta^T x_i) < 1 + \rho\sigma_i^T t$, which indicates that the data point i is on the left of the elbow of the hinge loss.
(ii) If $y_i(\beta_0 + \beta^T x_i) > 1 + \rho\sigma_i^T t$, that is, when the i-th data point is on the right of the elbow, then $\xi = 0$ and by equation (6) and (8) $\alpha_i = 0$.
(iii) If $y_i(\beta_0 + \beta^T x_i) = 1 + \rho\sigma_i^T t$, that data point is on the elbow. In this case we can only know $\alpha_i \in [0, 1]$.
(iv) It is obvious that at optimal, we have $t = |\beta|$.
Now for convenience let $f(x) = \beta_0 + \beta^T x$ and define the following index sets for the data points:

$$\mathcal{E} = \{i : y_i f(x_i) = 1 + \rho\sigma_i^T|\beta|, 0 \leq \alpha_i \leq 1\},\ \mathcal{E} \text{ for Elbow},$$

$$\mathcal{L} = \{i : y_i f(x_i) < 1 + \rho\sigma_i^T|\beta|, \alpha_i = 1\},\ \mathcal{L} \text{ for Left of the elbow},$$

$$\mathcal{R} = \{i : y_i f(x_i) > 1 + \rho\sigma_i^T|\beta|, \alpha_i = 0\},\ \mathcal{R} \text{ for Right of the elbow}.$$

Note that the above sets are similar to those defined in [2]. In addition, since the absolute value of β is involved, we also need to keep track of the following index sets for the parameter vector β:

$$\mathcal{V}_+ = \{j : \beta_j > 0, \nu_j = 0\},$$
$$\mathcal{V}_- = \{j : \beta_j < 0, \mu_j = 0\},$$
$$\mathcal{Z} = \{j : \beta_j = 0, \sum_{i=1}^{n} \alpha_i \sigma_{ji} \geq \mu_j + \nu_j\}.$$

2 Initialization

We will determine the initial state of the parameters and the sets defined above. Without loss of generality, we assume that $n_+ \geq n_- > 0$. We start with $\lambda = +\infty$ or equivalently $C = 0$. In this case, it is easy to see that $\beta = 0$.

Lemma 2.1. *Given $n_+ \geq n_- > 0$ and $\beta = 0$, the optimal value for the initial value of β_0 is $\beta_0 = 1$ and the loss is $\sum_{i=1}^{n} \xi_i = 2n_-$. If $n_+ > n_-$, then $\beta_0 = 1$.*

Proof. With $\beta = 0$, the primal problem becomes:

$$\begin{array}{rl} \min & \sum_{i=1}^{n} \xi_i, \\ \text{subject to} & \xi_i \geq 0,\ \xi_i \geq 1 - y_i\beta_0. \end{array}$$

Hence for $i \in \mathcal{I}_+$, $\xi_i \geq 1 - \beta_0$ and for $i \in \mathcal{I}_-$, $\xi_i \geq 1 + \beta_0$. At optimal, the equal sign must hold since we are minimizing the sum $\sum_{i=1}^{n} \xi_i$. Also, ξ_i's are nonnegative, therefore we have $1 - \beta_0 \geq 0$ and $1 + \beta_0 \geq 0$ and so $\beta_0 \in [-1, 1]$. Furthermore, the objective function can be written in terms of n_+ and n_-:

$$\sum_{i=1}^{n} \xi_i = n_+(1 - \beta_0) + n_-(1 + \beta_0) = (n_- - n_+)\beta_0 + (n_+ + n_-),$$

which is a linear function in β_0. If $n_- = n_+$, then the above function is simply a constant $2n_-$ and so β_0 can be arbitrary chosen in $[-1, 1]$; on the other hand if $n_+ > n_-$, the linear function is decreasing in β_0 and so to minimize the sum, we can pick β_0 on the right end point of $[-1, 1]$. So $\beta_0 = 1$ if $n_+ > n_-$.

From now on, we can safely assume that $n_+ > n_-$, in which case the initial state is $\beta = 0$, $\beta_0 = 1$, $\xi_i = 0$ for $i \in \mathcal{I}_+$ and $\xi_i = 2$ for $i \in \mathcal{I}_-$. Thus by optimality conditions, for any $i \in \mathcal{I}_-$, $\gamma_i = 0$ and so $\alpha_i = 1$. In addition, since $0 = \sum_{i=1}^{n} \alpha_i y_i = \sum_{i \in \mathcal{I}_+} \alpha_i - \sum_{i \in \mathcal{I}_-} \alpha_i$, we have $\sum_{i \in \mathcal{I}_+} \alpha_i = \sum_{i \in \mathcal{I}_-} \alpha_i = n_-$. The following lamma determines the initial value for the dual valuables:

Lemma 2.2. *Let $\tilde{\beta} = \sum_{i=1}^{n} \alpha_i y_i x_i + \nu - \mu$,*

$$\begin{array}{rl} (\alpha^*, \mu^*, \nu^*) = & \arg\min ||\tilde{\beta}||_2^2, \\ & \textit{subject to: } 0 \leq \alpha_i \leq 1, \forall i \in \mathcal{I}_+; \alpha_i = 1, \forall i \in \mathcal{I}_-; \\ & \sum_{i \in \mathcal{I}_+} \alpha_i = n_-, \rho\Sigma\alpha \geq \mu + \nu, \mu \geq 0, \nu \geq 0. \end{array}$$

and $c^* = \rho\Sigma\alpha^* - \mu^* - \nu^*$. Then for some λ_0 we have for all $\lambda > \lambda_0$, $(\alpha(\lambda), \mu(\lambda), \nu(\lambda), c(\lambda)) = (\alpha^*, \mu^*, \nu^*, c^*)$.

Proof. See [2].

Now let $\beta^* = \sum_{i=1}^n \alpha_i^* y_i x_i + \nu^* - \mu^*$ be the fixed coefficient direction corresponding to the initial (α^*, μ^*, ν^*). By lemma (XX) and the optimality conditions, we have for all $\lambda > \lambda_0$:

$$\beta(\lambda) = \frac{\sum_{i=1}^n \alpha_i^* y_i x_i + \nu^* - \mu^*}{\lambda}.$$

We now need to determine the point λ_0 that the dual variables (α, μ, ν, c) start to change. At the very beginning when $\lambda = +\infty$, there are two possible scenarios:

- There exist at least two data points in $\mathcal{I}_+$ with $0 < \alpha_i^* < 1$. Notice that there cannot be only one such element due to the integer constraint $\sum_{i \in \mathcal{I}_+} \alpha_i^* = n_-$.
- For all $i \in \mathcal{I}_+$, α_i^* is either 0 or 1.

For the first scenario, arbitrary pick an $i_+ \in \mathcal{I}_+$ such that $0 < \alpha_{i_+}^* < 1$. Since any of these points will stay in the elbow until a point in $\mathcal{I}_-$ enters the elbow, we can consider the element in $\mathcal{I}_-$ that first reaches its margin, i.e. $i_- = \arg\min_{i \in \mathcal{I}_-} x_i^T \beta^* + \rho \sigma_i^T |\beta^*|$. Therefore at $\lambda = \lambda_0$, as both i_+ and i_- are at their respective margin, by the definition of elbow, we have the following system of equations:

$$\beta_0 + \frac{1}{\lambda} x_{i_+}^T \beta^* = 1 + \frac{1}{\lambda} \rho \sigma_{i_+}^T |\beta^*|,$$

$$\beta_0 + \frac{1}{\lambda} x_{i_-}^T \beta^* = -1 - \frac{1}{\lambda} \rho \sigma_{i_-}^T |\beta^*|.$$

Solving for λ and β_0 yields:

$$\lambda_0 = \frac{1}{2}(\beta^{*T}(x_{i_+} - x_{i_-}) - \rho |\beta^*|^T (\sigma_{i_+} + \sigma_{i_-})),$$

$$\beta_0 = \frac{-\beta^{*T}(x_{i_+} + x_{i_-}) + \rho |\beta^*|^T (\sigma_{i_+} - \sigma_{i_-})}{\beta^{*T}(x_{i_+} - x_{i_-}) - \rho |\beta^*|^T (\sigma_{i_+} + \sigma_{i_-})}.$$

For the second scenario, for the initial parameter to change, a point in $\mathcal{I}_-$ and a point in $\mathcal{I}_+$ with $\alpha_i^* = 1$ must reach the margin simultaneously. Therefore we can let $i_+ = \arg\max_{i \in \mathcal{I}_+, \alpha_i = 1} x_i^T \beta^* - \rho \sigma_i^T |\beta^*|$ and obtain λ_0 and β_0 by solving the above equations.

Remarks: in [2], Hastie et.al. split the initialization step into two cases depending on whether $n_+ = n_-$ or $n_+ > n_-$. In their case $n_+ = n_-$, the initial state can be obtained efficiently without solving the quadratic programming problem. This does not seem to be the case in our problem set-up, as it also involves finding the initial values of the extra dual variables μ and ν. Therefore solving a quardratic problem for starting up the path algorithm cannot be avoided.

3 The Path

Our algorithm keeps track of the following events:

1. One or more points from $\mathcal{L}$ have just entered $\mathcal{E}$;
2. One or more points from $\mathcal{R}$ have just reentered $\mathcal{E}$;

3. One or more points in $\mathcal{E}$ have just left the set, to join either $\mathcal{R}$ or $\mathcal{L}$.

The above are events that are considered by [2] for deriving a path algorithm for standard SVM. Since our problem involves also the l_1 norm of the parameter β, we also need to keep track of the following event:

4. One or more indices of the parameter β enter $\mathcal{Z}$, that is, β_j becomes zero, initially being either positive or negative;
5. One or more indices of the parameter β enter $\mathcal{V}_+$ from $\mathcal{Z}$;
6. One or more indices of the parameter β enter $\mathcal{V}_-$ from $\mathcal{Z}$.

3.1 Piecewise Linearity between Events

By continuity, all of the index sets defined will stay the same until the next event occurs. Index by the superscript l the sets immediately after the l-th event occurs. Likewise, index all the parameters in the same manner, and let f^l be the classify function at this point. Also, for notation consistency define $\alpha_0 = \lambda\beta_0$ and so $\alpha_0^l = \lambda^l\beta_0^l$. Consider for $\lambda^l > \lambda > \lambda^{l+1}$, we have:

$$\begin{aligned}
f(x) &= [f(x) - \frac{\lambda^l}{\lambda} f^l(x)] + \frac{\lambda^l}{\lambda} f^l(x) \\
&= \frac{1}{\lambda}[x^T(\sum_{i=1}^{n} \alpha_i y_i x_i + \nu - \mu) + \alpha_0 - x^T(\sum_{i=1}^{n} \alpha_i^l y_i x_i + \nu^l - \mu^l) - \alpha_0^l + \lambda^l f^l(x)] \\
&= \frac{1}{\lambda}[x^T \sum_{i\in\mathcal{E}^l} (\alpha_i - \alpha_i^l) y_i x_i + x^T(\nu - \nu^l) - x^T(\mu - \mu^l) + (\alpha_0 - \alpha_0^l) + \lambda^l f^l(x)],
\end{aligned} \tag{14}$$

where the last equality follows from the fact that $\{1, ..., n\} = \mathcal{L} \cup \mathcal{E} \cup \mathcal{R}$, and $\alpha_i = 1$ on $\mathcal{L}$; $\alpha_i = 0$ on $\mathcal{R}$. The value of α_i for i not in the elbow remains unchanged between events. Since each data point in $\mathcal{E}^l$ is to stay in $\mathcal{E}$ for $\lambda \in (\lambda^{l+1}, \lambda^l)$, we have:

$$y_j f(x_j) = 1 + \rho\sigma_j^T|\beta|.$$

Plug in equation (14) for $f(x_j)$, the above formula becomes:

$$\frac{1}{\lambda}[\sum_{i\in\mathcal{E}^l} (\alpha_i - \alpha_i^l) y_j y_i x_j^T x_i + y_j x_j^T(\nu - \mu - (\nu^l - \mu^l)) + y_j(\alpha_0 - \alpha_0^l) + \lambda^l(1 + \rho\sigma_j^T|\beta^l|)] = 1 + \rho\sigma_j^T|\beta|. \tag{15}$$

We can expression the parameter β_i in terms of α. For any $i \in \mathcal{V}_+$, since $\nu_i = 0$, $\mu_i = \rho\sum_{k=1}^{n} \alpha_k\sigma_{ik}$. Plug all these parameters into the formula for β_i, we have

$$\beta_i = \frac{1}{\lambda}\sum_{k=1}^{n}(\alpha_i y_k x_{ik} - \mu_i) = \frac{1}{\lambda}\sum_{k=1}^{n}\alpha_k(y_k x_{ik} - \rho\sigma_{ik}) = \frac{1}{\lambda}\sum_{k=1}^{n}\alpha_k w_{ik}.$$

Similarly, for $i \in \mathcal{V}_-$, since $\mu_i = 0$, $\nu_i = \rho\sum_{k=1}^{n}\alpha_k\sigma_{ik}$ and so

$$\beta_i = \frac{1}{\lambda}\sum_{k=1}^{n}(\alpha_i y_k x_{ik} + \nu_i) = \frac{1}{\lambda}\sum_{k=1}^{n}\alpha_k(y_k x_{ik} + \rho\sigma_{ik}) = \frac{1}{\lambda}\sum_{k=1}^{n}\alpha_k z_{ik}.$$

Let $h_k = \sum_{i\in\mathcal{V}_+} \sigma_{ij} w_{ik} - \sum_{i\in\mathcal{V}_-} \sigma_{ij} z_{ik}$. By splitting β into positive and negative parts, we can get rid of the absolute value in $\sigma_j^T|\beta|$:

$$\begin{aligned}\sigma_j^T|\beta| &= \textstyle\sum_{i\in\mathcal{V}_+} \sigma_{ij}\beta_i - \sum_{i\in\mathcal{V}_-} \sigma_{ij}\beta_i \\ &= \textstyle\frac{1}{\lambda}\sum_{k=1}^n \alpha_k(\sum_{i\in\mathcal{V}_+} \sigma_{ij} w_{ik} - \sum_{i\in\mathcal{V}_-} \sigma_{ij} z_{ik}) \\ &= \textstyle\frac{1}{\lambda}\sum_{k=1}^n \alpha_k h_k\end{aligned}$$

Next, let us look at the term $x_j^T(\nu - \mu)$:

$$\begin{aligned}x_j^T(\nu-\mu) &= \textstyle\sum_{i\in\mathcal{V}_-} x_{ij}\nu_i - \sum_{i\in\mathcal{V}_+} x_{ij}\mu_i + \sum_{i\in Z} x_{ij}(\nu_i - \mu_i) \\ &= \textstyle\sum_{i\in\mathcal{V}_-} x_{ij}(\rho\sum_{k=1}^n \alpha_k\sigma_{ik}) - \sum_{i\in\mathcal{V}_+} x_{ij}(\rho\sum_{k=1}^n \alpha_k\sigma_{ik}) + \sum_{i\in Z} x_{ij}(\nu_i - \mu_i)\end{aligned}$$

Notice that for $i \in Z$, $\beta_i = 0$ and so $\sum_{k=1}^n \alpha_k y_k x_{ik} + \nu_i - \mu_i = 0$. Therefore $\nu_i - \mu_i = -\sum_{k=1}^n \alpha_k y_k x_{ik}$. Hence we have:

$$x_j^T(\nu-\mu) = \sum_{k=1}^n \alpha_k(\rho\sum_{i\in\mathcal{V}_-} x_{ij}\sigma_{ik} - \rho\sum_{i\in\mathcal{V}_+} x_{ij}\sigma_{ik} + \sum_{i\in Z} y_k x_{ik}) = \sum_{k=1}^n \alpha_k g_{jk},$$

where $g_{jk} = \rho\sum_{i\in\mathcal{V}_-} x_{ij}\sigma_{ik} - \rho\sum_{i\in\mathcal{V}_+} x_{ij}\sigma_{ik} + \sum_{i\in Z} y_k x_{ik}$. Now, let $\delta_k = \alpha_k^l - \alpha_k$. After some rearrangements, equation (15) becomes:

$$\sum_{k\in\mathcal{E}} \delta_k(y_j y_k x_j^T x_k + y_j g_{jk} - \rho h_{jk}) + y_j\delta_0 = \lambda^l - \lambda,$$

for all $j \in \mathcal{E}$. Thus by constructing a matrix $\mathbf{K}$ with entry $K_{jk} = y_j y_k x_j^T x_k + y_j g_{jk} - \rho h_{jk}$, the above system of equations can be represented by the following matrix form:

$$\mathbf{K}\boldsymbol{\delta} + \delta_0\mathbf{y}_l = (\lambda^l - \lambda)\mathbf{1},$$

where $\mathbf{y}_l$ is a vector of length m whose entries are y_i's for $i \in \mathcal{E}$. In addition, since $\sum_{i=1}^n \alpha_i y_i = 0$, we have $\sum_{k\in\mathcal{E}} y_k\delta_k = 0$ and so

$$\mathbf{y}_l^T\boldsymbol{\delta} = 0.$$

So all together we have $m+1$ unknown variables and $m+1$ linear equations. Now let:

$$\mathbf{A}_l = \begin{pmatrix} 0 & \mathbf{y}_l^T \\ \mathbf{y}_l & \mathbf{K}_l \end{pmatrix},\ \boldsymbol{\delta}^a = \begin{pmatrix} \delta_0 \\ \boldsymbol{\delta} \end{pmatrix},\ \mathbf{1}^a = \begin{pmatrix} 0 \\ \mathbf{1} \end{pmatrix}.$$

So the linear system can be written as

$$\mathbf{A}_l\boldsymbol{\delta}^a = (\lambda_l - \lambda)\mathbf{1}^a.$$

Let $\mathbf{b}^a = \mathbf{A}_l^{-1}\mathbf{1}^a$, then for $k = 0$ or $k \in \mathcal{E}_l$, we have

$$\alpha_k = \alpha_k^l - (\lambda^l - \lambda)b_k, \tag{16}$$

which shows linearity of α between two events, i.e. $\lambda^{l+1} < \lambda < \lambda^l$. To evaluate the function value of f at the data point x_j, we have

$$f(x_j) = \frac{\lambda^l}{\lambda}(f^l(x_j) - h^l(x_j)) + h^l(x_j), \tag{17}$$

where

$$h^l(x) = \sum_{k\in\mathcal{E}_l} b_j(y_k x_j^T x_k + g_{jk}) + b_0.$$

3.2 Finding the Next Event

The linear (or inverse linear) path established above continues until one of the following events occur:

1. One of the points (if any) on the elbow $\mathcal{E}_l$ is about to enter either $\mathcal{R}$ or $\mathcal{L}$. For the first case $\alpha_i = 0$, by solving equation (**??**) we can obtain a candidate breakpoint λ for the j-th data point:
$$\lambda = \frac{\lambda^l b_j - \alpha_j^l}{b_j}.$$
Similar, in the case where the point is about to enter $\mathcal{L}$ we have
$$\lambda = \frac{\lambda^l b_j - \alpha_j^l + 1}{b_j}.$$
Compute the above candidate break points for all data points that are currently on the elbow (so that we have $2m$ such candidates). Take the greatest such λ that is smaller than λ^l, denote that quantity by λ_a.

2. One of the points j in either $\mathcal{L}^l$ or $\mathcal{R}^l$ enters the elbow, that is $y_j f(x_j) = 1 + \rho\sigma_j^T|\beta|$. By equation (**??**), we have
$$\frac{\lambda^l}{\lambda}(f^l(x_j) - h^l(x_j))y_j + h^l(x_j)y_j = 1 + \frac{1}{\lambda}\sum_{i=1}^{n} \alpha_k h_k.$$
The right-hand-side of the above equation can be rewritten as
$$1 + \frac{1}{\lambda}\sum_{k\in\mathcal{L}} h_k + \frac{1}{\lambda}\sum_{k\in\mathcal{E}}(\alpha_k^l + (\lambda - \lambda^l)b_k)h_k.$$
So a break point candidate is
$$\lambda = \frac{\lambda^l(f^l(x_j) - h^l(x_{jk})) - y_j\sum_{k\in\mathcal{L}} h_{jk} - y_j\sum_{k\in\mathcal{E}}\alpha_k^l h_{jk} + \lambda^l\sum_{k\in\mathcal{E}} b_k h_{jk}}{y_j - h^l(x_j) + y_j\sum_{k\in\mathcal{E}} b_k h_{jk}}.$$

3. A nonzero component of β becomes zero. This means for each $i \in \mathcal{V}_+$, by equation (XX) we solve $\sum_{k=1}^n \alpha_k w_{ik} = 0$ and obtain a break point candidate
$$\lambda = \frac{\lambda^l\sum_{k\in\mathcal{E}} b_k w_{ik} - \sum_{k\in\mathcal{E}}\alpha_k^l w_{ik} - \sum_{k\in\mathcal{L}} w_{ik}}{\sum_{k\in\mathcal{E}} b_k w_{ik}}.$$
For $i \in \mathcal{V}_-$, simply replace w_{ik} by z_{ik}.

4. A zero component of β becomes nonzero. For a component $i \in Z$ to become positive, we require $c_i = 0$ and $\nu_i = 0$. By equation (XX), $\mu_i = \sum_{k=1}^n \alpha_k y_k x_{ik}$, together with the optimality condition (XX), we need to solve $\rho\sum_{k=1}^n \alpha_k\sigma_{ik} = \sum_{k=1}^n \alpha_k y_k x_{ik}$. Same analysis for $i \in Z$ to become negative. Hence we have
$$\lambda = \frac{\lambda^l\sum_{k\in\mathcal{E}} b_k w_{ik} - \sum_{k\in\mathcal{E}}\alpha_k^l w_{ik} - \sum_{k\in\mathcal{L}} w_{ik}}{\sum_{k\in\mathcal{E}} b_k w_{ik}}.$$

4 Termination

5 Computation Complexity

6 Extension to Kernal

(You need to transform the uncertainty in the original data space into the kernal space)

References

[1] Saharon Rosset, Ji Zhu. (2007)

[2] T Hastie, S Rosset, R Tibshirani… - The Journal of Machine …, 2004 - dl.acm.org

[3] L El Ghaoui, G Lanckeriet, G Natsoulis. (2003)

[4] Element of machine learning