电子科技大学

University of Electronic Science and Technology of China

Incremental SVM

Heng Zhang

Data Mining Lab, Big Data Research Center, UESTC
Email：junmshao@uestc.edu.cn
http://staff.uestc.edu.cn/shaojunming

Outline

1. Daul problem and KKT conditions of SVM
2. Incremental procedure
3. Regularization parameter perturbation
4. Kernel parameter perturbation

SVM classifiers of the form $f(x) = w \cdot \Phi(x) + b$ are learned from the data $\{(x_i, y_i) \in R^m * \{-1,1\} \forall\, i\, \{1, \ldots, N\}\}$ by minimizing

$$\min_{w,b,\xi} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N} \xi_i$$

subject to

$$y_i(w \cdot \Phi(x_i) + b) \geq 1 - \xi_i, \quad \xi_i \geq 0 \quad \forall i \in \{1, \ldots, N\}$$

Lagrange multiplier:

$$L(w,b,a)=\frac{1}{2}(w^T\cdot w)-\sum_{i=1}^{n}\alpha_i[y_i(w^T\cdot x+b)-1]$$

Partial derivatives = 0

$$\frac{\partial L}{\partial w}=w-\sum_{i=1}^{n}y_i\alpha_i x_i=0$$

$$\frac{\partial L}{\partial b}=\sum_{i=1}^{n}y_i\alpha_i=0$$

Daul form

$$\max_{0\le\alpha_i\le C} W = \frac{1}{2}\sum_{i,j=1}^{N}\alpha_i Q_{ij}\alpha_j - \sum_{i=1}^{N}\alpha_i + b\sum_{i=1}^{N} y_i\alpha_i$$

$$s.t., \alpha_i \ge 0, i = 1,\cdots,n$$

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

with Lagrange multiplier b and kernel function

$$Q_{ij} = y_i y_j \Phi(x_i)\cdot\Phi(x_j)$$

$$K(x,y) = \Phi(x)\cdot\Phi(y)$$

KKT conditions

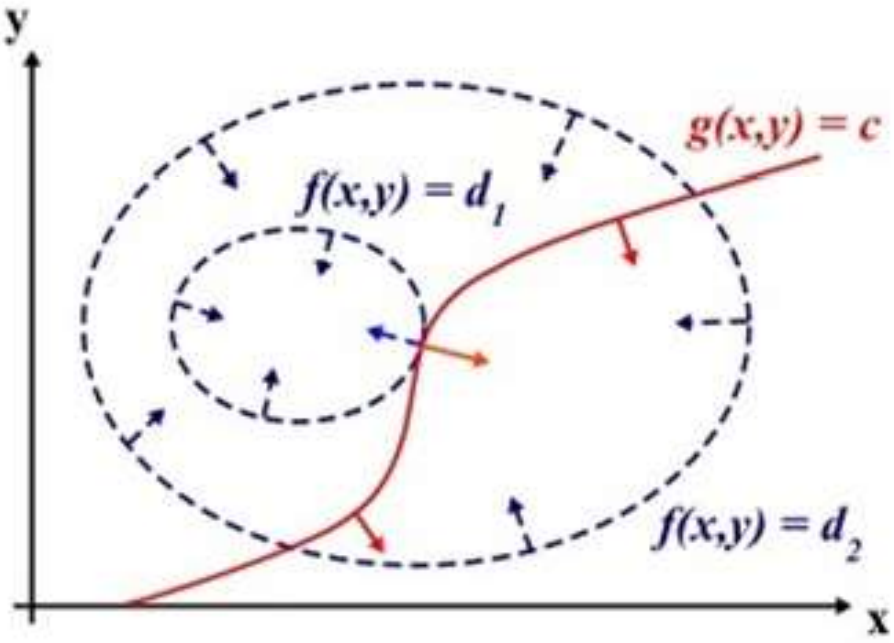
In order to solve the daul parameters $\{\alpha, b\}$

$$g_i = \frac{\partial W}{\partial \alpha_i} = \sum_j Q_{ij}\alpha_j + y_i b - 1 = y_i f(x_i) - 1 \begin{cases} > 0 & \alpha_i = 0 \\ = 0 & 0 < \alpha_i < C \\ > 0 & \alpha_i = C \end{cases}$$

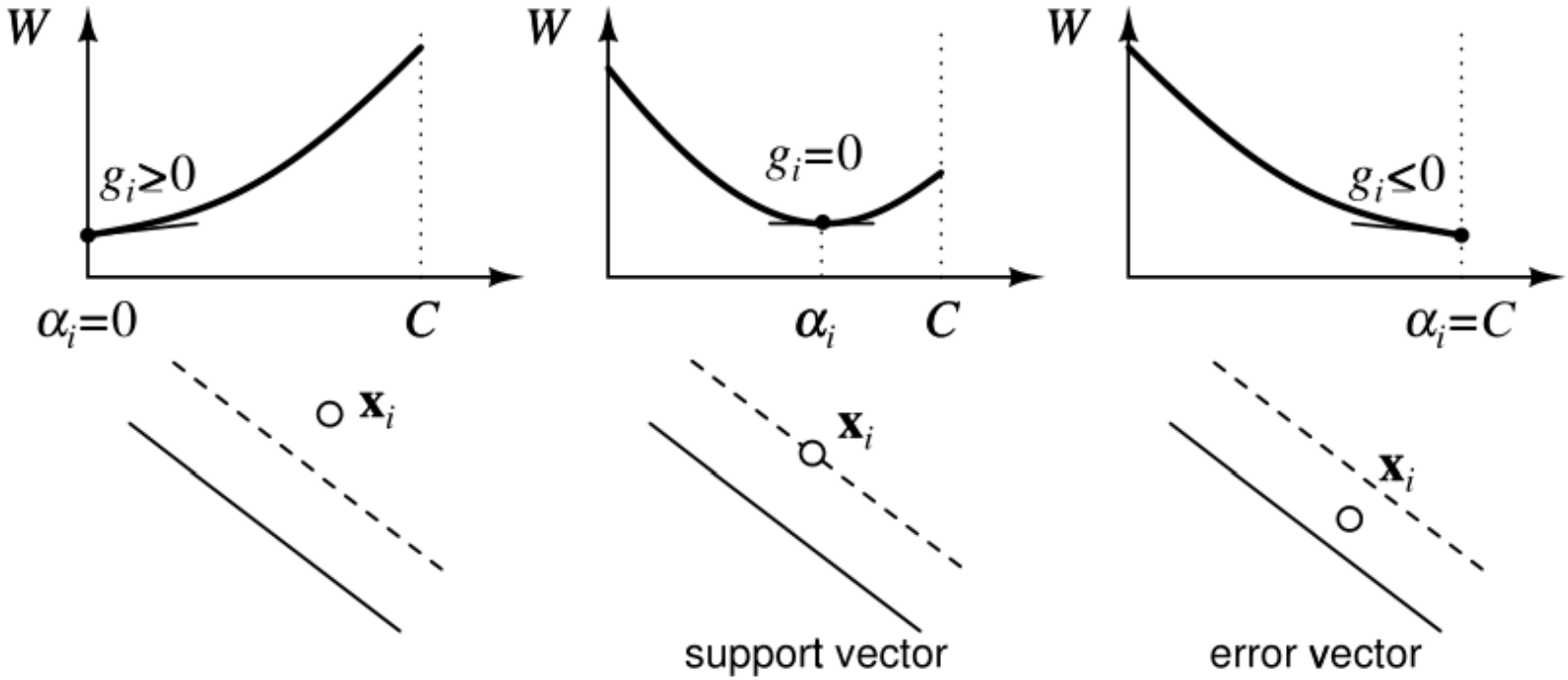
$$h = \frac{\partial W}{\partial b} = \sum_{j=1}^{N} y_j \alpha_j \equiv 0$$

KKT conditions

Why lagrange multipler and KKT could get the optimal value?



Picture from wiki

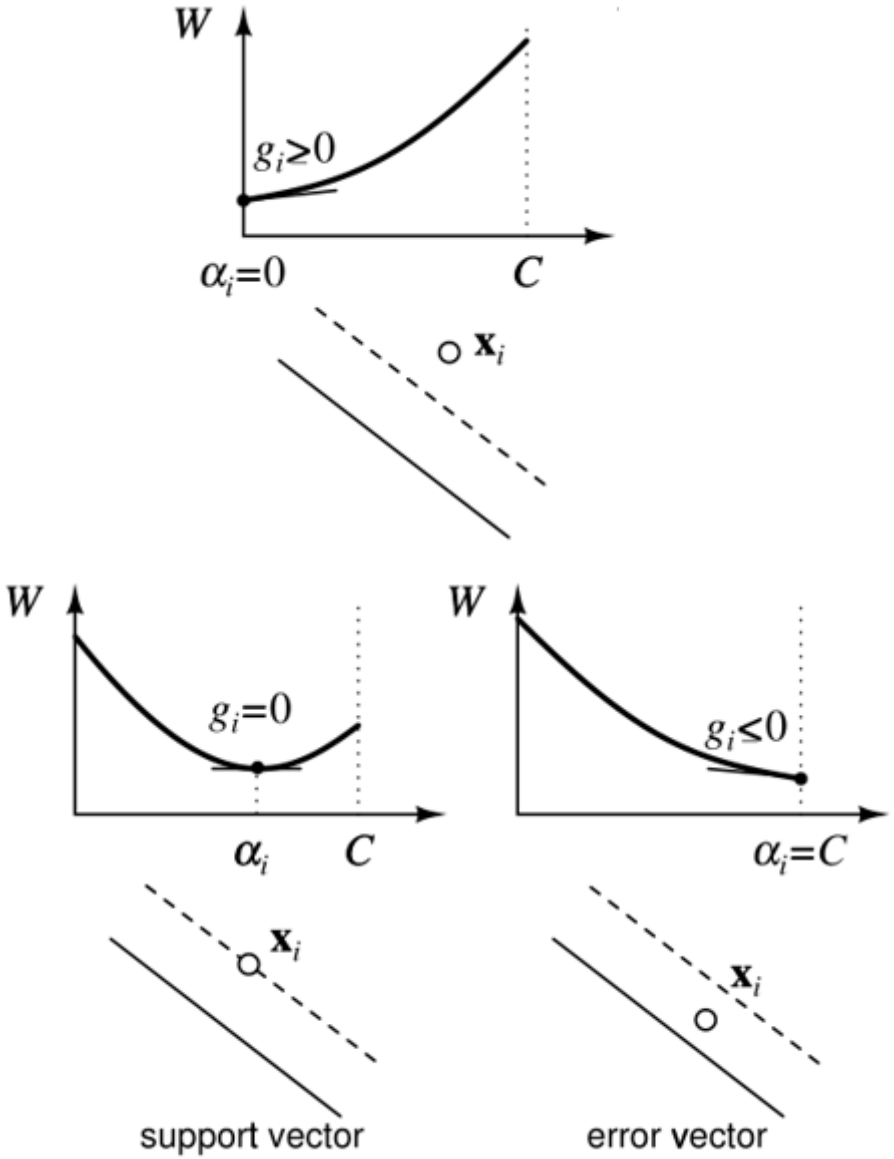


Based on the partial derivatives g_i ,the training examples can be partitioned into three different categories:

($g_i = 0$) the set **S** of margin support vectors on the margin

($g_i < 0$) the set **Ɛ** of error support vectors violating the margin

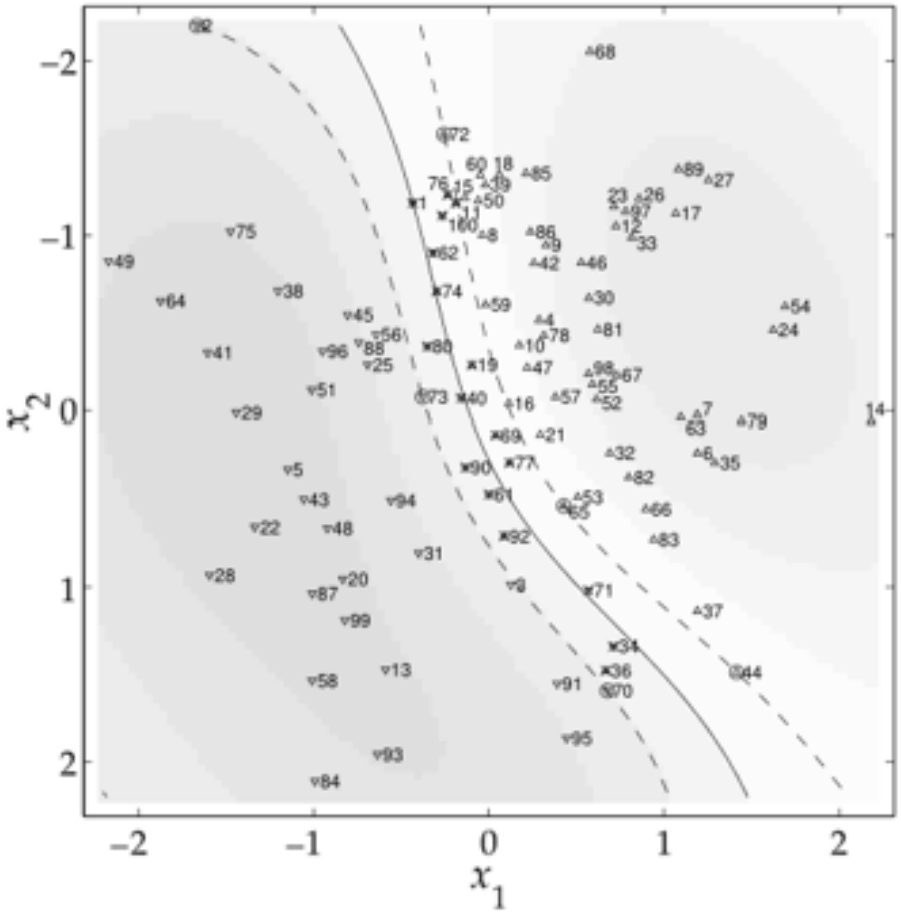
($g_i > 0$) the set **R** of reserve vectors exceeding the margin



R

U

Unlearned vector


−2
−1
0
1
2
x_2
−2
−1
0
1
2
x_1

Incremental Procedure

Margin vector coefficients change during each incremental ,simultaneously preserve the **KKT**:

$$\Delta g_i = \sum_{k \in S} Q_{ik} \Delta \alpha_k + \sum_{l \in U} Q_{il} \Delta \alpha_l + y_i \Delta b = 0 \quad \forall i \in S$$

$$\Delta h = \sum_{k \in S} y_k \Delta \alpha_k + \sum_{l \in U} y_l \Delta \alpha_l = 0$$

The overall perturbation process is controlled by a perturbation parameter p : varies from 0 to 1 as the SVM solution is perturbed from initial unlearned to final learned result. So let:

$$\Delta\alpha_k = \beta_k \Delta p \ (k \in S)$$

$$\Delta\alpha_l = \lambda_l \Delta p \ (l \in U)$$

$$\Delta b = \beta \Delta p$$

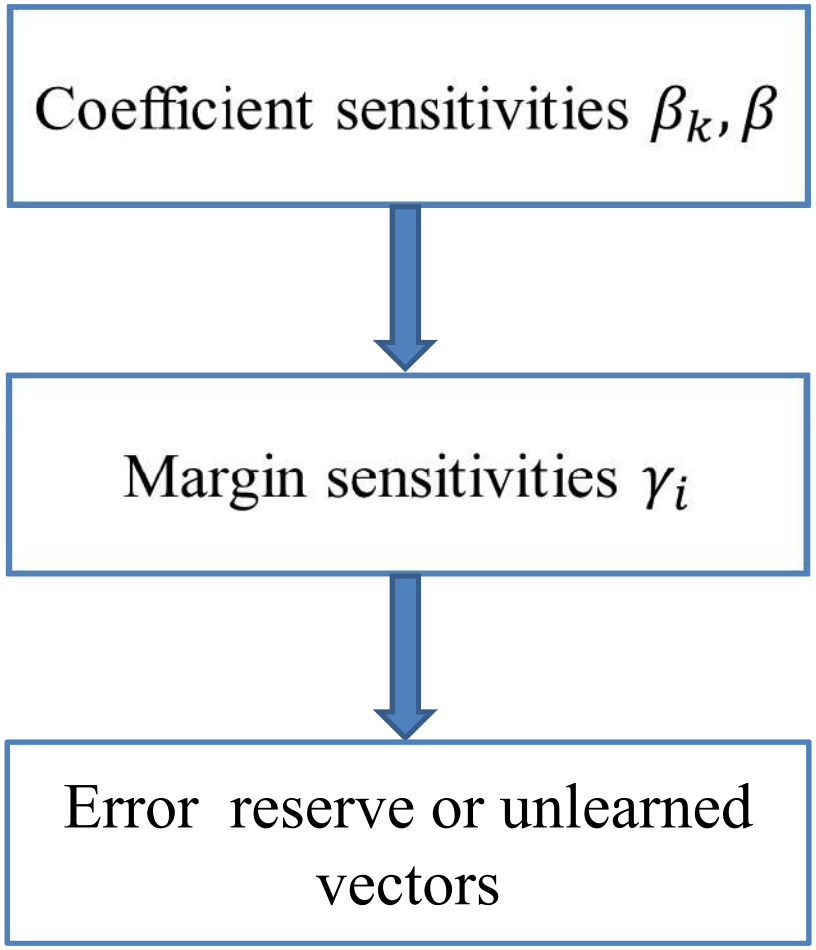
$$\gamma_i = \frac{\Delta g_i}{\Delta p} = \sum_{k \in S} Q_{ik}\beta_k + \sum_{l \in U} Q_{il}\lambda_l + y_i\beta = 0 \quad \forall i \in S$$

$$\frac{\Delta h}{\Delta p} = \sum_{k \in S} y_k\beta_k + \sum_{l \in U} y_l\lambda_l = 0$$

$$\{\lambda_l = C : \forall l \in U\}$$

$$\Delta p_{min} = \min_{c \in C} \Delta p_c$$

Common			
Initial Category	*New Category*	Δp	*Condition*
Margin	Reserve	$\frac{-\alpha_i}{\beta_i}$	$\beta_i < 0$
Error	Margin	$\frac{-g_i}{\gamma_i}$	$\gamma_i > 0$
Reserve	Margin	$\frac{-g_i}{\gamma_i}$	$\gamma_i < 0$
Incremental/Decremental Learning			
Margin	Error	$\frac{C-\alpha_i}{\beta_i}$	$\beta_i > 0$
Unlearned ($g_i < 0$)	Margin	$\frac{-g_i}{\gamma_i}$	$\gamma_i > 0$
Unlearned ($g_i < 0$)	Error	$\frac{C-\alpha_i}{\lambda_i}$	$\lambda_i > 0$
Regularization Parameter Perturbation			
Margin	Error	$\frac{C_t-\alpha_i}{\beta_i-\Delta C}$	$\beta_i > \Delta C$


Coefficient sensitivities β_k, β
Margin sensitivities γ_i
Error reserve or unlearned vectors

Coefficient sensitivities $\{\beta_k, \beta\}$

$$\gamma_i = \frac{\Delta g_i}{\Delta p} = \sum_{k \in S} Q_{ik} \beta_k + \sum_{l \in U} Q_{il} \lambda_l + y_i \beta = 0 \quad \forall i \in S$$

$$\frac{\Delta h}{\Delta p} = \sum_{k \in S} y_k \beta_k + \sum_{l \in U} y_l \lambda_l = 0$$

Coefficient sensitivities $\{\beta_k, \beta\}$

Transform the KKT:

$$Q\beta = -\sum_{l \in U} \lambda_l v_l$$

Where

$$\beta = \begin{bmatrix} \beta \\ \beta_{s1} \\ \vdots \\ \beta_{sn} \end{bmatrix} \quad v_l = \begin{bmatrix} y_l \\ Q_{s1l} \\ \vdots \\ Q_{snl} \end{bmatrix} \quad Q = \begin{bmatrix} 0 & y_{s1} & \cdots & y_{sn} \\ y_{s1} & Q_{s1s1} & \cdots & Q_{s1sn} \\ \vdots & \vdots & \ddots & \vdots \\ y_{sn} & Q_{sns1} & \cdots & Q_{snsn} \end{bmatrix}$$

And $n = |S|$ Q: symmetric but not positive-definite Jacobian

In order to compute the sensitivities require:

$$R = Q^{-1}$$

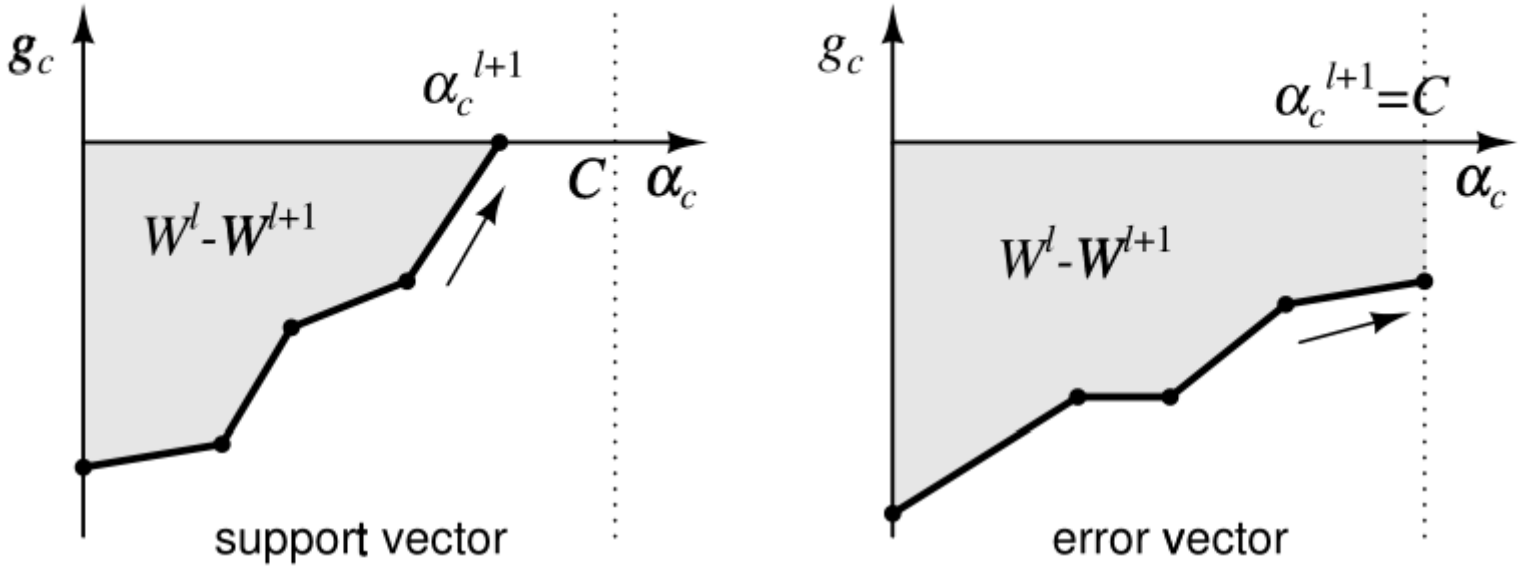
$$\beta = -\sum_{l \in U} \lambda_l R v_l$$

When added an example to S, R expands as:

$$R \leftarrow \begin{bmatrix} & & 0 \\ & R & \vdots \\ & & 0 \\ 0 & \cdots \quad 0 & 0 \end{bmatrix} + \frac{1}{\gamma_{s_{n+1}}^{s_{n+1}}} \begin{bmatrix} \beta \\ \beta_{s_1} \\ \vdots \\ \beta_{s_{ls}} \\ 1 \end{bmatrix} \begin{bmatrix} \beta & \beta_{s_1} & \cdots & \beta_{s_{ls}} & 1 \end{bmatrix}^T$$

$$\gamma_{s_{n+1}}^{s_{n+1}} = Q_{s_{n+1}s_{n+1}} + v_{s_{n+1}}^T \beta_{s_{n+1}}$$

Incremental procedure figure



Incremental learning. A new vector, initially for α_c=0 classified with negative margin $g_c < 0$, becomes a new margin or error vector.

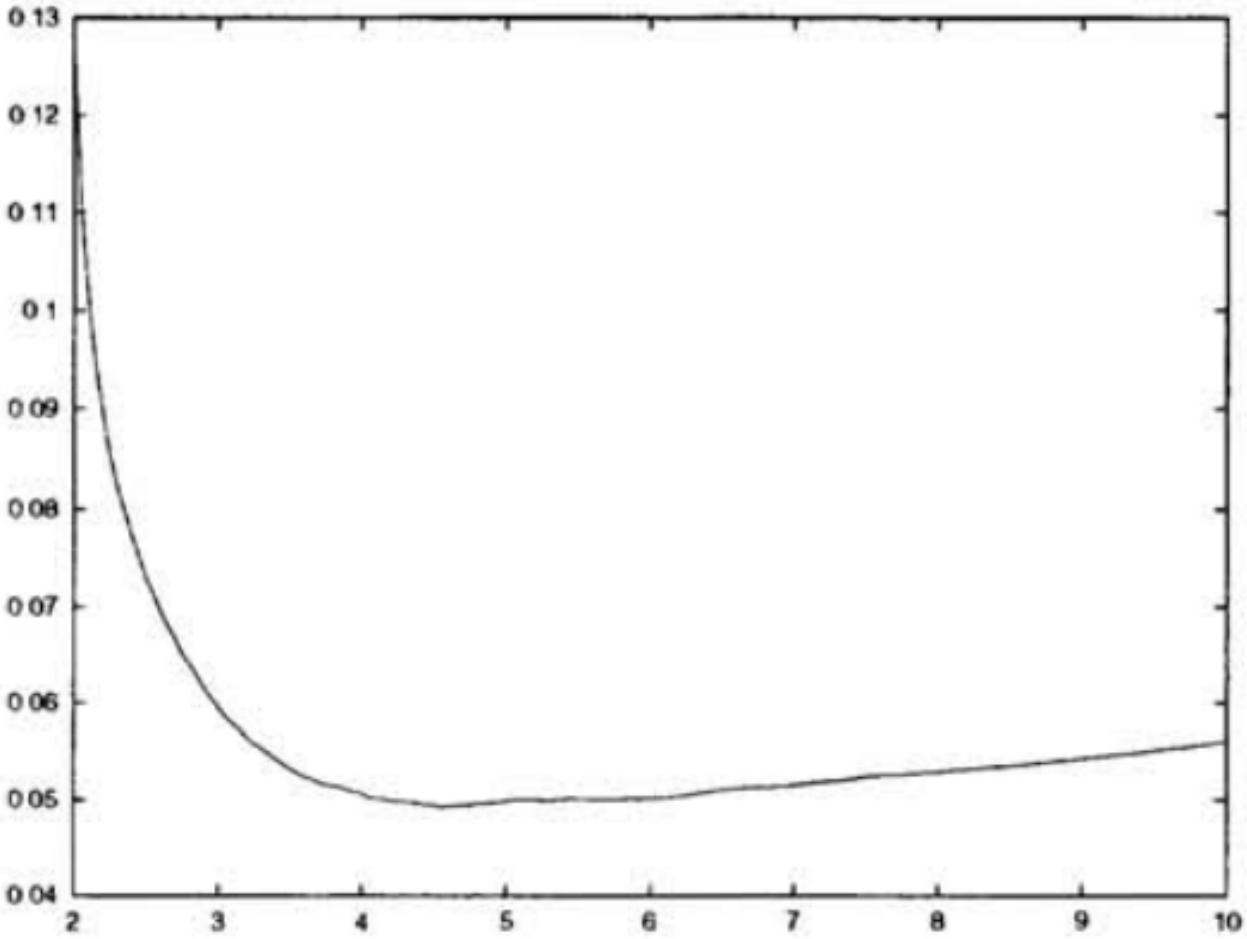
Incremental procedure

Algorithm 1 (Incremental Learning, $\ell \to \ell + 1$)

1. *Initialize α_c to zero;*
2. *If $g_c > 0$, terminate (c is not a margin or error vector);*
3. *If $g_c \leq 0$, apply the largest possible increment α_c so that (the first) one of the following conditions occurs:*
   - *(a) $g_c = 0$: Add c to margin set S, update $\mathcal{R}$ accordingly, and terminate;*
   - *(b) $\alpha_c = C$: Add c to error set E, and terminate;*
   - *(c) Elements of D^ℓ migrate across S, E, and R ("bookkeeping," section 2.3): Update membership of elements and, if S changes, update $\mathcal{R}$ accordingly.*

   and repeat as necessary.

Kernel parameter perturbation:



Generalization error (y-axis)as a function of σ (x-axis) for Gaussian kernels with zero training error and maximal margin over 10^5examples.

Kernel parameter perturbation:

Theorem: The margin γ of SVM depends smoothly on the kernel parameter σ.

Let $F(x, \bar{y})$ be a continuously differentiable function: $F: U \subseteq R \times V \subseteq R^p \rightarrow R$ and $(a, \bar{b})$ be a solution to the $F(x, \bar{y})=0$. Then near $(a, \bar{b})$ there exists one and only one function $\bar{y} = \bar{g}(x)$ such that $F(x, \bar{y})=0$ and such function is continuous.

$$W_\sigma(\bar{\alpha}) = \sum_{i=1}^{p} \bar{\alpha}_i - 1/2 \sum_{i,j} \bar{\alpha}_i \bar{\alpha}_j y_i y_j K(\sigma; x_i, x_j) + \lambda(\sum_i y_i \bar{\alpha}_i)$$

Kernel parameter perturbation:

Now consider the function $F(\sigma, \bar{\alpha}, \lambda)_i = (\frac{\partial W_\alpha}{\partial \bar{\alpha}})_{ji}$

$$\frac{\partial W_\sigma}{\partial \alpha_j} = 1 - y_i \sum_i \bar{\alpha}_i y_i K(\sigma; x_i, y_i) + \lambda y_i$$

and satisfied the equation $F\left(\sigma, \bar{\alpha}^0(\sigma), \lambda(\sigma)\right) = 0$;
$((\bar{\alpha}^0, \lambda) = \bar{g}(\sigma))$

So there exist a continuous function between σ *and* $\bar{\alpha}^0$

Kernel parameter perturbation:

As we known, the margin if $\|W\|^{-1}$ while the object function:

$$W(\alpha^0) = 0.5\|\mathbf{w}\|^2 = \sum_{i=1}^{m} \alpha_i^0 - \frac{1}{2}\sum_{i,j}^{m} \alpha_i^0 \alpha_j^0 y_i y_j k(\mathbf{x}_i, \mathbf{x}_j).$$

KKT conditions are satisfied ensuring that $\alpha_i^0 = 0$ while for support vector

$$y_i\left(\sum_{j=1}^{m} \alpha_j^0 y_j k(X_i, X_j) - b\right) = 1$$

Hence:

$$\sum_{i,j}^{m} \alpha_i^0 \alpha_j^0 y_i y_j k(\mathbf{x}_i, \mathbf{x}_j) = \sum_{i=1}^{m} \alpha_i^0 (1 + b y_i) = \sum_{i=1}^{m} \alpha_i^0,$$

since $\sum_{i=1}^{m} \alpha_i^0 y_i = 0$. ■

Kernel parameter perturbation:

$$\gamma^2 = (\sum_{i=1}^{p} \overline{\alpha}_i^0)^{-1}$$

Kernel Selection procedure:

1.Initialize σ to a very small value
2.Maximize the margin, then
    Observe the validation error
    Increase the kernel parameter: $\sigma \leftarrow \sigma + \Delta\sigma$
3.Stop when a predetermined value of σ is reached else repeat step2
4.Choose the minimum validation error one as the final σ.

Thanks

Heng Zhang

hengzhang64@gmail.com