The idea behind most of the SVM implementation algorithms is to reduce the training set. SMO takes this idea to the extreme: it updates the α's for two points at a time.

Analytical Solution for two points

So, once again we have a training set

$$S = \{x_1, \dots, x_n\} \; ; \; Y = \{y_1 \dots y_n\} \quad y_i = \pm 1$$

Let α_1, α_2 the multipliers for two of the training data, WLOG, x_1, x_2.

Recall that α_i, y_i must satisfy $\sum_{i=1}^{n} \alpha_i y_i = 0$

This means that

$$\alpha_1 y_1 + \alpha_2 y_2 + \sum_{i=3}^{n} \alpha_i y_i = 0$$

$$\Rightarrow \quad \alpha_1 y_1 + \alpha_2 y_2 = -\sum_{i=3}^{n} \alpha_i y_i = ct \quad \left(\text{Since } \alpha_i, i \geq 3 \text{ are unchanged}\right)$$

Thus it must be true that

$$\alpha_1^{old} y_1 + \alpha_2^{old} y_2 = \alpha_1^{new} y_1 + \alpha_2^{new} y_2 = ct \quad (*)$$

Very quickly notice from $(*)$ that if we update say α_2 to obtain α_2^{new}, α_1^{new} can also be calculated immediately.

2004/04/16

SMO-2

$$\alpha_1^{new} = (\alpha_1^{old} y_1 + \alpha_2^{old} y_2 - \alpha_2^{new} y_2)/y_1 =$$

$$= \alpha_1^{old} + \frac{y_2}{y_1}(\alpha_2^{old} - \alpha_2^{new})$$

$$\frac{y_2}{y_1} = \frac{y_2 y_1}{y_1^2} = \frac{y_2 y_1}{1} = y_2 y_1$$

$$\Rightarrow \boxed{\alpha_1^{new} = \alpha_1^{old} + y_2 y_1(\alpha_2^{old} - \alpha_2^{new})} \quad (A_1)$$

Introduce: $s = y_1 y_2$;

$f(\vec{x}_i)$ = the current hypothesis $= \sum_{j=1}^{n} \alpha_j y_j K(\vec{x}_i, \vec{x}_j) + b$

$E_i = f(\vec{x}_i) - y_i, \quad i = 1, 2$

E_i is the difference between the current output and the target classification of $\vec{x}_1$ and $\vec{x}_2$.

Let $A = \| \varphi(\vec{x}_1) - \varphi(\vec{x}_2) \| = K_{11} - 2K_{12} + K_{22}$

where φ is the implicit mapping corresponding to the kernel K: $K_{ij} = K(\vec{x}_i, \vec{x}_j)$.

THEOREM The maximum of the objective function $L(\vec{\alpha})$ from the optimization problem

Maximize $L(\vec{\alpha}) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j K(\vec{x}_i, \vec{x}_j)$

subject to $\sum_{i=1}^{n} \alpha_i y_i = 0$ and $\alpha_i \geq 0$ (eventually $\leq C$)

when only α_1 and α_2 are allowed to change is is attained by first changing α_i $i=1,2$ as follows:

(i) compute $\alpha_2^{new,unc} = \alpha_2^{old} + \frac{Y_2(E_1-E_2)}{A}$

(ii) clip $\alpha_2^{new,unc}$ to obtain $\alpha_2^{new} \in [L, H]$

(where L, H follow from the requirement $\alpha_j' \geq 0 \ (\leq C)$)

and

(iii) compute α_1^{new} from α_2^{new} using equation (A1)

Proof First define

$$V_i = \sum_{j=3}^{n} \alpha_j y_j K_{ij} = f(x_i) + b - \sum_{j=1}^{2} \alpha_j y_j K_{ij}$$

Consider $L(\vec{\alpha})$ as a function of α_1 and α_2 only. That is

$$L(\alpha_1, \alpha_2) = \alpha_1 + \alpha_2 - \frac{1}{2}\alpha_1^2 K_{11} - \frac{1}{2}\alpha_2^2 K_{22} - \frac{1}{2} 2 \alpha_1 \alpha_2 Y_1 Y_2 K_{12}$$

$$- \frac{1}{2} 2 Y_1 \alpha_1 V_1 - \frac{1}{2} 2 Y_2 \alpha_2 V_2 + \text{Constant}$$

Recall that $\sum_{i=1}^{n} \alpha_i^{old} Y_i = 0 = \sum_{i=1}^{n} \alpha_i Y_i = 0 \Rightarrow$

$$\alpha_1 + s\alpha_2 = \alpha_1^{old} + s\alpha_2^{old} = \gamma \ (\text{Constant}) \quad (\Gamma)$$

(easy $\alpha_1 Y_1 + \alpha_2 Y_2 = -\sum_{i=3}^{n} \alpha_i Y_i$. Multiply by Y_1

to obtain $\alpha_1 \underbrace{Y_1^2}_{=1} + \alpha_2 \underbrace{Y_1 Y_2}_{=s} = \underbrace{- Y_1 \sum_{i=1}^{n} \alpha_i Y_i}_{\gamma}$.

$\Rightarrow \quad \alpha_1 + s\alpha_2 = \gamma$)

So, now we can solve (Γ) for α_1 in terms of α_2

$\Rightarrow \alpha_1 = \gamma - s\alpha_2$

Substitute this in $L(\alpha_1, \alpha_2)$ to obtain $L(\alpha_2)$:

$$L(\alpha_2) = (\gamma - s\alpha_2) + \alpha_2 - \frac{1}{2}k_{11}(\gamma - s\alpha_2)^2 - \frac{1}{2}k_{22}\alpha_2^2$$

$$- sk_{12}(\gamma - \alpha_2)\alpha_2 - y_1(\gamma - s\alpha_2)v_1 - y_2\alpha_2 v_2 + \text{Constant}$$

$$= \gamma + (1-s)\alpha_2 - \frac{1}{2}k_{11}(\gamma - s\alpha_2)^2 - \frac{1}{2}k_{22}\alpha_2^2 -$$

$$- \gamma s k_{12}\alpha_2 - sk_{12}\alpha_2^2 - y_1\gamma v_1 - y_1 s v_1\alpha_2 - y_2 v_2\alpha_2 + \text{Const}$$

Take now derivative with respect to α_2 to obtain:

$$L'(\alpha_2) = (1-s) - \frac{1}{2}2k_{11}(\gamma - s\alpha_2)(-s) - \frac{1}{2}2k_{22}\alpha_2 -$$

$$- \gamma s k_{12} - 2sk_{12}\alpha_2 - y_1 s v_1 - y_2 v_2 =$$

$$= (1-s) + \gamma s k_{11} - k_{11}s^2\alpha_2 - k_{22}\alpha_2 - \gamma s k_{12}$$

$$- 2sk_{12}\alpha_2 - y_1 s v_1 - y_2 v_2 =$$

Use $\| s^2 = (y_1 y_2)^2 = 1$

$$= (1-s) + \gamma s k_{11} - k_{11}\alpha_2 - k_{22}\alpha_2 - \gamma s k_{12} - 2sk_{12}\alpha_2 - y_1 s v_1 - y_2 v_2$$

set $= 0$ Solve for α_2

$$-\alpha_2(k_{11} + k_{22} + 2\cancel{s}k_{12}) =$$

$$-1 + s + \gamma s(k_{11} - k_{12}) + y_1 s v_1 + y_2 v_2$$

$$= y_2(y_2 - y_1 + \gamma y_1(k_{11} - k_{12}) + v_1 - v_2$$

$$- \alpha_2 A y_2 = y_2 - y_1 + f(\vec{x}_1) - \sum_{j=1}^{2} \alpha_j y_j k_{1j} + \gamma y_1 k_{11} -$$

$$- f(\vec{x}_2) + \sum_{j=1}^{2} \alpha_j y_j k_{2j} - \gamma y_1 k_{12}$$

$$= y_2 - y_1 + f(\vec{x}_1) - f(\vec{x}_2) +$$

$$y_2 \alpha_2 k_{11} - y_2 \alpha_2 k_{12} + \alpha_2 \alpha_2 k_{22} - y_2 \alpha_2 k_{12} =$$

$$= -y_2 \alpha_2 A + (f(\vec{x}_1) - y_1) - (f(\vec{x}_2) - y_2)$$

Thus $$\alpha_2^{new} = \alpha_2^{old} \frac{-y_2 A}{-y_2 A} + \frac{E_1 - E_2}{-y_2 A}$$

$$= \alpha_2^{old} - \frac{y_2 (E_1 - E_2)}{A}$$

qed.

Let us look at the constraint on α_1, α_2

$\alpha_2^{new} \in [L, H]$ where

$$L = \begin{cases} \max(0, \alpha_2^{old} - \alpha_1^{old}) & \text{if } y_1 \neq y_2 \\ \max(0, \alpha_1^{old} + \alpha_2^{old} - C) & \text{if } y_1 = y_2 \end{cases} \quad H = \begin{cases} \min(C, C - \alpha_1^{old} + \alpha_2^{old}), & y_1 \neq y_2 \\ \min(C, \alpha_1^{old} + \alpha_2^{old}) & \text{if } y_1 = y_2 \end{cases}$$

These follow from the requirement $0 \leq \alpha_2^{new} \leq C$

How to select α_1 and α_2 ?

First choice : x_1 (α_1) : select x_1 which violates KKT conditions..

Second choice : x_2(α_2)

We look for the point whose update will cause highest increase in $L(\vec{\alpha})$. A quick approximation is to select x_2 such that

$|E_1 - E_2|$ is maximum.

Remarks

1) Bias (b): No prescription on how to select b. Because we use $E_1 - E_2$ b, whatever it is will be cancelled. So we can take $b=0$ and set it at the end based on KKT conditions. However b may be required in the stopping of the SMO algorithm (i.e ~~for~~ monitoring KKT conditions). An estimate of b can be computed from the updated points only