The Sequential Minimum Optimization (SMO) Algorithm

Let us take a break (sort of) from theory only and try to look at implementation issues.

As you can see in the algorithms for solving the SVM problem or the smallest (hyper)sphere problem we need to solve the associated optimization problem.

Recall that for the type of optimization problems we have, the KKT conditions are *necessary and sufficient*.

But the complexity of solving them in exponential in n, the size of the training set!

This means that using them is a real option only for a small training se.

However, this observation is at the same time very useful!
Why? Because it leads us to the idea that if we could select an appropriate small subset of the training set we could solve the problem.

QUESTION: How small a subset and what do we mean by "appropriate"?

ANSWER:

- The training subset can be very small: 2 elements
- By appropriate we mean that *the subset should be near the actual decision rule, that there should be at least one point from each class*. See figure 1 below.

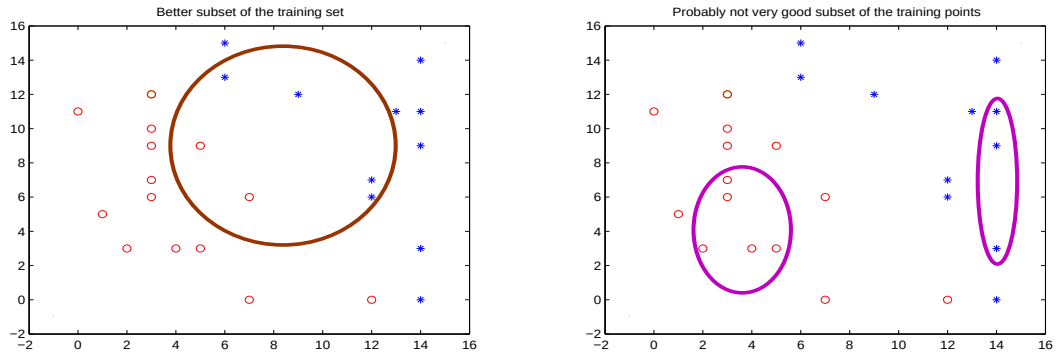

Figure 1: Examples of training subsets

Analytitcal Solution for Two Points

The dual optimization problem is

$$\begin{array}{ll} \text{maximize} & W(\alpha) = \sum \alpha_i - \frac{1}{2} \sum_{i,j=1}^{l} \alpha_i \alpha_j K(\mathbf{x}_i, \mathbf{x}_j) \\ \text{subject to} & \sum \alpha_i y_i = 0, \\ & 0 \leq \alpha_i \leq C, i = 1, \ldots, l \end{array} \tag{11}$$

where $C = \infty$ for the hard margin SVM.

At each step SMO chooses two elements α_i, α_j to jointly optimize, finds the optimal values for them given that *all others are fixed* and updates the vector α.

So, we have two issues:

1. Select the two elements: *heuristic*
2. optimize: *analytic*

WLOG assume the chosen multipliers are α_1 and α_2. Assume that we have l data points. Thus, $\alpha = (\alpha_1, \ldots, \alpha_l)$.

Recall the $\sum_{i=1}^{l} \alpha_i y_i = 0$. Thus, keeping α_3, α_l fixed, we have that

$$\alpha_1 y_1 + \alpha_2 y_2 = -\sum_{i=3}^{l} \alpha_i y_i = ct. = \alpha_1^{old} y_1 + \alpha_2^{old} y_2$$

Thus, in the two dimensional space (α_1, α_2), these two multipliers lie on a line.

Further, assume that

$$0 \leq \alpha_1, \alpha_2 \leq C \tag{12}$$

(where C is a constant that bounds the values that the multipliers take.
Now the SVM optimization problem is a one dimensional problem which can be solved analytically as follows:
In one step, we have:

$$\alpha_1^{new} y_1 + \alpha_2^{new} y_2 = \alpha_1^{old} y_1 + \alpha_2^{old} y_2$$

from which, taking into account 12 we obtain the following

$$U \leq \alpha_2^{new} \leq V \tag{13}$$

where

$$U = \max(0, \alpha_1^{old} - \alpha_2^{old}) \tag{14}$$
$$V = \min(C, C - \alpha_1^{old} - \alpha_2^{old}) \tag{15}$$

$$U = \max(0, \alpha_1^{old} - \alpha_2^{old}) \tag{16}$$
$$V = \min(C, C - \alpha_1^{old} - \alpha_2^{old}) \tag{17}$$

if $y_1 = y_2$.
Let

$$E_j = \sum_{i=1}^{l} \alpha_i y_i K(\mathbf{x}_i, \mathbf{x}_j) + b - y_j,\ j = 1, 2$$

That is, E_1 is the difference between the output of the classifier and the target output value y_1 for the training data $\mathbf{x}_1$, and similarly for E_2.

Remark 1 *The magnitude of E_j may be large even if the point is correctly classified (e.g., $Classifier\ output = 5$, $y_1 = 1$, then $E_1 = 4$)*

Remark 2 *The second derivative of the objective function along the diagonal is $-\mathcal{K}$ where*

$$\mathcal{K} = K(\mathbf{x}_1, \mathbf{x}_1) + K(\mathbf{x}_2, \mathbf{x}_2) - 2K(\mathbf{x}_1, \mathbf{x}_2)$$

Theorem 2 *The maximum of the objective funtion W when only α_1 and α_2 are allowed to change, is achieved by the following procedure:*

Step 1. *Compute* $\alpha_2^{new,unclipped} = \alpha_2^{old} + \frac{y_2(E_1 - E_2)}{\mathcal{K}}$.

Step 2. *Clip* $\alpha_2^{new,unclipped}$ *to enforce the constraint* $U \leq \alpha_2^{new} \leq V$*:*

$$\alpha_2^{new} = \begin{cases} V & \text{if } \alpha_2^{new,unclipped} > V \\ \alpha_2^{new,unclipped} & \text{if } U \leq \alpha_2^{new,unclipped} \leq V \\ U & \text{if } \alpha_2^{new,unclipped} < U \end{cases}$$

Step 3. *Obtain the value of* α_1^{new} *from* α_2^{new}*:*

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

Proof:

Let

$$v_i = \sum_{j=3}^{l} y_j \alpha_j K(\mathbf{x}_i, \mathbf{x}_j) = \sum_{j=1}^{l} y_j \alpha_j K(\mathbf{x}_i, \mathbf{x}_j) - \sum_{j=1}^{2} y_j \alpha_j K(\mathbf{x}_i, \mathbf{x}_j) \; i = 1, 2$$

Then we can rewrite the objective function as a function of α_1 and α_2 only as follows:

$$\begin{aligned} W(\alpha_1, \alpha_2) &= \alpha_1 + \alpha_2 - \tfrac{1}{2} K_{11} \alpha_1^2 - \tfrac{1}{2} K_{22} \alpha_2^2 \\ &\quad - y_1 y_2 K_{12} \alpha_1 \alpha_2 - y_1 \alpha_1 v_1 - y_2 \alpha_2 v_2 + CONSTANT \end{aligned}$$

where $K_{ij} = K(\mathbf{x}_i, \mathbf{x}_j)$ for $i, j = 1, 2$; The term CONSTANT contains only $\alpha_i, i = 3, \ldots, l$.

Furthermore, from the condition $\sum_{i=1}^{l} y_i \alpha_i = 0$ we can write

$$y_1 \alpha_1 + y_2 \alpha_2 = const.$$

Multiplying both sides by y_1, and taking into account that $y_i^2 = 1$, we obtain

$$y_1^2 \alpha_1 + y_1 y_2 \alpha_2 = \alpha_1 + s\alpha_2 = const. = \alpha_1^{old} + s\alpha_2^{old} = \gamma$$

where $s = y_1 y_2$.
That is, we obtain the constraint

$$\alpha_1 + s\alpha_2 = \gamma \tag{18}$$

We solve (18) for α_1 and plug its value in $W(\alpha_1, \alpha_2)$ to obtain the new objective functiion

$$\begin{aligned} W(\alpha_2) &= \gamma - s\alpha_2 + \alpha_2 - \tfrac{1}{2} K_{11} (\gamma - s\alpha_2)^2 - \tfrac{1}{2} K_{22} \alpha_2^2 \\ &\quad - s K_{12} \gamma - s\alpha_2 \alpha_2 - y_1 (\gamma - s\alpha_2) v_1 - y_2 \alpha_2 v_2 + CONSTANT \end{aligned}$$

We take the derivative of W with resp[ect to α_2 and set it equal to 0:

$$\begin{aligned} W'(\alpha_2) &= 1 - s + s K_{11} (\gamma - s\alpha_2) - K_{22} \alpha_2 \\ &\quad + K_{12} \alpha_2 - s K_{12} (gamma - s\alpha_2) + y_2 v_1 - y_2 v_2 \\ &= 0 \end{aligned}$$

Solving for α_2 we obtain:

$$\begin{aligned}\alpha_2^{new,unclipped}(K_{11}+K_{22}-2K_{12}) &= 1-s+\gamma s(K_{11}-K_{12})+y_2(v_1-v2)\\ &= y_2\left[y_2-y_1+\gamma y_1(K_{11}-K_{12})+v_1-v_2\right]\end{aligned}$$

Using $\mathcal{K}$ we obtain

$$\alpha_2^{new,unclipped}\mathcal{K}y_2 = y_2\alpha_2\mathcal{K}+E_1-E_2$$

and thus

$$\alpha_2^{new,unclipped}=\alpha_2^{old}+\frac{y_2(E_1-E_2)}{\mathcal{K}}$$

Clip according to U and V if necessary (when $C<\infty$).