Sequential Minimal Optimization (SMO)

Yuh-Jye Lee

Data Science and Machine Intelligence Lab
National Chiao Tung University

May 2, 2017

Sequential Minimal optimization (SMO)

- The SMO algorithm was proposed by John C. Platt in 1998 and became the fastest quadratic programming optimization algorithm, especially for linear SVM and sparse data performance.
- One of the best reference about SMO is "Sequential Minimal Optimization A Fast Algorithm for Training Support Vector Machines" written by John C. Platt.

Sequential Minimal optimization (SMO)

- Sequential
  - Not parallel
  - Optimize in sets of 2 Lagrange multipliers
- Minimal
  - Optimize smallest possible sub-problem at each step
- Optimization
  - Satisfy the constraints for the chosen pair of Lagrange multipliers

Sequential Minimal optimization (SMO)

- The Sequential Minimal Optimization (SMO) algorithm is derived by taking the idea of the decomposition method to its extreme and optimizing a minimal subset of just two points at each iteration.
- The power of this technique resides in the fact that the optimization problem for two data points admits an analytical solution, eliminating the need to use an iterative quadratic programming optimizer as part of the algorithm.
- The requirement that the condition $\sum_{i=1}^{\ell} y_i\alpha_i = 0$ is enforced throughout the iterations implies that the smallest number of multipliers that can be optimized at each step is 2: whenever one multiplier is updated, at least one other multiplier needs to be adjusted in order to keep the condition true.

Sequential Minimal optimization (SMO)

- At each step SMO chooses two elements α_i and α_j to jointly optimize, find the optimal values for those two parameters given that all the others are fixed, and updates the α vector accordingly.
- The choice of the two points is determined by a heuristic, while the optimization of the two multipliers is performed analytically.
- Despite needing more iterations to converge, each iteration uses so few operations that the algorithm exhibits an overall speed-up of some orders of magnitude.

Sequential Minimal optimization (SMO)

- Besides convergence time, other important features of the algorithm are that it does *not* to store the kernel matrix in memory, since no matrix operations are involved, that it dose not use other packages, and that it is fairly easy to implement.
- Note that since standard SMO does not use a cached kernel matrix, its introduction could be used to obtain a further speed-up, at the expense of increased space complexity.

1-norm Soft Margin

- **Primal Form**

$$\min_{\mathbf{w},b,\xi_i} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{\ell}\xi_i \tag{1}$$

subject to

$$y_i\left(\mathbf{w}^\top \mathbf{x}_i + b\right) + \xi_i - 1 \geqslant 0 \tag{2}$$

$$\xi_i \geqslant 0, \quad i = 1, 2, \ldots, \ell \tag{3}$$

1-norm Soft Margin

- **Dual Form**

$$\max_{\alpha} W(\alpha) = \sum_{i=1}^{\ell} \alpha_i - \frac{1}{2} \sum_{i=1}^{\ell} \sum_{j=1}^{\ell} \alpha_i \alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j) \qquad (4)$$

subject to

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

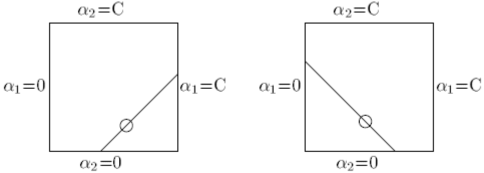
$$0 \leqslant \alpha_i \leqslant C \quad ; i = 1, 2, \ldots, \ell$$

where

$$K(\mathbf{x}_i, \mathbf{x}_j) \equiv \mathbf{x}_i^{\top} \mathbf{x}_j$$

- Without loss of generality we will assume that two elements, α_1, α_2 that have been chosen for updating to improve the objective value. In order to compute the new values for these two parameters, one can observe that in order not to violate the linear constraint $\sum_{i=1}^{\ell} \alpha_i y_i = 0$, The new values of the multipliers must be on a line,
- $$y_1\alpha_1^{(old)} + y_2\alpha_2^{(old)} = \text{constant} = y_1\alpha_1 + y_2\alpha_2 \qquad (5)$$

  in (α_1, α_2) space, and in the box defined by $0 \leqslant \alpha_1, \alpha_2 \leqslant C$ as shown in the following figure:



Sequential Minimal optimization (SMO)

- Without loss of generality, the algorithm first compute $\alpha_2^{(new)}$ and successively use it to obtain $\alpha_1^{(new)}$.
- The box constraint $0 \leqslant \alpha_1, \alpha_2 \leqslant C$ together with the linear equality constraint, provides a more restrictive constraint on the feasible values for $\alpha_2^{(new)}$:

$$U \leqslant \alpha_2^{(new)} \leqslant V, \tag{6}$$

where U and V are defined as follows:

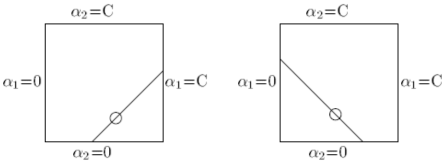
The Bound for α_2^{new}

- if $y_1 \neq y_2$

$$\begin{cases} U = \max\{0, \alpha_2^{(old)} - \alpha_1^{(old)}\}, \\ V = \min\{C, C - \alpha_1^{(old)} + \alpha_2^{(old)}\} \end{cases} \tag{7}$$

- if $y_1 = y_2$

$$\begin{cases} U = \max\{0, \alpha_1^{(old)} + \alpha_2^{(old)} - C\}, \\ V = \min\{C, \alpha_1^{(old)} + \alpha_2^{(old)}\} \end{cases} \tag{8}$$



$y_1 \neq y_2 \Rightarrow \alpha_1 - \alpha_2 = \gamma$ $y_1 = y_2 \Rightarrow \alpha_1 + \alpha_2 = \gamma$

Sequential Minimal optimization (SMO)

Theorem

The maximun of the objective function for the optimization problem

$$\max_{\alpha} W(\alpha) = \sum_{i=1}^{\ell} \alpha_i - \frac{1}{2} \sum_{i=1}^{\ell} \sum_{j=1}^{\ell} \alpha_i \alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j)$$

subject to

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

$$0 \leqslant \alpha_i \leqslant C \quad ; i = 1, 2, \ldots \ell$$

When only α_1 and α_2 are allowed to change, is achieved by first computing the quantity

$$\alpha_2^{(new,unc)} = \alpha_2^{(old)} + \frac{y_2\left\{E_2^{(old)} - E_1^{(old)}\right\}}{\kappa} \tag{9}$$

where

$$E_i \equiv f(\mathbf{x}_i) - y_i = \left(\sum_{j=1}^{\ell} \alpha_j y_j K_{ji} + b\right) - y_i \quad ; i = 1, 2, \tag{10}$$

and clipping it to enforce the constraint $U \leqslant \alpha_2^{(new)} \leqslant V$:

$$\alpha_2^{(new)} = \begin{cases} V, & \text{if } \alpha_2^{(new,unc)} > V \\ \alpha_2^{(new,unc)}, & \text{if } U \leqslant \alpha_2^{(new,unc)} \leqslant V \\ U, & \text{if } \alpha_2^{(new,unc)} < U \end{cases} \tag{11}$$

where U and V is defined by

- if $y_1 \neq y_2$

$$\begin{cases} U = \max\{0, \alpha_2^{(old)} - \alpha_1^{(old)}\}, \\ V = \min\{C, C - \alpha_1^{(old)} + \alpha_2^{(old)}\} \end{cases} \tag{12}$$

- if $y_1 = y_2$

$$\begin{cases} U = \max\{0, \alpha_1^{(old)} + \alpha_2^{(old)} - C\}, \\ V = \min\{C, \alpha_1^{(old)} + \alpha_2^{(old)}\} \end{cases} \tag{13}$$

and the value of $\alpha_1^{(new)}$ is obtained from $\alpha_2^{(new)}$ as

$$\alpha_1^{(new)} = \alpha_1^{(old)} + y_1 y_2 \left(\alpha_2^{(old)} - \alpha_2^{(new)}\right) \tag{14}$$

Sequential Minimal optimization (SMO)

Proof : For representation simplicity, let's define the following symbols for each element of matrix K

$$K(\mathbf{x}_i, \mathbf{x}_j) \equiv K_{ij} \tag{15}$$

$$f(\mathbf{x}_i) \equiv \sum_{j=1}^{\ell} \alpha_j y_j K_{ji} + b \tag{16}$$

$$v_i \equiv \sum_{j=3}^{\ell} \alpha_j y_j K_{ij} = f(\mathbf{x}_i) - \sum_{j=1}^{2} \alpha_j y_j K_{ij} - b \quad ; i = 1, 2, \tag{17}$$

$$E_i \equiv f(\mathbf{x}_i) - y_i = \left(\sum_{j=1}^{\ell} \alpha_j y_j K_{ji} + b \right) - y_i \quad ; i = 1, 2, \tag{18}$$

Sequential Minimal optimization (SMO)

Consider the objective as function of α_1 and α_2:

$$\begin{aligned}
W(\alpha_1, \alpha_2) = \alpha_1 + \alpha_2 - \frac{1}{2}K_{11}\alpha_1^2 - \frac{1}{2}K_{22}\alpha_2^2 - \alpha_1\alpha_2 y_1 y_2 K_{12} \\
- y_1\alpha_1 \sum_{j=3}^{\ell} y_j \alpha_j K_{1j} \\
- y_2\alpha_2 \sum_{j=3}^{\ell} y_j \alpha_j K_{2j} + \sum_{i=3}^{\ell} \alpha_i \\
- \frac{1}{2}\sum_{i=3}^{\ell}\sum_{j=3}^{\ell} \alpha_i \alpha_j y_i y_j K_{ij} \qquad (19)
\end{aligned}$$

Sequential Minimal optimization (SMO)

Substitute (16) (17) (18) and into (19) yields

$$W(\alpha_1,\alpha_2) = \alpha_1 + \alpha_2 - \frac{1}{2}K_{11}\alpha_1^2 - \frac{1}{2}K_{22}\alpha_2^2 - \alpha_1\alpha_2 y_1 y_2 K_{12} \\ - y_1\alpha_1 v_1 - y_2\alpha_2 v_2 + \text{constant} \tag{20}$$

Note also that the constraint $\sum_{i=1}^{\ell} \alpha_i^{(old)} y_i = \sum_{i=1}^{\ell} \alpha_i y_i = 0$, implies the condition

$$\alpha_1 + s\alpha_2 = \text{constant} = \alpha_1^{(old)} + s\alpha_2^{(old)} = \gamma \tag{21}$$

where $s = y_1 y_2$. The above equation demonstrates how $\alpha_1^{(new)}$ is computed from $\alpha_2^{(new)}$.

$$\alpha_1 = \gamma - s\alpha_2 \tag{22}$$

Sequential Minimal optimization (SMO)

Eliminating α_1 in (20), we have the objective function as α_2

$$\begin{aligned}
W(\alpha_2) &= \gamma - s\alpha_2 + \alpha_2 - \frac{1}{2}K_{11}(\gamma - s\alpha_2)^2 - \frac{1}{2}K_{22}\alpha_2^2 \\
&\quad - sK_{12}(\gamma - s\alpha_2)\alpha_2 - y_1(\gamma - s\alpha_2)v_1 - y_2\alpha_2 v_2 \\
&\qquad + \text{constant} \\
&= -\frac{1}{2}K_{11}\left(\gamma^2 - 2\gamma s\alpha_2 + \alpha_2^2\right) - \frac{1}{2}K_{22}\alpha_2^2 + s^2K_{12}\alpha_2^2 \\
&\quad + \left(1 - s - sK_{12}\gamma\right)\alpha_2 - y_1 v_1 y_2 v_2 + \text{constant} \\
&= \frac{1}{2}(2K_{12} - K_{11} - K_{22})\alpha_2^2 \\
&\quad + \left(1 - s + K_{11}s\gamma - K_{12}s\gamma + y_2 v_1 - y_2 v_2\right)\alpha_2 \\
&\quad + \text{constant}
\end{aligned} \tag{23}$$

Sequential Minimal optimization (SMO)

The stationary points satisfies

$$\begin{aligned}\frac{\mathrm{d}W(\alpha_2)}{\mathrm{d}\alpha_2} &= (2K_{12} - K_{11} - K_{22})\alpha_2 \\ &\quad + (1 - s + K_{11}s\gamma - K_{12}s\gamma + y_2 v_1 - y_2 v_2) \\ &= 0 \end{aligned} \tag{24}$$

This yields

$$\begin{aligned}\alpha_2^{(new,unc)}(K_{11} + K_{22} - 2K_{12}) &= 1 - s + K_{11}s\gamma - K_{12}s\gamma + y_2 v_1 - y_2 v_2 \\ &= 1 - s + (K_{11} - K_{12})s\gamma + y_2(v_1 - v_2) \end{aligned} \tag{25}$$

Sequential Minimal optimization (SMO)

multiplier (25) by y_2, it is easy to see

$$\begin{aligned}\alpha_2^{(new,unc)}\kappa y_2 &= y_2 - y_1 + (K_{11} - K_{12})y_1\gamma + v_1 - v_2\\ &= y_2 - y_1 + (K_{11} - K_{12})y_1\gamma + \left(f(\mathbf{x}_1) - \sum_{j=1}^{2} y_j\alpha_j K_{1j}\right)\\ &\quad - \left(f(\mathbf{x}_2) - \sum_{j=1}^{2} y_j\alpha_j K_{2j}\right) \qquad (26)\end{aligned}$$

and

$$y_1\gamma = y_1(\alpha_1 + s\alpha_2) = y_1\alpha_1 + y_2\alpha_2 \qquad (27)$$

Sequential Minimal optimization (SMO)

Since

$$\sum_{j=1}^{2} y_j\alpha_j K_{2j} - \sum_{j=1}^{2} y_j\alpha_j K_{1j} = y_1\alpha_1 K_{21} + y_2\alpha_2 K_{22} - y_1\alpha_1 K_{11} + y_2\alpha_2 K_{12} \tag{28}$$

Substitute(27) (28) into (26), we can find

$$\begin{aligned}
\alpha_2^{(new,unc)}\kappa y_2 &= y_2 - y_1 + (K_{11} - K_{12})(\alpha_1 y_1 + \alpha_2 y_2) + y_1\alpha_1 K_{21} \\
&\quad + y_2\alpha_2 K_{22} - y_1\alpha_1 K_{11} + y_2\alpha_2 K_{12} + f(\mathbf{x}_1) - f(\mathbf{x}_2) \\
&= y_2 - y_1 + f(\mathbf{x}_1) - f(\mathbf{x}_2) + y_2\alpha_2 K_{11} - y_2\alpha_2 K_{12} \\
&\quad + y_2\alpha_2 K_{22} - y_2\alpha_2 K_{12} \\
&= y_2\alpha_2\kappa + (f(\mathbf{x}_1) - y_1) - (f(\mathbf{x}_2) - y_2)
\end{aligned} \tag{29}$$

Sequential Minimal optimization (SMO)

So we have

$$\alpha_2^{(new)} = \alpha_2^{(old)} + \frac{y_2 \left\{ E_2^{(old)} - E_1^{(old)} \right\}}{\kappa}$$

where

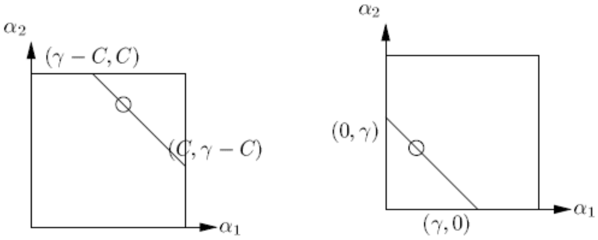
$$E_i \equiv f(\mathbf{x}_i) - y_i = \left(\sum_{j=1}^{\ell} \alpha_j y_j K_{ji} + b \right) - y_i \quad ; i = 1, 2,$$

Finally, we must clip $\alpha_2^{(new,unc)}$ if necessary to ensure it remains in the interval $[U, V]$.

Sequential Minimal optimization (SMO)

Discussion:

- If $s = y_1 y_2 = 1$ $(y_1 = y_2)$, then $\alpha_1 + \alpha_2 = \gamma$.
  1. If $\gamma > C$, then $\max \alpha_2 = C = \min V$ and $\min \alpha_2 = \gamma - C = \alpha_1 + \alpha_2 - C = \max U$ .See Figure 2.2
  2. If $\gamma < C$, then $\max \alpha_2 = \gamma = \alpha_1 + \alpha_2 = \min V$ and $\min \alpha_2 = 0 = \max U$ .See Figure 2.3



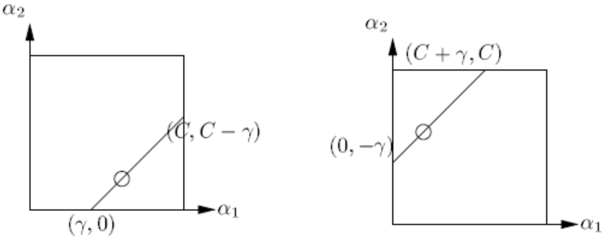
(a) Fig.2.2 $\gamma > C$

(b) Fig.2.3 $\gamma < C$

Figure: If $s = y_1 y_2 = 1$, then $\alpha_1 + \alpha_2 = \gamma$

Sequential Minimal optimization (SMO)

- If $s = y_1 y_2 = -1$ $(y_1 \neq y_2)$, then $\alpha_1 - \alpha_2 = \gamma$
  1. If $\gamma > 0$, then $\max \alpha_2 = C - \gamma = C - \alpha_1 + \alpha_2 = \min V$ and $\min \alpha_2 = 0 = \max U$. See Figure 2.4.
  2. If $\gamma < 0$, then $\max \alpha_2 = C = \min V$ and $\min \alpha_2 = -\gamma = -\alpha_1 + \alpha_2 = \max U$. See Figure 2.5.



(a) Fig.2.4 $\gamma > C$

(b) Fig.2.5 $\gamma < C$

Figure: If $s = y_1 y_2 = -1$, then $\alpha_1 - \alpha_2 = \gamma$

Sequential Minimal optimization (SMO)

From above discussion, we can find α_2 must lie in the following range to make sure it is clipped:

$$\max \alpha_2 = \min V \tag{30}$$

$$\min \alpha_2 = \max U \tag{31}$$

where U and V is given by

- if $y_1 \neq y_2$

$$\begin{cases} U = \max\{0, \alpha_2^{(old)} - \alpha_1^{(old)}\}, \\ V = \min\{C, C - \alpha_1^{(old)} + \alpha_2^{(old)}\} \end{cases} \tag{32}$$

- if $y_1 = y_2$

$$\begin{cases} U = \max\{0, \alpha_1^{(old)} + \alpha_2^{(old)} - C\}, \\ V = \min\{C, \alpha_1^{(old)} + \alpha_2^{(old)}\} \end{cases} \tag{33}$$

Sequential Minimal optimization (SMO)

and the value of $\alpha_1^{(new)}$ is obtained from $\alpha_2^{(new)}$ as

$$\alpha_1^{(new)} = \alpha_1^{(old)} + y_1 y_2 \left(\alpha_2^{(old)} - \alpha_2^{(new)}\right) \tag{34}$$

Clipping it to enforce the constraint $U \leqslant \alpha_2^{(new,clipped)} \leqslant V$:

$$\alpha_2^{(new,clipped)} = \begin{cases} V, & \text{if } \alpha_2^{(new,unc)} > V \\ \alpha_2^{(new,unc)}, & \text{if } U \leqslant \alpha_2^{(new,unc)} \leqslant V \\ U, & \text{if } \alpha_2^{(new,unc)} < U \end{cases} \tag{35}$$

□

Sequential Minimal optimization (SMO)

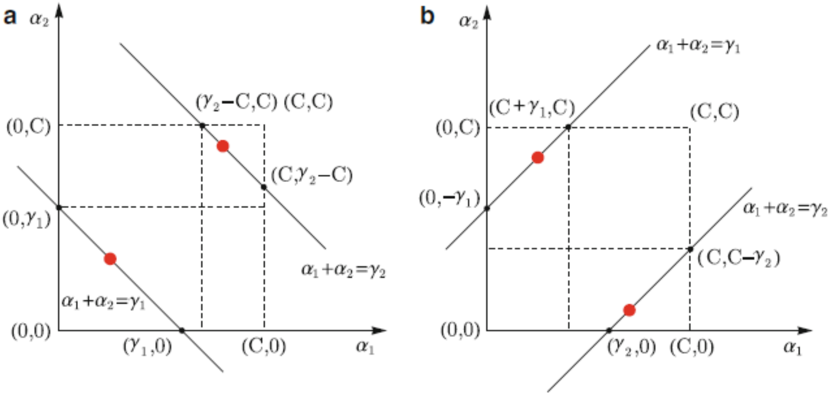


Figure: Two possible situations for the update of α_1 and α_2 in SMO. (a) $s = y_1 y_2 = 1$ and (b) $s = y_1 y_2 = -1$

Sequential Minimal optimization (SMO)

Remark

$$E_i \equiv f(\mathbf{x}_i) - y_i = \left(\sum_{j=1}^{\ell} \alpha_i y_i K_{ij} + b \right) - y_i \quad ; i = 1, 2, \qquad (36)$$

where $f(\mathbf{x})$ denote the current hypothesis determined by the value of α and b at a particular stage of learning, and E_i is the difference between function output and target classification on the training point $\mathbf{x}_1$ or $\mathbf{x}_2$. Note E_i can be large if a point is correctly classified.

For example if $y_1 = 1$,and the function output is $f(\mathbf{x}_1) = 5$, the classification is correct, but $E_1 = 4$.

Sequential Minimal optimization (SMO)

Remark

$$\frac{\mathrm{d}^2 W(\alpha_2)}{\mathrm{d}\alpha_2^2} = -K_{11} - K_{22} + 2K_{12} \equiv \kappa \leqslant 0$$

Proof.

$$\begin{aligned} \kappa \equiv -K_{11} - K_{22} + 2K_{12} &= -\mathbf{x}_1^\top \mathbf{x}_1 - \mathbf{x}_2^\top \mathbf{x}_2 - 2\mathbf{x}_1^\top \mathbf{x}_2 \qquad (37) \\ &= -(\mathbf{x}_2 - \mathbf{x}_1)^\top (\mathbf{x}_2 - \mathbf{x}_1) \\ &= -\left\|\mathbf{x}_2 - \mathbf{x}_1\right\|^2 \leqslant 0 \end{aligned}$$

□

Nello Cristianini (Author), John Shawe-Taylor (Author)
An Introduction to Support Vector Machines and Other Kernel-based Learning Methods.
Cambridge University Press, 2000