SVM Tutorial[1]

Yang Li

School of Computer Science and Technology
University of Science and Technology of China
Hefei China

May 15, 2018

[1] Updated on 15th May, 2018

Outline

1. Introduction
2. Linearly Separable
   - Hard Margin
   - Soft Margin
3. Lagrange Reformulation
4. Nonlinear SVM
5. SMO

1 Introduction

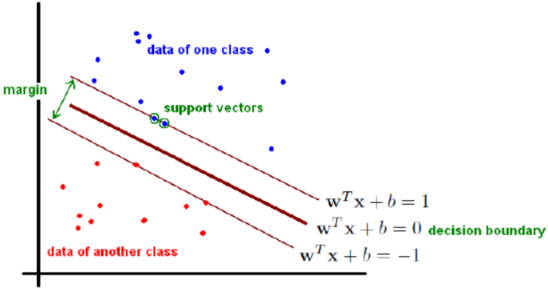
2 Linearly Separable

3 Lagrange Reformulation

4 Nonlinear SVM

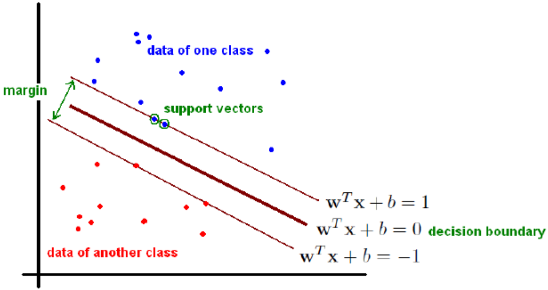
5 SMO

Motivation


data of one class
margin
support vectors
$\mathbf{w}^T\mathbf{x} + b = 1$
$\mathbf{w}^T\mathbf{x} + b = 0$
decision boundary
$\mathbf{w}^T\mathbf{x} + b = -1$
data of another class

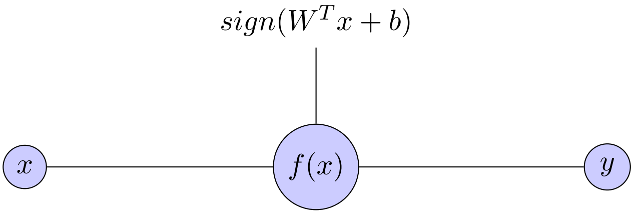
- Say the course instructors have observed that students get the most out of it if they are good at Math or Stats. Over time, they have recorded the scores of the enrolled students in these subjects. Also, for each of these students, they have a label depicting their performance in the ML course: Good or Bad.

Motivation


data of one class
margin
support vectors
$\mathbf{w}^T\mathbf{x} + b = 1$
$\mathbf{w}^T\mathbf{x} + b = 0$
decision boundary
$\mathbf{w}^T\mathbf{x} + b = -1$
data of another class

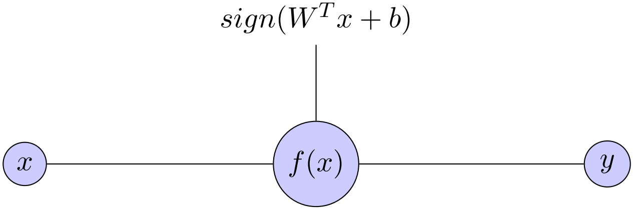
- The intuition is quite simple. We create a line for representations that capture their features of interest.
- Divide the points into two classes according to the sign of this linear function of the form $W^T x + b$.

Motivation



- We are dealing with a two-class, binary classification problem.
- A novel definition of margin.

Motivation



- We are dealing with a two-class, binary classification problem. (without saying)
- A novel definition of margin. (Why???)

1 Introduction

2 Linearly Separable
- Hard Margin
- Soft Margin

3 Lagrange Reformulation

4 Nonlinear SVM

5 SMO

1 Introduction

2 Linearly Separable
- Hard Margin
- Soft Margin

3 Lagrange Reformulation

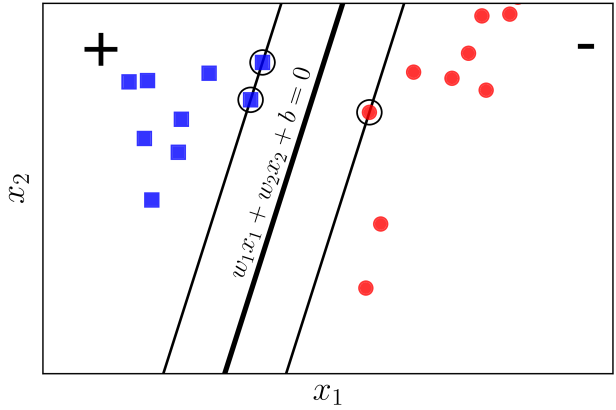
4 Nonlinear SVM

5 SMO

Basics

- Trying to construct a linear separating line for two class with as large margin as possible.
- Notations
    1. $x \in \Re^d =$ "real-valued feature vectors"
    2. $y \in \{-1, 1\} =$ "labels, classes, responses, predictions".
    3. $W \in \Re^d =$ "parameters of interest". A linear model is fully specified with it.
    4. $y^+ =$ " the set of x with positive labels".
    5. $y^- =$ " the set of x with negative labels".

Hard Margin



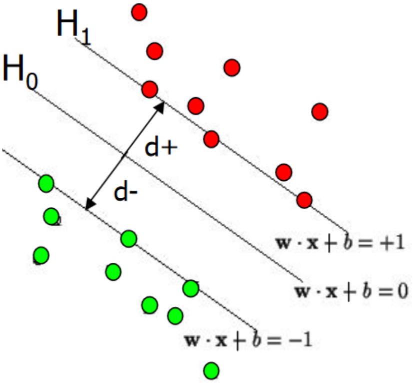
Tricks

- we already know:
  - $f(x) = sign(W^t x + b) = +1, \forall x \in y^+$
  - $f(x) = sign(W^t x + b) = -1, \forall x \in y^-$
- In summary, $y(W^T x + b) \geq 1$

Tricks

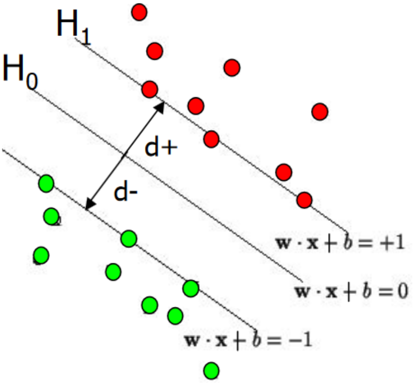
- we already know:
  - $f(x) = sign(W^t x + b) = +1, \forall x \in y^+$
  - $f(x) = sign(W^t x + b) = -1, \forall x \in y^-$
- In summary, $y(W^T x + b) \geq 1$
- Let's rescale the data such that anything on or above the boundary $w^T x + b = 1$ is of one class (with label 1), and anything on or below the boundary $w^T x + b = -1$ is of the other class (with label -1).

Margin Distance (aka. street width)


H_1
H_0
d+
d-
$\mathbf{w} \cdot \mathbf{x} + b = +1$
$\mathbf{w} \cdot \mathbf{x} + b = 0$
$\mathbf{w} \cdot \mathbf{x} + b = -1$

- Recall that a point (x_0, y_0) to a line $Ax + By + b = 0$ is $|Ax_0 + By_0 + b|/\sqrt{A^2 + B^2}$.

Margin Distance (aka. street width)


H1
H0
d+
d-
$\mathbf{w}\cdot\mathbf{x}+b=+1$
$\mathbf{w}\cdot\mathbf{x}+b=0$
$\mathbf{w}\cdot\mathbf{x}+b=-1$

- Recall that a point (x_0, y_0) to a line $Ax + By + b = 0$ is $|Ax_0 + By_0 + b|/\sqrt{A^2 + B^2}$.
- In our case, $|W^T x + b|/||W|| = 1/||W||$
- The total distance between two lines $\frac{2}{||W||}$.

Formal Objective

If we take $||W|| = \sqrt{W^T W}$, then we have

$$obj = \max_{W,b} \frac{2}{\sqrt{W^T W}}$$

s.t.

$$y_i(W^T x_i + b) \geq 1 \qquad \forall x_i$$

Formal Objective

If we take $||W|| = \sqrt{W^T W}$, then we have

$$obj = \max_{W,b} \frac{2}{\sqrt{W^T W}}$$

s.t.

$$y_i(W^T x_i + b) \geq 1 \qquad \forall x_i$$

- liar!!! not identical to the one in textbook.

Formal Objective

- After Changing a little bit. This quadratic programming problem is expressed as:

$$\min_{W,b} \frac{W^TW}{2}$$

s.t.

$$y_i(W^Tx_i + b) \geq 1 \qquad \forall x_i$$

1 Introduction

2 Linearly Separable
- Hard Margin
- Soft Margin

3 Lagrange Reformulation

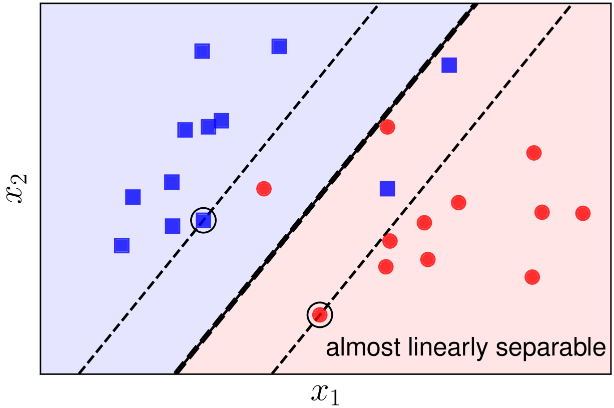
4 Nonlinear SVM

5 SMO

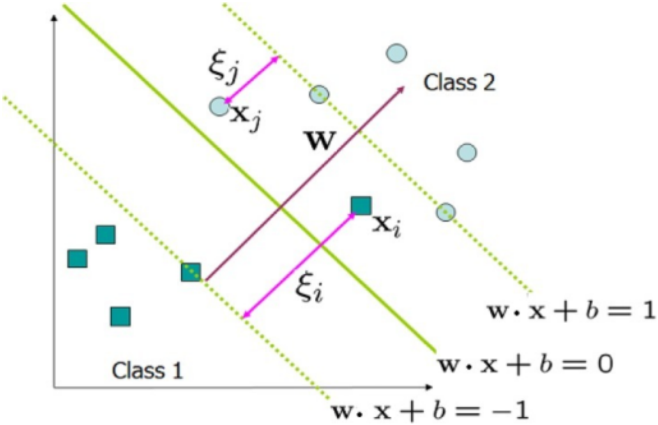
Soft Margin

- Consider the case that your data isn't perfectly linearly separable. Some cases are mingled together.
- Maybe you aren't guaranteed that all your data points are correctly labeled, so you want to allow some data points of one class to appear on the other side of the boundary.

Soft Margin



Soft Margin



ξ_j
$\mathbf{x}_j$
Class 2
$\mathbf{w}$
$\mathbf{x}_i$
ξ_i
$\mathbf{w} \cdot \mathbf{x} + b = 1$
$\mathbf{w} \cdot \mathbf{x} + b = 0$
Class 1
$\mathbf{w} \cdot \mathbf{x} + b = -1$

Formal Objective

- This quadratic programming problem is expressed as:

$$\min_{W,b} \frac{\sqrt{W^T W}}{2} + C \sum_i \epsilon_i$$

s.t.

$$y_i(W^T x_i + b) \geq 1 - \epsilon_i \qquad \forall x_i$$

$$\epsilon_i \geq 0 \qquad \forall i$$

1 Introduction

2 Linearly Separable

3 Lagrange Reformulation

4 Nonlinear SVM

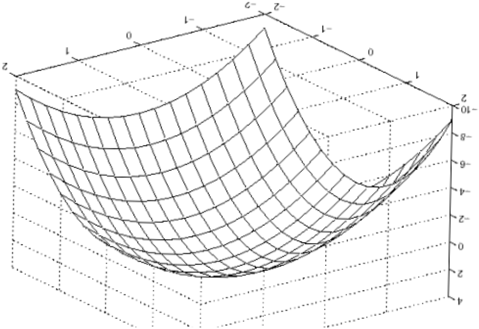
5 SMO

Convexity Check

- This is a constrained optimization problem.
  - Domain: $W \in \Re^d$
  - Convexity: Quadratic, the surface is a paraboloid, with just a single global minimum.

Convexity Check

- This is a constrained optimization problem.
  - Domain: $W \in \Re^d$
  - Convexity: Quadratic, the surface is a paraboloid, with just a single global minimum.
- Example: $2 + x^2 + 2y^2$



Basics for Convex Optimization

- Remove the constraint:
  - 1. Define auxiliary functions for constraints.
  - 2. Move constraints to objectives.
- Solve the updated objective, hopefully in a unconstrained space.

Basics for Convex Optimization

- Remove the constraint:
  - 1. Define auxiliary functions for constraints.
  - 2. Move constraints to objectives.
- Solve the updated objective, hopefully in a unconstrained space.

- Can we solve it definitely in the new space?
- Mostly Yes, but not definitely.

Auxiliary Functions for Constraints

- Two constraints:

$$
\begin{aligned}
&y_i(W^T x_i + b) \geq 1 - \epsilon_i \qquad \forall x_i \\
&\epsilon_i \geq 0 \qquad \forall i
\end{aligned}
$$

- Therefore,

$$\max_{\alpha_i \geq 0} \alpha_i (1 - y_i(W^T x_i + b))$$

Lagrange Form

- We find the Lagrange form for the above objective

$$\min_{W,b}[\frac{W^TW}{2} + \sum_i \max_{\alpha_i \geq 0} \alpha_i[1 - y_i(W^Tx_i + b)]]$$

- Or by linearity of max operator

$$J_0 = \sum_i \min_{W,b} \max_{\alpha_i \geq 0}[\frac{W^TW}{2} + \alpha_i[1 - y_i(W^Tx_i + b)]]$$

Lagrange Reformulation

- We change the oder of max and min at a loss

$$J_1 = \sum_i \max_{\alpha_i \geq 0} \min_{W,b} [\frac{W^T W}{2} + \alpha_i[1 - y_i(W^T x_i + b)]]$$

Lagrange Reformulation

- We change the oder of max and min at a loss

$$J_1 = \sum_i \max_{\alpha_i \geq 0} \min_{W,b} [\frac{W^T W}{2} + \alpha_i [1 - y_i (W^T x_i + b)]]$$

- sad effects $J_1 \leq J_0$
- equality holds with:

$$\alpha_i [1 - y_i (W^T x_i + b)] = 0, \qquad \forall i$$

Lagrange Reformulation

- We change the oder of max and min at a loss

$$J_1 = \sum_i \max_{\alpha_i \geq 0} \min_{W,b} [\frac{W^T W}{2} + \alpha_i [1 - y_i (W^T x_i + b)]]$$

- sad effects $J_1 \leq J_0$
- equality holds with:

$$\alpha_i [1 - y_i (W^T x_i + b)] = 0, \qquad \forall i$$

KKT condition

Solve J_1

- Solve $\frac{\partial J_1}{\partial w} = 0$.

$$W = \sum_i \alpha_i y_i x_i$$

- Solve $\frac{\partial J_1}{b}$

$$\sum_i \alpha_i y_i = 0$$

- After substituting and simplifying

$$\max_{\alpha_i \geq 0} [\sum_i \alpha_i - \frac{1}{2} \alpha_i \alpha_j y_i y_j x_i^T x_j]$$

$$s.t. \sum_i \alpha_i y_i = 0$$

Lagrange Form cont.

- In the case with slack variables, things become subtle

$$J_2 = \sum_i \max_{\alpha_i \geq 0} \min_{W,b} [\frac{W^T W}{2} + \alpha_i [1 - y_i(W^T x_i + b)] + C\epsilon_i]$$

$$s.t.$$

$$y_i(W^t x_i + b) \geq 1 - \epsilon_i$$

$$\epsilon_i \geq 0$$

- subtitle it into the above equation

$$J_2 = \sum_i \max_{\alpha_i \geq 0} \min_{W,b} [\frac{W^T W}{2} + \alpha_i (1 - y_i(W^T x_i + b) - \epsilon_i) - \tau_i \epsilon_i + C\epsilon_i]$$

Cont.

- The only difference:

$$\frac{\partial J_1}{\partial \epsilon_i} = -\alpha_i + C - \tau_i = 0$$

- $\tau_i = C - \alpha_i > 0 \Longrightarrow 0 \leq \alpha_i \leq C$

Cont.

- The only difference:

$$\frac{\partial J_1}{\partial \epsilon_i} = -\alpha_i + C - \tau_i = 0$$

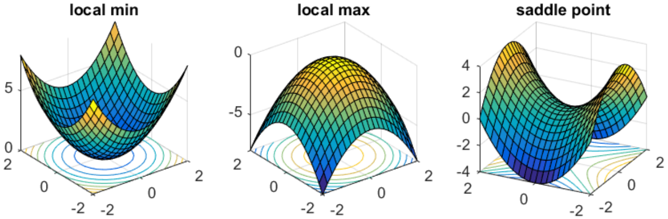
- $\tau_i = C - \alpha_i > 0 \Longrightarrow 0 \leq \alpha_i \leq C$
- In summary:

$$\max_{\alpha_i \geq 0} [\sum_i \alpha_i - \frac{1}{2} \alpha_i \alpha_j y_i y_j x_i^T x_j]$$

$$s.t. \sum_i \alpha_i y_i = 0$$

$$\alpha_i < C$$

Saddle Point



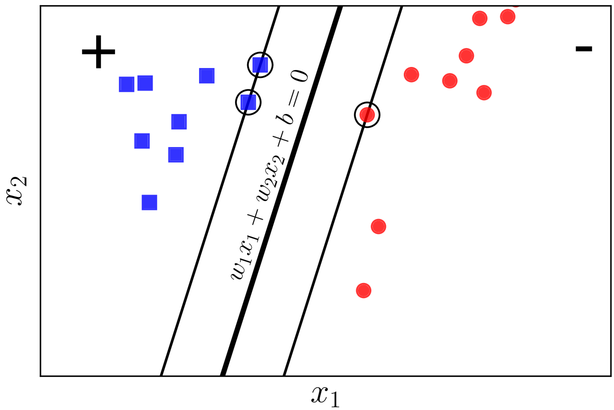
Sparsity

- Revisit the most important condition:

$$\alpha_i[1 - y_i(W^T x_i + b)] = 0, \qquad \forall i$$

- Discuss different situations:
  - $y_i(W^T x_i + b) > 1 \Longleftrightarrow \alpha_i = 0$
  - $y_i(W^T x_i + b) = 1 \Longleftrightarrow \alpha_i > 0$
- $W = \sum_{i \in sup} \alpha_i y_i x_i$
- Sparsity is nothing but the KKT condition.

Revisit the condition



1 Introduction

2 Linearly Separable

3 Lagrange Reformulation

4 Nonlinear SVM

5 SMO

Kernalization

- The above analysis holds with linear cases!!!
- Extending to nonlinear cases is nontrivial.

Kernalization

- The above analysis holds with linear cases!!!
- Extending to nonlinear cases is nontrivial.

- Kernalization:
  - Extend linear analysis to nonlinear cases
  - Theoretically and practically sound: No one knows what really happened in the target space, if it did not perform well, no one knows how to improve it.
  - Easy to use

Kernalization

- Let's suppose we magically solve the optimization problem and get $\alpha_i, \forall i$:

$$W = \sum_i \alpha_i y_i x_i$$

- Suppose we transform x_i into a new space where they are linearly separable

$$W = \sum_i \alpha_i y_i \phi(x_i)$$

- The classification boundary becomes

$$y = \sum_i \alpha_i y_i \phi(x_i) \phi(x) + b$$

- Define $K(\cdot, \cdot) = \langle \phi(\cdot), \phi(\cdot) \rangle$

$$y = \sum_i \alpha_i y_i K(x_i, x) + b$$

Pros and Cons of Kernalization

- Advantages
  1. Preserve correspondence between mapped points and original ones.
  2. Be able to deal with nonlinear ones.
  3. Not increase computation burden significantly.
- Disadvantages:
  1. require inner product form
  2. No one knows what happens in the target space theoretically and practically.

1. Introduction
2. Linearly Separable
3. Lagrange Reformulation
4. Nonlinear SVM
5. SMO

Solving

- Revisiting the objective function:

$$\underbrace{min}_{\alpha} \frac{1}{2} \sum_{i=1,j=1}^{m} \alpha_i \alpha_j y_i y_j K(x_i, x_j) - \sum_{i=1}^{m} \alpha_i$$

$$s.t. \sum_{i=1}^{m} \alpha_i y_i = 0$$

$$0 \leq \alpha_i \leq C$$

- KKT condition:

$$\alpha_i^* (y_i (w^T x_i + b) - 1 + \xi_i^*) = 0$$

   - $\alpha_i^* = 0 \Rightarrow y_i(w^* \bullet \phi(x_i) + b) \geq 1$
   - $0 < \alpha_i^* < C \Rightarrow y_i(w^* \bullet \phi(x_i) + b) = 1$
   - $\alpha_i^* = C \Rightarrow y_i(w^* \bullet \phi(x_i) + b) \leq 1$

SMO

- Only care about α_1 and α_2
-

$$\underbrace{min}_{\alpha_1,\alpha_2} \frac{1}{2}K_{11}\alpha_1^2 + \frac{1}{2}K_{22}\alpha_2^2 + y_1y_2K_{12}\alpha_1\alpha_2 - (\alpha_1 + \alpha_2)$$

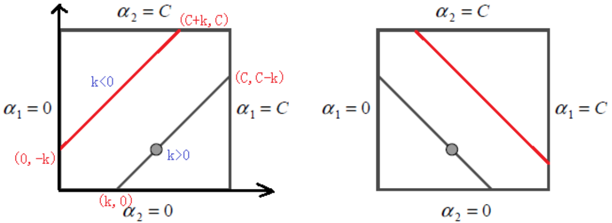
$$+y_1\alpha_1\sum_{i=3}^{m} y_i\alpha_iK_{i1} + y_2\alpha_2\sum_{i=3}^{m} y_i\alpha_iK_{i2}$$

$$s.t.\ \ \alpha_1y_1 + \alpha_2y_2 = -\sum_{i=3}^{m} y_i\alpha_i = \varsigma$$

$$0 \leq \alpha_i \leq C \ \ i = 1, 2$$

Two-variable Optimization

- Following $\alpha_1 y_1 + \alpha_2 y_2 = k$ and $\alpha_i \in [0, C]$.
- $y_i \in \{1, -1\}$



$y_1 \neq y_2 \Rightarrow \alpha_1 - \alpha_2 = k$

$y_1 = y_2 \Rightarrow \alpha_1 + \alpha_2 = k$

- Omit the condition $\alpha_i \in [0, C]$.
- Define some variables:
  - $g(x) = w^* \bullet \phi(x) + b^\star = \sum_{j=1}^{m} \alpha_j^* y_j K(x, x_j) + b^*$
  - $v_i = \sum_{i=3}^{m} y_j \alpha_j K(x_i, x_j) = g(x_i) - \sum_{i=1}^{2} y_j \alpha_j K(x_i, x_j) - b$
- The simplified objective

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

SMO

- substitute $\alpha_1 = y_1(\varsigma - \alpha_2 y_2)$, we obtain

$$
\begin{aligned}
J(\alpha_2) =& \frac{1}{2}K_{11}(\varsigma - \alpha_2 y_2)^2 + \frac{1}{2}K_{22}\alpha_2^2 + y_2 K_{12}(\varsigma - \alpha_2 y_2)\alpha_2 \\
& - (y_1(\varsigma - \alpha_2 y_2) + \alpha_2) + (\varsigma - \alpha_2 y_2)v_1 + y_2\alpha_2 v_2
\end{aligned}
$$

- Solve $\frac{\partial J}{\partial \alpha_2}$

$$
\begin{aligned}
\frac{\partial J}{\partial \alpha_2} =& K_{11}\alpha_2 + K_{22}\alpha_2 - 2K_{12}\alpha_2 - \\
& K_{11}\varsigma y_2 + K_{12}\varsigma y_2 + y_1 y_2 - 1 - v_1 y_2 + y_2 v_2 = 0
\end{aligned}
$$

- Arrange and we get

$$
\begin{aligned}
&(K_{11}+K_{22}-2K_{12})\alpha_2 \\
=&y_2(y_2-y_1+\varsigma K_{11}-\varsigma K_{12}+v_1-v_2) \\
=&y_2(y_2-y_1+\varsigma K_{11}-\varsigma K_{12}+(g(x_1)-\sum_{j=1}^{2}y_j\alpha_j K_{1j}-b) \\
&-(g(x_2)-\sum_{j=1}^{2}y_j\alpha_j K_{2j}-b))
\end{aligned}
$$

$$\Longrightarrow$$

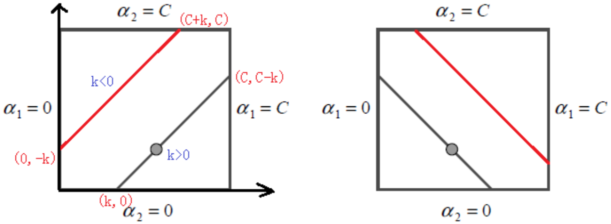
$$
\begin{aligned}
&(K_{11}+K_{22}-2K_{12})\alpha_2^{new} \\
=&y_2((K_{11}+K_{22}-2K_{12})\alpha_2^{old}y_2+y_2-y_1+g(x_1)-g(x_2)) \\
=&(K_{11}+K_{22}-2K_{12})\alpha_2^{old}+y2(E_1-E_2)
\end{aligned}
$$

$$\Longrightarrow$$

$$\alpha_2^{new}=\alpha_2^{old}+\frac{y_2(E_1-E_2)}{K_{11}+K_{22}-2K_{12})}$$

SMO

- Following $\alpha_1 y_1 + \alpha_2 y_2 = k$ and $\alpha_i \in [0, C]$.
- $y_i \in \{1, -1\}$



$y_1 \neq y_2 \Rightarrow \alpha_1 - \alpha_2 = k$ $\qquad$ $y_1 = y_2 \Rightarrow \alpha_1 + \alpha_2 = k$

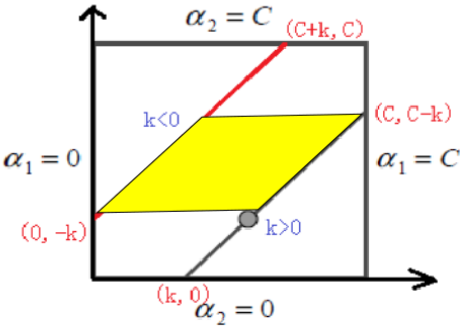
- left:

$$L = \max(0, \alpha_2^{old} - \alpha_1^{old})$$
$$H = \min(C, C + \alpha_2^{old} - \alpha_1^{old})$$

- right:

$$L = \max(0, \alpha_2^{old} + \alpha_1^{old} - C)$$
$$H = \min(C, \alpha_2^{old} + \alpha_1^{old})$$

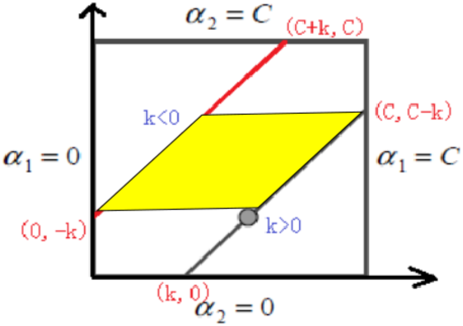
SMO



- $k = \alpha_2^{old} - \alpha_1^{old}$
- $\alpha_1 \in [-k, C - k]$
- $\alpha_2 \in [k, C + k]$

$$L = \max(0, \alpha_2^{old} - \alpha_1^{old})$$
$$H = \min(C, C + \alpha_2^{old} - \alpha_1^{old})$$

SMO



Projection operator:

$$\alpha_2^{new} = \begin{cases} H & \alpha_2^{new} > H \\ \alpha_2^{new} & L \leq \alpha_2^{new} \leq H \\ L & \alpha_2^{new} < L \end{cases}$$

SMO

Update b

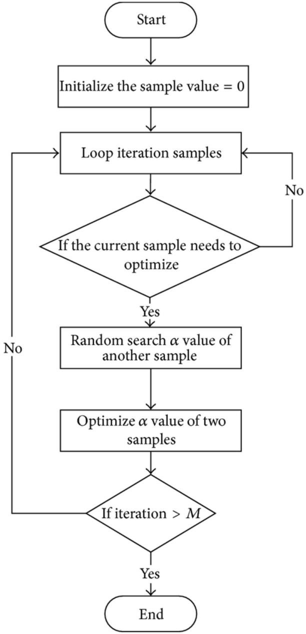
- $0 < \alpha_i^* < C \Longrightarrow y_i(w^* \bullet \phi(x_i) + b^*) = 1$
- transform a little

$$w^* \bullet \phi(x_i) + b^* = y_i$$

$$b^* = y_i - \sum_j \alpha_j^* y_j K(x_j, x_i)$$

- we get an update equation

$$b^{new} = y_1 - \sum_{i=3}^{m} \alpha_i y_i K_{i1} - \alpha_1^{new} y_1 K_{11} - \alpha_2^{new} y_2 K_{21}$$



Start

Initialize the sample value = 0

Loop iteration samples

If the current sample needs to optimize

No

Yes

Random search α value of another sample

Optimize α value of two samples

If iteration > M

No

Yes

End

Conclusion

- The idea of large margin
- Slack Variable
- Dual Problem and its solver
- Kernelization
- SMO

Questions?