Theoretical principles of Support Vector Machines

Aim: We try to understand the kernel trick from a mathematical perspective.

book: Support Vector Machines for Pattern Classification (Chapter 2); Abe, S. (2010)

1. Set the scene

Let us start with understanding duality. Later on we want to bring the constrained problem of finding the optimal separating hyperplane into its dual form.

Suppose we want to minimize

$$f(x,y) = x^2 + 2y$$

subject to an equality constraint

$$3x + 2y + 1 = 0.$$

We define a new function called the Lagrangian:

$$g(x,y,\lambda) = f(x,y) - \underset{\text{Lagrangian multiplier}}{\lambda} \cdot [3x+2y+1]$$

$$= [x^2+2y] - \lambda \cdot [3x+2y+1]$$

Let us compute the partial derivatives:

$$\frac{\partial}{\partial x} g(x,y,\lambda) = 2x - 3\lambda \overset{!}{=} 0$$

$$\frac{\partial}{\partial y} g(x,y,\lambda) = 2 - 2\lambda \overset{!}{=} 0$$

$$\frac{\partial}{\partial \lambda} g(x,y,\lambda) = -[3x+2y+1] \overset{!}{=} 0$$

$$\Rightarrow -3x = 1+2y$$

$$\Rightarrow 2x - 3\lambda = 2 - 2\lambda$$

$$\Rightarrow \lambda = 2[x-1]$$

$$\Rightarrow \lambda =$$

$$x = -\frac{1+2y}{3}$$

$$\Rightarrow \lambda = 2\left[-\frac{1}{3} - \frac{2}{3}y - 1\right]$$

$$\lambda = -\frac{8}{3} - \frac{4}{3}y$$

$$\Rightarrow 2 - 2\left[-\frac{8}{3} - \frac{4}{3}y\right] = 0$$

$$\Rightarrow \frac{6}{3} + \frac{16}{3} + \frac{8}{3}y = 0$$

$$\Rightarrow -\frac{22}{8} = y \Rightarrow \boxed{y = -\frac{11}{4}}$$

$$\Rightarrow x = -\frac{1}{3} - \frac{2}{3}\left[-\frac{11}{4}\right]$$

$$= -\frac{1}{3} + \frac{22}{12} = -\frac{4}{12} + \frac{22}{12} = \frac{\cancel{18}^{\,3}}{\cancel{12}_{\,2}} =$$

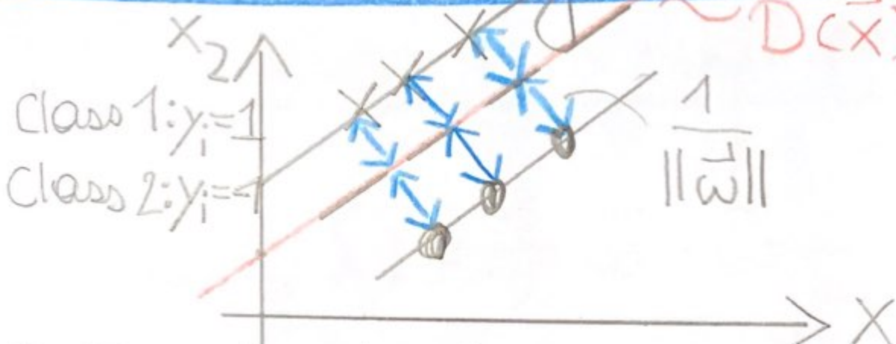
$$\Rightarrow \boxed{x = \frac{3}{2}} \quad \Rightarrow \lambda = 2\left[\frac{3}{2} - \frac{3}{3}\right] =$$

$$= 2\left[\frac{9}{6} - \frac{6}{6}\right] = \boxed{1} \quad 3$$

This method can be generalized to inequality constraints (Karush-Kuhn-Tucker (KKT) conditions)

2. Hard-Margin SVM

$D(\vec{x}) = \vec{\omega}^T \cdot \vec{x} + \vec{b}$



Decision function:

$$D(\vec{x}) = \vec{\omega}^T \cdot \vec{x} + \underbrace{b}_{\text{bias term}} \quad ; \ \underline{M} \text{ training data}$$

If linearly separable:

$$\vec{\omega}^T \vec{x}_i + b \begin{cases} \geq 1 & \text{for } y_i = 1 \\ \leq -1 & \text{for } y_i = -1 \end{cases}$$

$$\Leftrightarrow \boxed{y_i \left[\vec{\omega}^T \cdot \vec{x}_i + b\right] \geq 1}, \quad i = 1 \ldots M$$

training sample $\vec{x}$, distance to hyperplane

$$\frac{|D(\vec{x})|}{\|\vec{\omega}\|} \Rightarrow \text{optimal separating hyperplane, which minimizes } \|\omega\|$$

$$\boxed{\begin{aligned}&\text{minimize } Q(\vec{\omega}, b) = \frac{1}{2}\|\vec{\omega}\|^2 \\ &\text{subject to } y_i[\vec{\omega}^T \cdot \vec{x}_i + b] \geq 1\end{aligned}}$$

optimal separating hyperplane

has a unique solution!

2.1 Convert into dual problem

$$Q(\vec{\omega}, b, \vec{\lambda}) = \frac{1}{2}\vec{\omega}^T \cdot \vec{\omega} - \sum_{i=1}^{M} \lambda_i \{ y_i(\vec{\omega}^T \cdot \vec{x}_i + b) - 1 \} \quad (1)$$

$\vec{\lambda} = (\lambda_1, \dots, \lambda_M)^T$ non-negative Lagrangian multiplier

Karush-Kuhn-Tucker (KKT) conditions:

$$\frac{\partial Q}{\partial \vec{\omega}} = 0 \quad (2)$$

$$\frac{\partial Q}{\partial b} = 0 \quad (3)$$

$$\lambda_i \{ y_i(\vec{\omega}^T \cdot \vec{x}_i + b) - 1 \} = 0 \quad (4)$$

$$\lambda_i \geq 0 \quad (5)$$

$$\lambda_i \neq 0 \Rightarrow y_i(\vec{\omega}^T \cdot \vec{x}_i + b) = 1$$

$\Rightarrow \vec{x}_i$ with $\lambda_i \neq 0$ the support vectoo

from (1) and (3):

$$\sum_{i=1}^{M} \lambda_i y_i = 0$$

from (1) and (2): $\left(\frac{1}{2}\vec{\omega} + \frac{1}{2}\vec{\omega} - \sum \lambda_i y_i x_i\right)$

$$\vec{\omega} = \sum_{i=1}^{M} \lambda_i y_i \vec{x}_i$$

Dual problem:

$$\boxed{Q(\lambda) = \sum_{i=1}^{M} \lambda_i - \frac{1}{2} \sum_{i,j=1}^{M} \lambda_i y_i \lambda_j y_j \vec{x}_i^T \vec{x}_j \quad \text{with} \sum_{i=1}^{M} \lambda_i y_i = 0 \; ; \; \lambda_i \geq 0, \; i = 1..M}$$

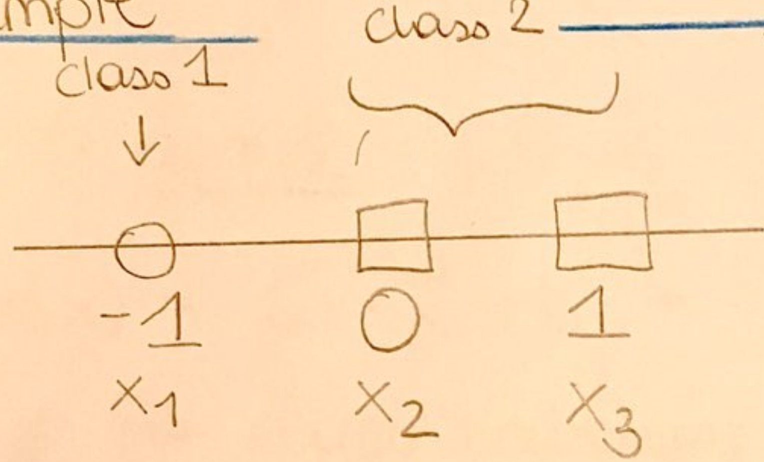
Decision function: $D(\vec{x}) = \sum_{i \in S} \lambda_i y_i \vec{x}_i^T \vec{x} + b$

Then

$$D(\vec{x}) = \sum_{i\in S} \alpha_i y_i \vec{x}_i^{T} \cdot \vec{x} + b$$

$$b = \sum_{i\in S} \left(y_i - \vec{\omega}^{T} \cdot \vec{x}_i\right) \frac{1}{|S|}$$

2.2. Undestand it better via an example

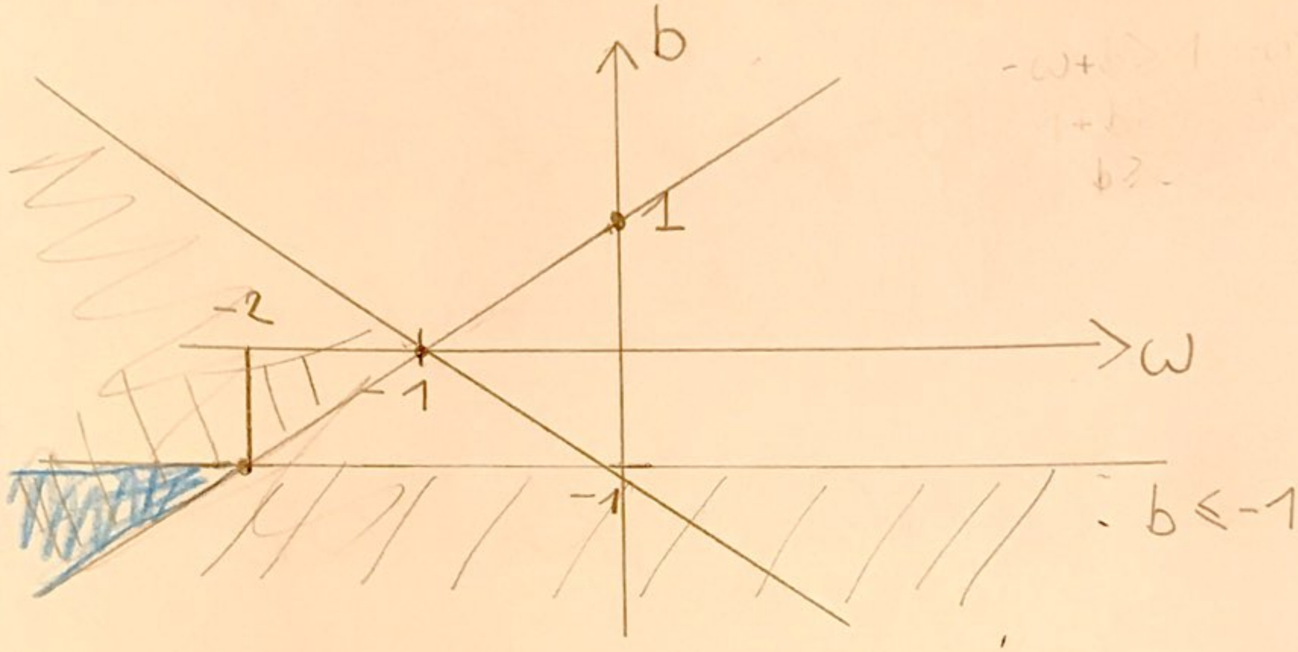


$$y_i \left[\vec{\omega}^{T} \cdot \vec{x}_i + b\right] \geq 1 \quad , \quad i = 1 \ldots M$$

$$\Rightarrow \quad \left.\begin{array}{r} \omega \cdot (-1) + b \geq -1 \end{array}\right\} (y_i = 1)$$

$$\left.\begin{array}{r} (-1) \cdot b \geq 1 \\ (-1)\left[\omega + b\right] \geq 1 \end{array}\right\} (y_i = -1)$$

Our solution has to minimize $\|\vec{\omega}\|^2$



$\Rightarrow$ $\underline{\omega = -2}$ $\quad$ $\underline{b = -1}$

$$D(x) = \vec{\omega}^T \cdot \vec{x} + b = -2 \cdot x - 1$$

$\Rightarrow$ class boundary at $x = -\frac{1}{2}$

$x = 0$ and $x = -1$ are support vectors

Dual problem:

$$Q(\vec{\lambda}) = \lambda_1 + \lambda_2 + \lambda_3 - \frac{1}{2}\Big[\lambda_1 \cdot 1 \cdot \lambda_1 \cdot 1 (-1) \cdot (-1)$$
$$+ \lambda_1 \cdot 1 \cdot \lambda_2 \cdot (-1) \cdot (-1)(0) + \dots\Big]$$

$$= \lambda_1 + \lambda_2 + \lambda_3 - \frac{1}{2}\left[\lambda_1^2 + 2\lambda_1\lambda_3 + \lambda_3^2\right]$$

subject to $\lambda_1 - \lambda_2 - \lambda_3 = 0$, $\alpha_i \geq 0$

$$\Rightarrow \lambda_2 = \lambda_1 - \lambda_3$$

$$\Rightarrow Q(\lambda) = 2\lambda_1 - \frac{1}{2}\left[\lambda_1 + \lambda_3\right]^2 ; \alpha_i \geq 0$$

$$\Rightarrow \lambda_3 = 0 \quad Q(\lambda) = 2\lambda_1 - \frac{1}{2}\lambda_1^2$$

$$= -\frac{1}{2}\left[\lambda_1 - 2\right]^2 + 2$$

$$\Rightarrow \lambda_1 = 2$$

$$\Rightarrow \underbrace{\lambda_1 = 2 , \lambda_2 = 2}_{\text{support vector}} , \lambda_3 = 0$$

$$\vec{\omega} = \sum \lambda_i y_i \vec{x}_i = \lambda_1 y_1 x_1 + \lambda_2 y_2 x_2 + \lambda_3 y_3 x_3$$

$$= (2) \cdot (1) \cdot (-1) = -2$$

$b = -1$

If $x_3 = 1$ would be in Class 1 then this problem is not linearly separable in this space:

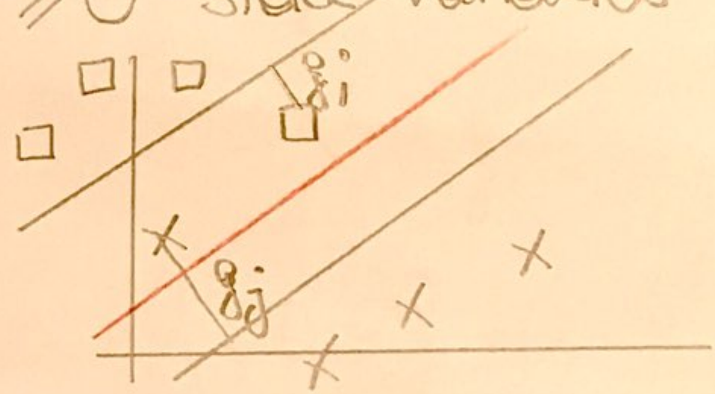
$$(-1)\,[\omega + b] \geq 1$$
$$\downarrow$$
$$(+1)\,[\omega + b] \geq 1$$

2.3. Soft-Margin Support Vector Machines

To allow unseparable solutions (when hard-margin SVM is unsolvable):

$$y_i\,[\vec{\omega}^T \cdot \vec{x}_i + b] \geq 1 - \xi_i \,, \quad i=1..M$$

$\xi_i \geq 0$ slack variables



With ξ_i feasible solutions always exist.

$$\text{minimize } Q(\vec{\omega}, b, \xi) = \frac{1}{2}\|\vec{\omega}\|^2 + \frac{C}{p}\sum_{i=1}^{M} \xi_i^p$$

$$\text{subject to } y_i\left[\vec{\omega}^T \cdot \vec{x}_i + b\right] \geq 1 - \xi_i$$

$p = 1$ L1 norm
$p = 2$ L2 norm

C is the margin parameter (trade-off between maximization of margin and minimization of classification error)

$$Q(\vec{\omega}, b, \vec{\xi}, \vec{\alpha}, \vec{\beta}) = \frac{1}{2}\|\vec{\omega}\|^2 + C\sum \xi_i -$$

(L1 norm)

$$- \sum \alpha_i \left(y_i\left[\vec{\omega}^T \cdot \vec{x}_i + b\right] - 1 - \xi_i\right)$$

$$- \sum \beta_i \xi_i$$

$$\frac{\partial Q}{\partial \omega} = 0, \quad \frac{\partial Q}{\partial b} = 0, \quad \frac{\partial Q}{\partial \xi} = 0, \quad \alpha_i\left(y_i\left[\ldots\right]\right) = 0$$
$$\beta_i \xi_i = 0$$

$$\Rightarrow \vec{\omega} = \sum_{i=1}^{M} \alpha_i y_i \vec{x}_i$$

$$\sum_{i=1}^{M} \alpha_i y_i = 0, \quad \alpha_i + \beta_i = C \quad i = 1 \ldots M$$

Dual Problem:

maximize $Q(\vec{\alpha}) = \sum \alpha_i - \frac{1}{2} \sum \alpha_i \alpha_j y_i y_j \vec{x}_i^T \vec{x}_j$

subject to $\sum_{i=1}^{M} y_i \alpha_i = 0 \quad \underbrace{C \geq \alpha_i \geq 0}_{\text{only difference to hard-margin SVM}}$

$0 < \alpha_i < C$ (unbounded) $\forall$ support vectors

$\alpha_i = C$ bounded support vector

$D(\vec{x}) = \sum_{i \in S} \alpha_i y_i \vec{x}_i^T \vec{x} + b$; $i \in S \leftarrow$ support vector $\quad b = \frac{1}{|U|} \sum_{i \in U} (y_i - \vec{w}^T \cdot \vec{x}_i)$

3. Kernel Tricks

$$[x_1 \dots x_M]^T \longmapsto [\phi_1(\vec{x}), \dots, \phi_\ell(\vec{x})]^T$$

$$D(\vec{x}) = \vec{\omega}^T \cdot \vec{\phi}(\vec{x}) + b$$

Note: $\vec{\omega}$ is now $[\omega_1, \dots, \omega_\ell]$ ℓ-dim !

Hilbert-Schmidt theory

Let us assume (kernel trick)

$$\underbrace{K(\vec{x}, \vec{x}')}_{\text{kernel (no need to treat high-dim. feature space explicitly)}} = \phi^T(\vec{x})\, \phi(\vec{x}')$$

Dual problem:

$$\max \quad Q(\vec{\lambda}) = \sum_{i=1}^{M} \lambda_i - \frac{1}{2} \sum_{i=1}^{M} \lambda_i \lambda_j y_i y_j K(\vec{x}_i, \vec{x}_j)$$

$$\text{subject to} \quad \sum_{i=1}^{M} y_i \lambda_i = 0 \qquad 0 \leq \lambda_i \leq C$$

KKT - conditions :

$$\lambda_i \left(y_i \left[\sum y_j \lambda_j K(\vec{x}_i, \vec{x}_j) + b \right] - 1 - \xi_i \right) = 0$$

$$(C - \lambda_i)\xi_i = 0$$

$$(\alpha_i \geq 0), \ (\xi_i \geq 0)$$

Decision function:

$$D(\vec{x}) = \sum \lambda_i y_i K(\vec{x}_i, \vec{x}) + b$$

$$b = y_j - \sum \alpha_i y_i K(\vec{x}_i, \vec{x}_j)$$

Linear kernel $K(\vec{x}, \vec{x}') = \vec{x}^T \cdot \vec{x}'$

Polynomial kernel $K(\vec{x}, \vec{x}') = [\vec{x}^T \cdot \vec{x}' + 1]^d$

$d=2,\ m=2$

$$\vec{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \qquad \vec{x}' = \begin{bmatrix} x_1' \\ x_2' \end{bmatrix}$$

$$K(\vec{x},\vec{x}') = \left[\begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} x_1' \\ x_2' \end{bmatrix} + 1 \right]^2 =$$

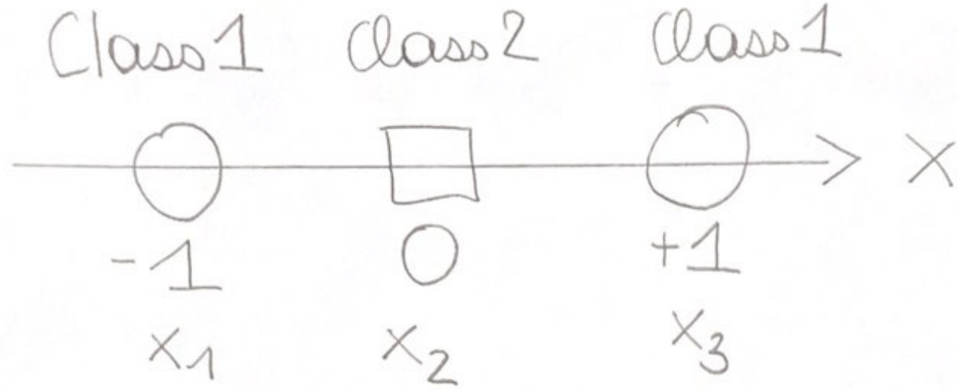
$$= \left[x_1 x_1' + x_2 x_2' + 1 \right]^2 =$$

$$= (x_1 x_1')^2 + 2 x_1 x_1' x_2 x_2' + 1 + (x_2 x_2')^2$$

$$+ 2 x_1 x_1' + 2 x_2 x_2' = \phi^T(\vec{x})\,\phi(\vec{x}')$$

$$= \underbrace{\left[1, \sqrt{2}\,x_1, \sqrt{2}\,x_2, \sqrt{2}\,x_1 x_2, x_1^2, x_2^2 \right]}_{\phi^T(\vec{x})} \underbrace{\begin{bmatrix} 1 \\ \sqrt{2}\,x_1' \\ \sqrt{2}\,x_2' \\ \sqrt{2}\,x_1' x_2' \\ x_1'^2 \\ x_2'^2 \end{bmatrix}}_{\phi(\vec{x}')}$$

3.1 Our Example



$$Q(\vec{\alpha}) = \alpha_1 + \alpha_2 + \alpha_3 - \frac{1}{2}\sum_{i=1}^{3}\sum_{j=1}^{3} \alpha_i \alpha_j y_i y_j \left[\vec{x}_i^T \cdot \vec{x}_j + 1\right]^2$$

$i=1, j=1 \quad \alpha_1\alpha_1(1)(1)\left[(-1)^2+1\right]^2 = \alpha_1^2 \cdot 4$

$i=1, j=2 \quad \alpha_1\alpha_2(1)(-1)\left[0+1\right]^2 = -\alpha_1\alpha_2$

$i=1, j=3 \quad \alpha_1\alpha_3(1)(1)\left[(-1)(1)+1\right]^2 = 0$

$i=2, j=1 \quad \alpha_2\alpha_1(-1)(1)\left[0+1\right]^2 = -\alpha_1\alpha_2$

$i=2, j=2 \quad \alpha_2^2$

$i=2, j=3 \quad \alpha_2\alpha_3(-1)(1)\left[0+1\right]^2 = -\alpha_2\alpha_3$

$i=3, j=1 \quad 0$

$i=3, j=2 \quad -\alpha_2\alpha_3$

$i=3, j=3 \quad \alpha_3\alpha_3(1)(1)\left[1+1\right]^2 = 4\alpha_3^2$

$$Q(\vec{\alpha}) = \alpha_1 + \alpha_2 + \alpha_3$$
$$- \frac{1}{2}\left[\alpha_1^2 \cdot 4 + \alpha_3^2 \cdot 4 + \alpha_2^2 - 2\alpha_1\alpha_2 - 2\alpha_2\alpha_3\right] \quad (1)$$

subject to

$$\sum y_i \alpha_j = \alpha_1 - \alpha_2 + \alpha_3 \overset{!}{=} 0 \quad (2)$$
$$C \geq \alpha_i \geq 0$$

From (2): $\alpha_2 = \alpha_1 + \alpha_3$

into (1):

$$Q(\vec{\alpha}) = 2\alpha_1 + 2\alpha_3 - \left[2\alpha_1^2 + 2\alpha_3^2 - \frac{1}{2}(\alpha_1 + \alpha_3)^2\right]$$

$$\frac{\partial Q}{\partial \alpha_1} = 2 - 4\alpha_1 + \frac{1}{2} 2(\alpha_1 + \alpha_3) \overset{!}{=} 0 \quad (1)$$
$$= 2 - 3\alpha_1 + \alpha_3$$

$$\frac{\partial Q}{\partial \alpha_3} = 2 - 4\alpha_3 + \alpha_1 + \alpha_3 \quad (2)$$
$$= 2 - 3\alpha_3 + \alpha_1 = 0$$

$$\overset{(1)}{\Rightarrow} \quad 3\alpha_1 = 2 + \alpha_3$$

$$\alpha_1 = \frac{2}{3} + \frac{\alpha_3}{3}$$

(2) $$2 - 3\alpha_3 + \frac{2}{3} + \frac{\alpha_3}{3} = 0$$

$$\frac{8}{3} - \frac{8}{3}\alpha_3 = 0$$

$$\Rightarrow \underline{\alpha_3 = 1} \quad \Rightarrow \underline{\alpha_1 = 1}$$

$$\underline{\alpha_2 = \alpha_1 + \alpha_3 = 2}$$

$C \geq 2$, $x = -1, 0, 1$ are support vectors

$$b = y_j - \sum_{i \in S} \alpha_i y_i \underbrace{K(\vec{x}_i, \vec{x}_j)}_{[\vec{x}_i^T x_j + 1]^2}$$

$$= (1) - \alpha_1 (1) [1]^2 - \alpha_2 (-1) + \dots$$

$$= -1$$

$$D(\vec{x}) = \sum_i \alpha_i y_i \left[\vec{x}_i^T \cdot \vec{x} + 1\right]^2 + b$$

$$= (1)(1)\left[-\vec{x} + 1\right]^2 + 1$$
$$+ (2)(-1)\left[\ 1\right]^2 +$$
$$+ (1)(1)\left[\vec{x} + 1\right]^2 - 1$$

$$= \left[x - 1\right]^2 + \left[x + 1\right]^2 - 3$$
$$= 2x^2 - 1$$

boundaries $x = \pm \frac{1}{\sqrt{2}}$