Support Vector Machines

464.561A Models and Technologies for Information Services

Jonghun Park

jonghun@snu.ac.kr

Dept. of Industrial Eng.

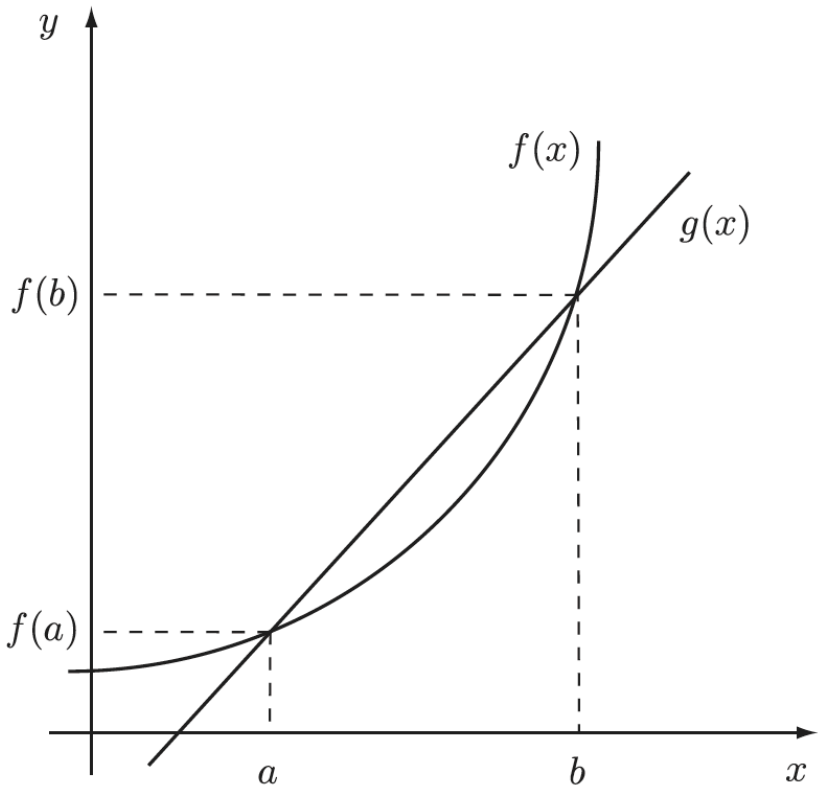
Seoul National University

convex optimization problem

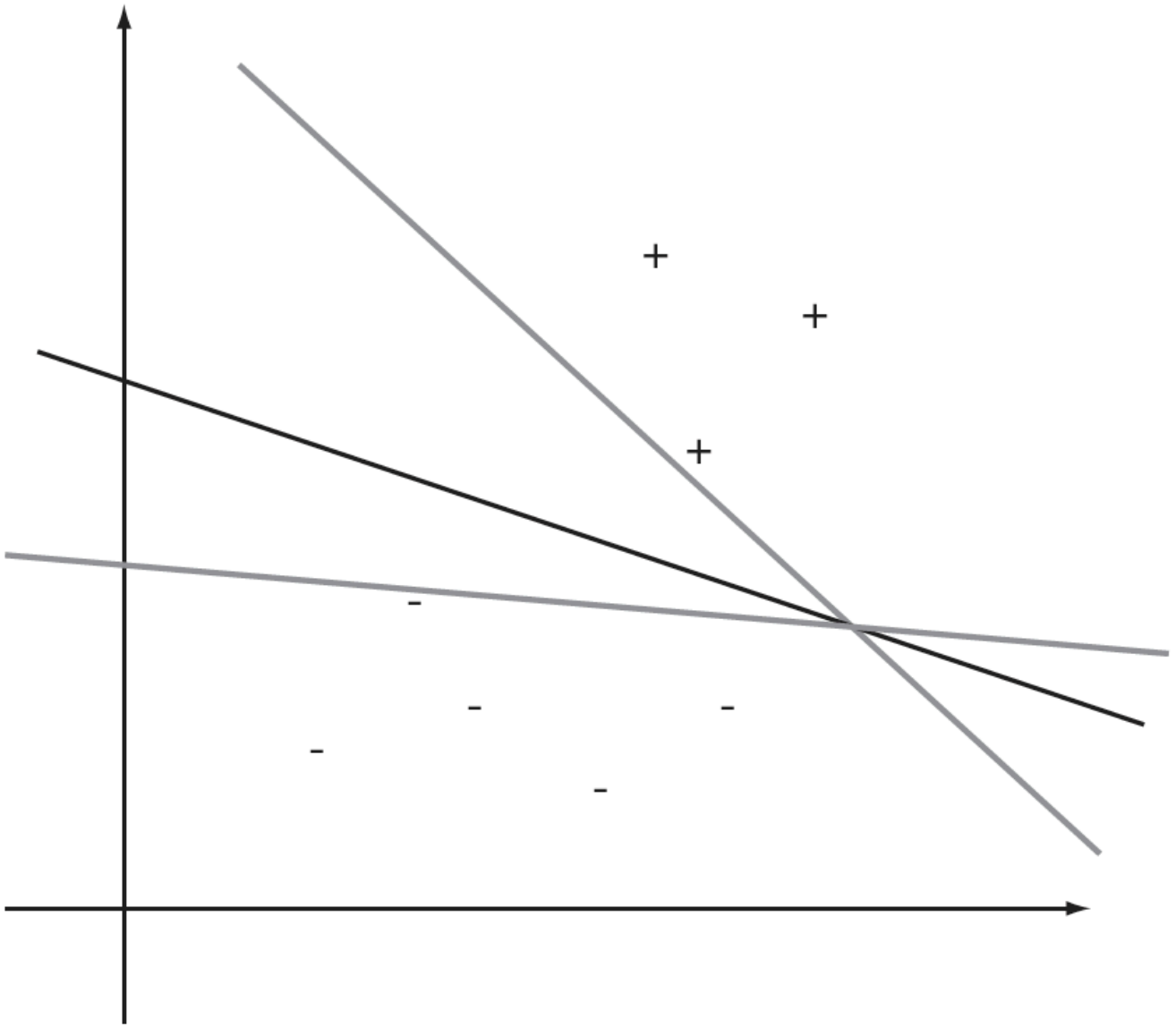
$$\min_{\mathbf{x}} \phi(\mathbf{x})$$

s.t.

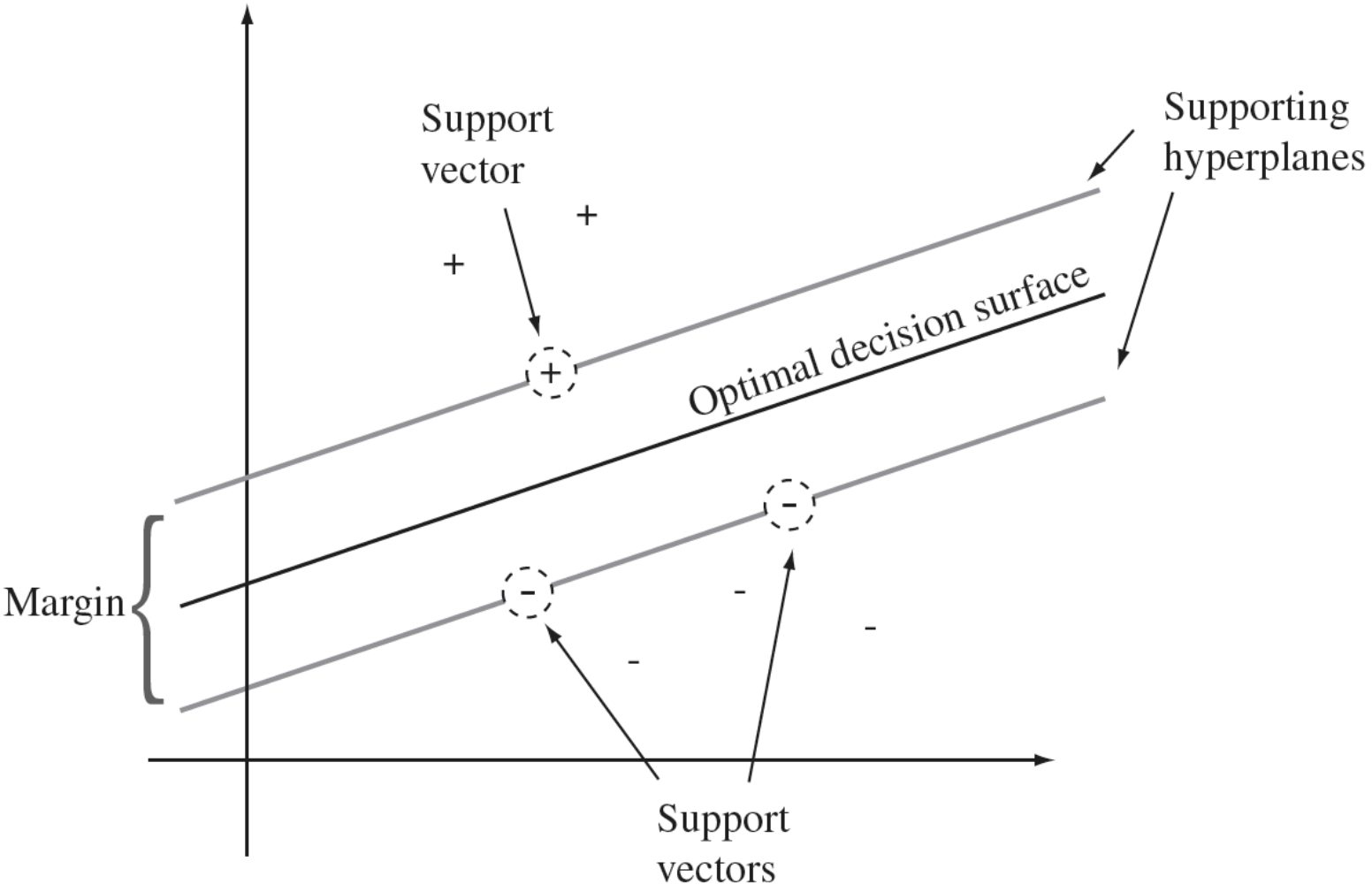
$$h_i(\mathbf{x}) \geq c_i, i = 1, \ldots, l$$



binary classification problem



optimal decision surface



maximum margin

given a linearly separable training set and the optimal decision surface

$$D = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_l, y_l)\} \subseteq \mathbb{R}^n \times \{+1, -1\}$$

$$\mathbf{w}^* \cdot \mathbf{x} = b^*$$

the maximum margin is given by

$$m^* = \max \phi(\mathbf{w}, b) = \phi(\mathbf{w}^*, b^*)$$

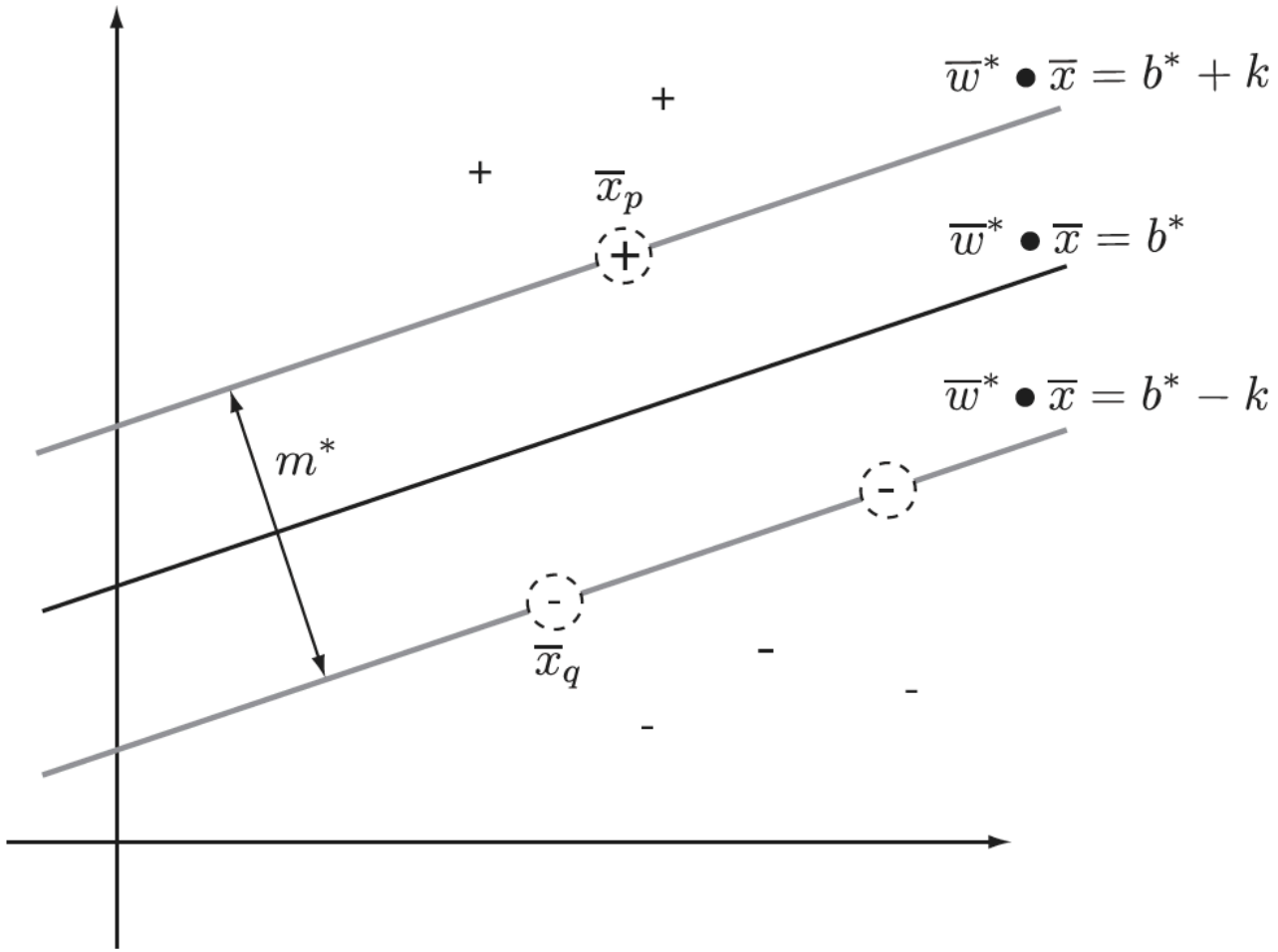
computation of maximum margin

$$\mathbf{w}^* \cdot \mathbf{x}_p = b^* + k$$

$$(\mathbf{x}_p, +1) \in D$$

$$\mathbf{w}^* \cdot \mathbf{x}_q = b^* - k$$

$$(\mathbf{x}_q, -1) \in D$$



computation of maximum margin

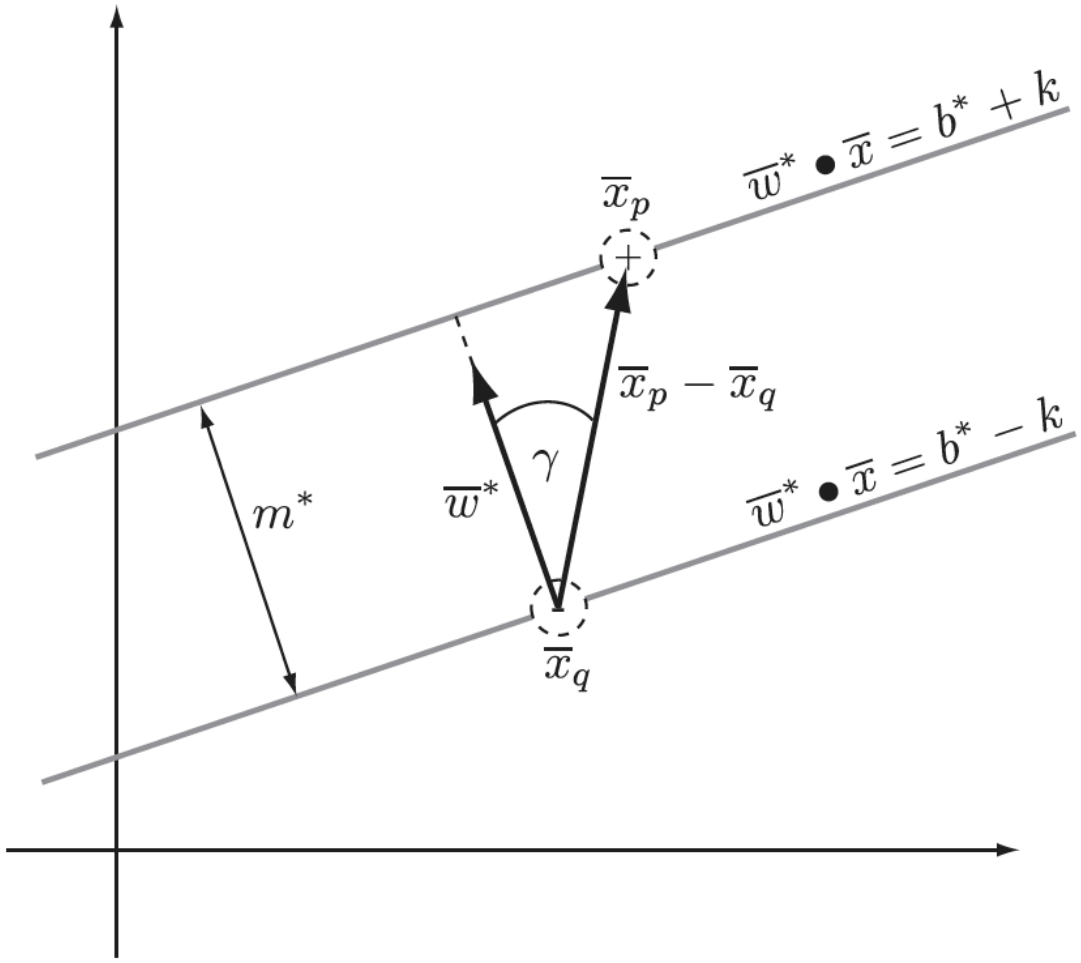
$$m^* = |\mathbf{x}_p - \mathbf{x}_q| \cos\gamma$$

$$= \frac{\mathbf{w}^* \cdot (\mathbf{x}_p - \mathbf{x}_q)}{|\mathbf{w}^*|}$$

$$= \frac{\mathbf{w}^* \cdot \mathbf{x}_p - \mathbf{w}^* \mathbf{x}_q}{|\mathbf{w}^*|}$$

$$= \frac{(b^* + k) - (b^* - k)}{|\mathbf{w}^*|}$$

$$= \frac{2k}{|\mathbf{w}^*|}$$

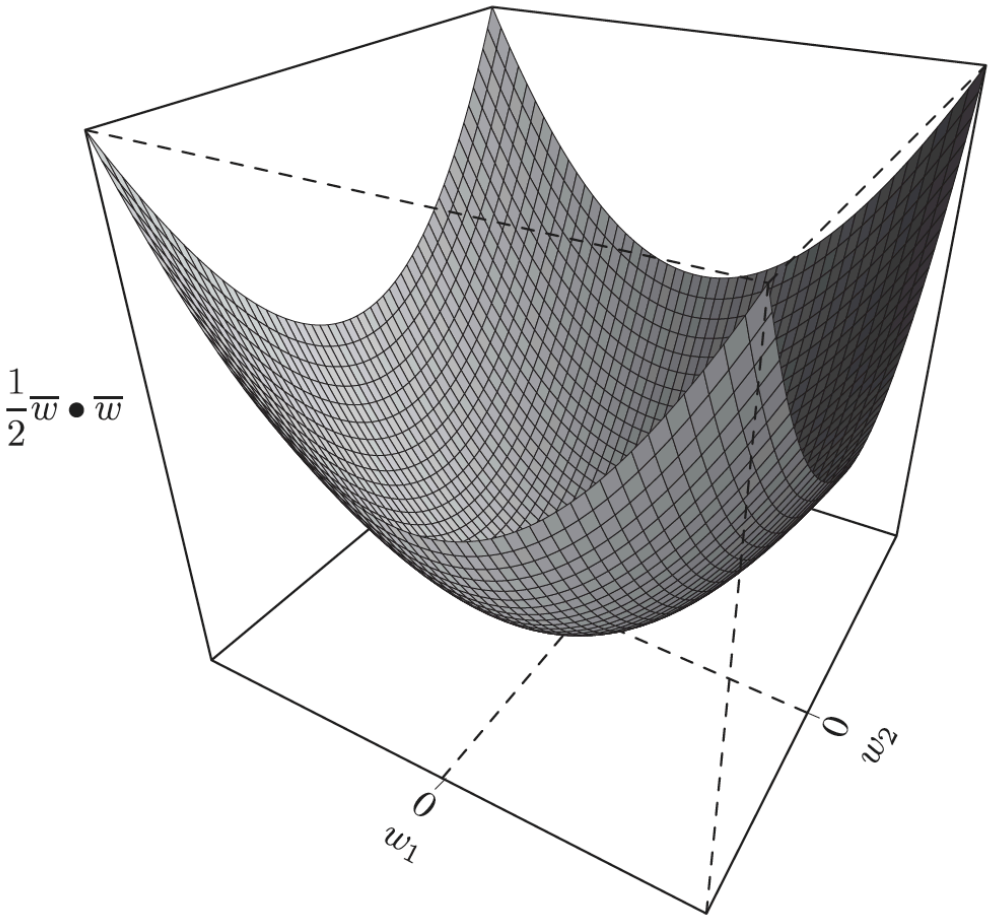


objective function

$$m^* = \max \frac{2k}{|\mathbf{w}|}$$

$$= \min \frac{|\mathbf{w}|}{2k}$$

$$\Rightarrow \phi(\mathbf{w}, b) = \frac{1}{2}\mathbf{w} \cdot \mathbf{w}$$



constraints

$$\mathbf{w}^* \cdot \mathbf{x}_i \geq b^* + k,\ \forall(\mathbf{x}_i, y_i) \in D \text{ s.t. } y_i = +1$$

$$\mathbf{w}^* \cdot \mathbf{x}_i \leq b^* - k,\ \forall(\mathbf{x}_i, y_i) \in D \text{ s.t. } y_i = -1$$

$$\Rightarrow$$

$$\mathbf{w} \cdot \mathbf{x}_i \geq 1 + b,\ \forall(\mathbf{x}_i, y_i) \in D \text{ s.t. } y_i = +1$$

$$\mathbf{w} \cdot (-\mathbf{x}_i) \geq 1 - b,\ \forall(\mathbf{x}_i, y_i) \in D \text{ s.t. } y_i = -1$$

$$\Rightarrow$$

$$\mathbf{w} \cdot (y_i \mathbf{x}_i) \geq 1 + y_i b,\ \forall(\mathbf{x}_i, y_i) \in D$$

maximum margin classifier

$$\min \phi(\mathbf{w}, b) = \min_{\mathbf{w},b} \frac{1}{2}\mathbf{w} \cdot \mathbf{w}$$

s.t.

$$\mathbf{w} \cdot (y_i \mathbf{x}_i) \geq 1 + y_i b, \ \forall (\mathbf{x}_i, y_i) \in D$$

$$D = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_l, y_l)\} \subseteq \mathbb{R}^n \times \{+1, -1\}$$

quadratic programming

$$(\mathbf{w}^*, b^*) = \arg_{\mathbf{w},b} \min \left(\frac{1}{2} \mathbf{w}^T \mathbf{Q} \mathbf{w} - \mathbf{q} \cdot \mathbf{w} \right)$$

$$\text{s.t.} \quad \mathbf{X}^T \mathbf{w} \geq \mathbf{c}$$

$$\text{where } \mathbf{Q} = \mathbf{I}, \mathbf{q} = \mathbf{0}$$

$$\mathbf{X} = \begin{pmatrix} y_1 x_1^1 & \dots & y_i x_i^1 & \dots & y_l x_l^1 \\ \vdots & & \vdots & & \vdots \\ y_1 x_1^n & \dots & y_i x_i^n & \dots & y_l x_l^n \end{pmatrix}$$

$$\mathbf{c} = \begin{pmatrix} 1 + y_1 b \\ 1 + y_2 b \\ \vdots \\ 1 + y_l b \end{pmatrix} \qquad \mathbf{x}_i = (x_i^1, \dots, x_i^n)$$

Lagrangian optimization problem

$$\min_{\mathbf{x}} \phi(\mathbf{x})$$

s.t.

$$g_i(\mathbf{x}) \geq 0, i = 1, \ldots, l$$

$$\mathbf{x} \in \mathbb{R}^n$$

$$\max_{\boldsymbol{\alpha}} \min_{\mathbf{x}} L(\boldsymbol{\alpha}, \mathbf{x}) = \max_{\boldsymbol{\alpha}} \min_{\mathbf{x}} \left(\phi(\mathbf{x}) - \sum_{i=1}^{l} \alpha_i g_i(\mathbf{x}) \right)$$

s.t.

$$\alpha_i \geq 0, i = 1, \ldots, l$$

nested optimization

$$\max_{\boldsymbol{\alpha}} L(\boldsymbol{\alpha}, \mathbf{x}^*) = \max_{\boldsymbol{\alpha}} \left(\phi(\mathbf{x}^*) - \sum_{i=1}^{l} \alpha_i g_i(\mathbf{x}^*) \right)$$

$$\min_{\mathbf{x}} L(\boldsymbol{\alpha}^*, \mathbf{x}) = \min_{\mathbf{x}} \left(\phi(\mathbf{x}) - \sum_{i=1}^{l} \alpha_i^* g_i(\mathbf{x}) \right)$$

$$\Rightarrow \frac{\partial L}{\partial \mathbf{x}}(\boldsymbol{\alpha}, \mathbf{x}^*) = \mathbf{0}$$

KKT (Karush-Kuhn-Tucker) condition

$\boldsymbol{\alpha}^*$ and $\mathbf{x}^*$ s.t.

$$\max_{\boldsymbol{\alpha}} \min_{\mathbf{x}} L(\boldsymbol{\alpha}, \mathbf{x}) = L(\boldsymbol{\alpha}^*, \mathbf{x}^*) = \phi(\mathbf{x}^*) - \sum_{i=1}^{l} \alpha_i^* g_i(\mathbf{x}^*)$$

$\mathbf{x}^*$ is a solution to the primal objective function iff

$$\frac{\partial L}{\partial \mathbf{x}}(\boldsymbol{\alpha}^*, \mathbf{x}^*) = \mathbf{0}$$

$$\alpha_i^* g_i(\mathbf{x}^*) = 0$$

$$g_i(\mathbf{x}^*) \geq 0$$

$$\alpha_i^* \geq 0$$

complementary condition

$$\alpha_i^* g_i(\mathbf{x}^*) = 0$$

$$\Rightarrow L(\boldsymbol{\alpha}^*, \mathbf{x}^*) = \phi(\mathbf{x}^*)$$

Lagrangian dual

$$\frac{\partial L}{\partial \mathbf{x}}(\boldsymbol{\alpha}, \mathbf{x}^*) = \mathbf{0} \Rightarrow L(\boldsymbol{\alpha}, \mathbf{x}^*) = \phi'(\boldsymbol{\alpha})$$

$$\max_{\boldsymbol{\alpha}} \phi'(\boldsymbol{\alpha})$$

$$\text{s.t.} \quad \alpha_i \geq 0, i = 1, \ldots, l$$

$$\Rightarrow$$

$$\max_{\boldsymbol{\alpha}} \phi'(\boldsymbol{\alpha}) = \phi'(\boldsymbol{\alpha}^*) = L(\boldsymbol{\alpha}^*, \mathbf{x}^*) = \phi(\mathbf{x}^*)$$

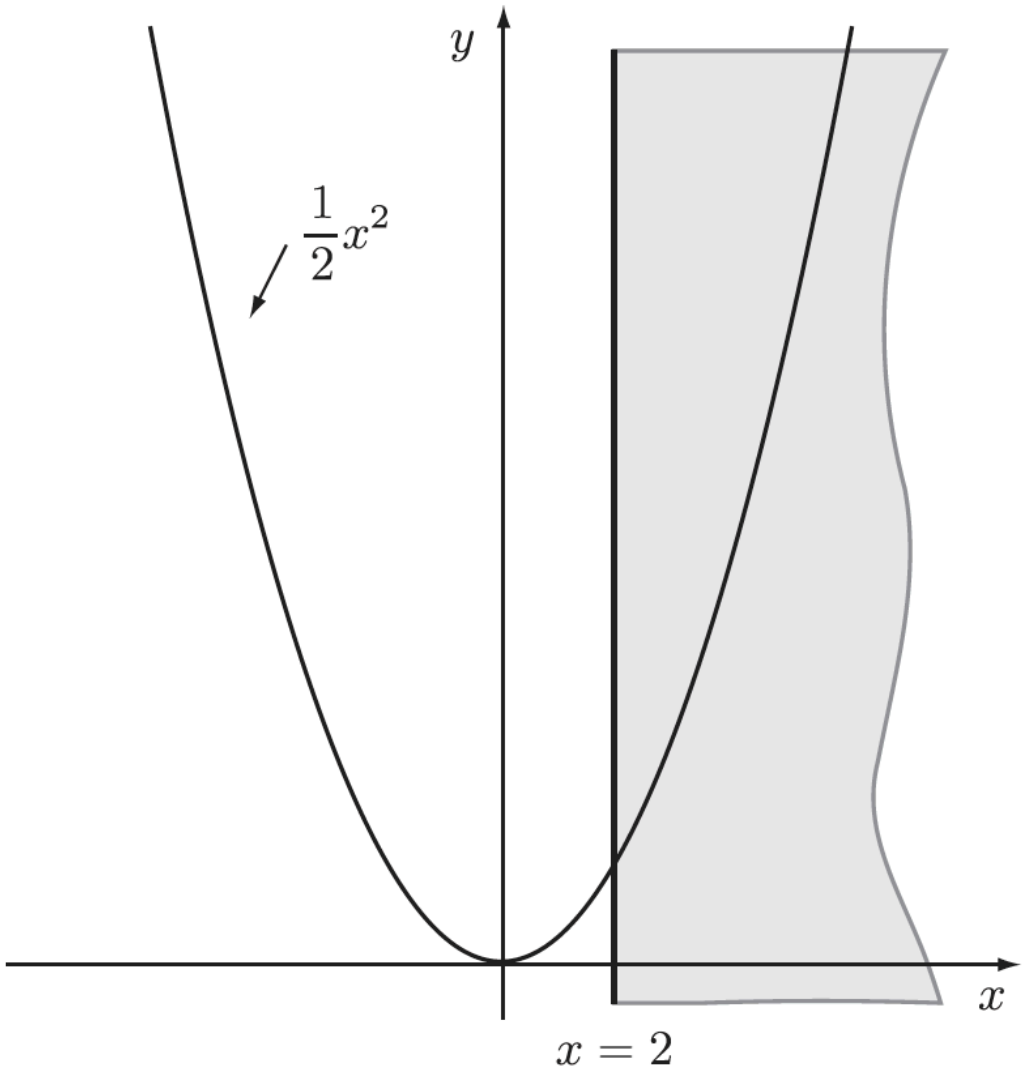
s.t. $\mathbf{x}^*$ and $\boldsymbol{\alpha}^*$ satisfy KKT conditions.

example

$$\min \phi(x) = \min \frac{1}{2}x^2$$

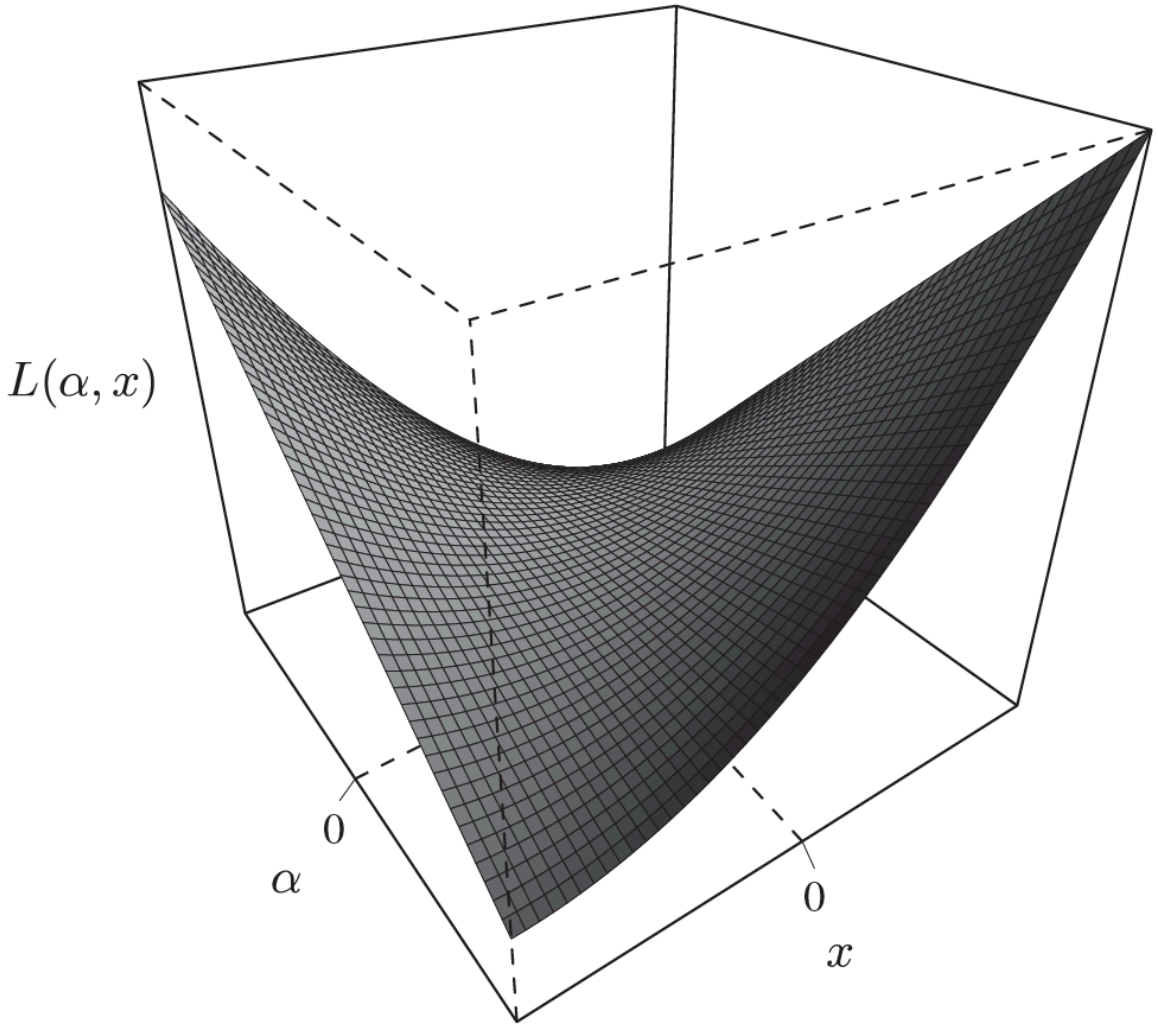
s.t.

$$g(x) = x - 2 \geq 0$$



example

$$L(\alpha, x) = \frac{1}{2}x^2 - \alpha(x-2)$$



example

$$\frac{\partial L}{\partial x}(\alpha, x^*) = x^* - \alpha = 0$$

$$\Rightarrow x^* = \alpha$$

$$\Rightarrow L(\alpha, x^*) = 2\alpha - \frac{1}{2}\alpha^2$$

$$\Rightarrow \max_{\alpha} \phi'(\alpha) = \max_{\alpha} \left(2\alpha - \frac{1}{2}\alpha^2\right) \text{s.t. } \alpha \geq 0$$

$$\Rightarrow \frac{d\phi'}{d\alpha}(\alpha^*) = 2 - \alpha^* = 0$$

$$\Rightarrow x^* = \alpha^* = 2, \alpha^* g(x^*) = \alpha^*(x^* - 2) = 0$$

dual maximum margin optimization

$$\min_{\mathbf{w},b} \phi(\mathbf{w},b) = \min_{\mathbf{w},b} \frac{1}{2}\mathbf{w}\cdot\mathbf{w}$$

$$\text{s.t. } g_i(\mathbf{w},b) = y_i(\mathbf{w}\cdot\mathbf{x}_i - b) - 1 \geq 0, i = 1,\ldots,l$$

$$\max_{\boldsymbol{\alpha}} \min_{\mathbf{w},b} L(\boldsymbol{\alpha},\mathbf{w},b) = \phi(\mathbf{w},b) - \sum_{i=1}^{l} \alpha_i g_i(\mathbf{w},b)$$

$$= \frac{1}{2}\mathbf{w}\cdot\mathbf{w} - \sum_{i=1}^{l} \alpha_i y_i \mathbf{w}\cdot\mathbf{x}_i + b\sum_{i=1}^{l} \alpha_i y_i + \sum_{i=1}^{l} \alpha_i$$

$$\text{s.t. } \alpha_i \geq 0, i = 1,\ldots,l$$

KKT conditions

$$\boldsymbol{\alpha}^*, \mathbf{w}^*, \text{ and } b^* \text{ s.t.} \max_{\boldsymbol{\alpha}} \min_{\mathbf{w},b} L(\boldsymbol{\alpha}, \mathbf{w}, b) = L(\boldsymbol{\alpha}^*, \mathbf{w}^*, b^*)$$

will satisfy

$$\frac{\partial L}{\partial \mathbf{w}}(\boldsymbol{\alpha}^*, \mathbf{w}^*, b^*) = \mathbf{0}$$

$$\frac{\partial L}{\partial b}(\boldsymbol{\alpha}^*, \mathbf{w}^*, b^*) = 0$$

$$\alpha_i^*(y_i(\mathbf{w}^* \cdot \mathbf{x}_i - b^*) - 1) = 0$$

$$y_i(\mathbf{w}^* \cdot \mathbf{x}_i - b^*) - 1 \geq 0$$

$$\alpha_i^* \geq 0$$

$$i = 1, \ldots, l$$

derivation of Lagrangian dual

$$\frac{\partial L}{\partial \mathbf{w}}(\boldsymbol{\alpha}, \mathbf{w}^*, b) = \mathbf{w}^* - \sum_{i=1}^{l} \alpha_i y_i \mathbf{x}_i = \mathbf{0}$$

$$\Rightarrow \mathbf{w}^* = \sum_{i=1}^{l} \alpha_i y_i \mathbf{x}_i$$

$$\frac{\partial L}{\partial b}(\boldsymbol{\alpha}, \mathbf{w}, b^*) = \sum_{i=1}^{l} \alpha_i y_i = 0$$

$\Rightarrow$ becomes a constraint

maximum margin Lagrangian dual

$$\max_{\boldsymbol{\alpha}} \phi'(\boldsymbol{\alpha}) = \max_{\boldsymbol{\alpha}} L(\boldsymbol{\alpha}, \mathbf{w}^*, b^*)$$

$$= \max_{\boldsymbol{\alpha}} \left(\sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} \alpha_i \alpha_j y_i y_j \mathbf{x}_i \cdot \mathbf{x}_j \right)$$

$$\text{s.t.} \quad \sum_{i=1}^{l} \alpha_i y_i = 0$$

$$\alpha_i \geq 0$$

$$i = 1, \ldots, l$$

KKT complementary condition

$$\alpha_j^* > 0, \text{ for } (\mathbf{x}_j, y_j) \in D$$

$$\Rightarrow \mathbf{w}^* \cdot \mathbf{x}_j = b^* + 1, \text{ if } y_j = +1$$

$$\mathbf{w}^* \cdot \mathbf{x}_j = b^* - 1, \text{ if } y_j = -1$$

$$\alpha_j^* = 0, \text{ for } (\mathbf{x}_j, y_j) \in D$$

$$\Rightarrow \mathbf{w}^* \cdot \mathbf{x}_j > b^* + 1, \text{ if } y_j = +1$$

$$\mathbf{w}^* \cdot \mathbf{x}_j < b^* - 1, \text{ if } y_j = -1$$

implications

- The primal maximum-margin optimization computes the supporting hyperplanes whose margin is limited by the support vectors
- The dual maximum-margin optimization computes the support vectors that limit the size of the margin of the supporting hyperplanes

computation of b^*

pick a support vector with nonzero Lagrangian multipliers

$(\mathbf{x}_{sv+}, y_{sv+})$ s.t. $y_{sv+} = +1$

$$b^* = \mathbf{w}^* \cdot \mathbf{x}_{sv+} - 1 = \sum_{i=1}^{l} \alpha_i^* y_i \mathbf{x}_i \cdot \mathbf{x}_{sv+} - 1$$

$\vdots$

$$\mathbf{w}^* \cdot \mathbf{x}_j = b^* + 1, \text{ if } y_j = +1$$

$$\mathbf{w}^* = \sum_{i=1}^{l} \alpha_i y_i \mathbf{x}_i$$

dual classification function

optimal decision surface: $\mathbf{w}^* \cdot \mathbf{x} = b^*$

$$\Rightarrow \hat{f}(\mathbf{x}) = sgn(\mathbf{w}^* \cdot \mathbf{x} - b^*)$$

$$\Rightarrow \hat{f}(\mathbf{x}) = sgn\left(\sum_{i=1}^{l} \alpha_i^* y_i \mathbf{x}_i \cdot \mathbf{x} - \sum_{i=1}^{l} \alpha_i^* y_i \mathbf{x}_i \cdot \mathbf{x}_{\mathbf{sv}+} + 1\right)$$

$\because$

$$\mathbf{w}^* = \sum_{i=1}^{l} \alpha_i y_i \mathbf{x}_i$$

$$b^* = \sum_{i=1}^{l} \alpha_i^* y_i \mathbf{x}_i \cdot \mathbf{x}_{sv+} - 1$$

linear SVM: problem

target function $f : \mathbb{R}^n \rightarrow \{+1, -1\}$

labeled, linearly separable training set:

$D = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_l, y_l)\} \subseteq \mathbb{R}^n \times \{+1, -1\}$

where

$\mathbf{x}_i \in \mathbb{R}^n, y_i = f(\mathbf{x}_i), i = 1, \ldots, l$

compute a classifier $\hat{f} : \mathbb{R}^n \rightarrow \{+1, -1\}$ using D s.t.

$\hat{f}(\mathbf{x}) \cong f(\mathbf{x}), \forall \mathbf{x} \in \mathbb{R}^n$

linear SVM: training

$$\boldsymbol{\alpha}^* = \arg_{\boldsymbol{\alpha}} \max \left(\sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} \alpha_i \alpha_j y_i y_j \mathbf{x}_i \cdot \mathbf{x}_j \right)$$

$$\text{s.t.} \quad \sum_{i=1}^{l} \alpha_i y_i = 0$$

$$\alpha_i \geq 0, i = 1, \ldots, l$$

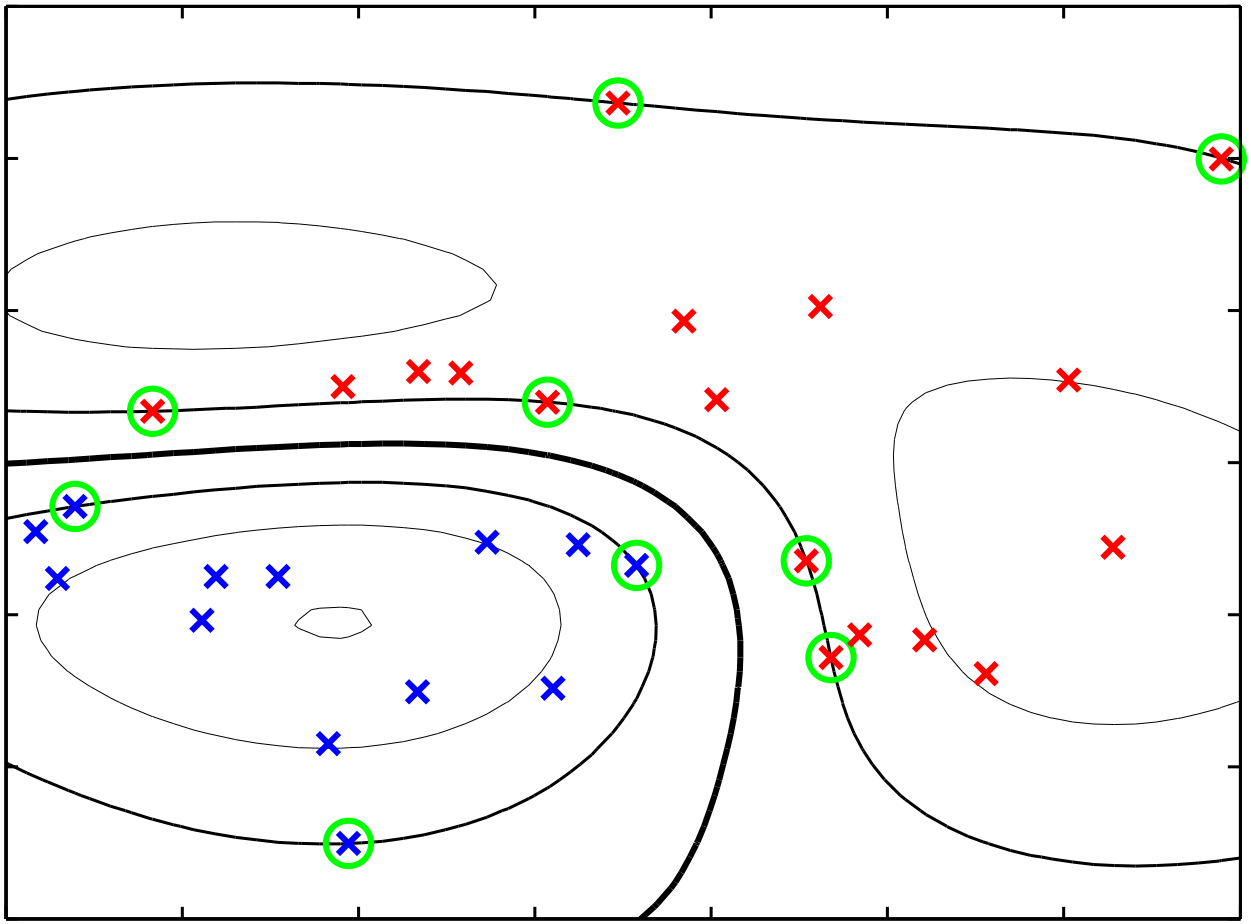
linear SVM: classification

$$\hat{f}(\mathbf{x}) = sgn\left(\sum_{i=1}^{l} \alpha_i^* y_i \mathbf{x}_i \cdot \mathbf{x} - \sum_{i=1}^{l} \alpha_i^* y_i \mathbf{x}_i \cdot \mathbf{x}_{sv+} + 1\right)$$

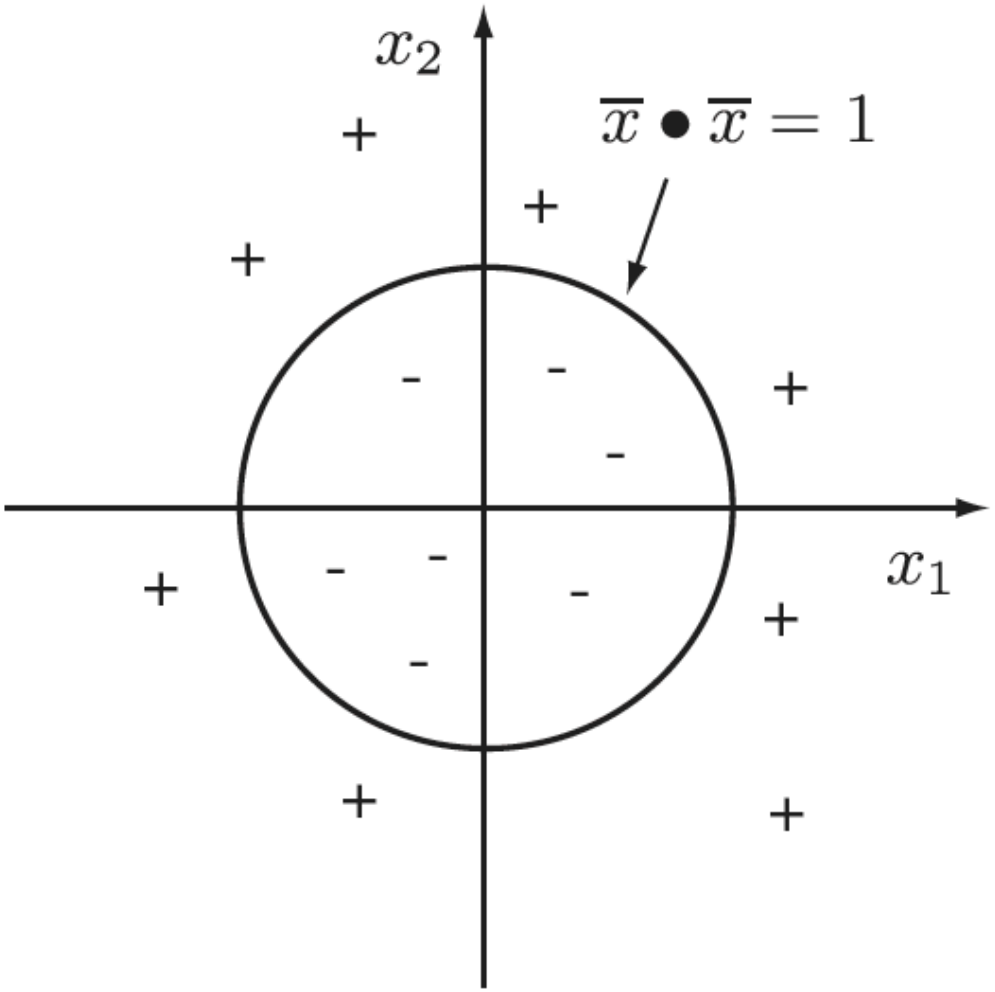
where

$$(\mathbf{x}_{sv+}, +1) \in \{(\mathbf{x}_i, +1) | (\mathbf{x}_i, +1) \in D \text{ and } \alpha_i^* > 0\}$$

SVM for non-linearly separable data



nonlinear SVM example



$$\mathbf{w} \cdot \mathbf{x} = b$$

$$\mathbf{x} \cdot \mathbf{x} = 1$$

transformation into feature space

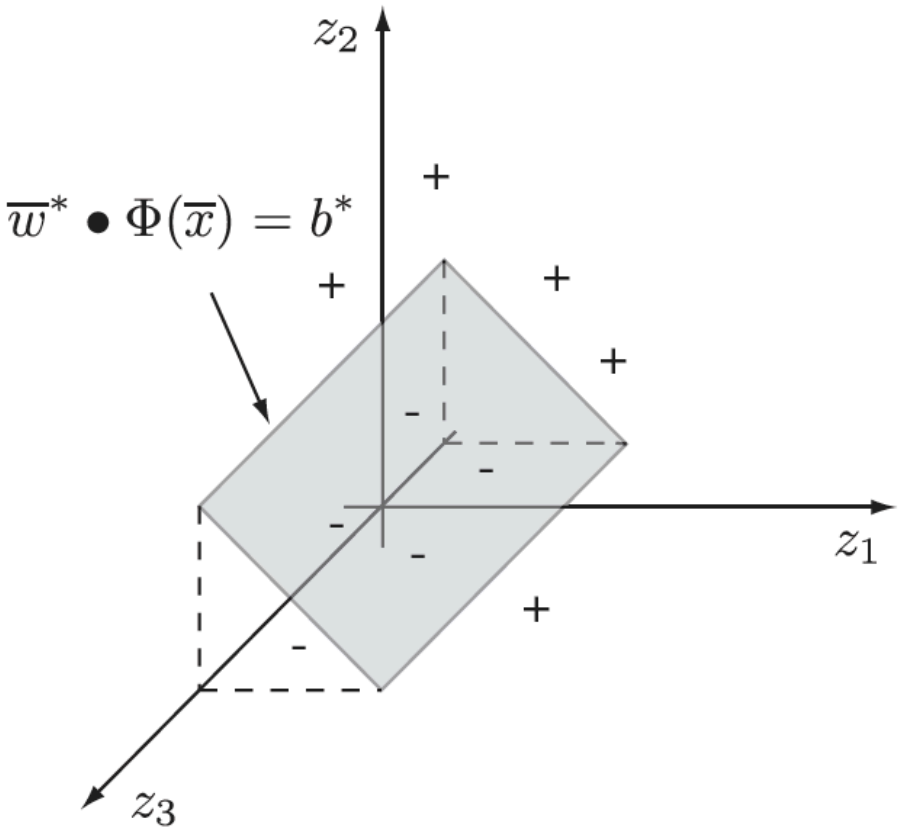
$$\Phi(\mathbf{x}) = \Phi(x_1, x_2) = (x_1^2, x_2^2, \sqrt{2}x_1x_2) = (z_1, z_2, z_3) = \mathbf{z}$$

where $\mathbf{x} \in \mathbb{R}^2, \mathbf{z} \in \mathbb{R}^3$

$$\mathbf{w}^* \cdot \Phi(\mathbf{x}) = b^*$$

$$\mathbf{w}^* \triangleq (1, 1, 0), b^* \triangleq 1$$

$$\Rightarrow z_1 + z_2 = 1$$



classification in feature space

$$\hat{f}(\mathbf{x}) = sgn(\mathbf{w}^* \cdot \Phi(\mathbf{x}) - b^*)$$
$$= sgn(\mathbf{w}^* \cdot \mathbf{z} - b^*)$$
$$= sgn\left(\sum_{i=1}^{d} w_i^* z_i - b^*\right)$$

dual representation

$$\mathbf{w}^* = \sum_{i=1}^{l} \alpha_i^* y_i \Phi(\mathbf{x}_i)$$

$$\hat{f}(\mathbf{x}) = sgn(\mathbf{w}^* \cdot \Phi(\mathbf{x}) - b^*)$$

$$= sgn(\sum_{i=1}^{l} \alpha_i^* y_i \Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}) - b^*)$$

$$\rightarrow sgn(\sum_{i=1}^{l} \alpha_i^* y_i (\mathbf{x}_i \cdot \mathbf{x})^2 - b^*)$$

example kernel function

$$\Phi(\mathbf{x}) = (x_1^2, x_2^2, \sqrt{2}x_1x_2)$$

$$\Phi(\mathbf{x}) \cdot \Phi(\mathbf{y}) = (x_1^2, x_2^2, \sqrt{2}x_1x_2) \cdot (y_1^2, y_2^2, \sqrt{2}y_1y_2)$$

$$= x_1^2y_1^2 + x_2^2y_2^2 + 2x_1x_2y_1y_2$$

$$= (x_1y_1 + x_2y_2)(x_1y_1 + x_2y_2)$$

$$= (\mathbf{x} \cdot \mathbf{y})(\mathbf{x} \cdot \mathbf{y})$$

$$= (\mathbf{x} \cdot \mathbf{y})^2$$

kernel function

$$k(\mathbf{x}, \mathbf{y}) = \Phi(\mathbf{x}) \cdot \Phi(\mathbf{y})$$

where $\Phi : \mathbb{R}^n \rightarrow \mathbb{R}^m$ s.t. $m \geq n$

$$\hat{f}(\mathbf{x}) = sgn\left(\sum_{i=1}^{l} \alpha_i^* y_i k(\mathbf{x}_i, \mathbf{x}) - b^*\right)$$

standard kernels

kernel name	kernel function	free parameters
linear	$k(\mathbf{x}, \mathbf{y}) = \mathbf{x} \cdot \mathbf{y}$	none
homogeneous polynomial	$k(\mathbf{x}, \mathbf{y}) = (\mathbf{x} \cdot \mathbf{y})^d$	$d \geq 2$
non-homogeneous polynomial	$k(\mathbf{x}, \mathbf{y}) = (\mathbf{x} \cdot \mathbf{y} + c)^d$	$d \geq 2, c > 0$
Gaussian	$k(\mathbf{x}, \mathbf{y}) = e^{-(\lvert\mathbf{x}-\mathbf{y}\rvert^2/2\sigma^2)}$	$\sigma > 0$

computation of b^*

$$
\begin{aligned}
b^* &= \mathbf{w}^* \cdot \Phi(\mathbf{x}_{sv+}) - 1 \\
&= \sum_{i=1}^{l} \alpha_i^* y_i \Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}_{sv+}) - 1 \\
&= \sum_{i=1}^{l} \alpha_i^* y_i k(\mathbf{x}_i, \mathbf{x}_{sv+}) - 1
\end{aligned}
$$

nonlinear SVM: training

$$\boldsymbol{\alpha}^* = \arg_{\boldsymbol{\alpha}} \max \left(\sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} \alpha_i \alpha_j y_i y_j k(\mathbf{x}_i, \mathbf{x}_j) \right)$$

$$\text{s.t.} \quad \sum_{i=1}^{l} \alpha_i y_i = 0$$

$$\alpha_i \geq 0, i = 1, \ldots, l$$

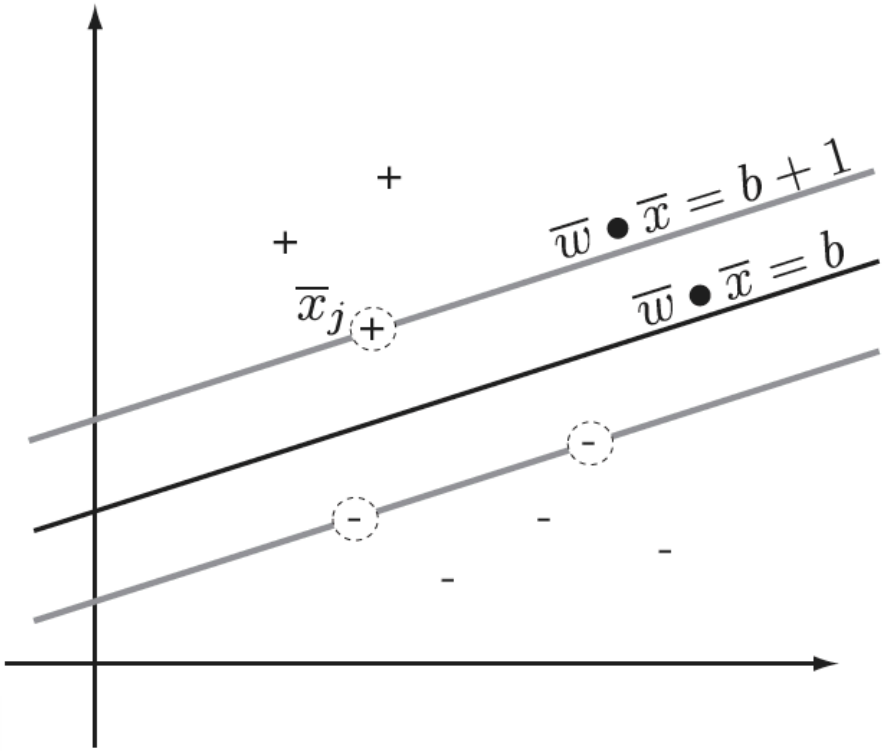
maximum margin classifier

maximum-margin classifier: $\hat{f}(\mathbf{x}) = sgn(\mathbf{w}^* \cdot \mathbf{x} - b^*)$

primal optimization problem: $\min \phi(\mathbf{w}, b) = \min \frac{1}{2}\mathbf{w} \cdot \mathbf{w}$

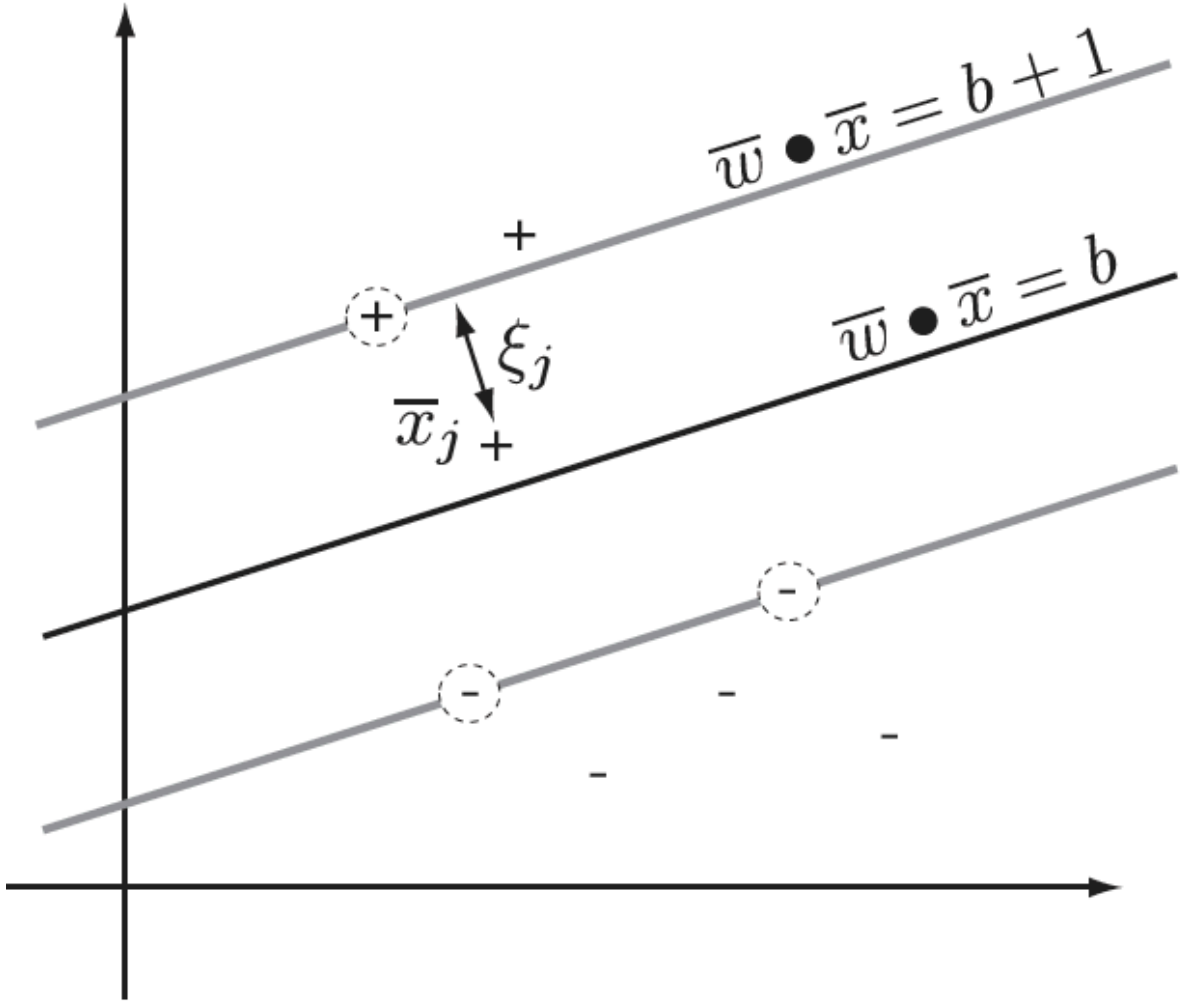
s.t. $y_i(\mathbf{w} \cdot \mathbf{x}_i - b) - 1 \geq 0, i = 1, \ldots, l$

$(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_l, y_l) \in \mathbb{R}^n \times \{+1, -1\}$



introducing slack variables

$$y_i(\mathbf{w} \cdot \mathbf{x}_i - b) + \xi_i - 1 \geq 0, i = 1, \ldots, l$$



soft-margin optimization problem

$$\min_{\mathbf{w},\boldsymbol{\xi},b} \phi(\mathbf{w},\boldsymbol{\xi},b) = \min_{\mathbf{w},\boldsymbol{\xi},b} \left(\frac{1}{2}\mathbf{w}\cdot\mathbf{w} + C\sum_{i=1}^{l}\xi_i \right)$$

$$\text{s.t. } y_i(\mathbf{w}\cdot\mathbf{x}_i - b) + \xi_i - 1 \geq 0, i = 1,\ldots,l$$

$$\xi_i \geq 0, i = 1,\ldots,l$$

$$\boldsymbol{\xi} = (\xi_1,\ldots,\xi_l), C > 0$$

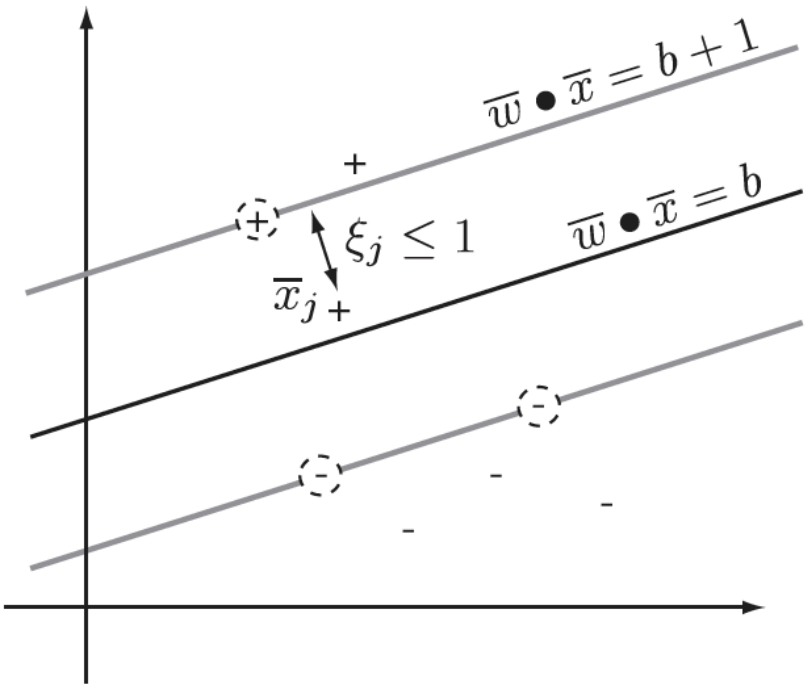
$$(\mathbf{x}_1, y_1),\ldots,(\mathbf{x}_l, y_l) \in \mathbb{R}^n \times \{+1,-1\}$$

$$\text{note: } \hat{f}(\mathbf{x}) = sgn(\mathbf{w}^*\cdot\mathbf{x} - b^*)$$

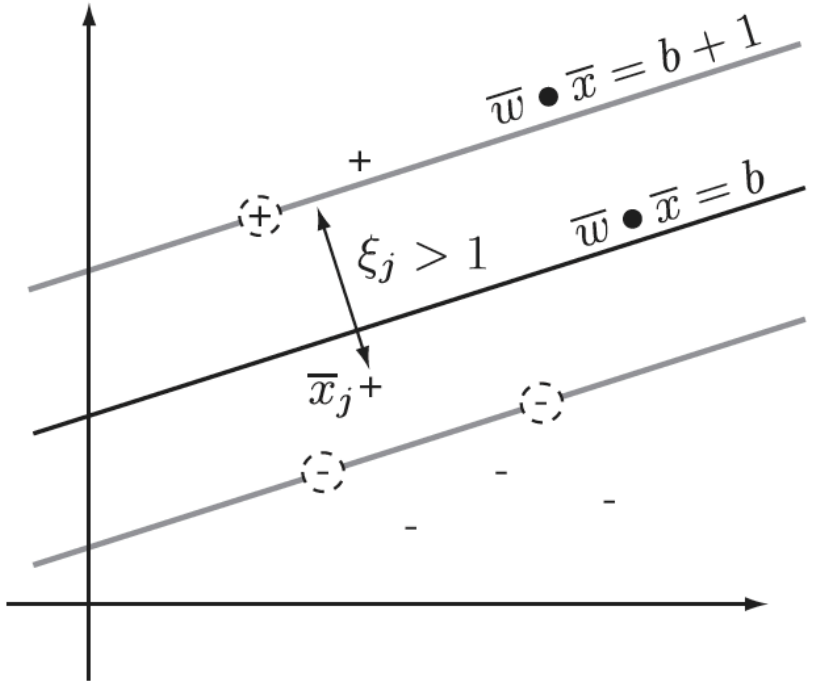
soft-margin misclassifications

for $(\mathbf{x}_j, +1),\ \mathbf{w} \cdot \mathbf{x}_j = b + (1 - \xi_j)$

$\xi_j \leq 1$



$\xi_j > 1$



dual setting for soft-margin classifiers

$$L(\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{w}, \boldsymbol{\xi}, b) = \frac{1}{2}\mathbf{w} \cdot \mathbf{w} + C \sum_{i=1}^{l} \xi_i - \sum_{i=1}^{l} \alpha_i (y_i(\mathbf{w} \cdot \mathbf{x}_i - b) + \xi_i - 1) - \sum_{i=1}^{l} \beta_i \xi_i$$

Lagrangian optimization problem

$$\max_{\boldsymbol{\alpha},\boldsymbol{\beta}} \min_{\mathbf{w},\boldsymbol{\xi},b} L(\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{w}, \boldsymbol{\xi}, b)$$

$$\text{s.t. } \alpha_i \geq 0$$

$$\beta_i \geq 0$$

$$i = 1, \ldots, l$$

KKT conditions

$\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*, \mathbf{w}^*, \boldsymbol{\xi}^*,$ and b^* will satisfy for $i = 1, \ldots, l$

$$\frac{\partial L}{\partial \mathbf{w}}(\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{w}^*, \boldsymbol{\xi}, b) = \mathbf{0}$$

$$\frac{\partial L}{\partial \xi_i}(\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{w}, \boldsymbol{\xi}^*, b) = 0$$

$$\frac{\partial L}{\partial b}(\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{w}, \boldsymbol{\xi}, b^*) = 0$$

$$\alpha_i^*(y_i(\mathbf{w}^* \cdot \mathbf{x}_i - b^*) + \xi_i^* - 1) = 0$$

$$\beta_i^* \xi_i^* = 0$$

$$y_i(\mathbf{w}^* \cdot \mathbf{x}_i - b^*) + \xi_i^* - 1 \geq 0$$

$$\alpha_i^* \geq 0, \beta_i^* \geq 0, \xi_i^* \geq 0$$

optimal value

$$\max_{\boldsymbol{\alpha},\boldsymbol{\beta}} \min_{\mathbf{w},\boldsymbol{\xi},b} L(\boldsymbol{\alpha},\boldsymbol{\beta},\mathbf{w},\boldsymbol{\xi},b) = L(\boldsymbol{\alpha}^*,\boldsymbol{\beta}^*,\mathbf{w}^*,\boldsymbol{\xi}^*,b^*)$$

$$= \frac{1}{2}\mathbf{w}^* \cdot \mathbf{w}^* + C\sum_{i=1}^{l} \xi_i^*$$

Langrangian dual

$$\frac{\partial L}{\partial \mathbf{w}}(\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{w}^*, \boldsymbol{\xi}, b) = \mathbf{w}^* - \sum_{i=1}^{l} \alpha_i y_i \mathbf{x}_i = \mathbf{0}$$

$$\Rightarrow \mathbf{w}^* = \sum_{i=1}^{l} \alpha_i y_i \mathbf{x}_i$$

$$\frac{\partial L}{\partial b}(\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{w}, \boldsymbol{\xi}, b^*) = \sum_{i=1}^{l} \alpha_i y_i = 0 \quad \Rightarrow \sum_{i=1}^{l} \alpha_i y_i = 0$$

$$\frac{\partial L}{\partial \xi_i}(\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{w}, \boldsymbol{\xi}^*, b) = C - \alpha_i - \beta_i = 0$$

$$\Rightarrow \alpha_i = C - \beta_i \quad \Rightarrow 0 \leq \alpha_i \leq C$$

soft margin Lagrangian dual

$$\max_{\boldsymbol{\alpha}} \phi'(\boldsymbol{\alpha}) = \max_{\boldsymbol{\alpha}} \left(\sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} \alpha_i \alpha_j y_i y_j \mathbf{x}_i \cdot \mathbf{x}_j \right)$$

$$\text{s.t.} \quad \sum_{i=1}^{l} \alpha_i y_i = 0$$

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

$$i = 1, \ldots, l$$

computation of b^*

pick a support vector with a zero-valued slack variable

$$(\mathbf{x}_{sv+}, +1) \text{ s.t. } \xi^*_{sv+} = 0 \Rightarrow 0 < \alpha^*_{sv+} < C$$

$$b^* = \mathbf{w}^* \cdot \mathbf{x}_{sv+} - 1 + \xi_{sv+}$$

$$= \sum_{i=1}^{l} \alpha_i^* y_i \mathbf{x}_i \cdot \mathbf{x}_{sv+} - 1$$

$$\because \quad \mathbf{w}^* = \sum_{i=1}^{l} \alpha_i y_i \mathbf{x}_i$$

readings

- Searching Social Media Streams on the Web
- Finding Advertising Keywords on Web Pages
- Optimizing Search Engines using Clickthrough Data
- Hierarchical Document Categorization with Support Vector Machines
- Predicting Structured Objects with Support Vector Machines
- Hidden Markov Support Vector Machines

references

- N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*, Cambridge University Press, 2000.
- L. Hamel, *Knowledge Discovery with Support Vector Machines*, Wiley, 2009.