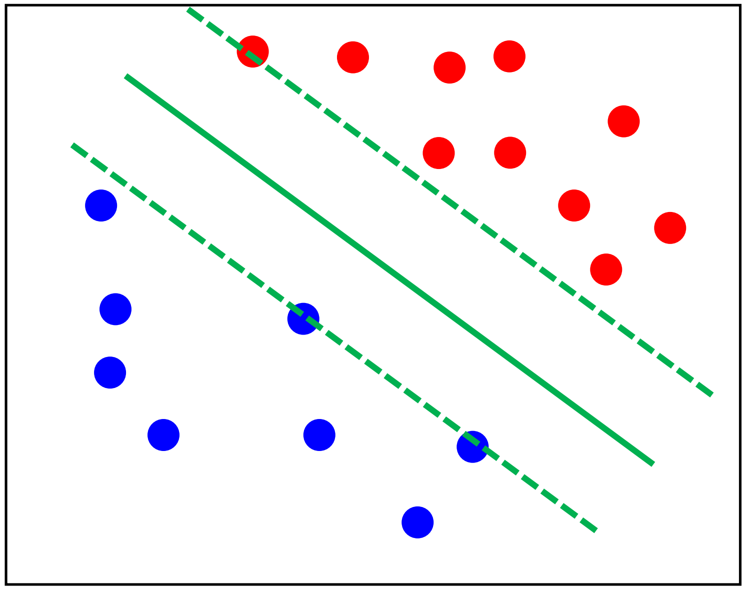
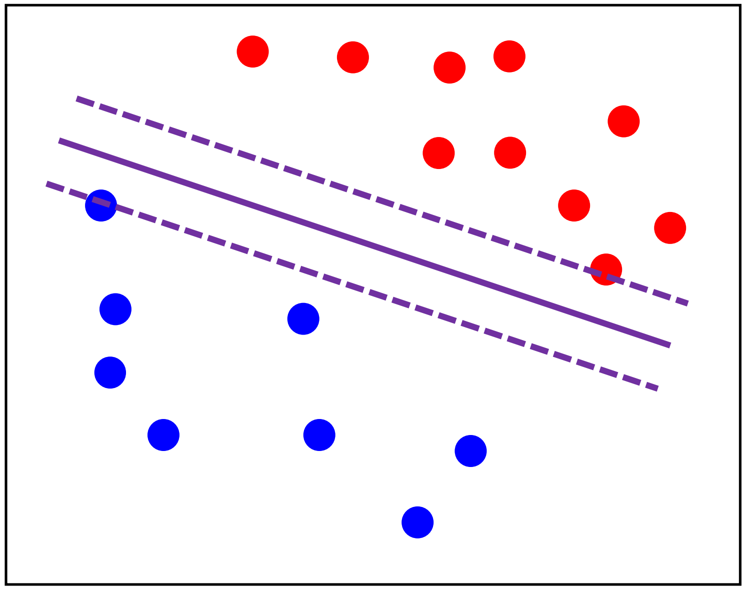
Support Vector Machine

Hanwool Jeong

hwjeong@kw.ac.kr

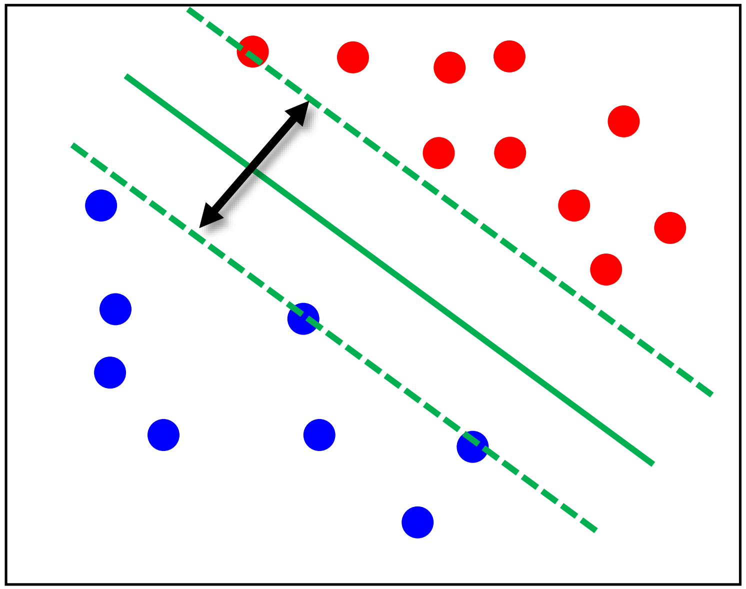
Large Margin vs. Small margin Classification

- Which is better? The optimal boundary for classification?
- The sample closest to the boundary in the dataset determines the boundary. Any other sample does not have effect.
  ➔Sample determining the boundary is called **Support vector**

Key Idea of
Support Vector Machine (SVM)

- Key idea of SVM is to maximize the margin, which generally improves the accuracy of classification model.
- Then, we should formulate the margin then find the optimal parameter that maximize the margin.



1) Mathematically expressing the margin
2) Maximize the expressed term

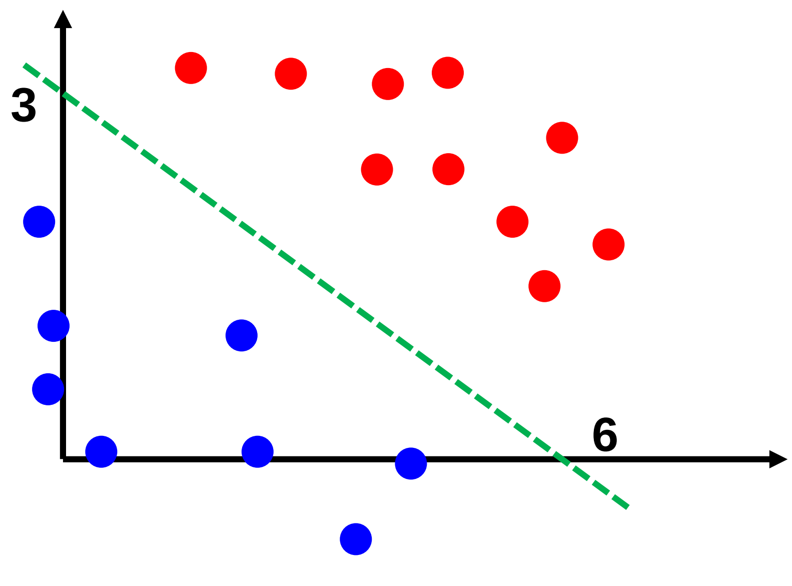
Boundary for Binary Classification

- First, let's assume linearly separable situation.
- Boundary hyperplane equation:

$$h(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + b = 0$$

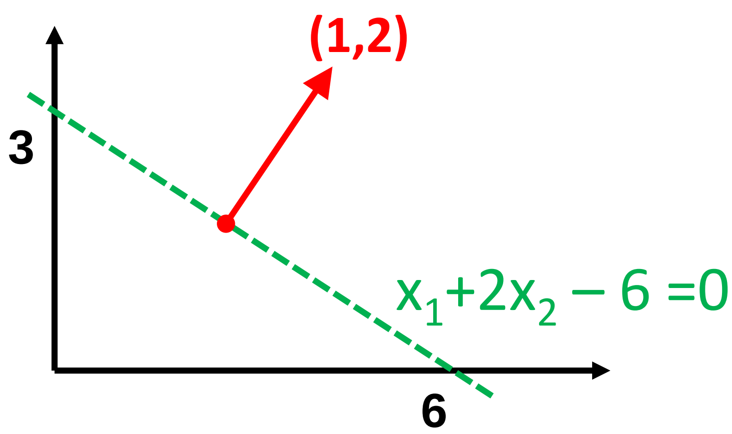
where $\mathbf{x}$ is D dimensional point in the feature space.

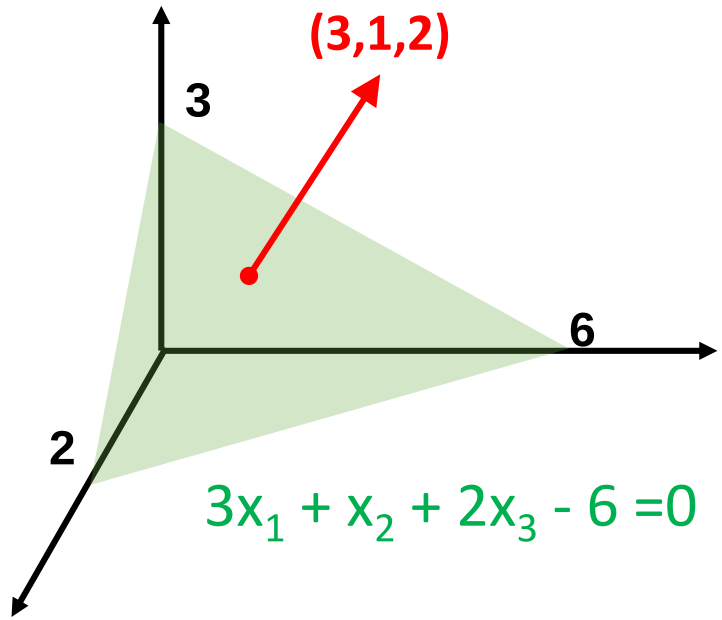
- We do classify as:



Looking into Hyperplane Expression $\mathbf{w}^T\mathbf{x} + b = 0$

1) Graphical/geometrical meaning of **w**
2) Is the mathematical expression unique?





General Form of Hyperplane

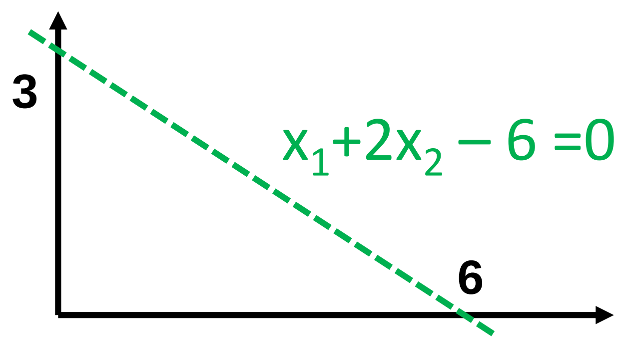
- General form : h(**x**) = 0 where h(**x**) = $ax_1 + bx_2 + c$
- h(**x**) is not unique for the given hyperplane.

$$h(x) = kx_1 + 2kx_2 - 6k$$

- We can freely choose h(**x**)=0 for mathematical convenience.

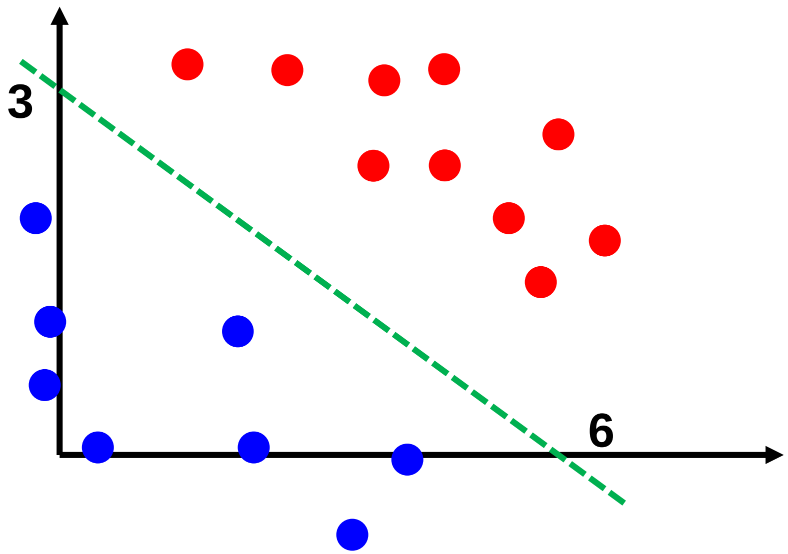
$$x_1+2x_2 – 6 =0 \quad \textbf{vs.} \quad 2x_1+4x_2 – 12 =0$$

- We can set any one condition of $h(x_0)=h_0$, where x_0 is not on the hyperplane, for the mathematical convenience.



Now, What do We have to do? Don't Forget Our Goal

- Mathematically expressing the margin.
- We should find out the distance from hyperplane to arbitrary point **x**.



Distance Point to Hyperplane

- Boundary hyperplane equation:

$$h(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + b = 0$$

- How can you derive the distance from the hyperplane to arbitrary point **x**?
- Considering **w** means the normal vector of the plane, we can utilize it. Say the strategy for the derivation.

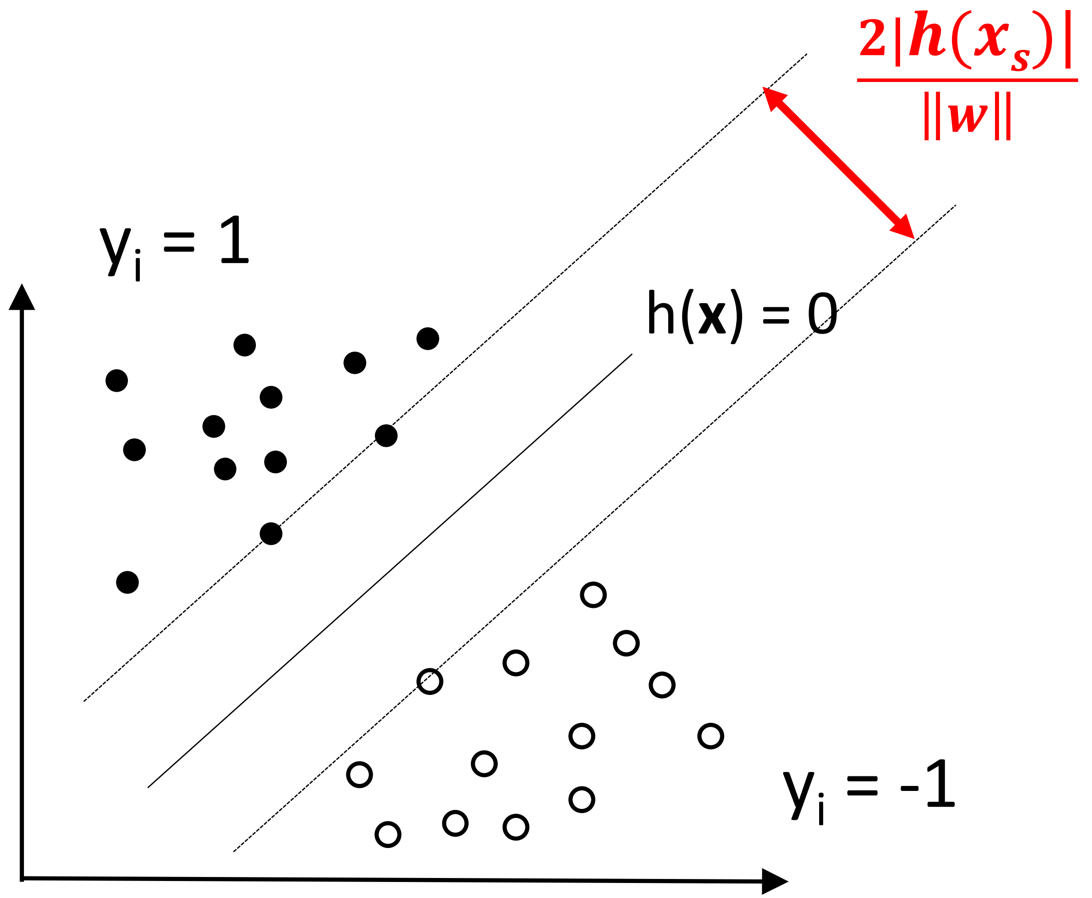
$$d = \frac{|h(\boldsymbol{x}')|}{\|\boldsymbol{w}\|}$$

Example

- Derive (1,-2, 4) to the plane $(3, 2, 6)^T \bullet (x_1, x_2, x_3) = 5$

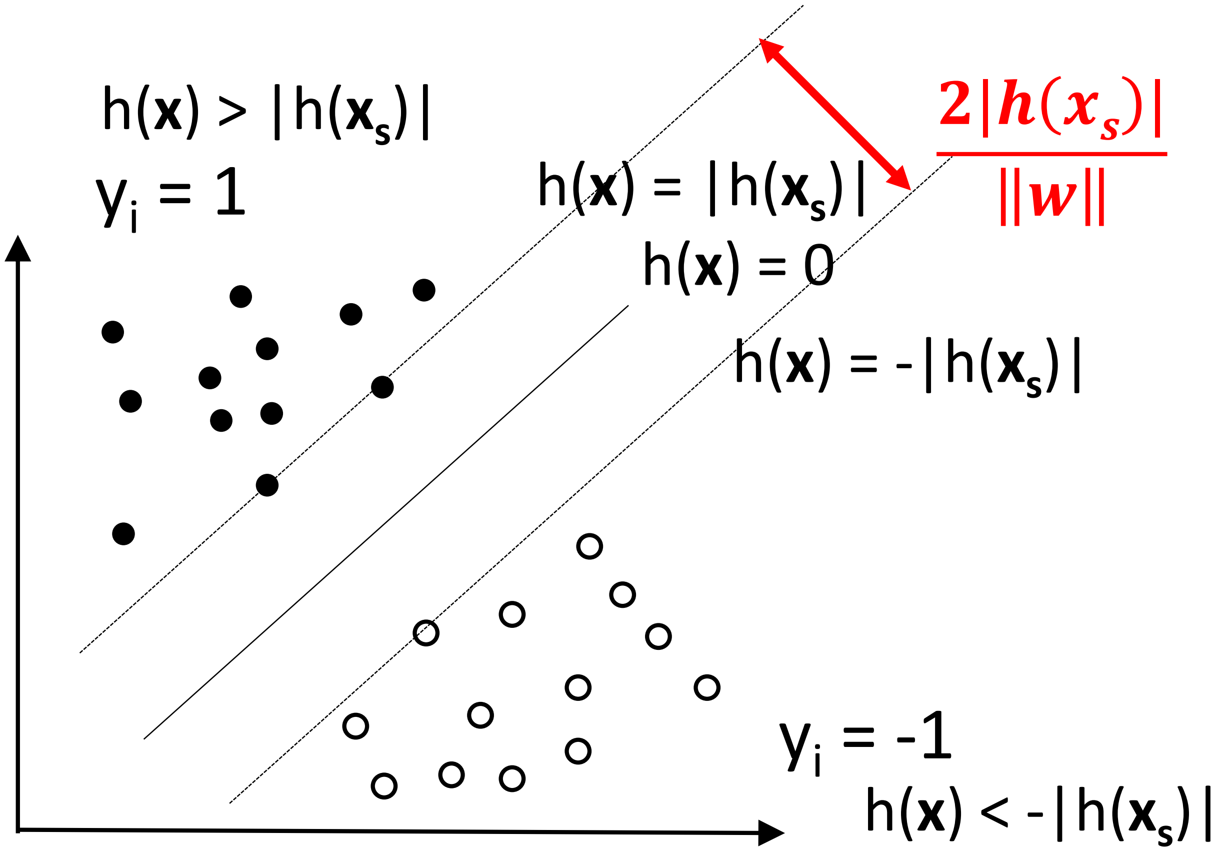
Let's See Training Set

- Maximizing the margin, is that all?
- Is there any additional **constraint** for the hyperplane?



Yes! Boundary Should Properly Classify the Dataset

- Can you express it mathematically? **"Classify properly"**
- Inserting all data into h(**x**), then check…

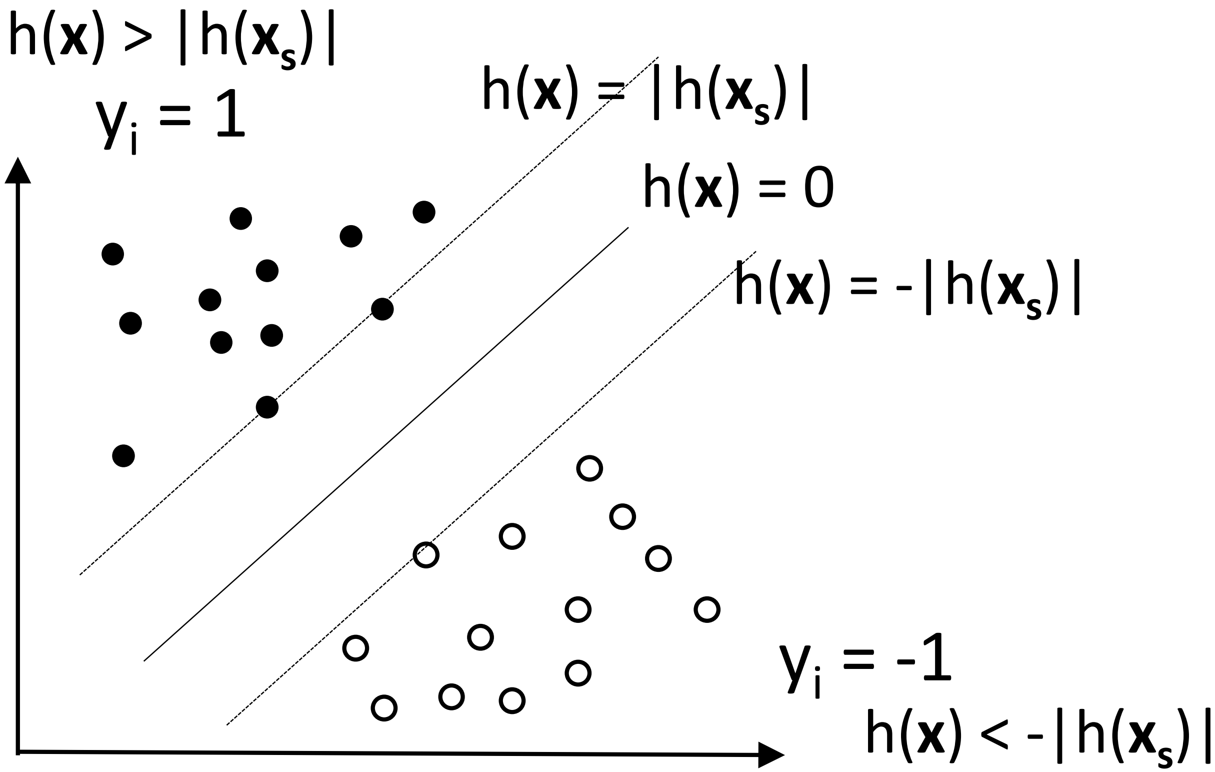


Mathematical Expression for Proper Classification Constraint

- Revisit the constraints

$h(\mathbf{x_i}) = \mathbf{w}^T\mathbf{x_i} + b \geq |h(\mathbf{x_s})|$ for $\mathbf{x_i}$ if $y_i = 1$
$h(\mathbf{x_i}) = \mathbf{w}^T\mathbf{x_i} + b \leq -|h(\mathbf{x_s})|$ for $\mathbf{x_i}$ if $y_i = -1$

$$\Rightarrow \quad y_i(\mathbf{w}^T\mathbf{x_i} + b) \geq |h(\mathbf{x_s})|$$



Now We Complete the Optimization Problem

- Maximize the margin
- With the constraint of the proper classification

$$\text{maximize} \quad d = \frac{2|h(x_s)|}{\|\boldsymbol{w}\|}$$

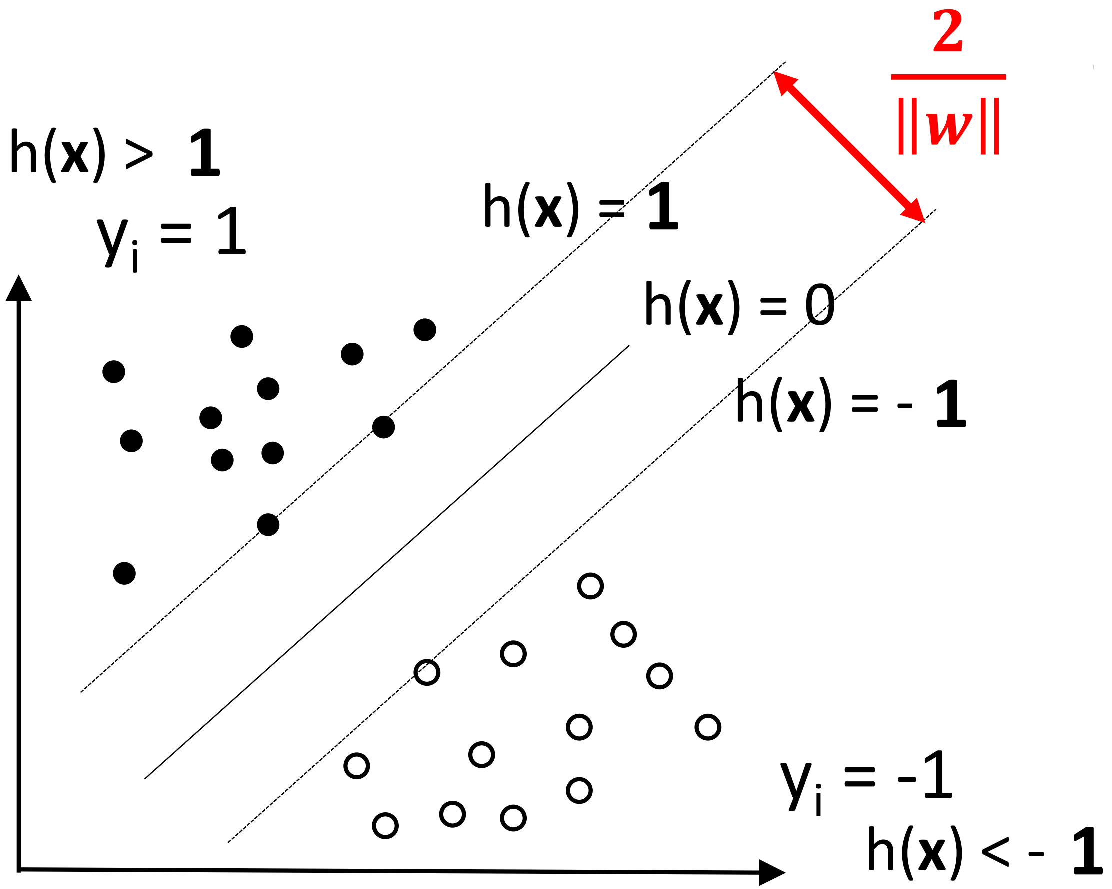
$$\text{subject to} \quad (\mathbf{w_i}^T\mathbf{x_i} + b)y_i \geq |h(\mathbf{x_s})|$$

- Don’t forget we can choose any $h(x_0) = h_0$, where x_0 is not on the hyperplane, for the mathematical convenience.
- How about setting

$$h(\mathbf{x_s}) = 1$$

Let $h(x_s) = 1$

- To uniquely decide the form of h(**x**)



Finally Decided Optimized Problem

- Optimization for SVM boundary hyperplane.

$$\text{Maximize} \quad d = \frac{2}{||\boldsymbol{w}||}$$

$$\text{subject to} \quad (\mathbf{w_i}^T\mathbf{x_i} + b)y_i \geq \mathbf{1}$$

- Oh, Lagrange multiplier! You remember?

Revisit KKT Condition w/ Lagrange Multiplier

- Suppose that

$$\text{Minimize } f(\mathbf{x})$$

$$\text{subject to } \mathbf{g_i(x) \leq 0} \quad \text{for } i = 1, \ldots, M$$

- Auxiliary function

$$L = f(\mathbf{x}) + \sum_i^M \mu_i g(\mathbf{x})$$

- Then solve not only $\boldsymbol{\nabla_{x,\mu} L = 0}$ but also

$$\boldsymbol{\mu_i g(x^*) = 0}$$

$$\boldsymbol{\mu_i \geq 0}$$

Applying Lagrange Multiplier

- For the mathematical convenience , it can be reduced to

Maximize $\frac{2}{\|w\|}$ subject to $(\mathbf{w_i}^T\mathbf{x_i} + b)y_i \geq 1$ ➡ Minimize $\frac{\|w\|^2}{2}$ subject to $1 - y_i(\mathbf{w}^T\mathbf{x_i} + b) \leq 0$

- Lagrange auxiliary function is:

$$L(\mathbf{w},b,\boldsymbol{\mu}) = \frac{\|\boldsymbol{w}\|^2}{2} + \sum_{i=1}^{N} \mu_i g(\mathbf{x_i}) = \frac{\|\boldsymbol{w}\|^2}{2} - \sum_{i=1}^{N} \mu_i \{y_i(\mathbf{w}^T\mathbf{x_i} + b) - 1\}$$

- What is the parameter we should determine? **w** and b
- Then, we need to find **w** and b that satisfy $\nabla_{w,b,\mu} L = 0$ with $\mu_i\{y_i(\mathbf{w}^T\mathbf{x_i} + b) - 1\}=0$, and $\mu_i \geq 0$ (KKT condition)

With L(w,b,μ) = $\frac{\|\boldsymbol{w}\|^2}{2}$ -$\sum_{i=1}^{N}$ μ_i {y_i(wTxi + b) − 1},

- Differentiating **L** with ***w*** then equalizing it to 0,

$$\boldsymbol{w} = \sum_{i=1}^{N} \mu_i \, y_i \mathbf{x_i} \qquad (1)$$

- Differentiating **L** with b then equalizing it to 0,

$$\sum_{i=1}^{N} \mu_i \, y_i = 0 \qquad (2)$$

- Before differentiating with μ, let's see KKT condition first,

$$\mu_i \geq 0 \qquad (3)$$

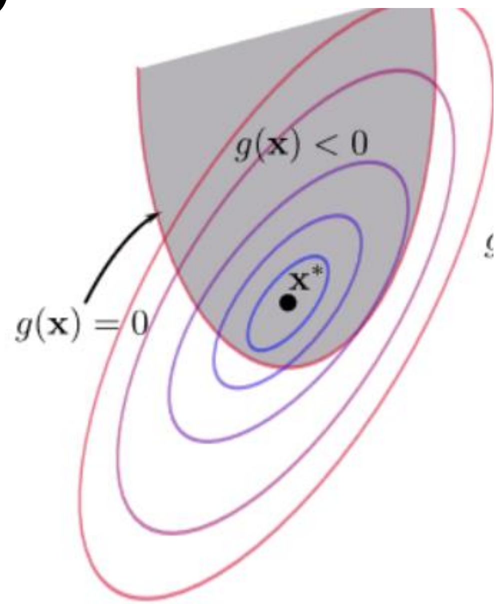
- Do you remember the meaning of $\mu_i = 0$ and $\mu_i > 0$? For i that correspond to support vectors, what would μ_i be?

$$\mu_i\{y_i(\mathbf{w}^T\mathbf{x_i} + b) - 1\} = 0 \qquad (4)$$

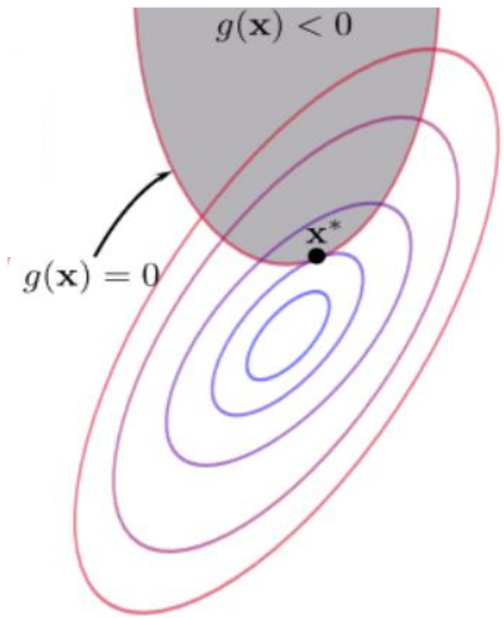
Revisit Meaning of μ·g(x) = 0

1) g(x) ≤ 0 constraint has no meaning ➔ μ = 0

   $g(\mathbf{x}) = 1 - y_i(\mathbf{w}^T\mathbf{x_i} + b)$



2) Optimal point is g(**x**) = 0



Let's Eliminate w & b in L

(1) $\boldsymbol{w} = \sum_{i=1}^{N} \mu_i\, y_i x_i$ (2) $\sum_{i=1}^{N} \mu_i\, y_i = 0$ (3) $\mu_i \geq 0$ (4) $\mu_i\{y_i(\mathbf{w}^T\mathbf{x_i} + b) - 1\} = 0$

$$L(\mathbf{w}, b, \boldsymbol{\mu}) = \frac{\|\boldsymbol{w}\|^2}{2} - \sum_{i=1}^{N} \mu_i \{y_i(\mathbf{w}^T\mathbf{x_i} + b) - 1\}$$

$$= \frac{\|\boldsymbol{w}\|^2}{2} - \sum_{i=1}^{N} \mu_i\, (y_i\mathbf{w}^T\mathbf{x_i} - 1) = \frac{\|\boldsymbol{w}\|^2}{2} - \sum_{i=1}^{N} \mu_i\, (y_i\mathbf{w}^T\mathbf{x_i}) + \sum_{i=1}^{N} \mu_i$$

$$= \sum_{i=1}^{N} \mu_i - \frac{\sum_{i=1}^{N}\sum_{j=1}^{N} \mu_i\mu_j y_i y_j \mathbf{x_i}^T\mathbf{x_j}}{2}$$

$$\frac{\|\boldsymbol{w}\|^2}{2} = \frac{\boldsymbol{w}^T\boldsymbol{w}}{2} = \frac{\boldsymbol{w}^T \sum_{j=1}^{N} \mu_j y_j \mathbf{x_j}}{2} = \frac{\sum_{j=1}^{N} \mu_j y_j \boldsymbol{w}^T \mathbf{x_j}}{2}$$

$$= \frac{\sum_{j=1}^{N} \mu_j\, y_j \left(\sum_{i=1}^{N} \mu_i\, y_i \mathbf{x_i}^T\right)\mathbf{x_j}}{2} = \frac{\sum_{i=1}^{N}\sum_{j=1}^{N} \mu_i\mu_j\, y_i y_j \mathbf{x_i}^T\mathbf{x_j}}{2}$$

Lagrange Auxiliary Function L(μ) & Determining w & b

$$L(\mu) = \sum_{i=1}^{N} \mu_i - \frac{\sum_{i=1}^{N}\sum_{j=1}^{N} \mu_i\mu_j y_i y_j \mathbf{x_i}^T\mathbf{x_j}}{2}$$

- We should find μ making $\nabla_\mu L = 0$. From now, it is not our job.
- Does this maximize or minimize L(μ)?
- Then, the problem is changed into finding **μ**

Maximize L(**μ**)

Subject to $\sum_{i=1}^{N} \mu_i y_i = 0$ and $\mu_i \geq 0$

Determining w & b from Optimal μ

- Then we can find **w***

$$\mathbf{w}^* = \sum_{i=1}^{N} \mu_i y_i \mathbf{x}_i$$

- Note that only for support vectors, $\mu_i > 0$ otherwise $\mu_i = 0$
- How about b? You remember? $\mu_i\{y_i(\mathbf{w}^T\mathbf{x}_i + b) - 1\}=0$

$$\sum_{i=S.V} \{y_i(\mathbf{w}*^T\mathbf{x}_i + b) - 1\}=0 \Rightarrow (1/N_{S.V}) \sum_{i=S.V} (y_i - \mathbf{w}*^T\mathbf{x}_i) = b$$

Prediction?

- With determined **w** and b,

$$\mathbf{w^*} = \sum_{i=1}^{N} \mu_i \, y_i \mathbf{x_i}$$

$$b = (1/N_{S.V}) \sum_{i=S.V.} (y_i - \mathbf{w}*^T \mathbf{x_i})$$

Checkpoints

- Expressing the margin (to be maximized) mathematically
- Expressing the constraint mathematically
- Applying Lagrange multiplier to find out the optimal SVM boundary
- Coming up next : soft boundary & kernel trick