

Machine Learning

Lecture 3: Support Vector Machine

Asst. Prof. Dr. Santitham Prom-on

Department of Computer Engineering, Faculty of Engineering
King Mongkut's University of Technology Thonburi

Topics

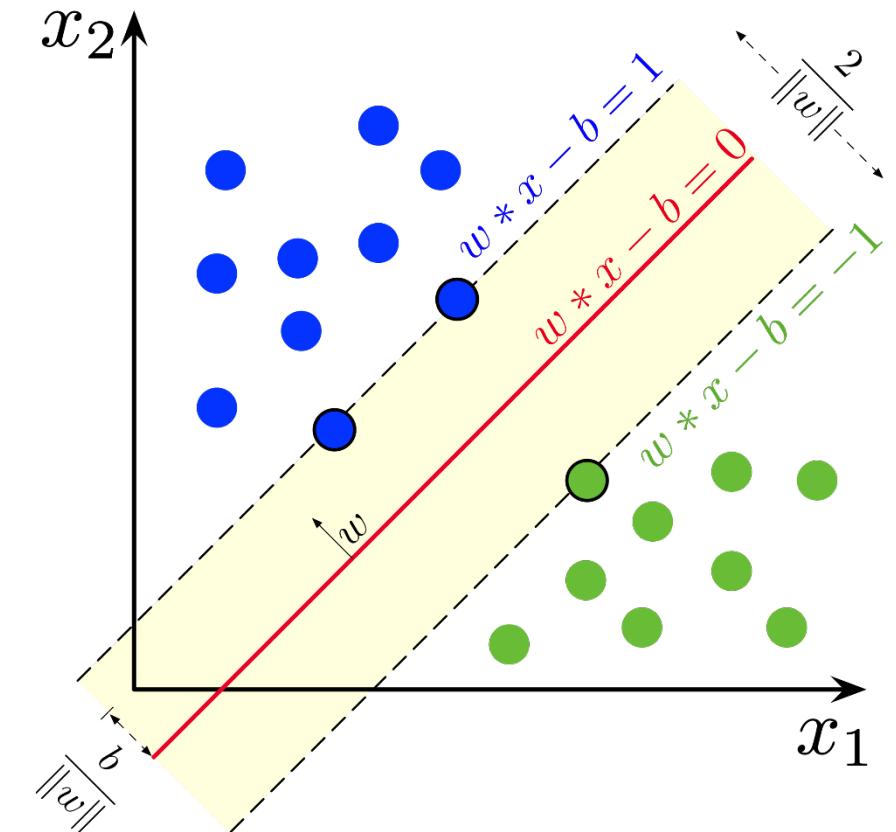
- Linear SVM classification
- Soft margin
- Training SVM
- Nonlinear SVM classification
- Kernel trick
- Hinge Loss

<https://colab.research.google.com/drive/1N4g0QDmCV9r0Hg6eLD7fdcRXXHJLSfM?usp=sharing>

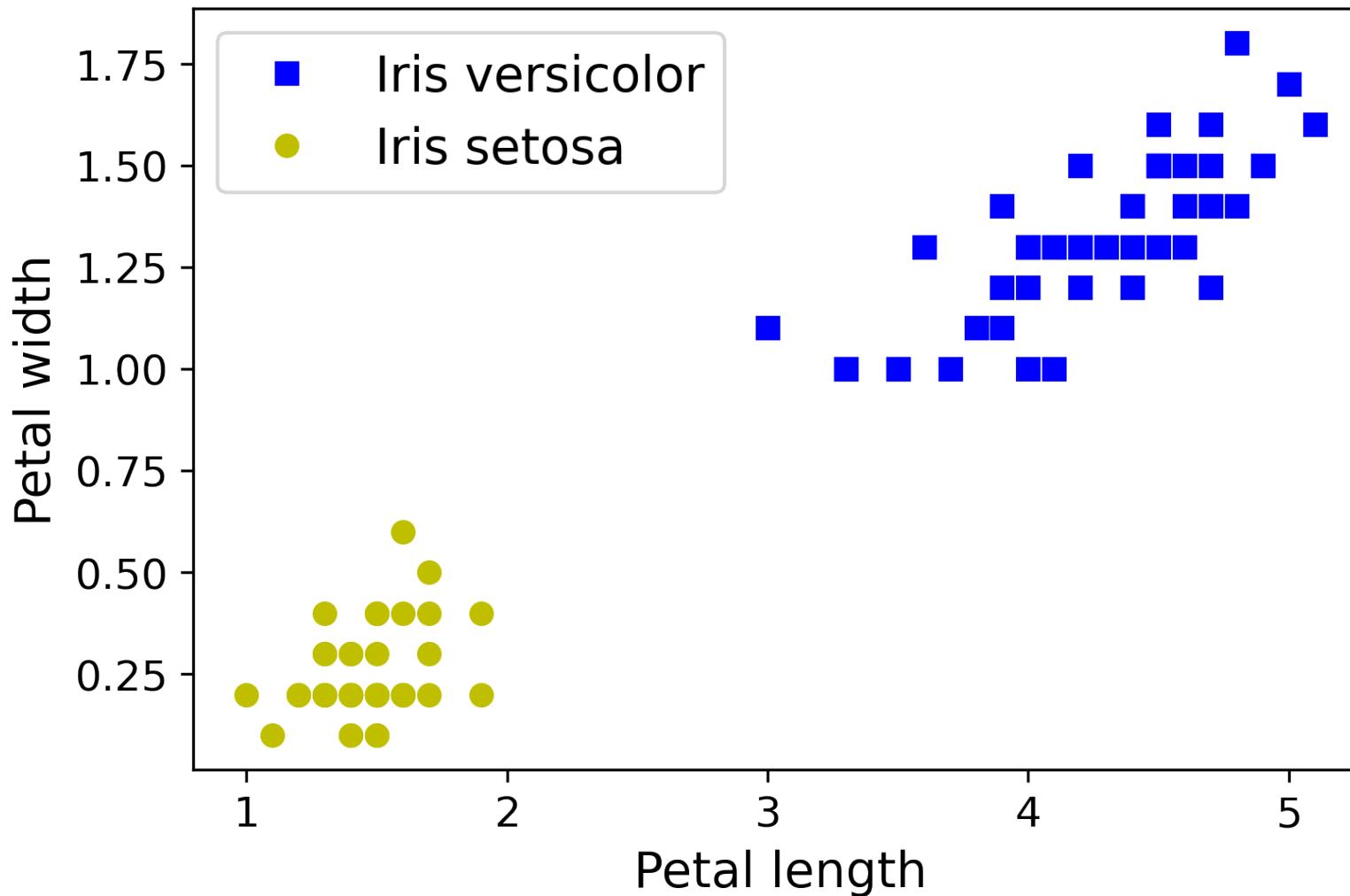
Support Vector Machine (SVM)

- A Support Vector Machine (SVM) is a very powerful and versatile Machine Learning model, capable of performing linear or nonlinear classification, regression, and even outlier detection.
- SVMs are particularly well suited for classification of complex but small- or medium-sized datasets.

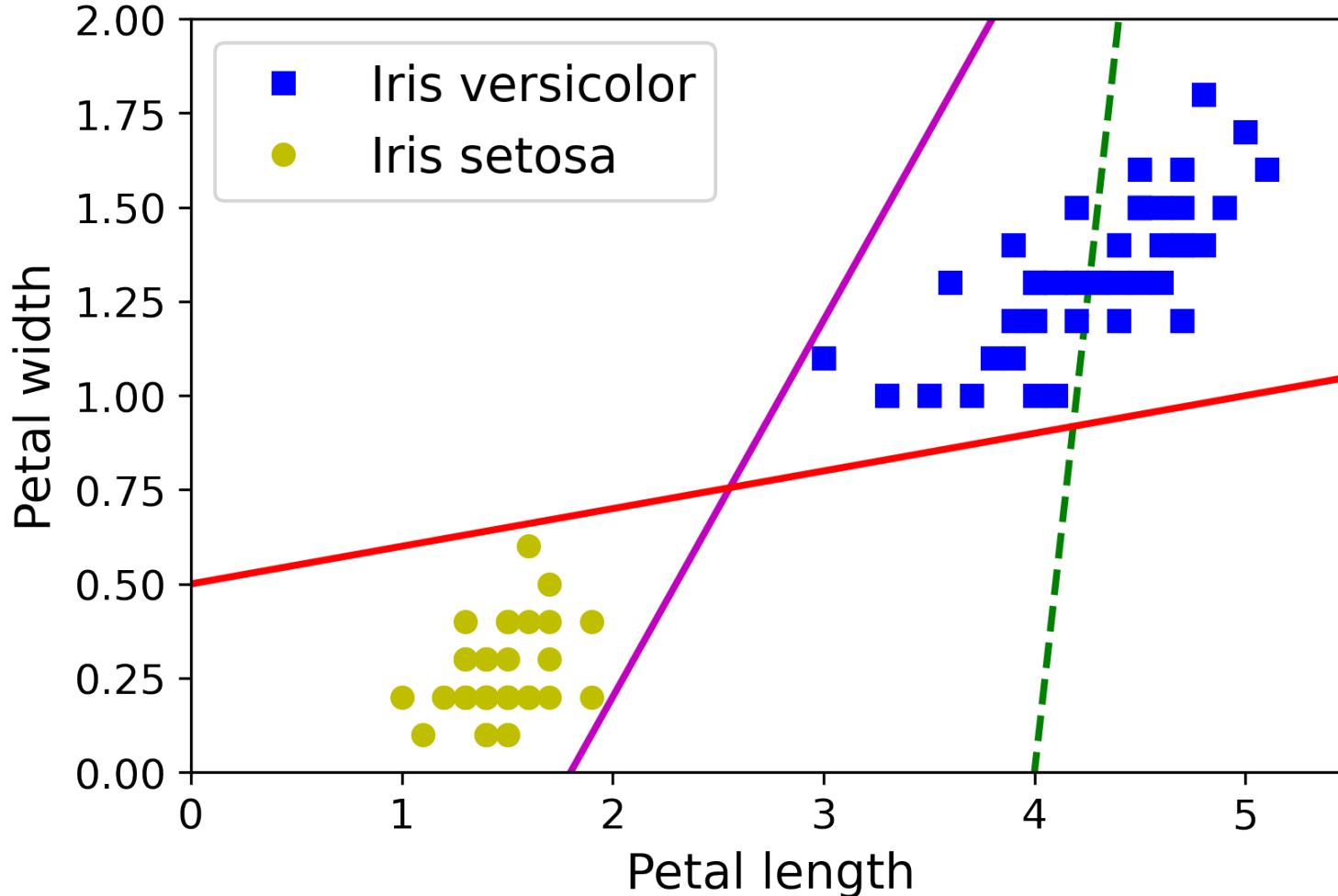
Data / Signal → Feature Extraction → ML → output



Linearly separable



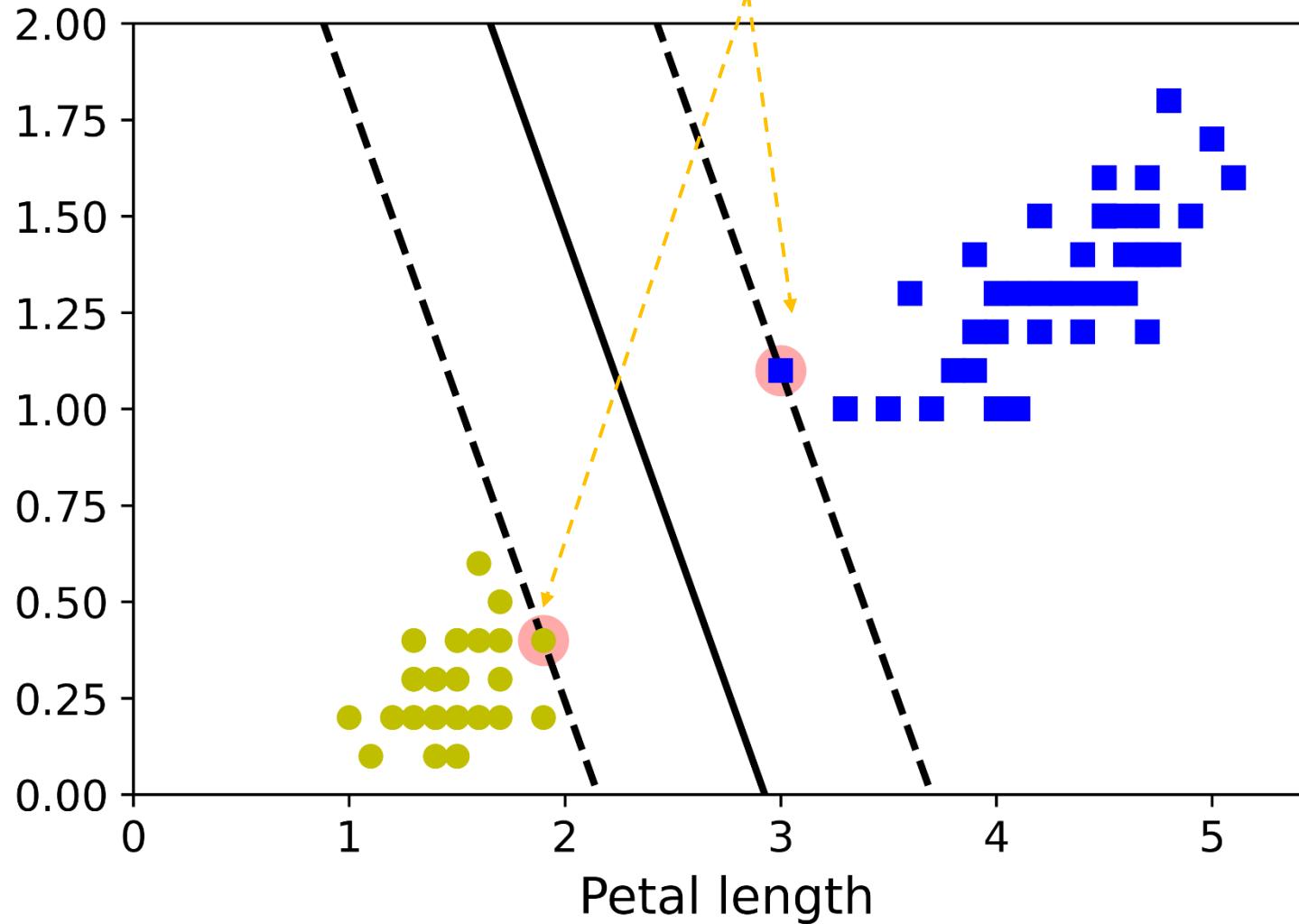
Which line?



Support vectors

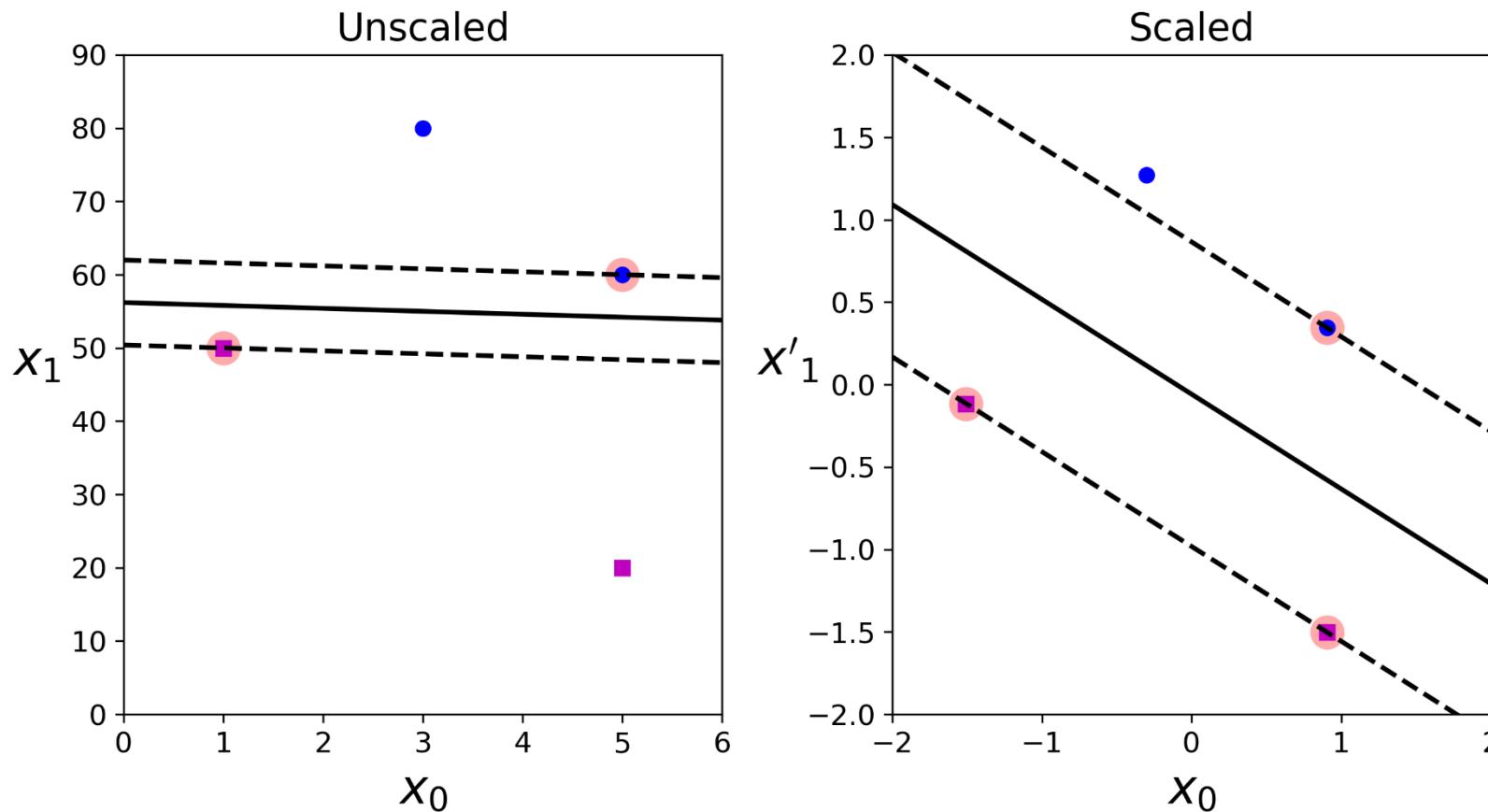
$$\begin{bmatrix} x_i \\ y_i \end{bmatrix}$$

Support vectors



Feature scaling

StandardScaler - data กว้างๆ ก็ต้อง $\rightarrow z\text{ score}$
min-max scaler เมื่อมาบีบ data ที่กว้างๆ จายตัว $\rightarrow 0\dots 1$
MaxAbsScaler $\rightarrow -1\dots +1$
RobustScaler - non-linear \rightarrow คิด IQR



Decision function: linear SVM

$$\hat{y} = \begin{cases} 0 & \text{if } \mathbf{w}^T \mathbf{x} + b < 0, \\ 1 & \text{if } \mathbf{w}^T \mathbf{x} + b \geq 0 \end{cases}$$

$$\mathbf{w} = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \end{bmatrix}$$

- $h = 0$
- $h = \pm 1$

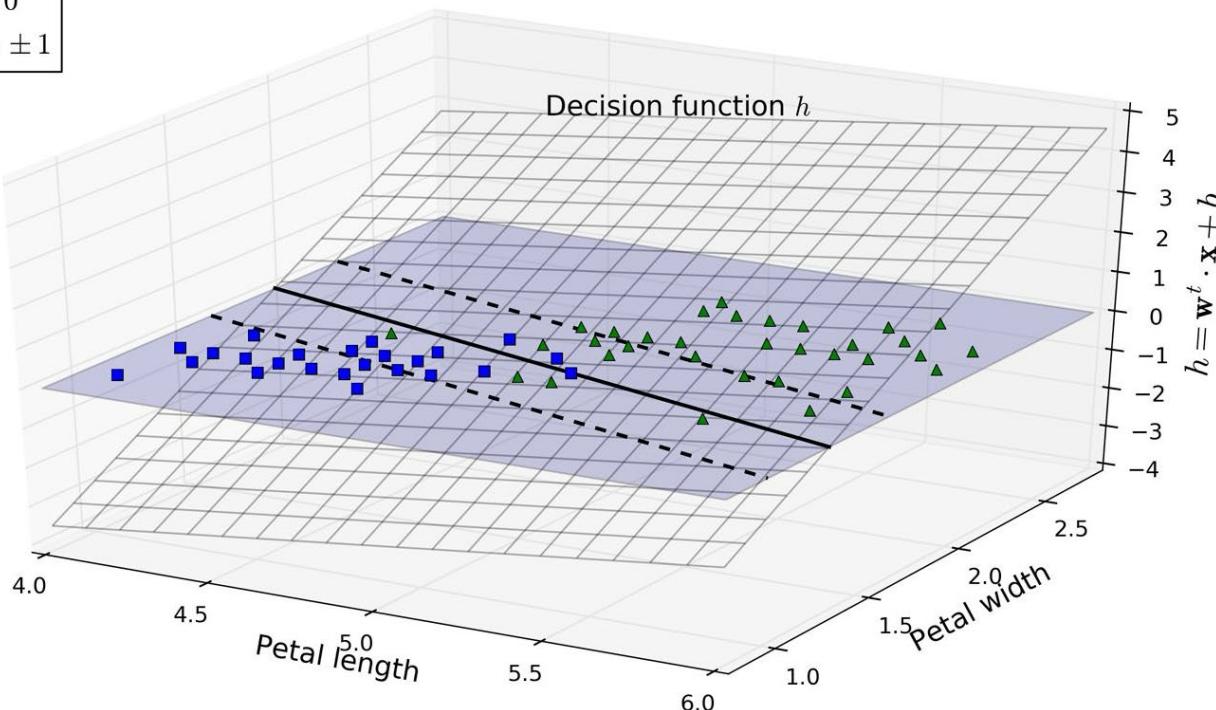


Figure 5-12. Decision function for the iris dataset

Training objective

Regularization
- Ridge
- Lasso

- The smaller the weight vector, the larger the margin

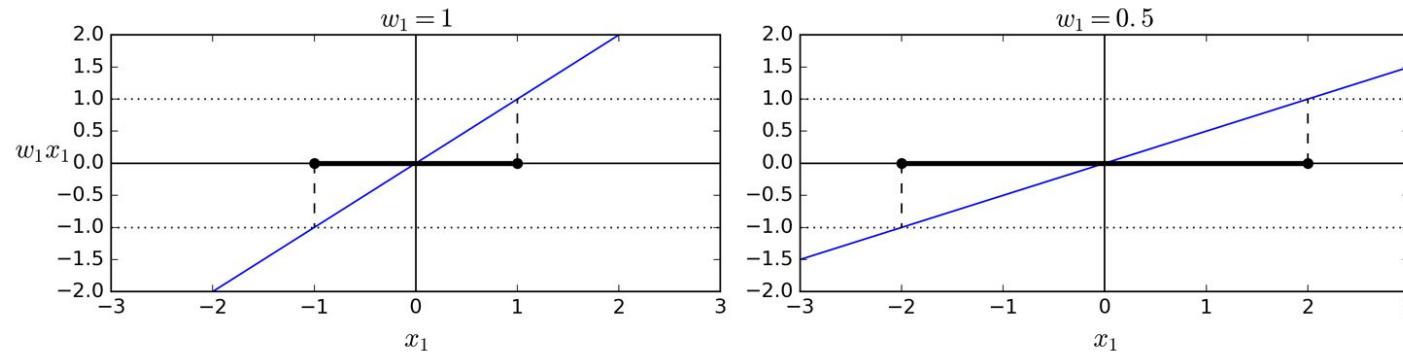


Figure 5-13. A smaller weight vector results in a larger margin

- misclassified point*
- Margin violation. Let $t^{(i)} = -1$ for negative instance (if $y^{(i)} = 0$) and $t^{(i)} = 1$ for positive instance (if $y^{(i)} = 1$)

$$t^{(i)}(\mathbf{w}^T \mathbf{x}^{(i)} + b) \geq 1$$

Hard margin linear classifier

Objective function

$$\hat{y} = \begin{cases} 0; w^T x + b < 0 \\ 1; w^T x + b \geq 0 \end{cases}$$

objective function

$$\underset{\mathbf{w}, b}{\text{minimize}} \quad \frac{1}{2} \mathbf{w}^T \mathbf{w}$$

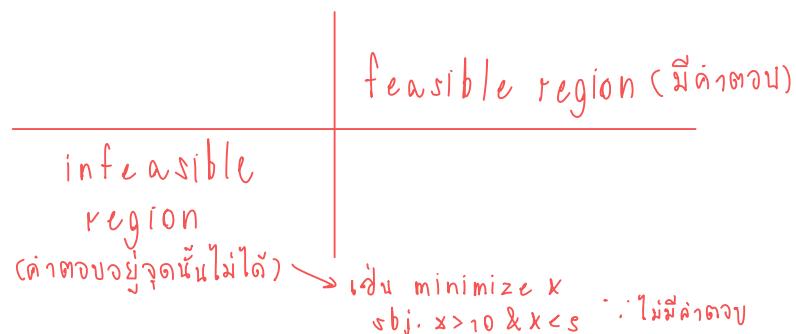
large margin (long weight)
ห้ามห่างใกล้

subject to

$$t^{(i)}(\mathbf{w}^T \mathbf{x}^{(i)} + b) \geq 1 \quad \text{for } i = 1, 2, \dots, m$$

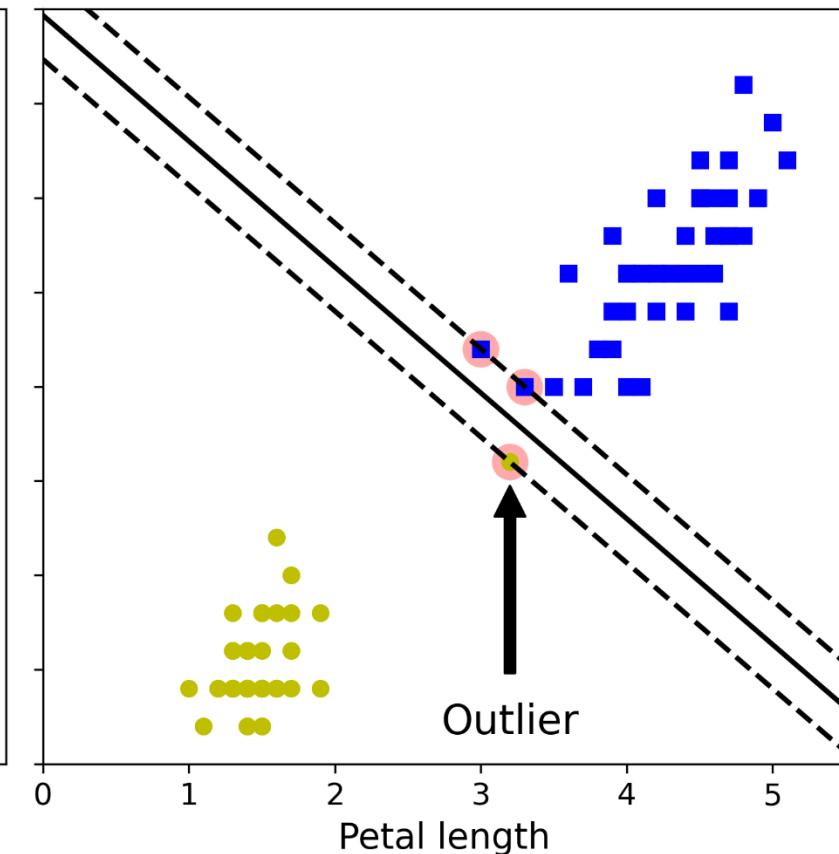
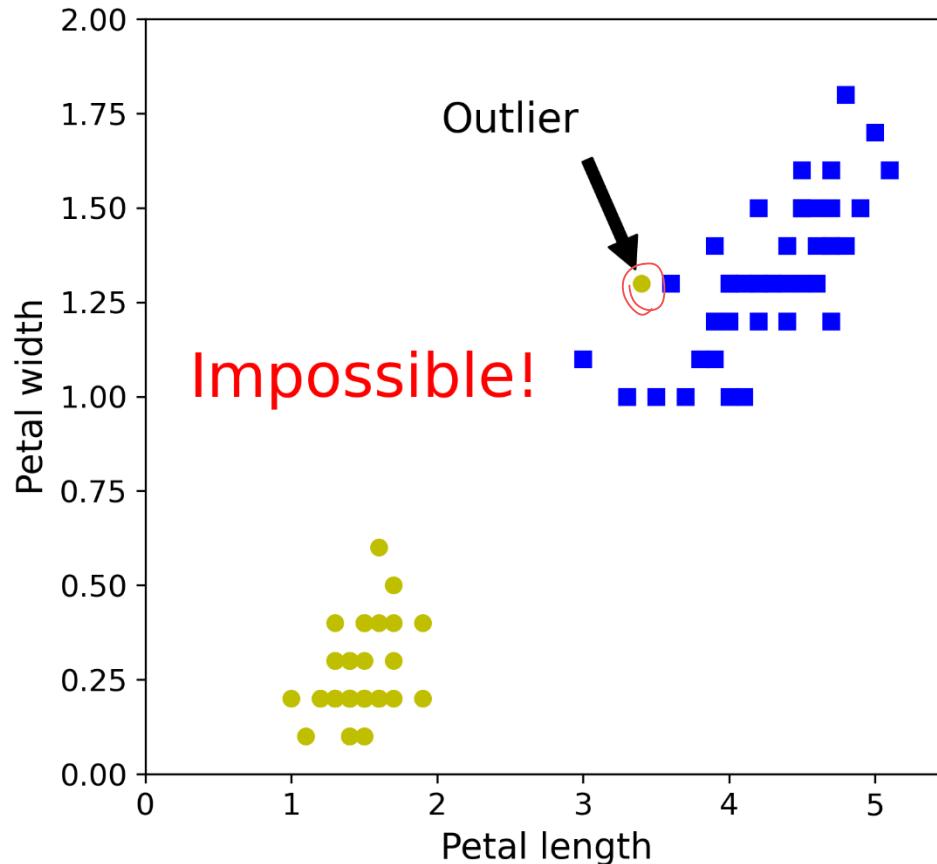
constraints

margin violation

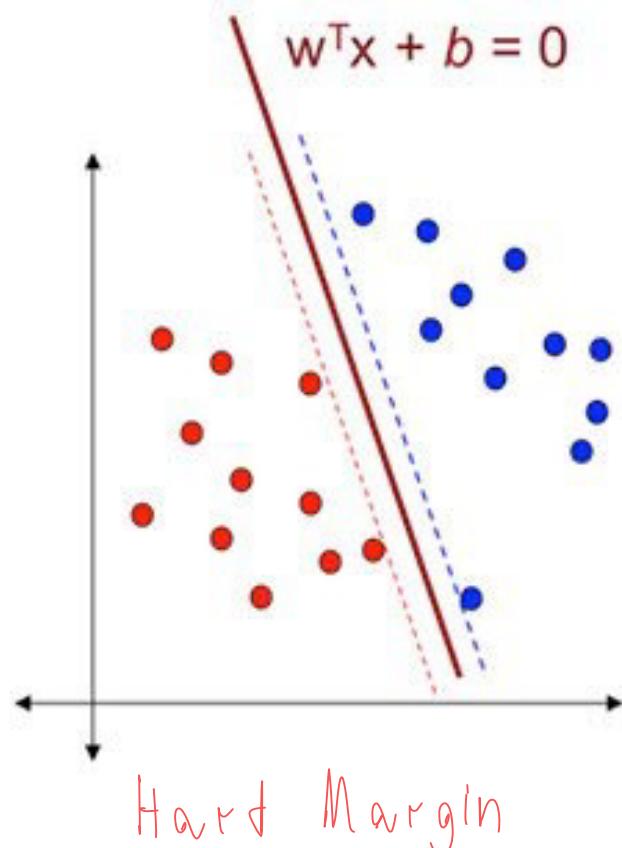


Sensitivity to outliers

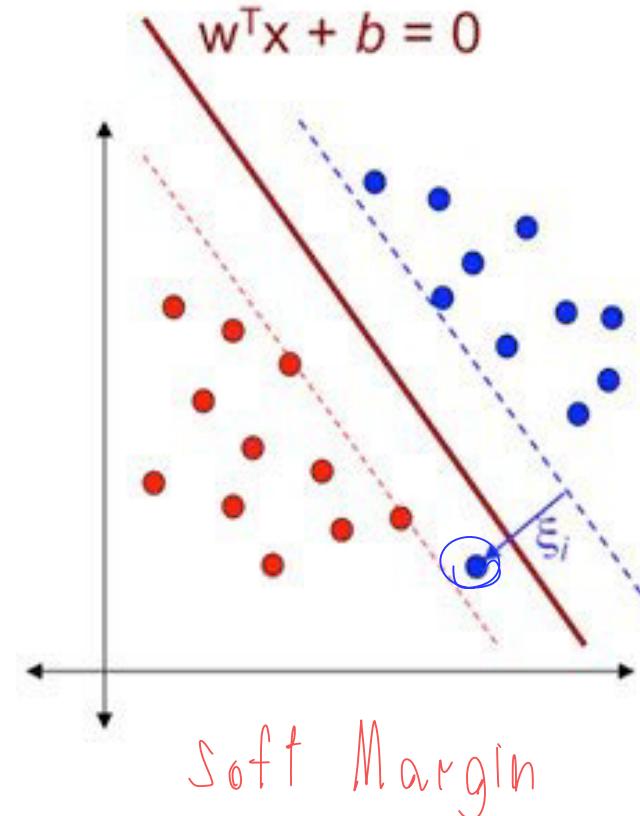
∴ Hard margin



Soft margin classification



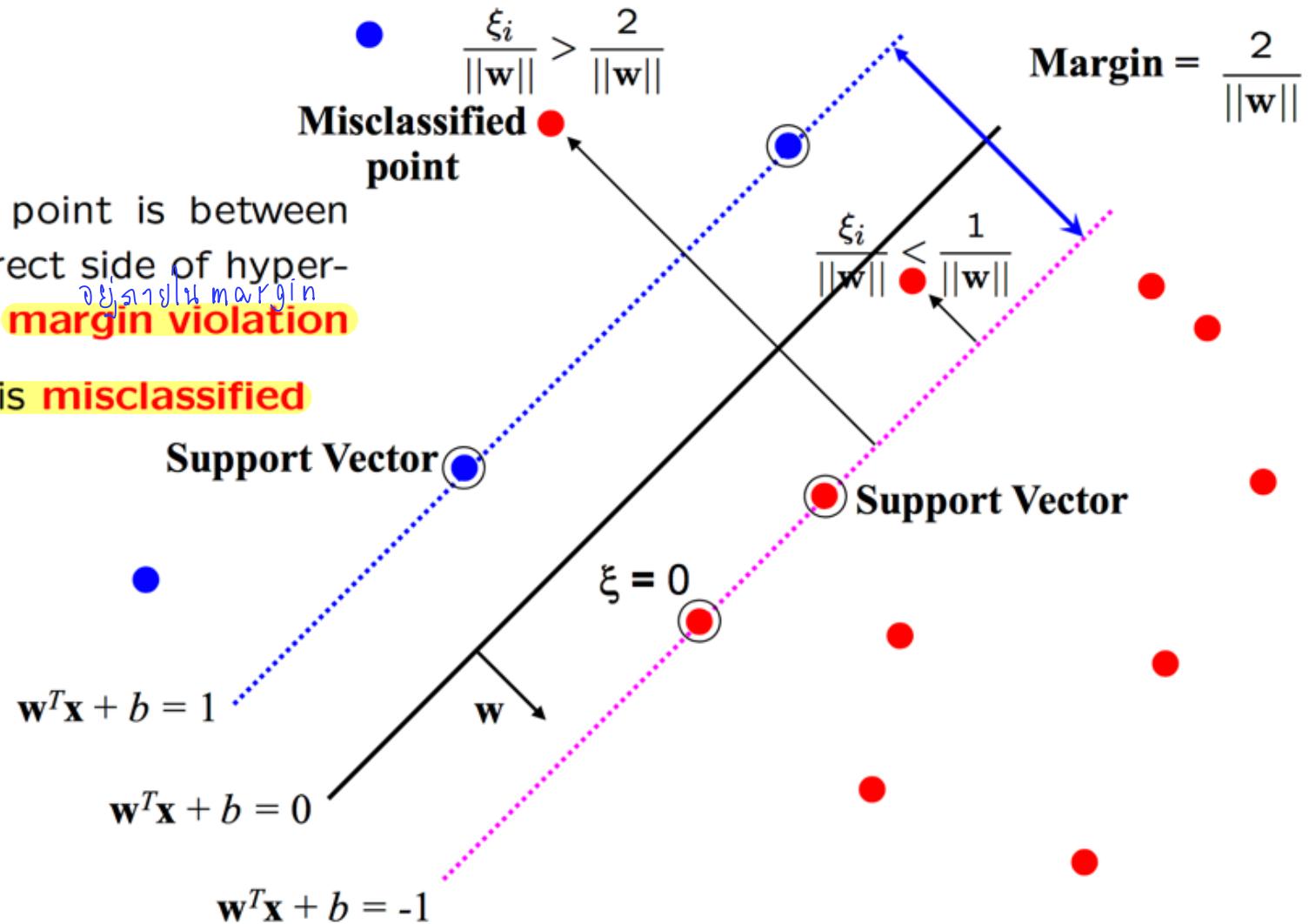
չ օւ ի՞նչ error ըստ



Error variables

$$\xi_i \geq 0$$

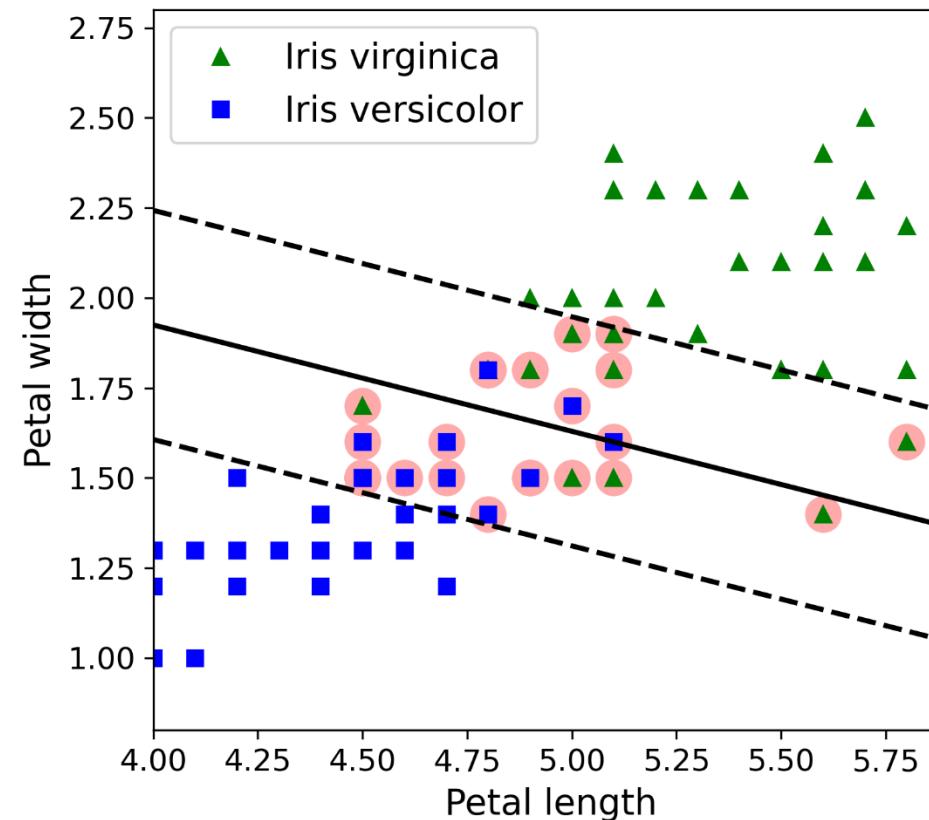
- for $0 < \xi \leq 1$ point is between margin and correct side of hyperplane. This is a *margin violation*
- for $\xi > 1$ point is **misclassified**



Regularization via margin

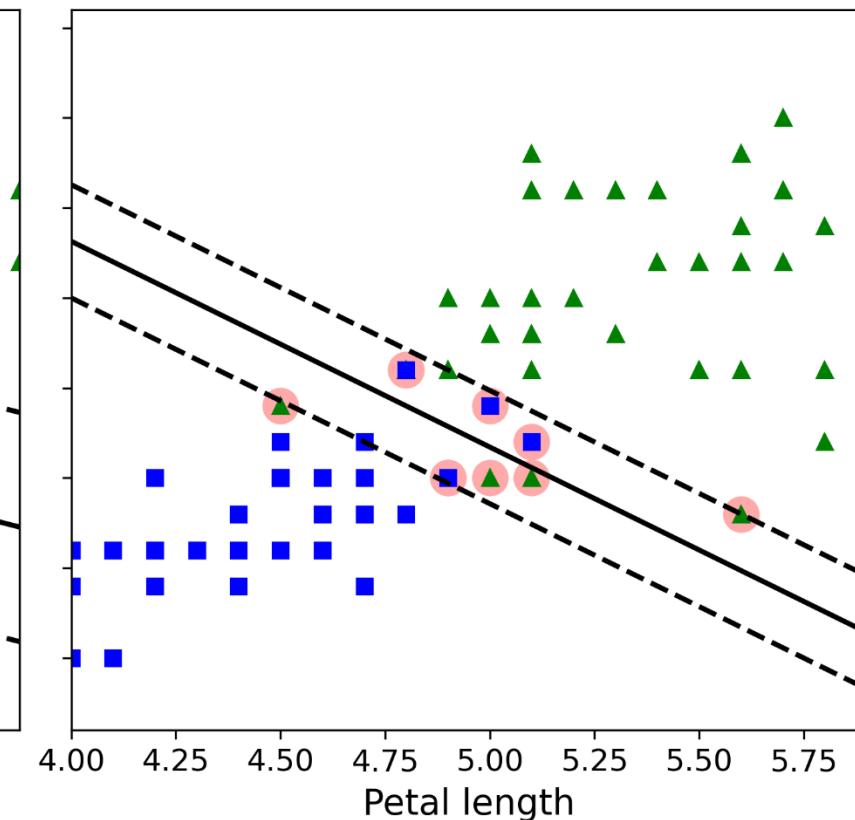
ແກບຢືນກວ່າງ ໃນໄຈລີ
(generalize)

$C = 1$



model ຍື່ງເຂົ້າ
margin

$C = 100$



Soft margin linear classifier

Objective function

$$\underset{\mathbf{w}, b, \zeta}{\text{minimize}} \quad \frac{1}{2} \mathbf{w}^T \mathbf{w} + C \sum_{i=1}^m \zeta^{(i)}$$

margin

error

เพื่อนำไป → subject to $t^{(i)}(\mathbf{w}^T \mathbf{x}^{(i)} + b) \geq 1 - \zeta^{(i)}$ and $\zeta^{(i)} \geq 0$ for $i = 1, 2, \dots, m$

(-1, 1)

(-1, 1)

- C is a regularization parameter:

- small C allows constraints to be easily ignored → large margin

- large C makes constraints hard to ignore → narrow margin

- $C = \infty$ enforces all constraints: hard margin

Solving hard and soft linear SVM Quadratic programming problem

Equation 5-5. Quadratic Programming problem

$$\text{Minimize}_{\mathbf{p}} \quad \frac{1}{2} \mathbf{p}^T \mathbf{H} \mathbf{p} + \mathbf{f}^T \mathbf{p}$$

subject to $\mathbf{A}\mathbf{p} \leq \mathbf{b}$

where $\begin{cases} \mathbf{p} & \text{is an } n_p\text{-dimensional vector } (n_p = \text{number of parameters}), \\ \mathbf{H} & \text{is an } n_p \times n_p \text{ matrix,} \\ \mathbf{f} & \text{is an } n_p\text{-dimensional vector,} \\ \mathbf{A} & \text{is an } n_c \times n_p \text{ matrix } (n_c = \text{number of constraints}), \\ \mathbf{b} & \text{is an } n_c\text{-dimensional vector.} \end{cases}$

Linear Programming
Integer programming

Simplex

Dual problem

Lagrange multiplier

ຮາກສ້າງ

- Transform constrained optimization objective into an unconstrained one
- For example,

$$\text{minimize } f(x,y) = x^2 + 2y$$

$$\text{subjective } 3x + 2y + 1 = 0 \quad (\text{equality constrain})$$

- Lagrangian

$$\text{minimize } g(x,y,\alpha) = f(x,y) - \alpha(3x + 2y + 1)$$

Dual problem Generalized Lagrangian

$\alpha^{(i)}$ as Karush-Kuhn-Tucker (KKT) multiplier

Equation C-1. Generalized Lagrangian for the hard margin problem

$$\mathcal{L}(\mathbf{w}, b, \alpha) = \frac{1}{2}\mathbf{w}^T\mathbf{w} - \sum_{i=1}^m \alpha^{(i)}(t^{(i)}(\mathbf{w}^T\mathbf{x}^{(i)} + b) - 1)$$

with $\alpha^{(i)} \geq 0$ for $i = 1, 2, \dots, m$

Partial derivative of generalized Lagrangian

Equation C-2. Partial derivatives of the generalized Lagrangian

$$\nabla_{\mathbf{w}} \mathcal{L}(\mathbf{w}, b, \alpha) = \mathbf{w} - \sum_{i=1}^m \alpha^{(i)} t^{(i)} \mathbf{x}^{(i)}$$

gradient

$\frac{\partial \mathcal{L}}{\partial \mathbf{w}}$

$$\frac{\partial}{\partial b} \mathcal{L}(\mathbf{w}, b, \alpha) = - \sum_{i=1}^m \alpha^{(i)} t^{(i)}$$

Stationary points, set to extrema

Equation C-3. Properties of the stationary points

$$\hat{\mathbf{w}} = \sum_{i=1}^m \hat{\alpha}^{(i)} t^{(i)} \mathbf{x}^{(i)}$$

$$\sum_{i=1}^m \hat{\alpha}^{(i)} t^{(i)} = 0$$

Substitute to generalized Lagrangian Loss function

Equation C-4. Dual form of the SVM problem

$$\mathcal{L}(\hat{\mathbf{w}}, \hat{b}, \alpha) = \frac{1}{2} \sum_{i=1}^m \sum_{j=1}^m \alpha^{(i)} \alpha^{(j)} t^{(i)} t^{(j)} \underbrace{\mathbf{x}^{(i)T} \mathbf{x}^{(j)}}_{\text{kernel}} - \sum_{i=1}^m \alpha^{(i)}$$

with $\alpha^{(i)} \geq 0$ for $i = 1, 2, \dots, m$

Dual problem Objective function

Equation 5-6. Dual form of the linear SVM objective

$$\underset{\alpha}{\text{minimize}} \frac{1}{2} \sum_{i=1}^m \sum_{j=1}^m \alpha^{(i)} \alpha^{(j)} t^{(i)} t^{(j)} \mathbf{x}^{(i)T} \mathbf{x}^{(j)} - \sum_{i=1}^m \alpha^{(i)}$$

subject to $\alpha^{(i)} \geq 0$ for $i = 1, 2, \dots, m$

ແກ້ໄຂ α ພາວພເຕອ weight

Primal solution

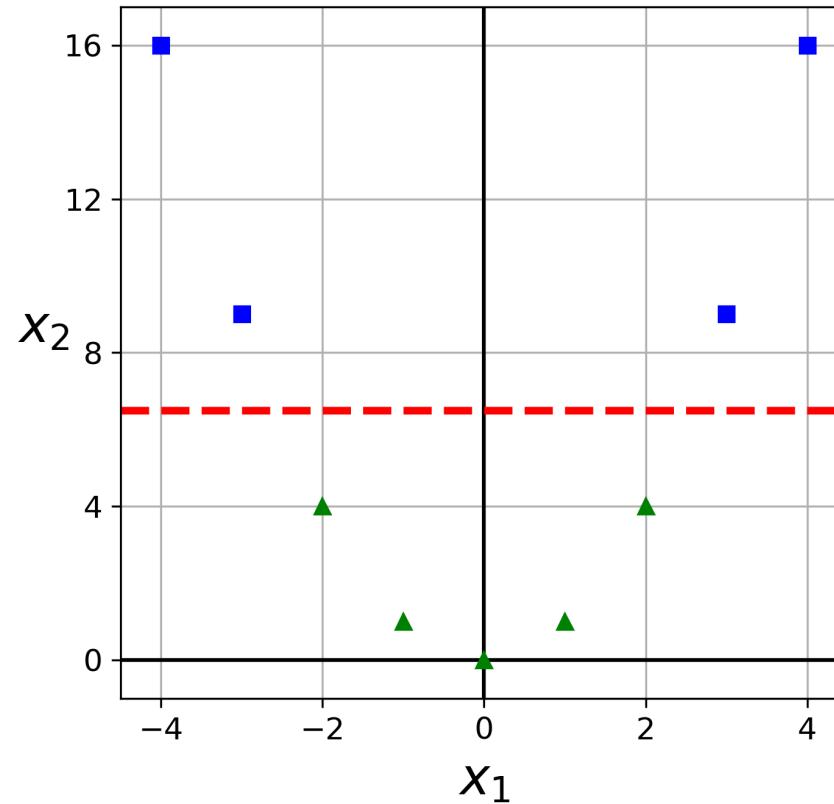
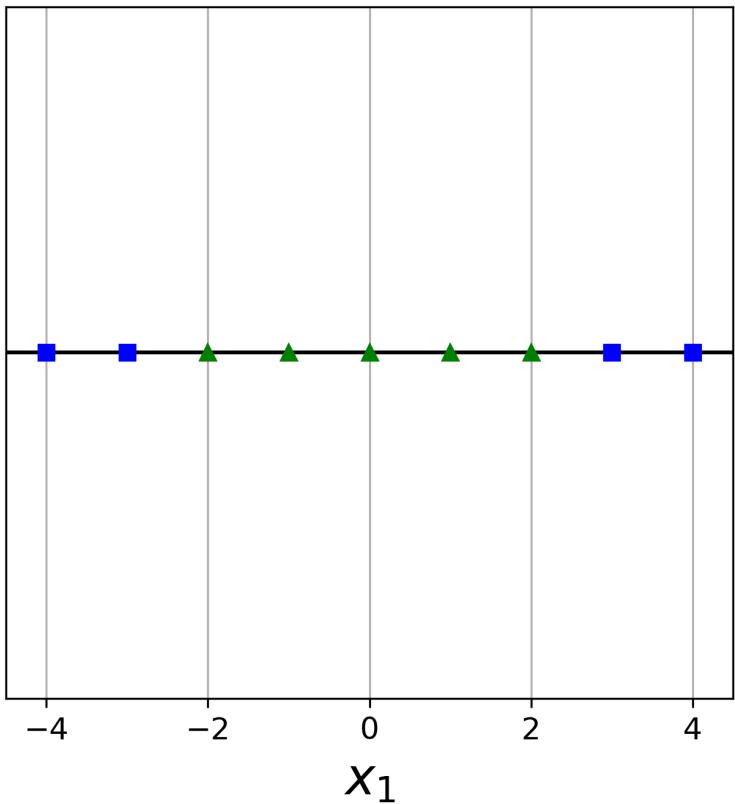
- After obtaining the solution of dual problem, $\alpha^{(i)}$, we can use them to compute the solution of primal problem as follows

$$\hat{\mathbf{w}} = \sum_{i=1}^m \hat{\alpha}^{(i)} t^{(i)} \mathbf{x}^{(i)}$$

$$\hat{b} = \frac{1}{n_s} \sum_{\substack{i=1 \\ \hat{\alpha}^{(i)} > 0}}^m (t^{(i)} - \hat{\mathbf{w}}^T \mathbf{x}^{(i)})$$

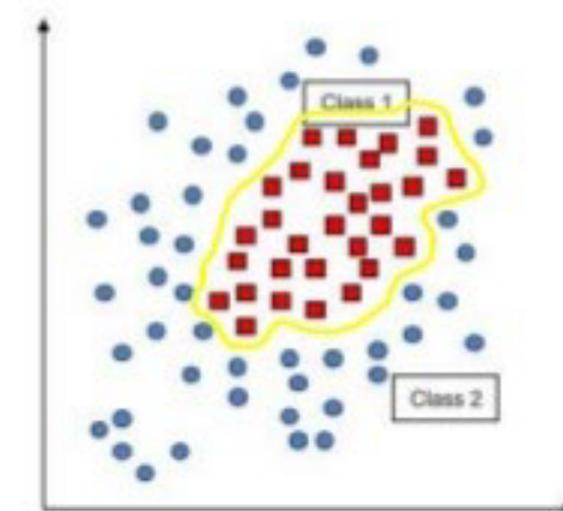
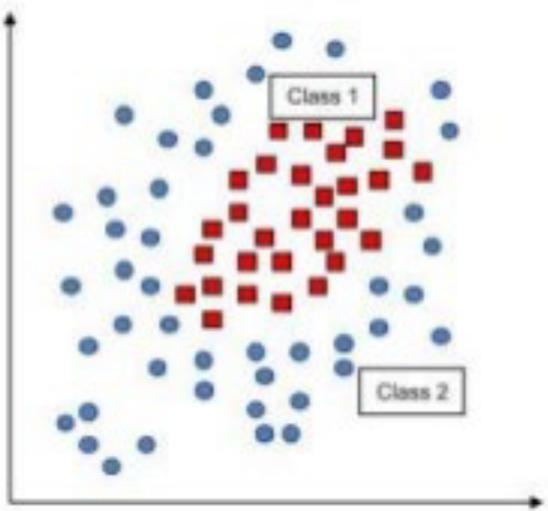
Nonlinear SVM

Adding features to make a dataset linearly separable

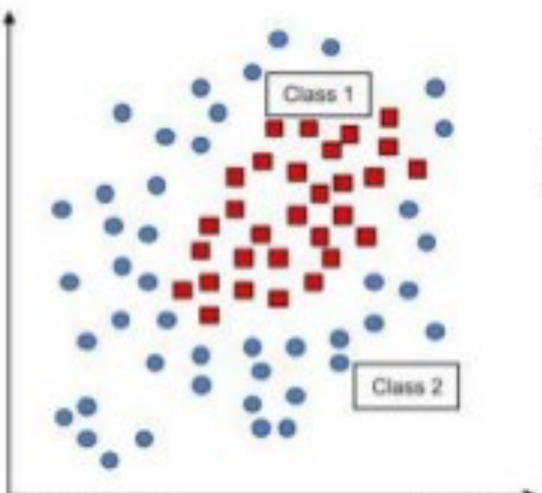


$$x_2 = (x_1)^2$$

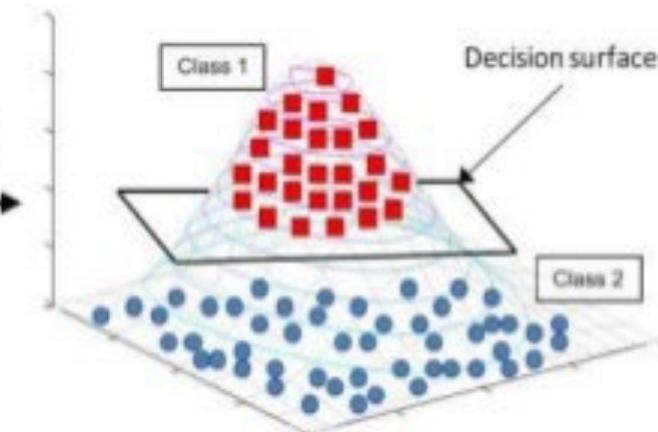
SVM Kernel



Non Linear
Decision
Boundary



kernel



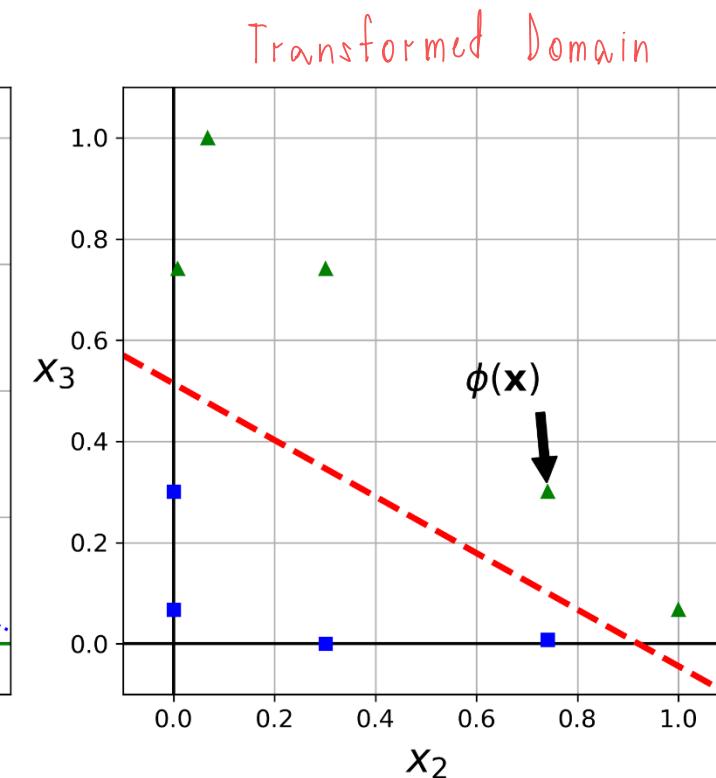
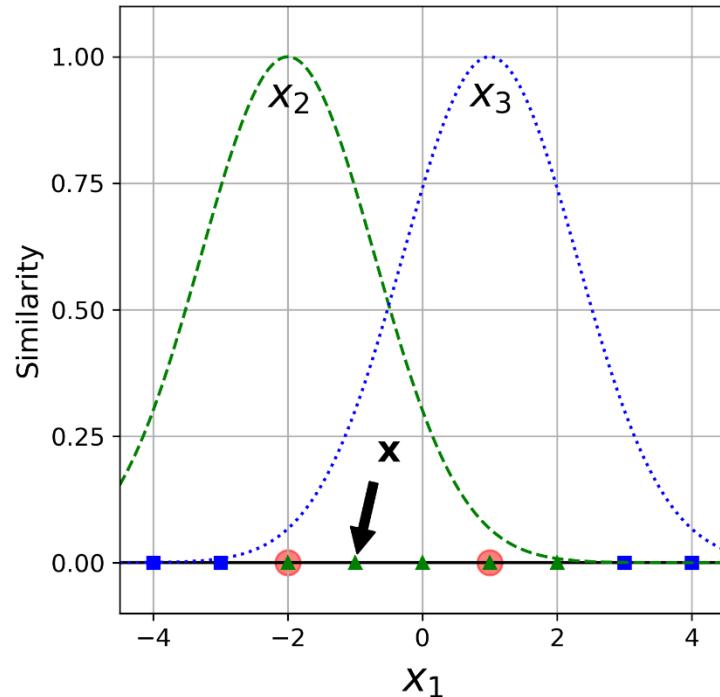
Kernel
method

Adding similarity feature

Radial Basis
Equation 5-1. Gaussian RBF

$$\phi_{\gamma}(\mathbf{x}, \ell) = \exp(-\gamma \|\mathbf{x} - \ell\|^2)$$

x_2, x_3 is systematically finds in SVM.

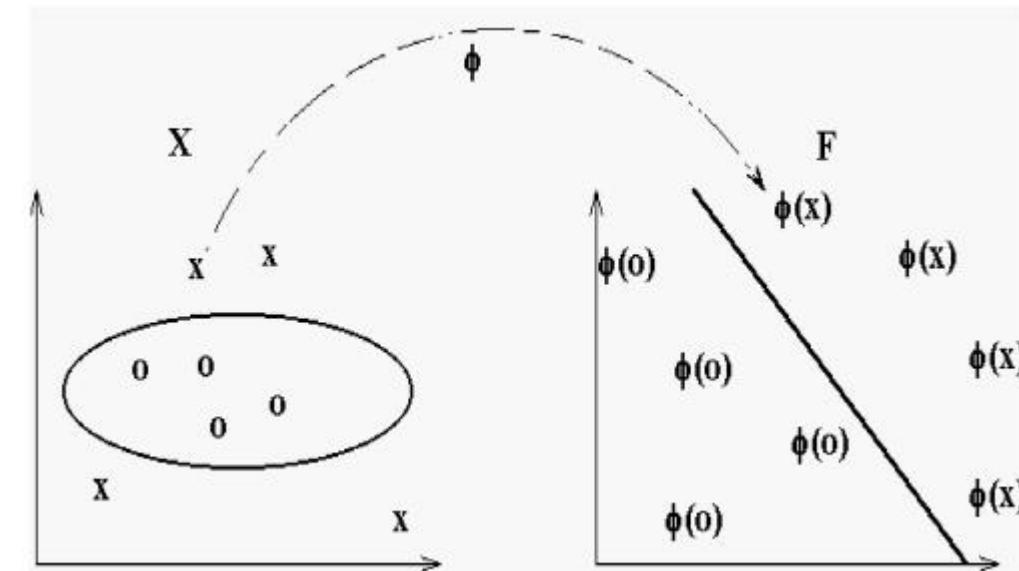


Kernelized SVM

Second-degree polynomial

Equation 5-8. Second-degree polynomial mapping

$$\phi(\mathbf{x}) = \phi\left(\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}\right) = \begin{pmatrix} x_1^2 \\ \sqrt{2}x_1x_2 \\ x_2^2 \end{pmatrix}$$



Kernel trick

Equation 5-9. Kernel trick for a 2nd-degree polynomial mapping

$$\begin{aligned}\phi(\mathbf{a})^T \phi(\mathbf{b}) &= \begin{pmatrix} a_1^2 \\ \sqrt{2} a_1 a_2 \\ a_2^2 \end{pmatrix}^T \begin{pmatrix} b_1^2 \\ \sqrt{2} b_1 b_2 \\ b_2^2 \end{pmatrix} = a_1^2 b_1^2 + 2a_1 b_1 a_2 b_2 + a_2^2 b_2^2 \\ &= (a_1 b_1 + a_2 b_2)^2 = \left(\begin{pmatrix} a_1 \\ a_2 \end{pmatrix}^T \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} \right)^2 = (\mathbf{a}^T \mathbf{b})^2\end{aligned}$$

Recall the objective function

Equation 5-6. Dual form of the linear SVM objective

$$\underset{\alpha}{\text{minimize}} \frac{1}{2} \sum_{i=1}^m \sum_{j=1}^m \alpha^{(i)} \alpha^{(j)} t^{(i)} t^{(j)} \underline{\mathbf{x}^{(i)T} \mathbf{x}^{(j)}} - \sum_{i=1}^m \alpha^{(i)}$$

subject to $\alpha^{(i)} \geq 0$ for $i = 1, 2, \dots, m$

Replacing with $\phi(\mathbf{x}^{(i)})^T \phi(\mathbf{x}^{(j)})$

For 2-nd degree polynomial kernel, $\phi(\mathbf{x}^{(i)})^T \phi(\mathbf{x}^{(j)}) = (\mathbf{x}^{(i)T} \mathbf{x}^{(j)})^2$

Common kernel

Equation 5-10. Common kernels

Linear: $K(\mathbf{a}, \mathbf{b}) = \mathbf{a}^T \mathbf{b}$

Polynomial: $K(\mathbf{a}, \mathbf{b}) = (\gamma \mathbf{a}^T \mathbf{b} + r)^d$

Gaussian RBF: $K(\mathbf{a}, \mathbf{b}) = \exp(-\gamma \|\mathbf{a} - \mathbf{b}\|^2)$

Sigmoid: $K(\mathbf{a}, \mathbf{b}) = \tanh(\gamma \mathbf{a}^T \mathbf{b} + r)$

Prediction with kernelized SVM

Equation 5-11. Making predictions with a kernelized SVM

$$\begin{aligned} h_{\hat{\mathbf{w}}, \hat{b}}(\phi(\mathbf{x}^{(n)})) &= \hat{\mathbf{w}}^T \phi(\mathbf{x}^{(n)}) + \hat{b} = \left(\sum_{i=1}^m \hat{\alpha}^{(i)} t^{(i)} \phi(\mathbf{x}^{(i)}) \right)^T \phi(\mathbf{x}^{(n)}) + \hat{b} \\ &= \sum_{i=1}^m \hat{\alpha}^{(i)} t^{(i)} (\phi(\mathbf{x}^{(i)})^T \phi(\mathbf{x}^{(n)})) + \hat{b} \\ &= \sum_{\substack{i=1 \\ \hat{\alpha}^{(i)} > 0}}^m \hat{\alpha}^{(i)} t^{(i)} K(\mathbf{x}^{(i)}, \mathbf{x}^{(n)}) + \hat{b} \end{aligned}$$

No need of $\mathbf{w}!!$

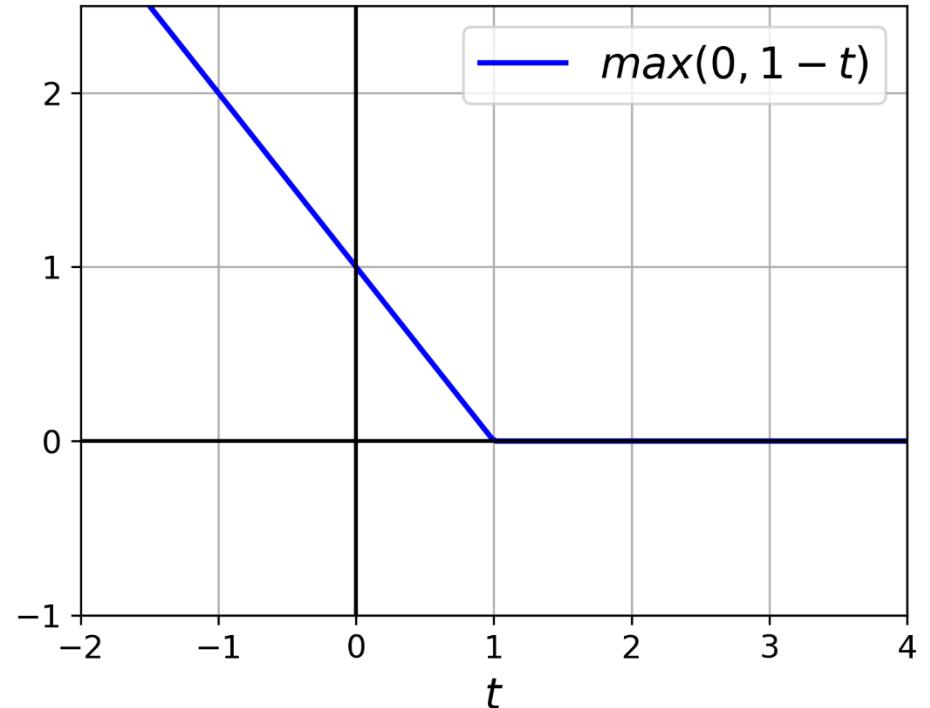
Bias term for kernel trick

Equation 5-12. Computing the bias term using the kernel trick

$$\begin{aligned}\hat{b} &= \frac{1}{n_s} \sum_{\substack{i=1 \\ \hat{\alpha}^{(i)} > 0}}^m (t^{(i)} - \hat{\mathbf{w}}^T \phi(\mathbf{x}^{(i)})) = \frac{1}{n_s} \sum_{\substack{i=1 \\ \hat{\alpha}^{(i)} > 0}}^m \left(t^{(i)} - \left(\sum_{j=1}^m \hat{\alpha}^{(j)} t^{(j)} \phi(\mathbf{x}^{(j)}) \right)^T \phi(\mathbf{x}^{(i)}) \right) \\ &= \frac{1}{n_s} \sum_{\substack{i=1 \\ \hat{\alpha}^{(i)} > 0}}^m \left(t^{(i)} - \sum_{j=1}^m \hat{\alpha}^{(j)} t^{(j)} K(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}) \right)\end{aligned}$$

Hinge Loss

- Derivative equals to -1 when $t^{(i)} < 1$
- This makes it suitable for using with Gradient Descent in online learning



$$J(\mathbf{w}, b) = \frac{1}{2} \mathbf{w}^T \mathbf{w} + C \sum_{i=1}^m \max(0, 1 - t^{(i)}(\mathbf{w}^T \mathbf{x}^{(i)} + b))$$

End of Lecture 3

Question?