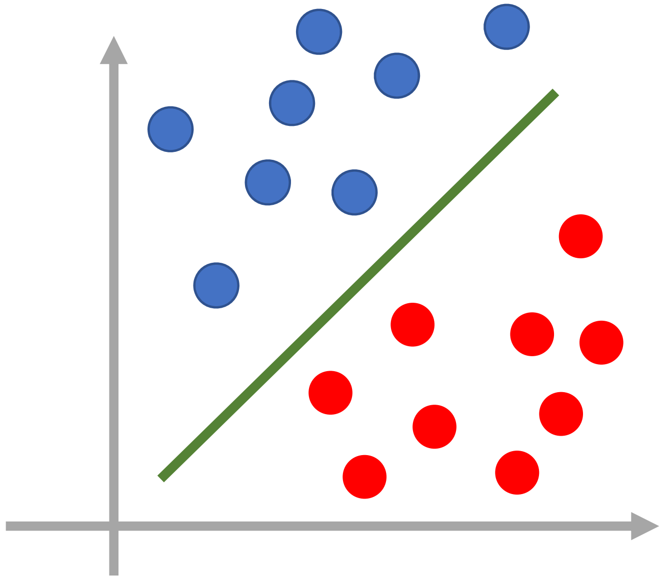
3 Support Vector Machine

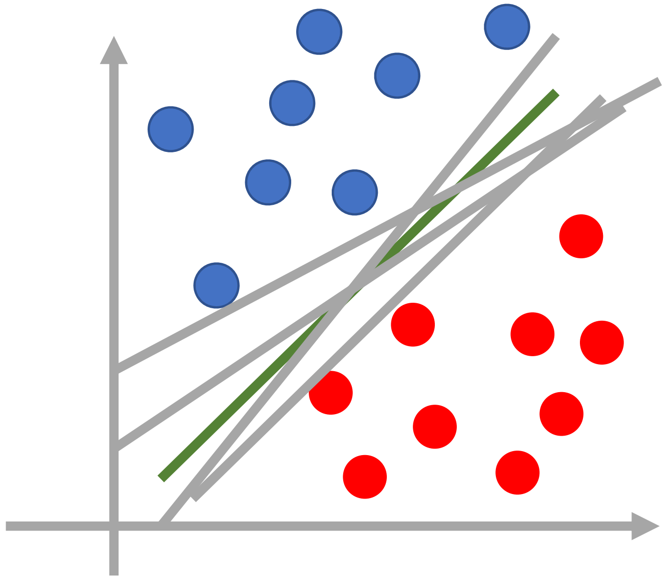
Support Vector Machine

◈Linear Model：try to find a hyperplane which can separate the samples belonging to different classes



Think

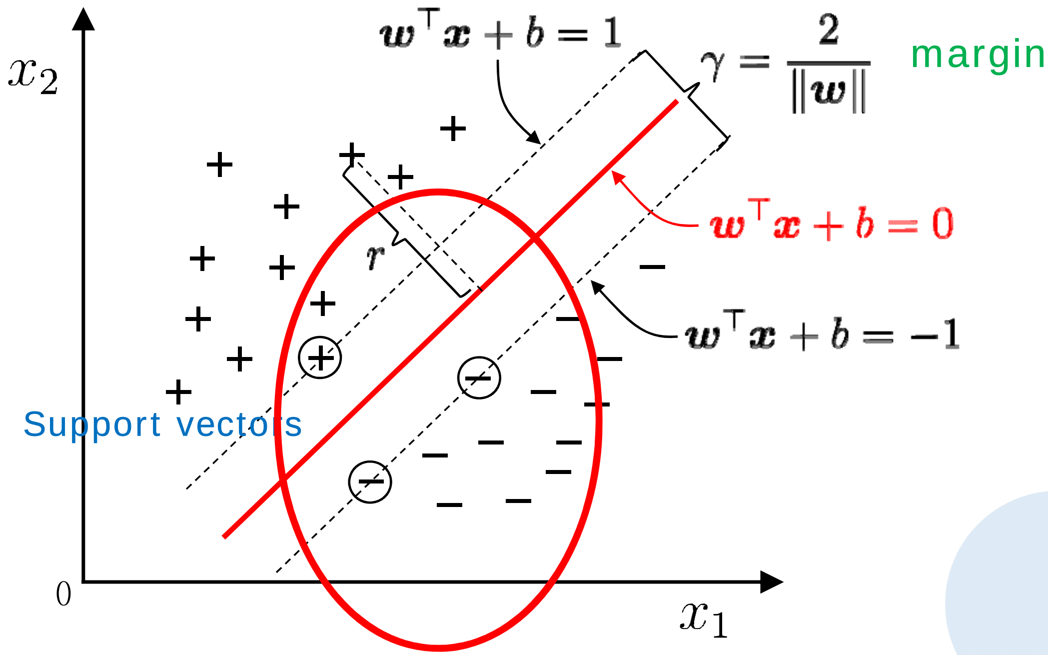
- ◇Which of the following possible hyperplanes is the best?
- ◇The green one due to high tolerance, Robustness, and better generalization.



Green is best

Margin and Support Vector

Hyperplane: $\boldsymbol{w}^\top \boldsymbol{x} + b = 0$



Dual Problem

- Lagrange multipliers
  - Step1 ：Use lagrange multipliers $\alpha_i \geq 0$ and derive lagrange function:

$$L(\boldsymbol{w}, b, \boldsymbol{\alpha}) = \frac{1}{2}\|\boldsymbol{w}\|^2 - \sum_{i=1}^{m} \alpha_i \left(y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b) - 1\right)$$

  - Step 2 ：the gradient of $L(\boldsymbol{w}, b, \boldsymbol{\alpha})$ with respect $\boldsymbol{w}$ and b should be 0 =>

$$\boxed{\boldsymbol{w} = \sum_{i=1}^{m} \alpha_i y_i \boldsymbol{x}_i,} \quad \sum_{i=1}^{m} \alpha_i y_i = 0.$$

  - Step 3 ：Plugging the above equations into $L(\boldsymbol{w}, b, \boldsymbol{\alpha})$

$$\min_{\boldsymbol{\alpha}} \quad \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j \boldsymbol{x}_i^\top \boldsymbol{x}_j - \sum_{i=1}^{m} \alpha_i$$

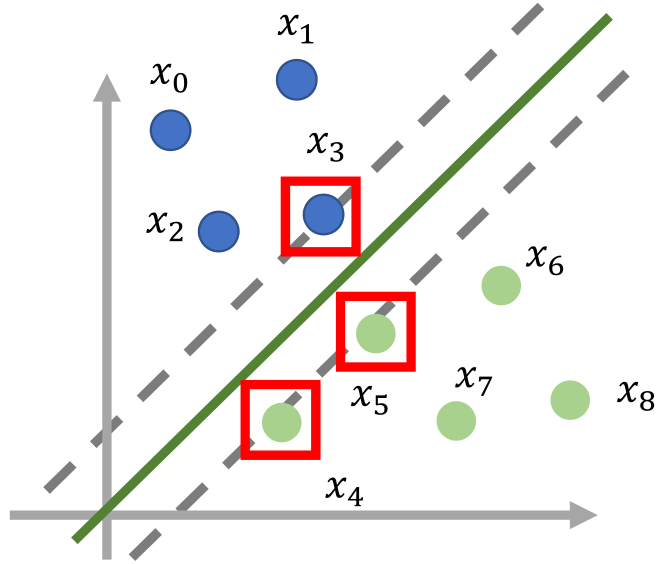
$$\text{s.t.} \quad \sum_{i=1}^{m} \alpha_i y_i = 0, \ \alpha_i \geq 0, \ i = 1, 2, \ldots, m.$$

Support Vector

$$w\sum_{k=0}^{n}(\alpha_k x_k) = w\,(\alpha_0 x_0 + \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_3 + \cdots + \alpha_n x_n)$$

$$w\sum_{k=0}^{n}(\alpha_k x_k) = w\,(0 \times x_0 + 0 \times x_1 + 0 \times x_2 + \alpha_3 x_3 + \alpha_4 x_4 + \alpha_5 x_5 + 0 \times x_6 + 0 \times x_7 + 0 \times x_8)$$

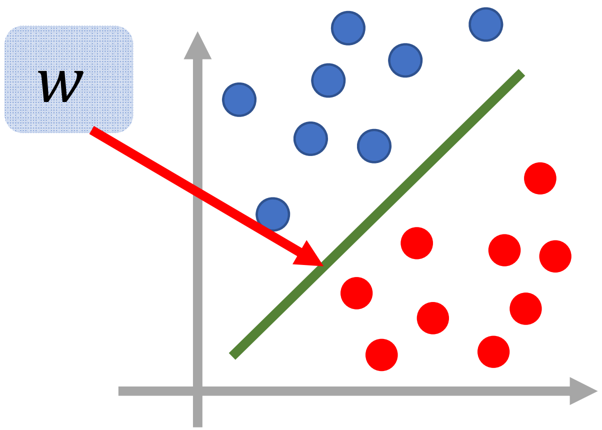
$$w\sum_{k=0}^{n}(\alpha_k x_k) = w\,(\alpha_3 x_3 + \alpha_4 x_4 + \alpha_5 x_5)$$



Margin

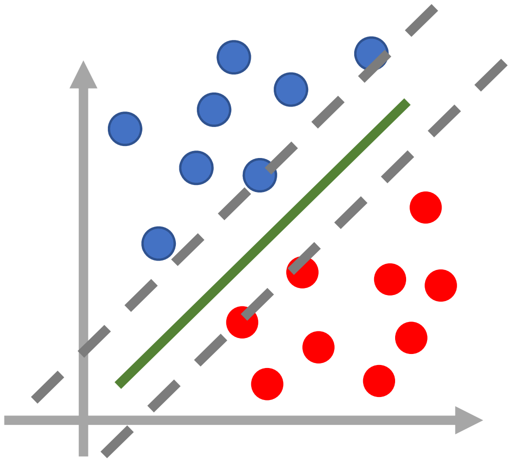
Distance of x to the hyperplane

$$\frac{|w^T x + b|}{\|w\|}$$



We hope that the data is outside the scope of margin

$$\frac{y_i(w^T x + b)}{\|w\|} \geq \gamma$$



Margin

We hope that the data is outside the scope of margin

$$\frac{y_i\left(w^T x + b\right)}{\|w\|} \geq \gamma$$

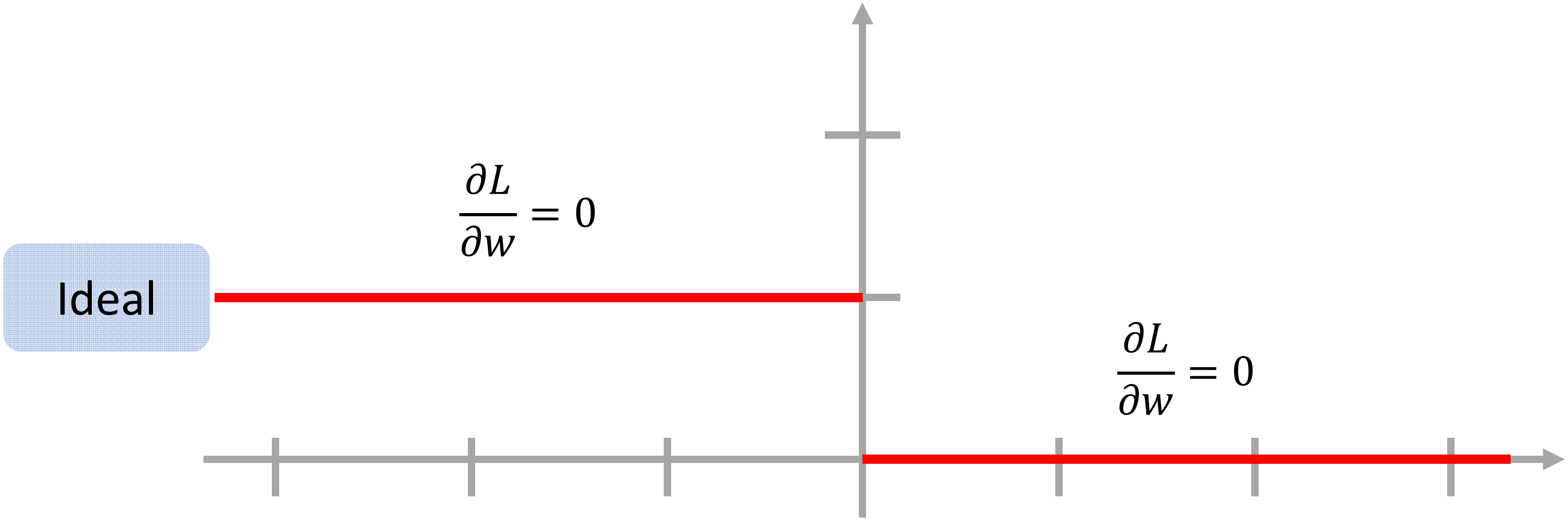
Because there are many different solutions, we fixed $\gamma\|w\|=1$

$$y_i\left(w^T x + b\right) \geq \gamma\|w\| = 1$$

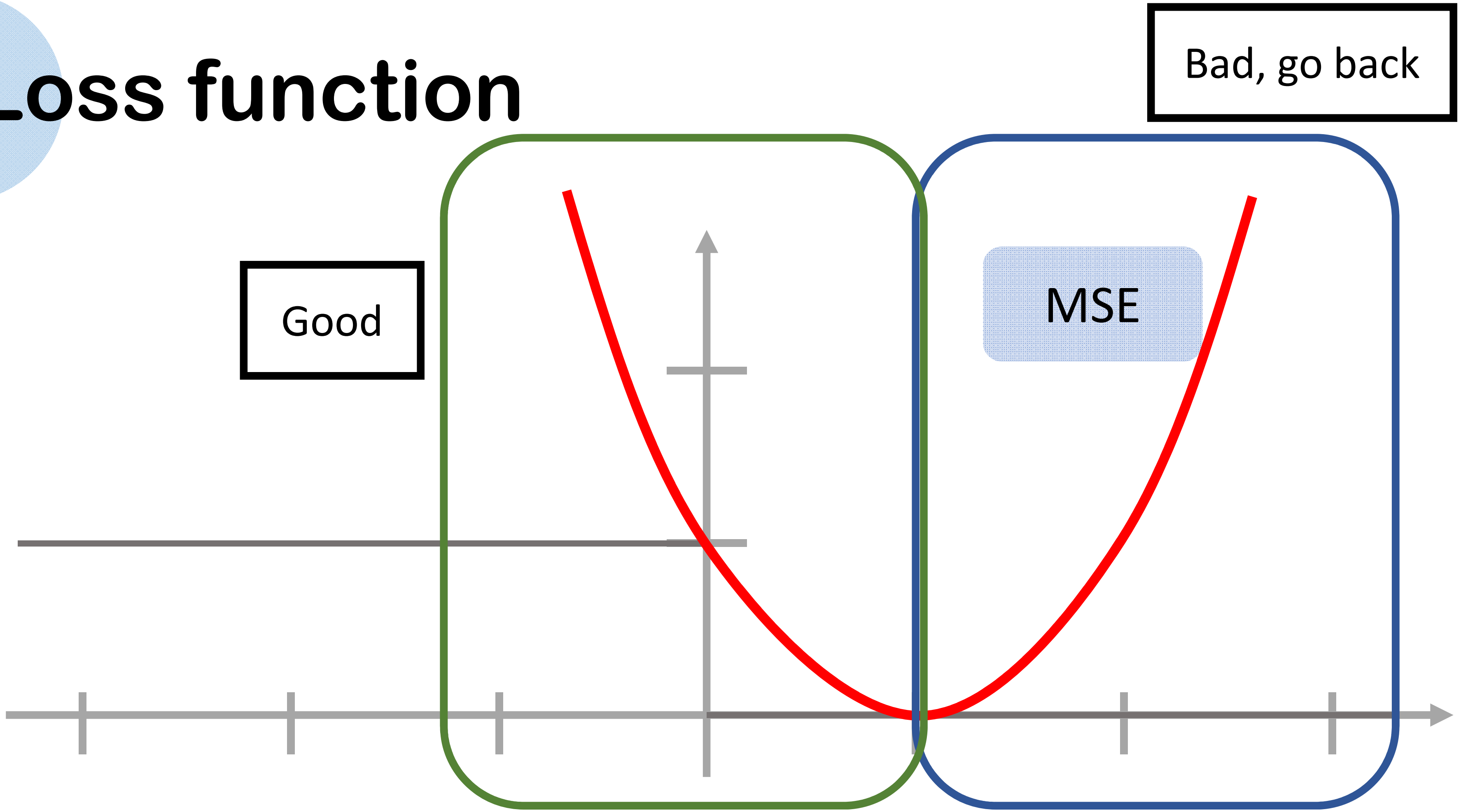
$$\arg\min_{w,b} \frac{1}{2}\|w\|^2 \quad \Rightarrow \quad s.t.\ y_i\left(w^T x + b\right) \geq 1, i = 1,2,\dots,m.$$

Loss function

We can’t use gradient descent in this loss funtion



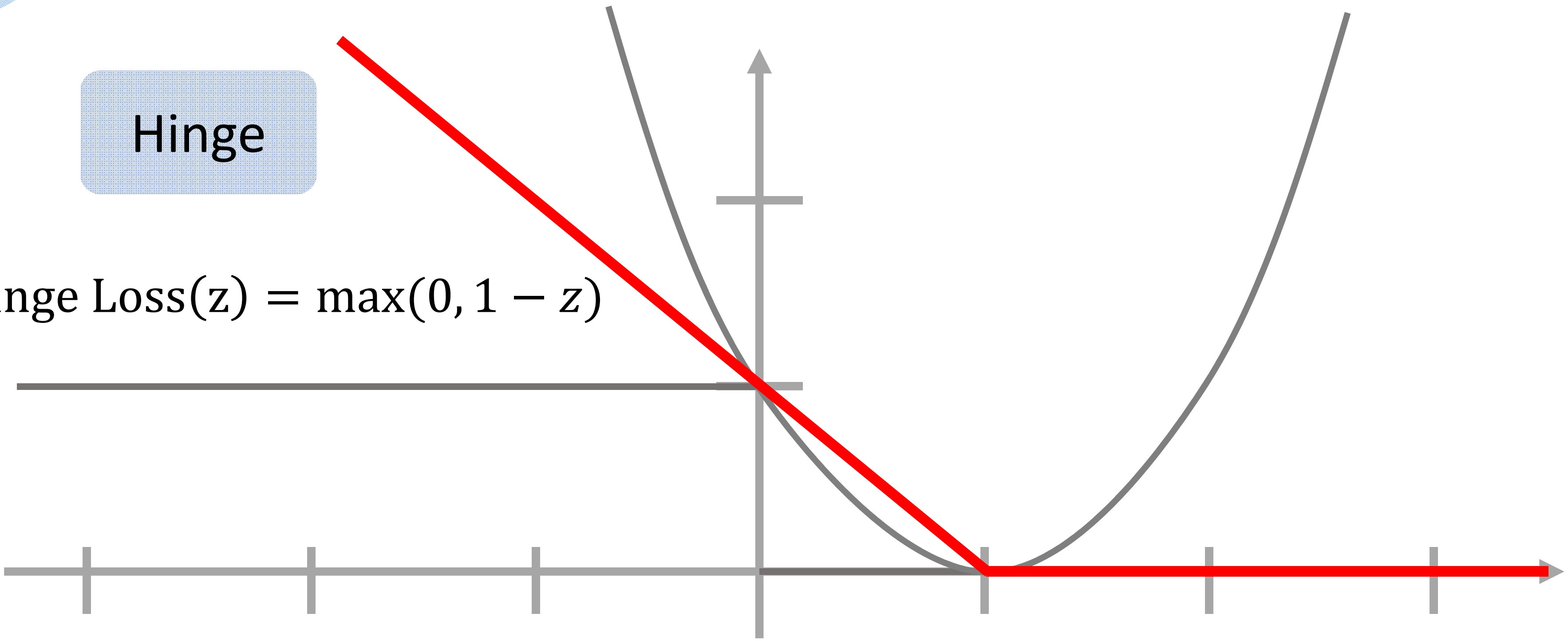
Loss function



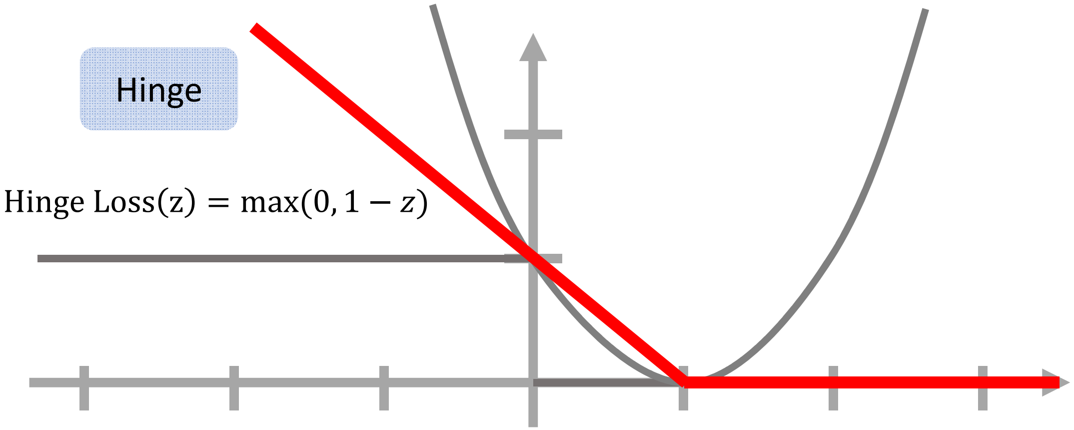
Loss function

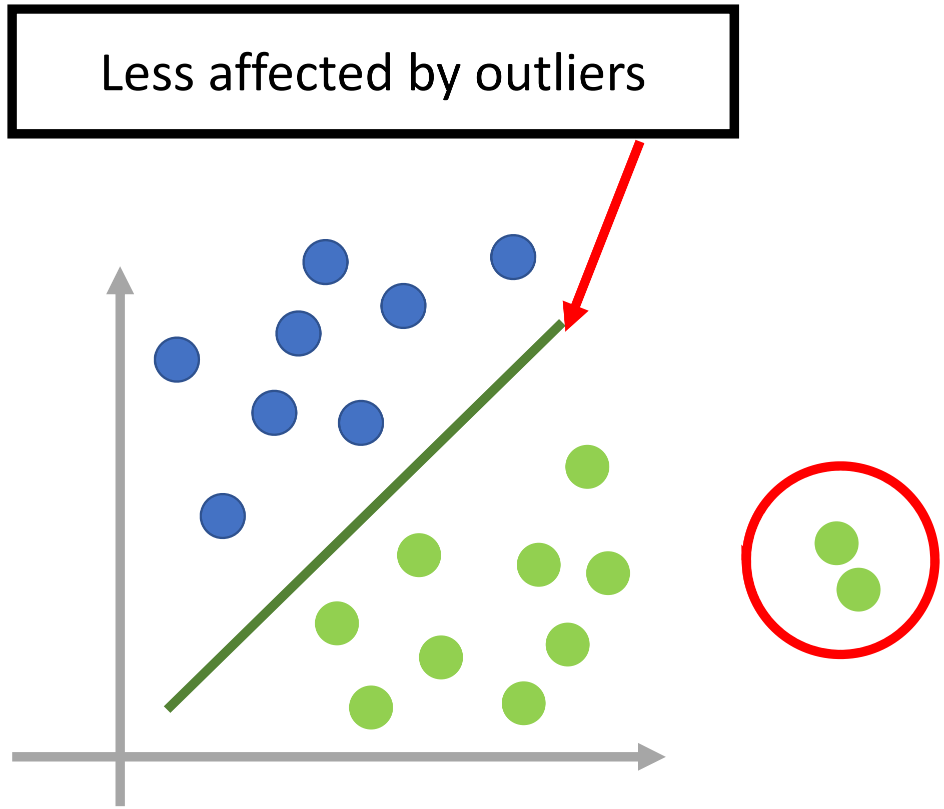
Hinge

$$\text{Hinge Loss}(z) = \max(0, 1 - z)$$

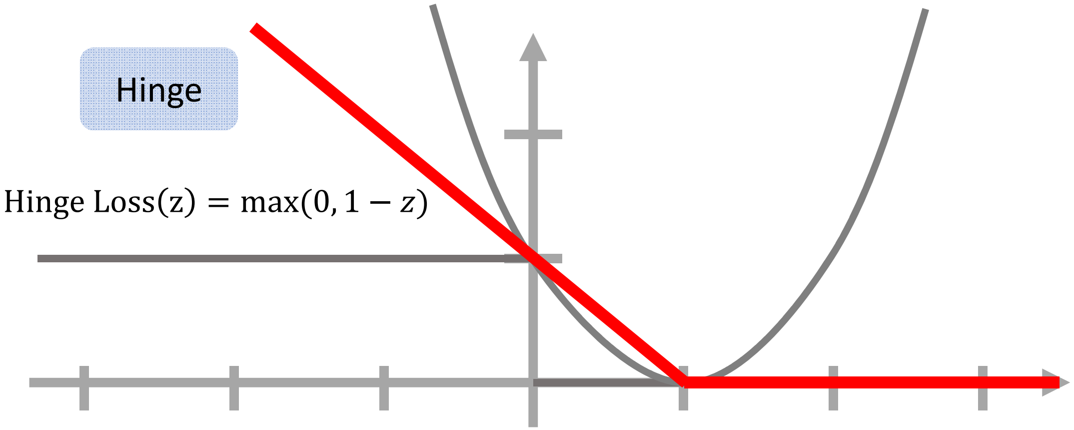


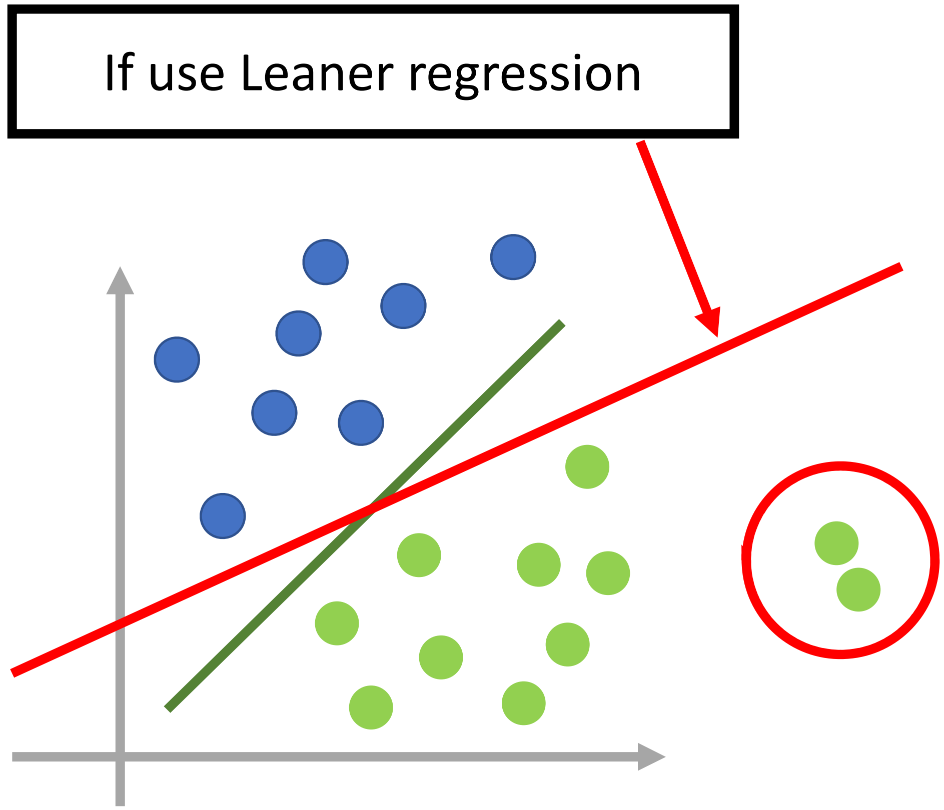
Loss function



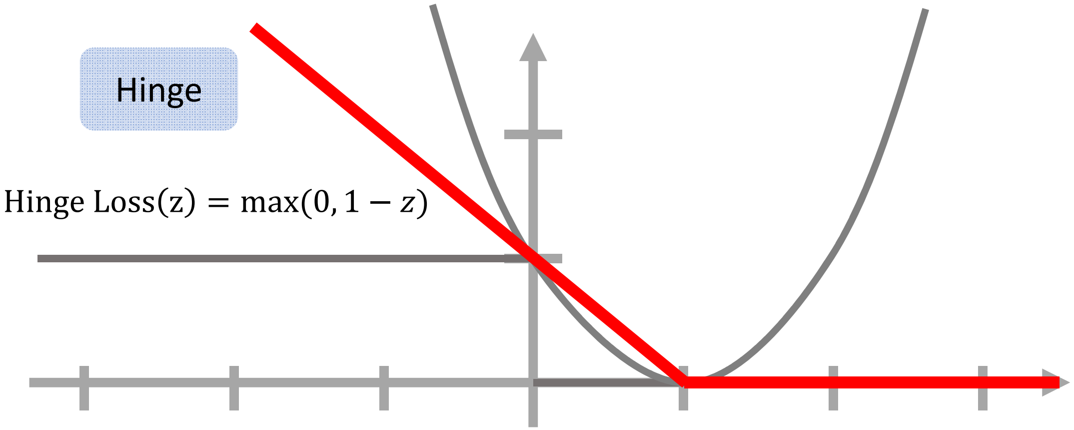


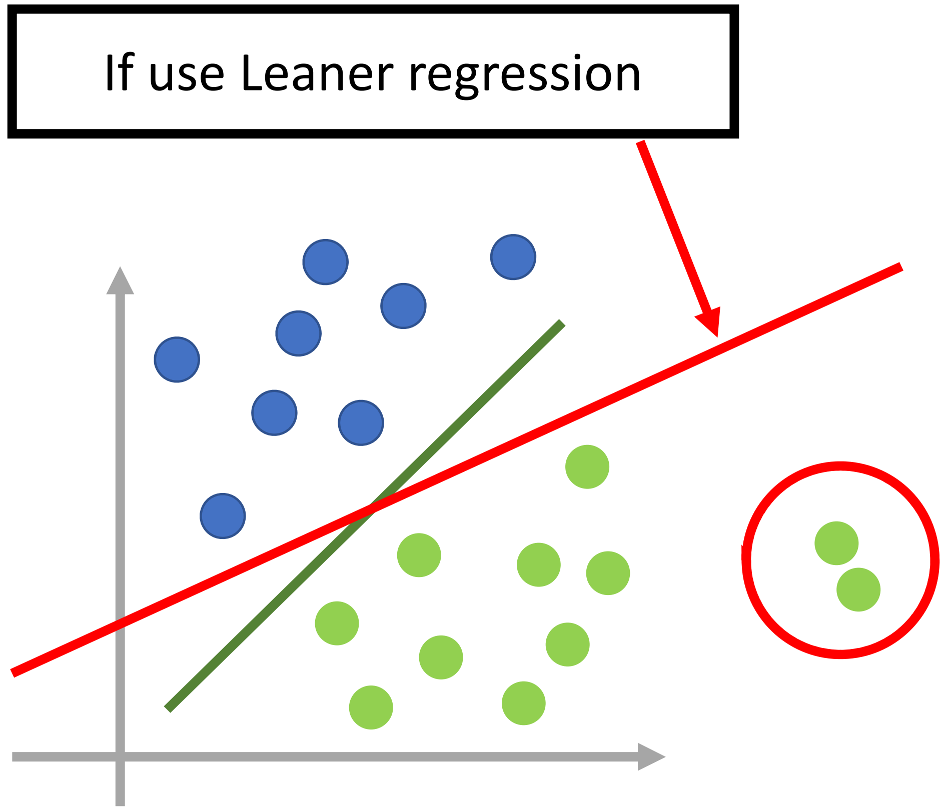
Loss function





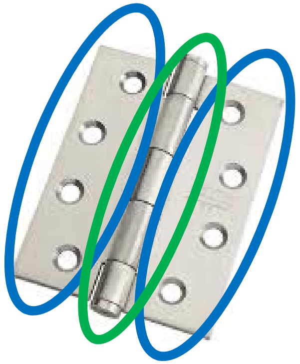
Hinge loss to SVM

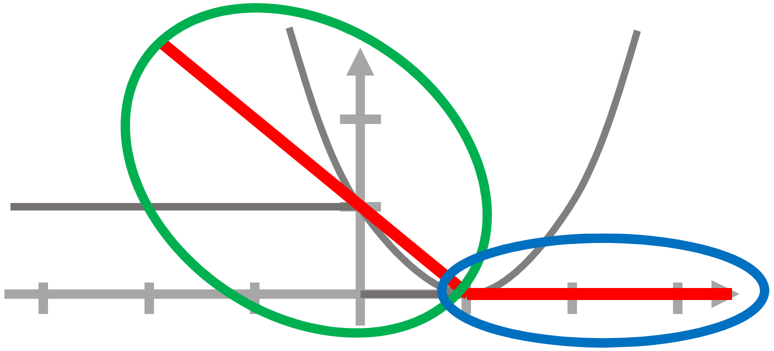




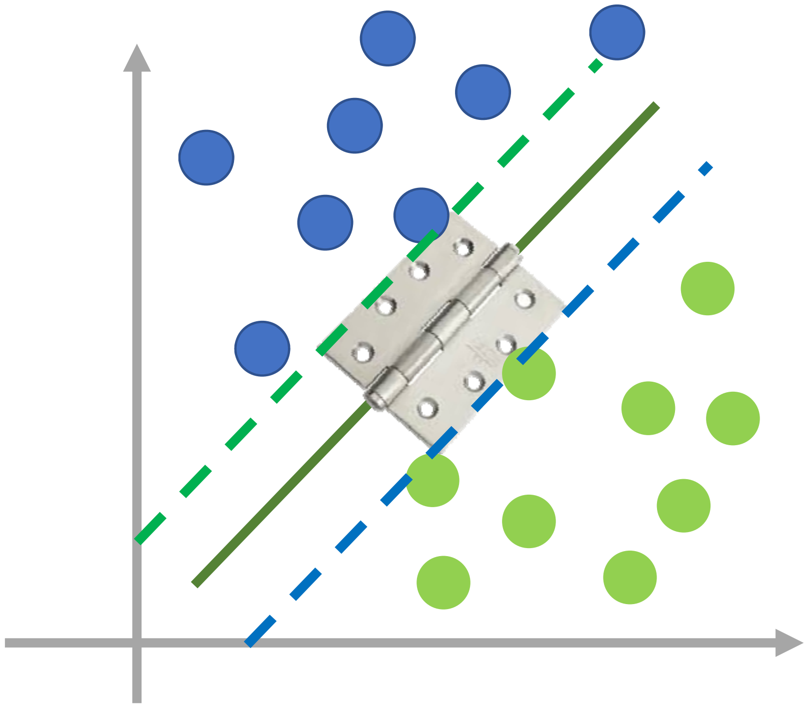
Hinge loss to SVM

Hinge





$$\text{Hinge Loss}(z) = \max(0, 1 - z)$$

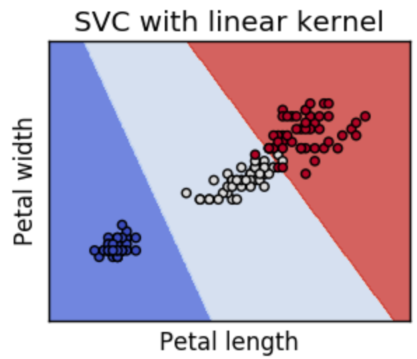


Kernel function 介紹

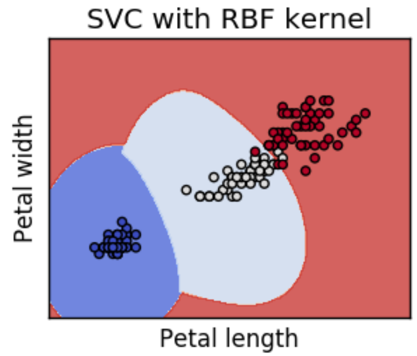
$$f(x) = w^T x + b = \sum_{i=1}^{m} \alpha_i y_i \boxed{x_i^T x_j} + b$$

$$k(x_i, x_j) = x_i^T x_j$$

$$k(x_i, x_j) = x_i^T x_j$$

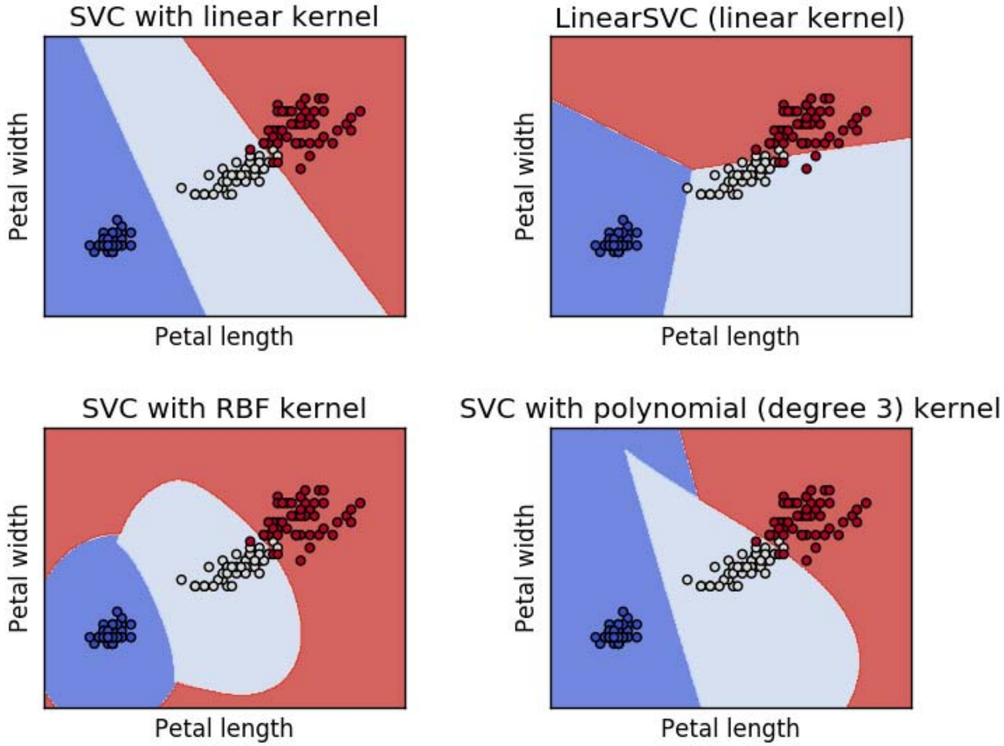


$$k(x_i, x_j) = \exp\left(-\frac{\left\|x_i - x_j\right\|^2}{2\delta^2}\right)$$



Different Kernel function

$$k(x_i, x_j) = x_i^T x_j$$



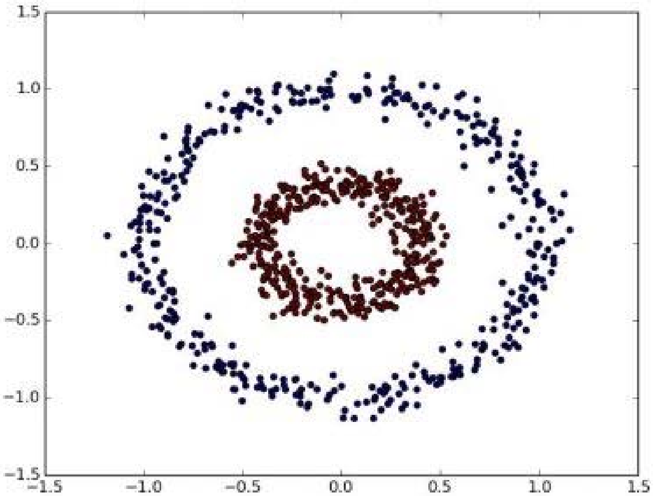
$$k(x_i, x_j) = \exp(-\frac{\|x_i - x_j\|^2}{2\delta^2})$$

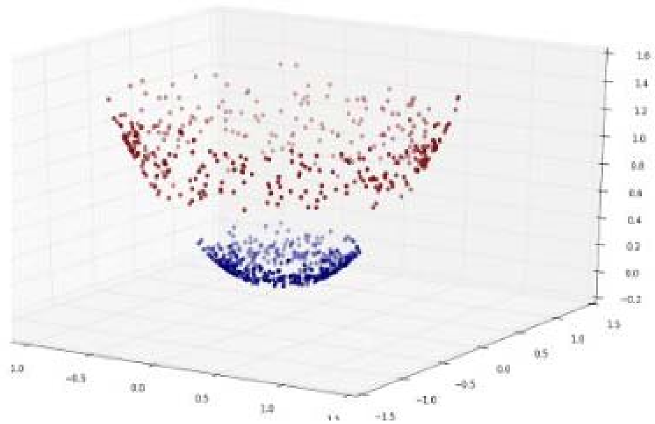
$$k(x_i, x_j) = (x_i^T x_j)^d$$

Reference : http://dataaspirant.com/2017/01/25/svm-classifier-implemenation-python-scikit-learn/

Kernel Trick

- Linear decision boundary does not work well here. But we can project up to 3-dimension surface.





Reference : https://codingmachinelearning.wordpress.com/2016/08/02/svm-visualizing-the-kernel-function/