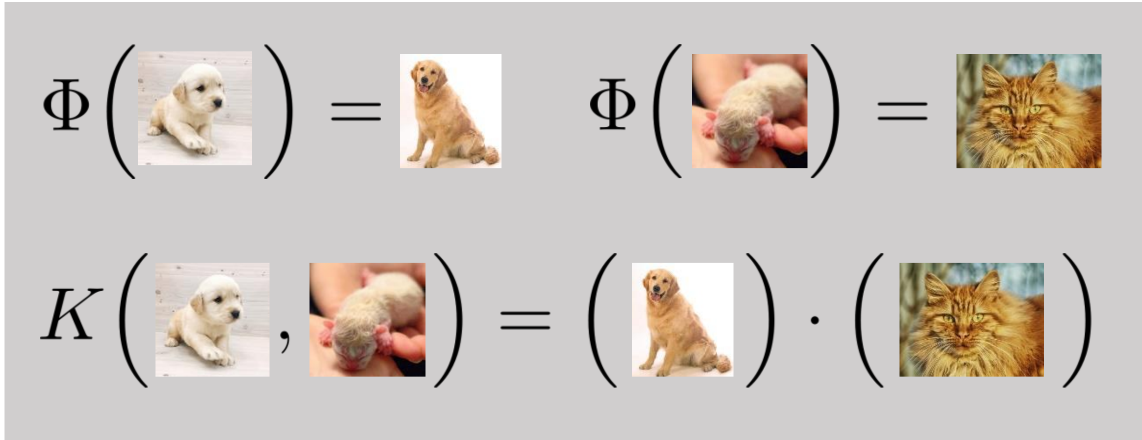


Kernel-based Learning:
Support Vector Machine – Linear & Hard Margin

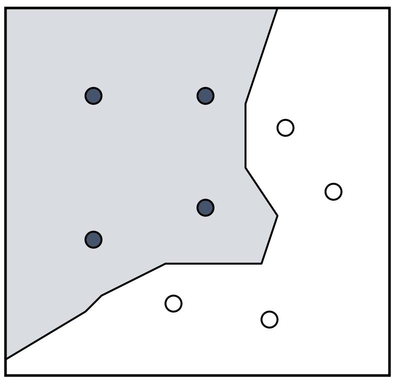
Pilsung Kang

School of Industrial Management Engineering

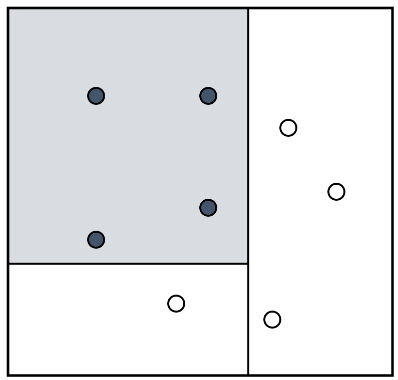
Korea University

Discriminant Function

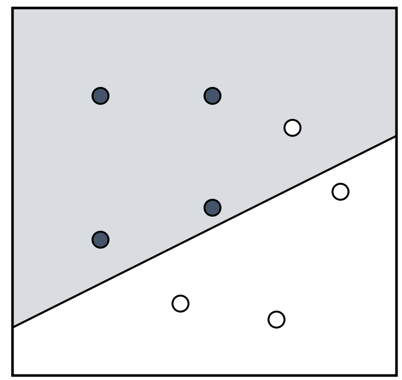
- Discriminant functions in classification



Nearest
Neighbor

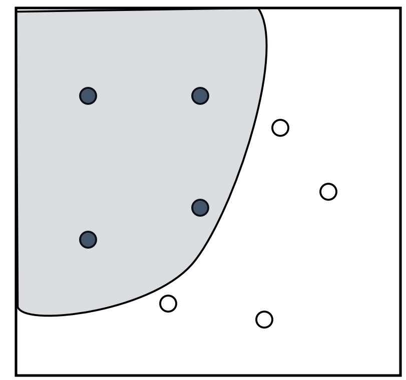


Decision
Tree



Linear
Functions

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



Nonlinear
Functions

Linear Classification

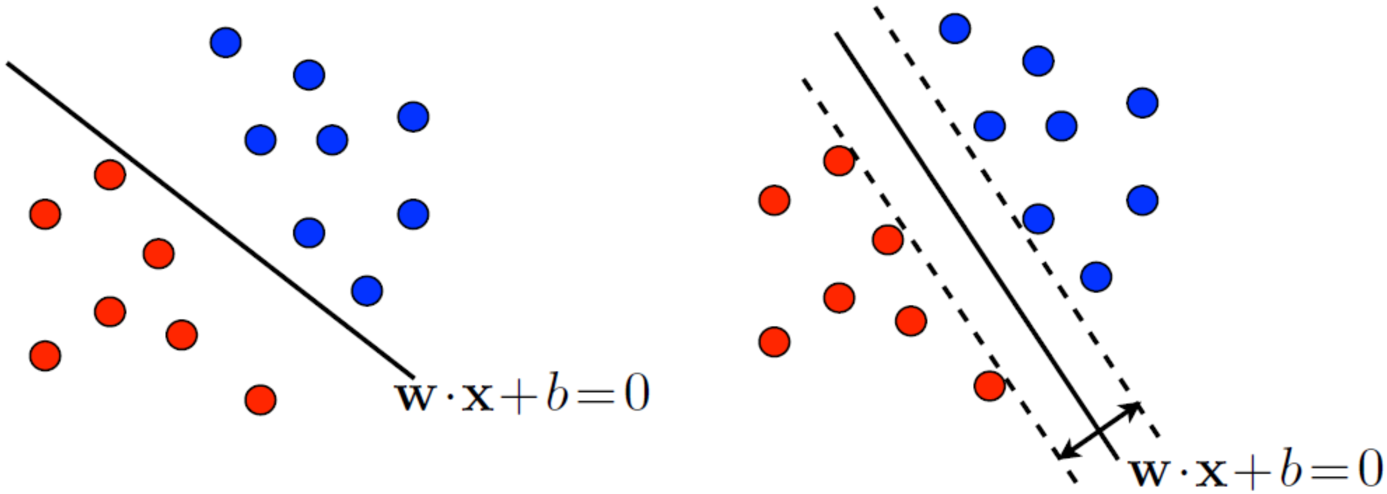
- Binary Classification Problem
  - ✓ Training data: sample drawn i.i.d. from set $X \in R^d$ according to some distribution D

$$S = \Big((x_1, y_1), ..., (x_n, y_n)\Big) \in X \times \{-1, +1\}$$

  - ✓ Problem: find hypothesis $h : X \to \{-1, +1\}$ in *H* (classifier) with small generalization error $R_D(h)$
  - ✓ Linear classifier
    - Hypothesis based on hyperplanes
    - Linear separation in high-dimensional data

Linear Classification

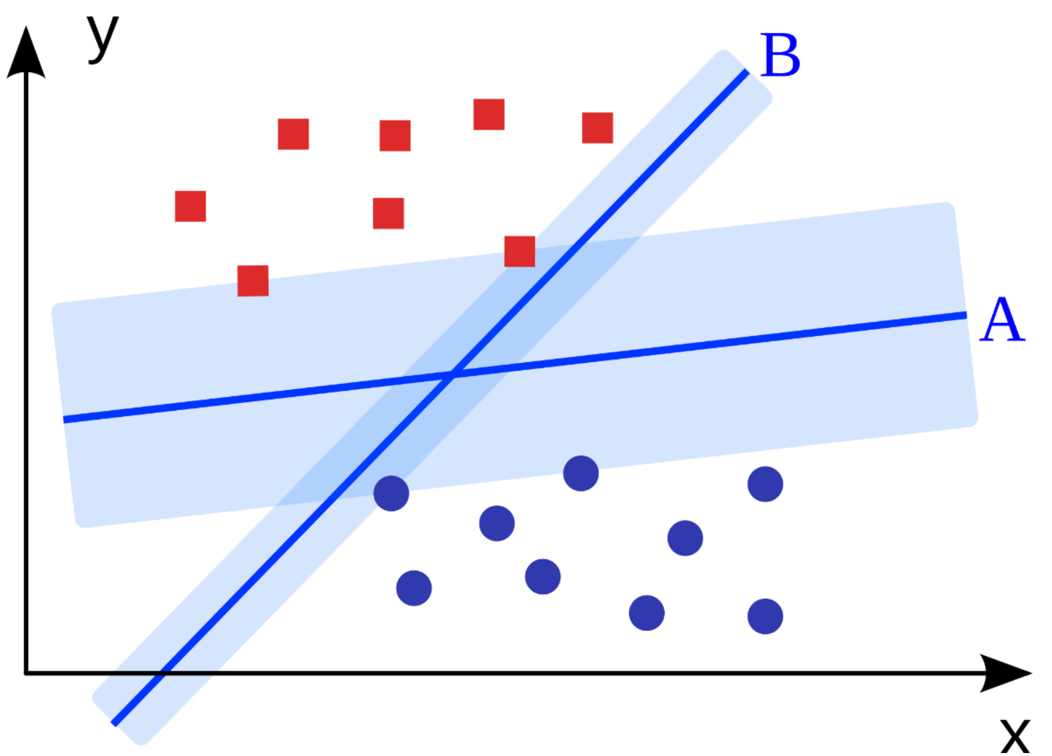
- Binary classification problem



$$H = \{\mathbf{x} \to sign(\mathbf{w} \cdot \mathbf{x} + b : \ \mathbf{w} \in R^d, \ b \in R\}$$

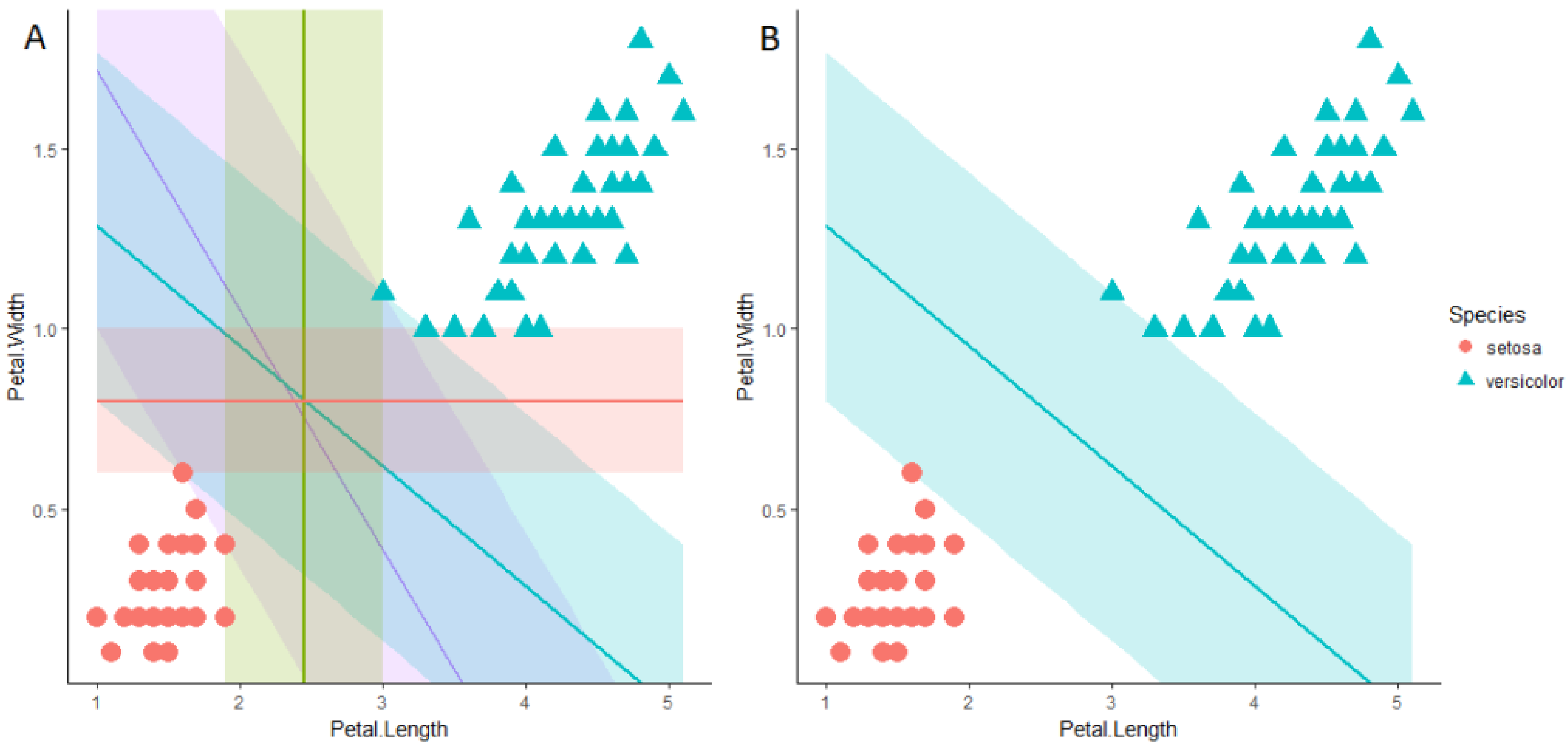
Local Optimum? Global Optimum!

- Which classification boundary is better?
  - ✓ How do you define "better"?



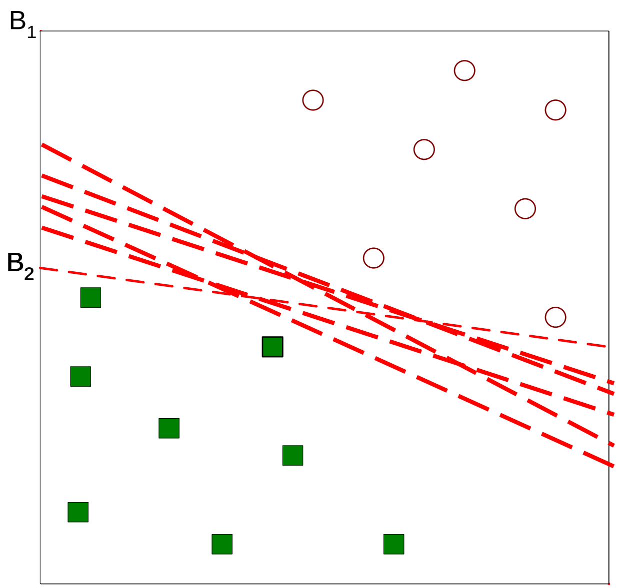
Local Optimum? Global Optimum!

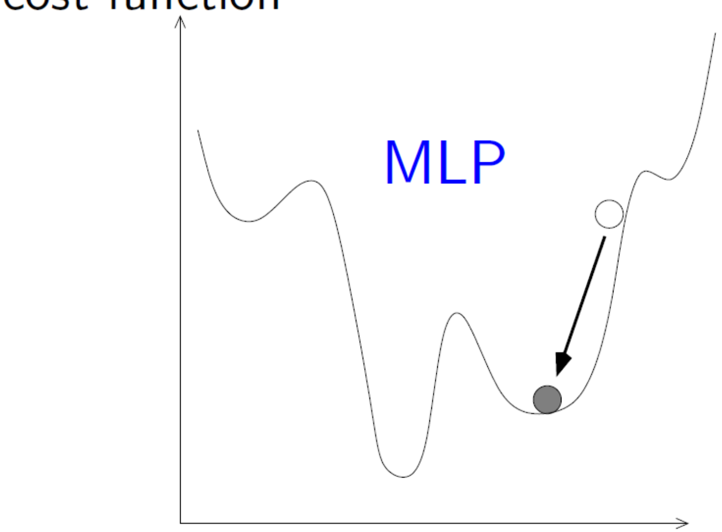
- Which classification boundary is better?
  - ✓ Find the hyperplane that maximizes the margin!



Local Optimum? Global Optimum!

- Artificial neural network (ANN)
  - ✓ Universal approximation of continuous nonlinear functions
  - ✓ Learning from input-output patterns
  - ✓ Parallel network architecture, multiple inputs and outputs
  - ✓ But, existence of many local optimums!



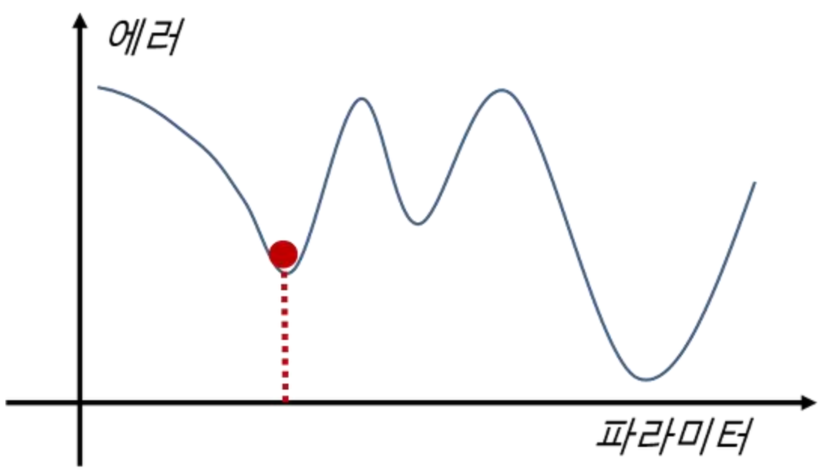


Local Optimum? Global Optimum!

- Artificial neural network (ANN) (cont')
- But!!

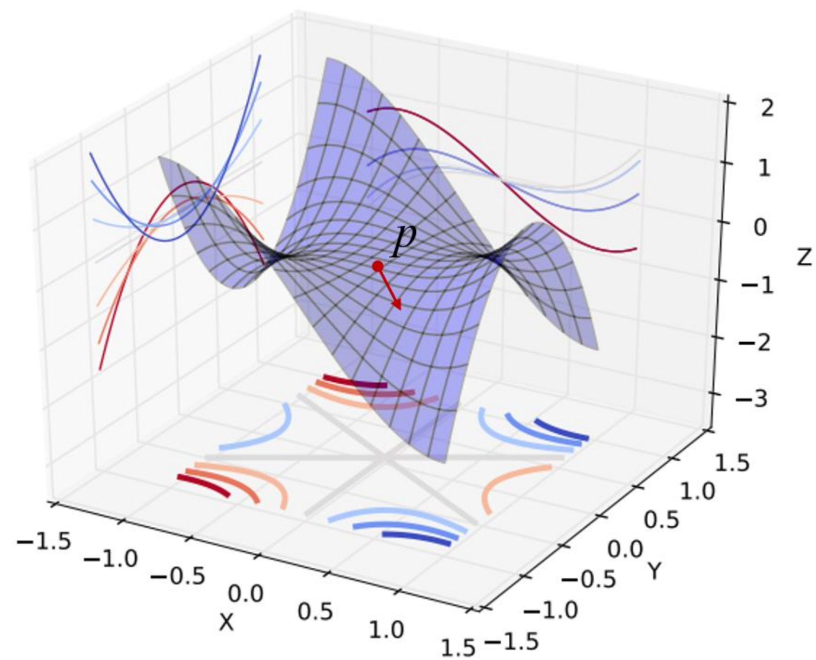
Dauphin, Y. N., Pascanu, R., Gulcehre, C., Cho, K., Ganguli, S., & Bengio, Y. (2014). Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. In *Advances in neural information processing systems* (pp. 2933-2941).

이럴 줄 알았는데…



https://darkpgmr.tistory.com/148

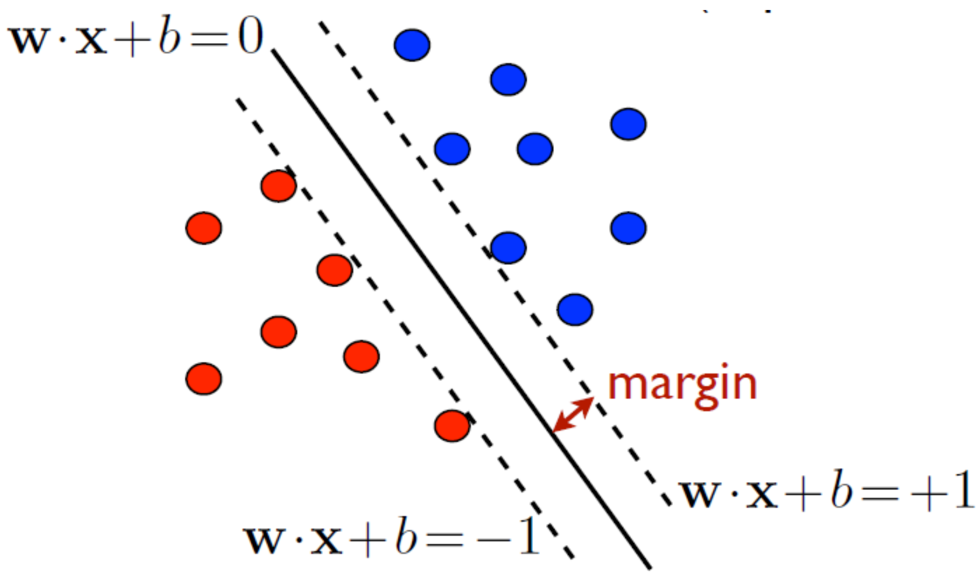
고차원에서 모든 방향으로 gradient가 0인 경우는 거의 없더라…



Support Vector Machine: Formulation

Burges (1998)

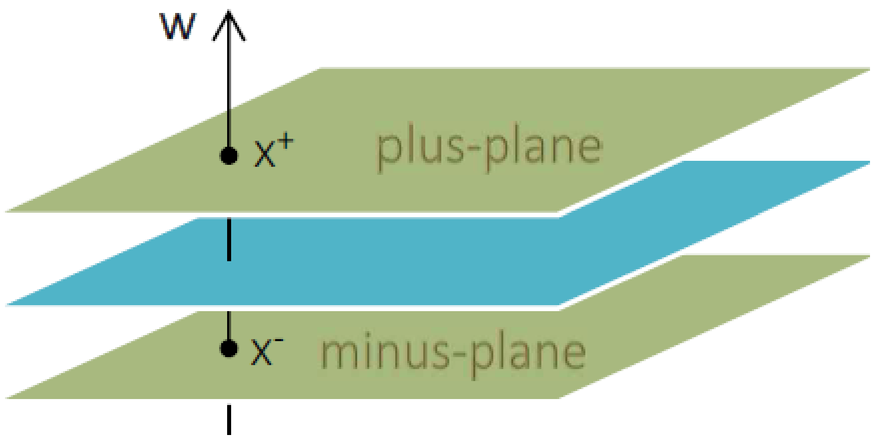
- Optimal hyperplane: Maximize the margin



- Canonical hyperplane: $\mathbf{w}$ and b chosen such that for closest points $|\mathbf{w}\cdot\mathbf{x}+b|=1$.

Support Vector Machine: Formulation

- How to compute the margin?



$$\boldsymbol{w}^T\boldsymbol{x} + b = 1$$

$$\boldsymbol{w}^T\boldsymbol{x} + b = 0$$

$$\boldsymbol{w}^T\boldsymbol{x} + b = -1$$

$$\text{margin} = \frac{1}{||\mathbf{w}||^2}$$

Margin and VC Dimension

- Recall the relationship between margin and VC dimension

The VC dimension of a separating hyperplane with a margin Δ is bounded as follows

$$h \leq min\left(\left\lceil \frac{R^2}{\Delta^2} \right\rceil, D\right) + 1$$

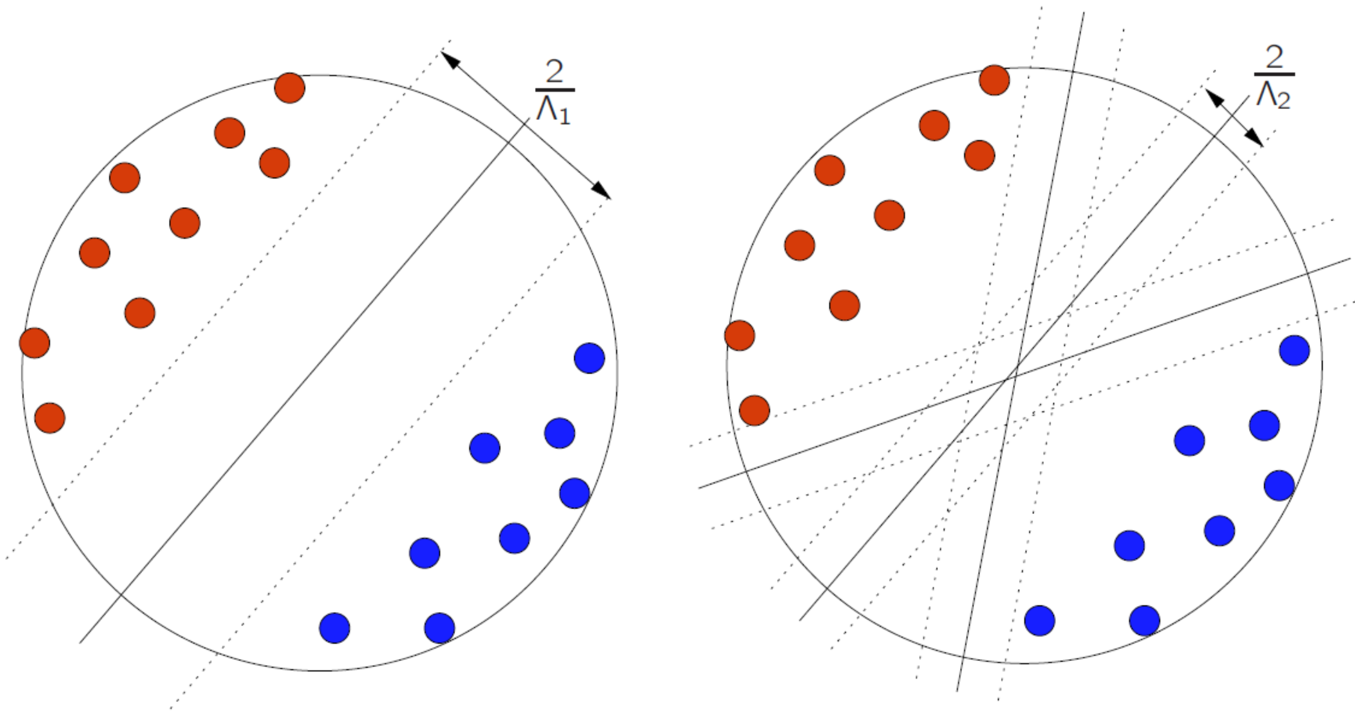
where D is the dimensionality of the input space, and R is the radius of the smallest sphere containing all the input vectors

- Maximizing the margin → Minimizing the VC dimension → Minimizing the Expected Risk

$$R[f] \leq R_{emp}[f] + \sqrt{\frac{h\left(ln\frac{2n}{h} + 1\right) - \ln\left(\frac{\delta}{4}\right)}{n}}$$

Margin and VC Dimension

- An illustrative example



✓ If we choose a hyperplane with a large margin (left), there is only a small number of possibilities to separate the data → lower VC dimension

Support Vector Machine: Cases

- Support Vector Machine Formulation

	Hard margin?	Soft margin?
Linearly separable?	Basic form (Case 1)	Introduce penalty terms (Case 2)
Linearly non-separable?	Utilize Kernel Trick	Introduce penalty terms Utilize Kernel Trick (Case 3)

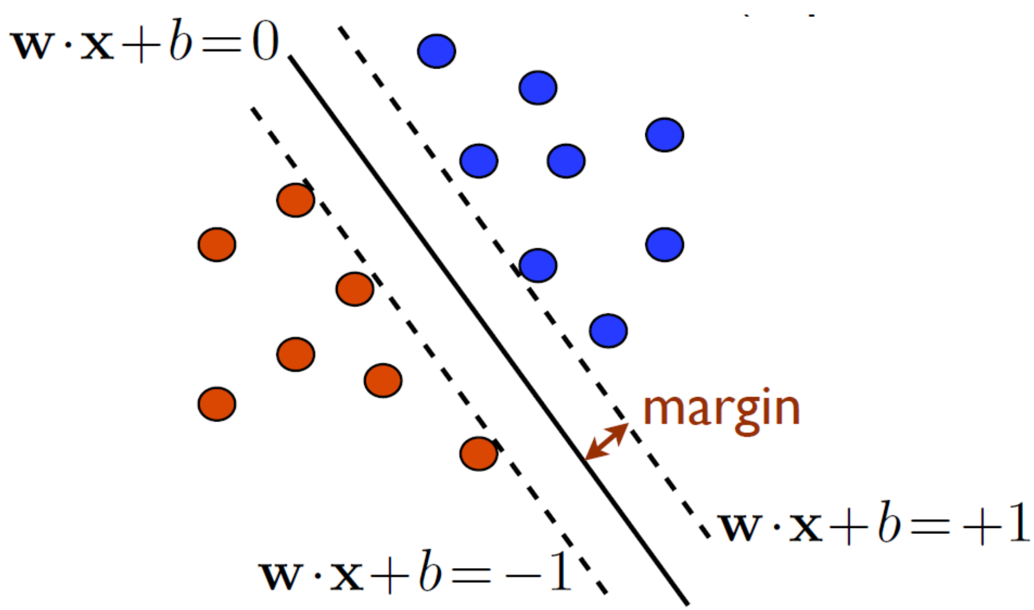
SVM Case 1: Linear Case & Hard Margin

- Optimization Problem
  - ✓ Objective function

$$\min \quad \frac{1}{2}\|\mathbf{w}\|^2$$

  - ✓ Constraints

$$s.t. \quad y_i(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i + b) \geq 1$$



SVM Case 1: Linear Case & Hard Margin

- Optimization Problem
  - ✓ Lagrangian Problem

$$\min \quad L_P(\mathbf{w}, b, \alpha_i) = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{N} \alpha_i \big(y_i \big(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i + b\big) - 1\big)$$

$$s.t. \quad \alpha_i \geq 0$$

  - ✓ KKT conditions

$$\frac{\partial L_p}{\partial \mathbf{w}} = 0 \quad \Rightarrow \quad \mathbf{w} = \sum_{i=1}^{N} \alpha_i \, y_i \mathbf{x}_i \qquad\qquad \frac{\partial L_p}{\partial b} = 0 \quad \Rightarrow \quad \sum_{i=1}^{N} \alpha_i y_i = 0$$

SVM Case 1: Linear Case & Hard Margin

- From Primal to Dual

$$\min \quad L_P(\mathbf{w}, b, \alpha_i) = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{N} \alpha_i\big(y_i\big(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i + b\big) - 1\big)$$
$$s.t. \quad \alpha_i \geq 0$$

$$\max \quad L_D(\alpha_i) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j \mathbf{x}_i^{\mathrm{T}} \mathbf{x}_j$$
$$s.t. \sum_{i=1}^{N} \alpha_i y_i = 0 \quad and \quad \alpha_i \geq 0$$

- Solution

$$f(\mathbf{x}_{new}) = sign\left(\sum_{i=1}^{N} \alpha_i y_i \, \mathbf{x}_i^{\mathrm{T}} \mathbf{x}_{new} + b\right)$$

SVM Case 1: Linear Case & Hard Margin

- From KKT condition, we know that

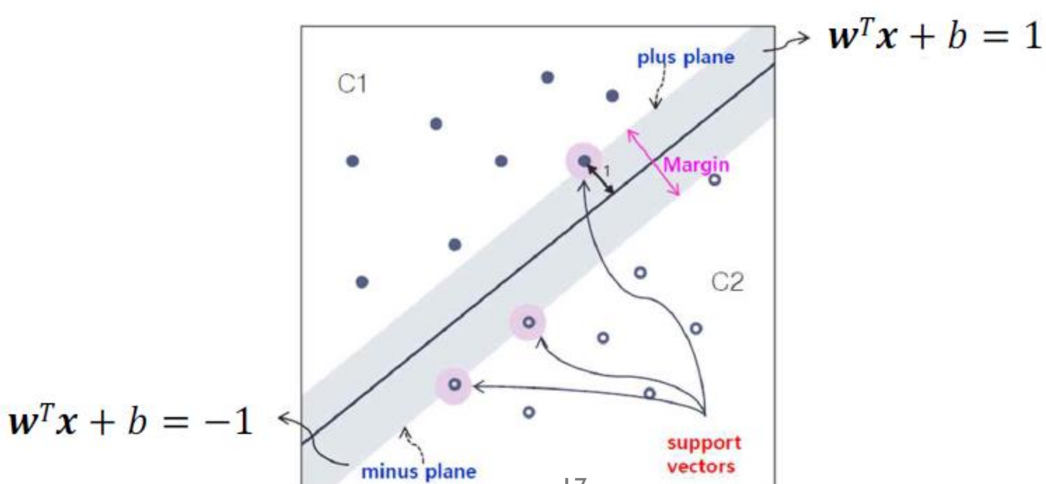
$$\alpha_i\left(y_i\left(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i + b\right) - 1\right) = 0$$

  ✓ Thus, the only support vectors have $\alpha_i \neq 0$

  ✓ The solution has the form

$$\mathbf{w} = \sum_{i=1}^{N} \alpha_i\, y_i \mathbf{x}_i = \sum_{i \in SV}^{N} \alpha_i\, y_i \mathbf{x}_i$$

  ✓ b can be computed by $y_i\left(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i + b\right) - 1 = 0$ with a support vector $\mathbf{x}_i$



SVM Case 1: Linear Case & Hard Margin

- Compute the margin
  - ✓ Since the SVs lie on the marginal hyperplanes, for any support vector $\mathbf{x}_i$, $\mathbf{w}^{\mathrm{T}}\mathbf{x}_i + b = y_i$

$$b = y_i - \sum_{i=1}^{N} \alpha_i y_i (\mathbf{x}_j, \mathbf{x}_i)$$

  - ✓ Multiplying both sides by $\alpha_i y_i$ and taking the sum leads to

$$\sum_{i=1}^{N} \alpha_i y_i b = \sum_{i=1}^{N} \alpha_i y_i^2 - \sum_{i,j=1}^{N} \alpha_i \alpha_j y_i y_j (\mathbf{x}_i, \mathbf{x}_j)$$

  - ✓ Using the fact that $y_i^2 = 1$

$$0 = \sum_{i=1}^{N} \alpha_i - \mathbf{w}^{\mathrm{T}}\mathbf{w}$$

$$\rho^2 = \frac{1}{\|\mathbf{w}\|_2^2} = \frac{1}{\sum_{i=1}^{N} \alpha_i} = \frac{1}{\|\boldsymbol{\alpha}\|_1}$$


Q & A

References

Research Papers

- Burges, C.J.C. (1998). A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery 2: 121-167.
- Müller, K., Mika, S., Rätsch, G., Tsuda, K., and Schölkopf, B. (2001). An introduction to kernel-based learning algorithms. IEEE Transactions on Neural Networks 12(2): 181-201.
- Zhang, C., Bengio, S., Hardt, M., Recht, B., & Vinyals, O. (2016). Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*.