

Machine Learning Basics & Support Vector Machine (SVM) 1



한양대학교 ERICA
소프트웨어융합대학
COLLEGE OF COMPUTING

인공지능학과
Department of
Artificial Intelligence

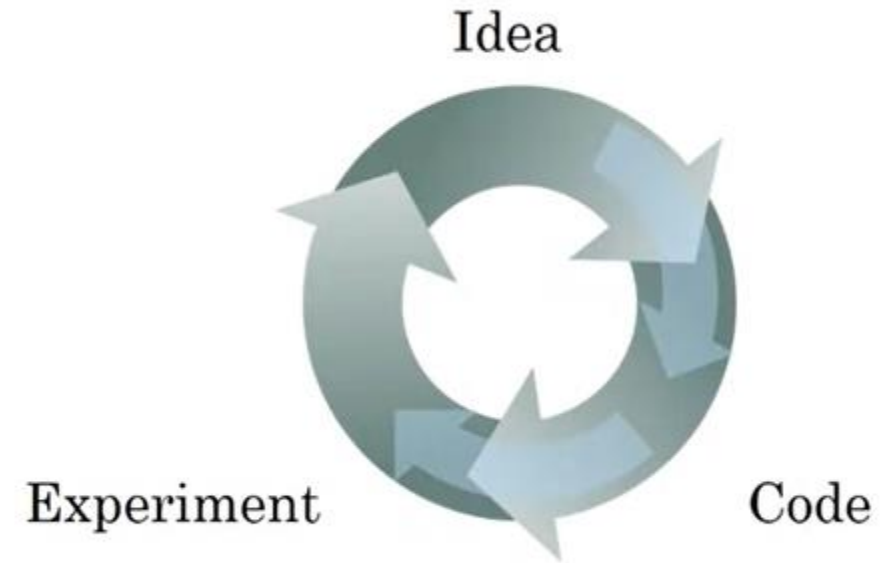
정 우 환 (whjung@hanyang.ac.kr)

Fall 2021

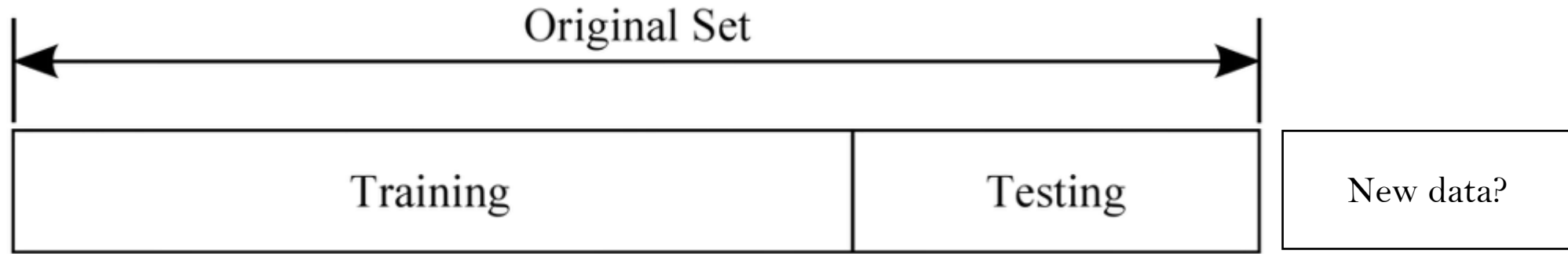
Train, Validation, Test datasets

Applied ML is a highly iterative process

- Hyper parameters
 - # layers
 - # hidden units
 - Learning rates
 - Activation functions
 - ...

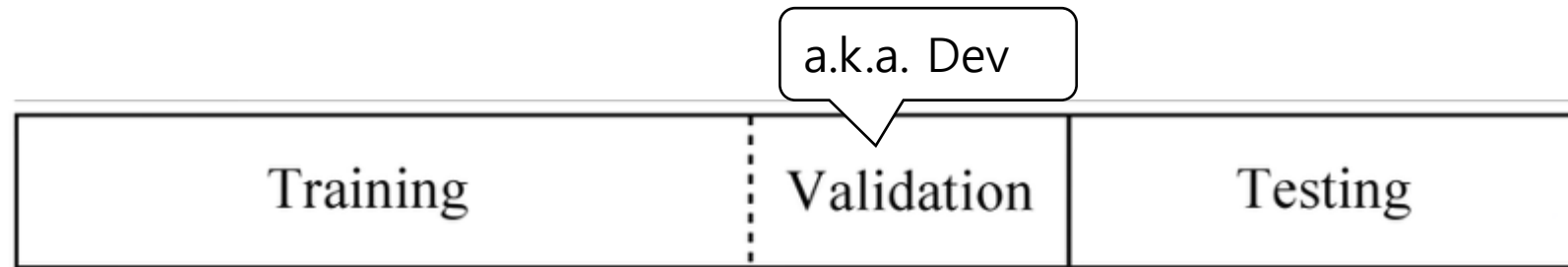


Train, Valid, Test datasets



Learn parameters

Tune hyperparameters



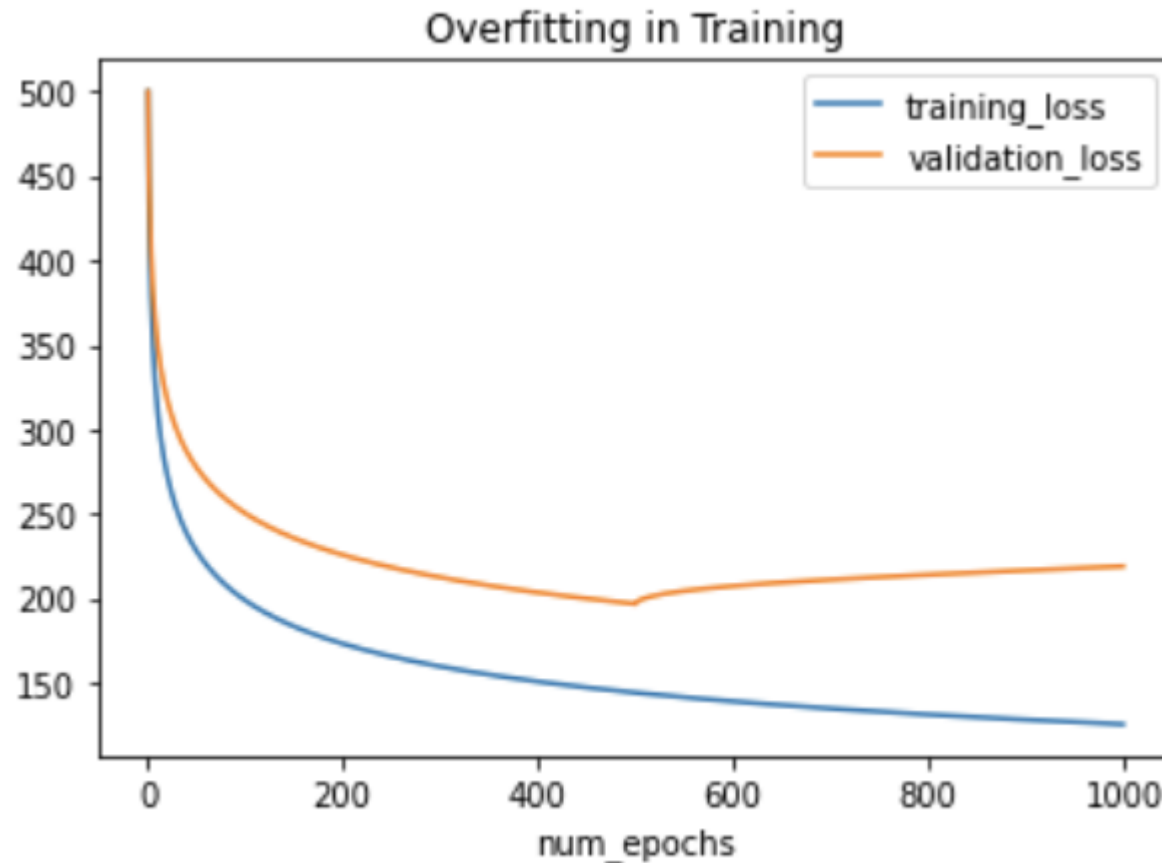
Learn parameters

Tune hyperparameters

Final evaluation

Early Stopping

“The number of epochs” is a hyperparameter!



Mismatched train/test distribution

Training set:

Cat pictures from
webpages

Dev/test sets:

Cat pictures from users
using your app

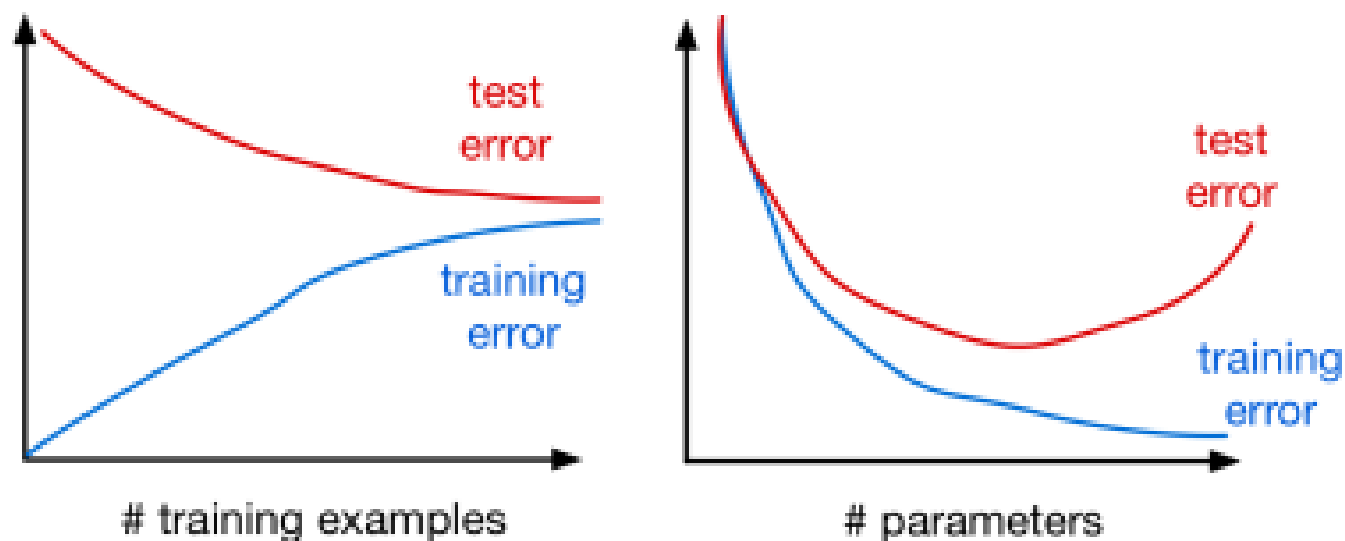
Test

Dev

Test

Generalization Performance

The *generalization performance* of a learning algorithm refers to the performance on out-of-sample data of the models learned by the algorithm.

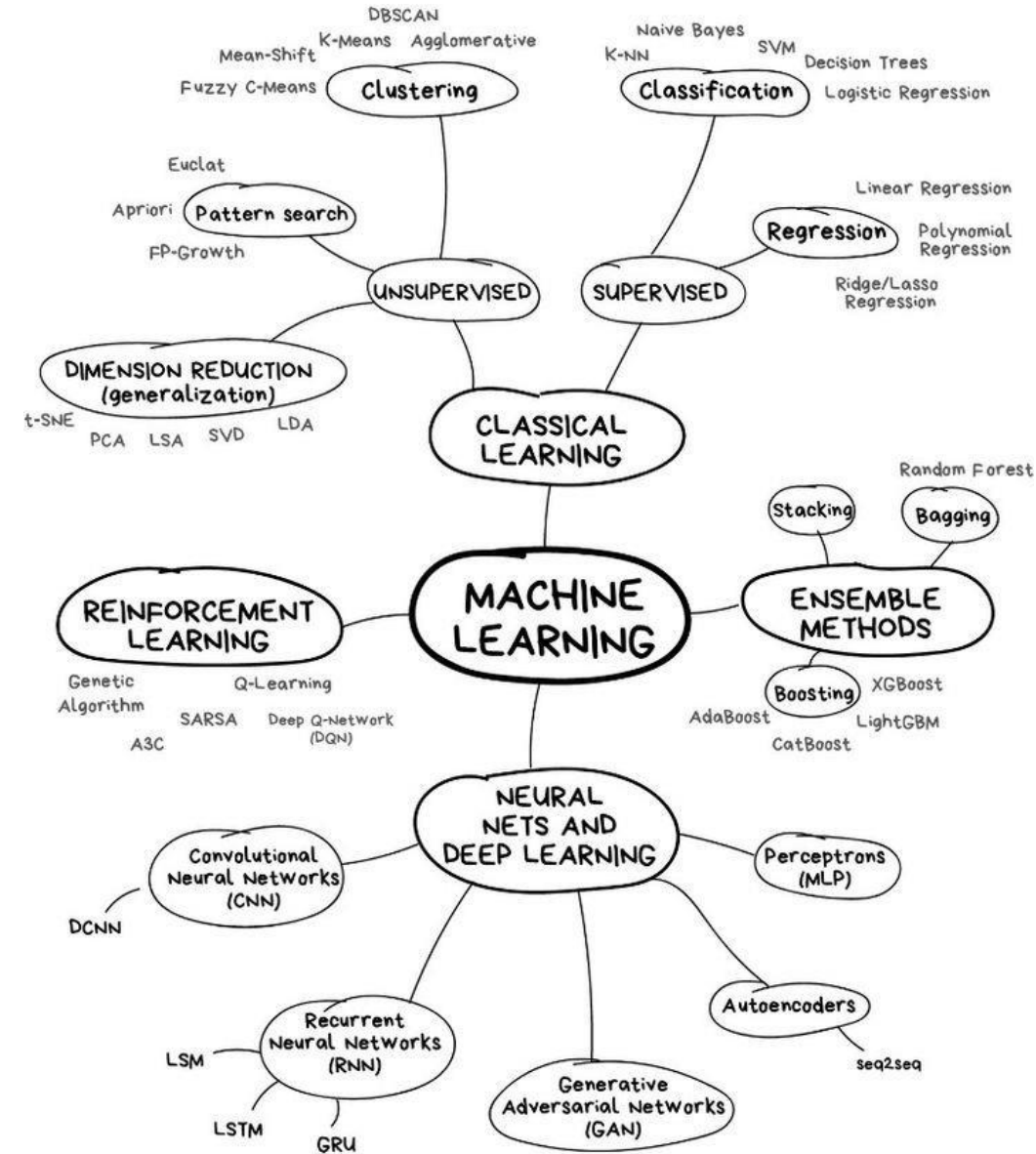


Improving Generalization Performance :Avoiding Overfitting

- Regularization (Weight decay)
- Early stopping
- Reducing model size
- ...

Non-neural classification algorithms

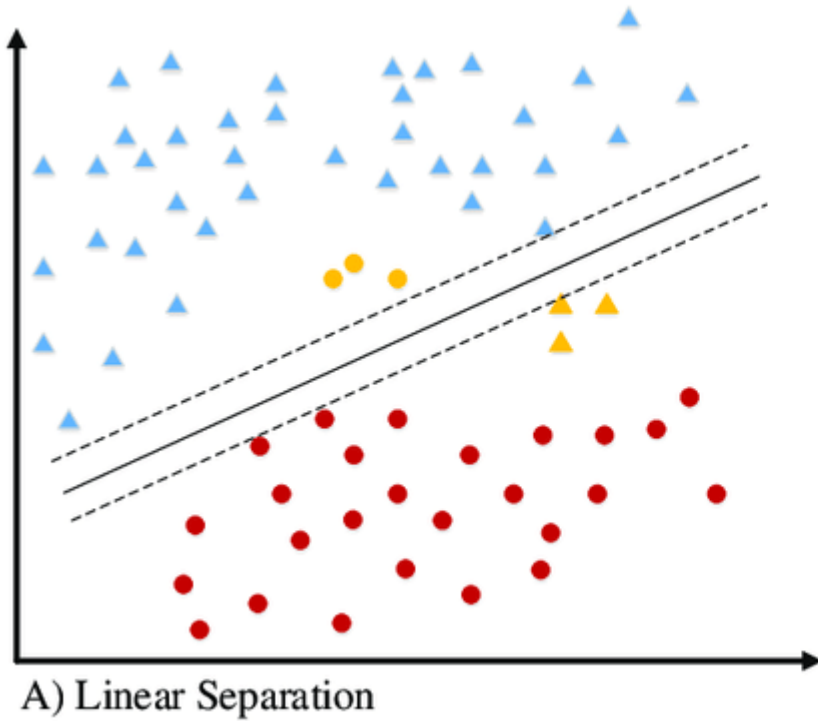
- K-nearest neighbor (k-NN) classifier
- Naïve Bayes classifiers
- Decision trees
- Support Vector Machine (SVM)



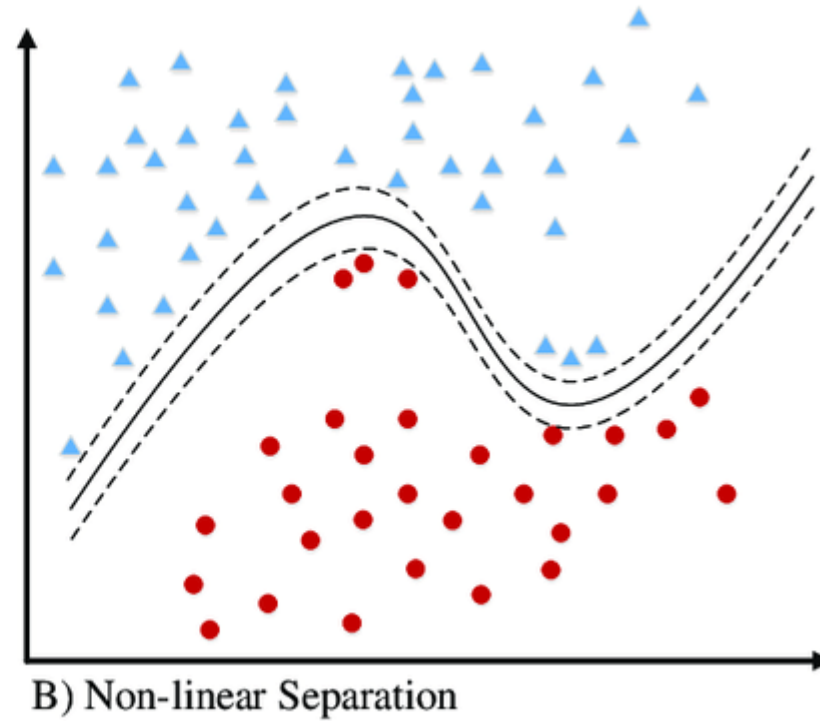
SVM: Support Vector Machines

- A classification method for both **linear** and **nonlinear** data developed by Vapnik et al. (1992)
- Advantages:
 - Good **generalization performance**
 - Effective in high dimensional spaces (Even when # dims > # samples)
- Disadvantages:
 - Not suitable for large data sets.
 - SVMs do not directly provide probability estimates
- Applications:
 - handwritten digit recognition, object recognition, speaker identification, benchmarking time-series prediction tests,...

SVM: Support Vector Machines

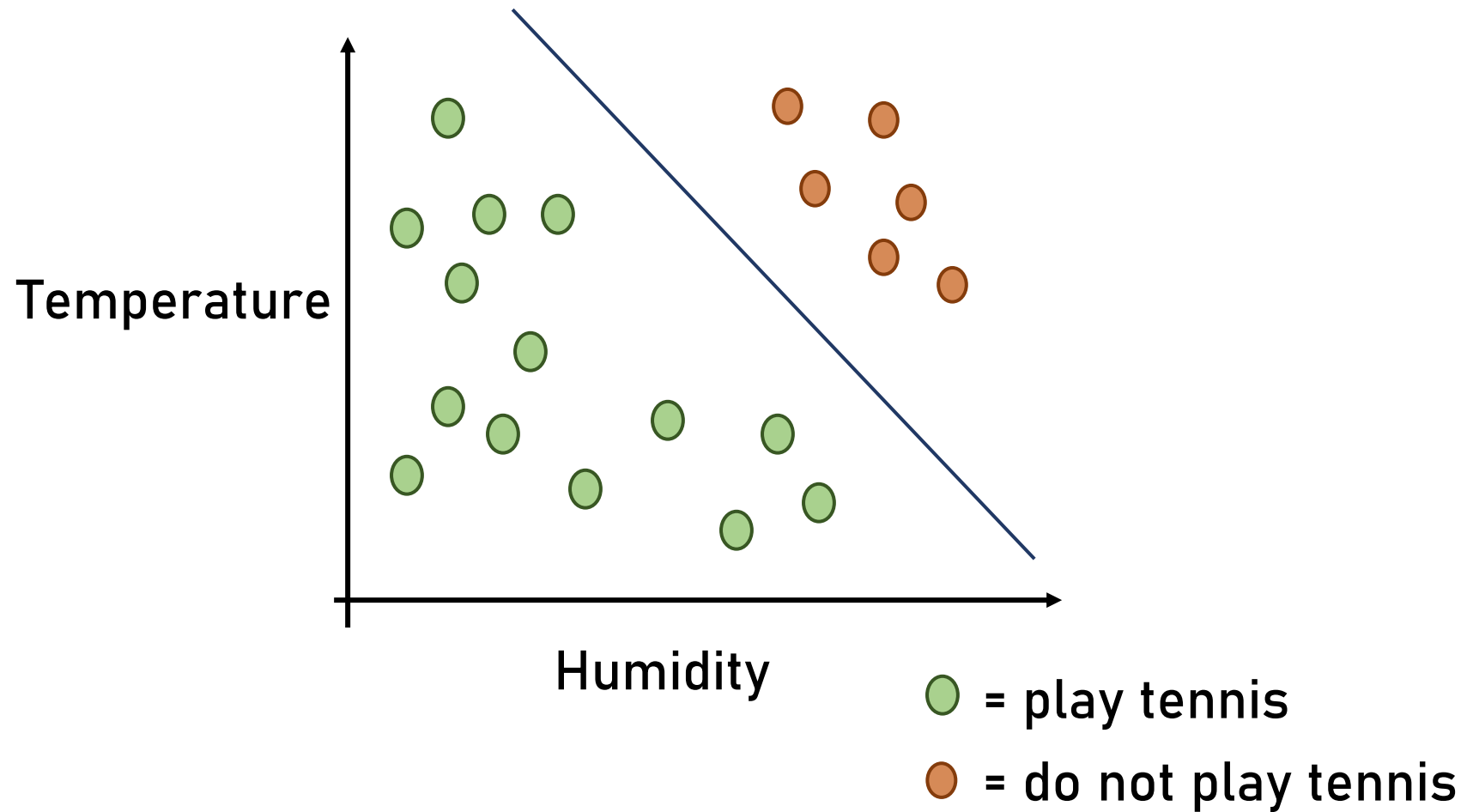


Linear SVM

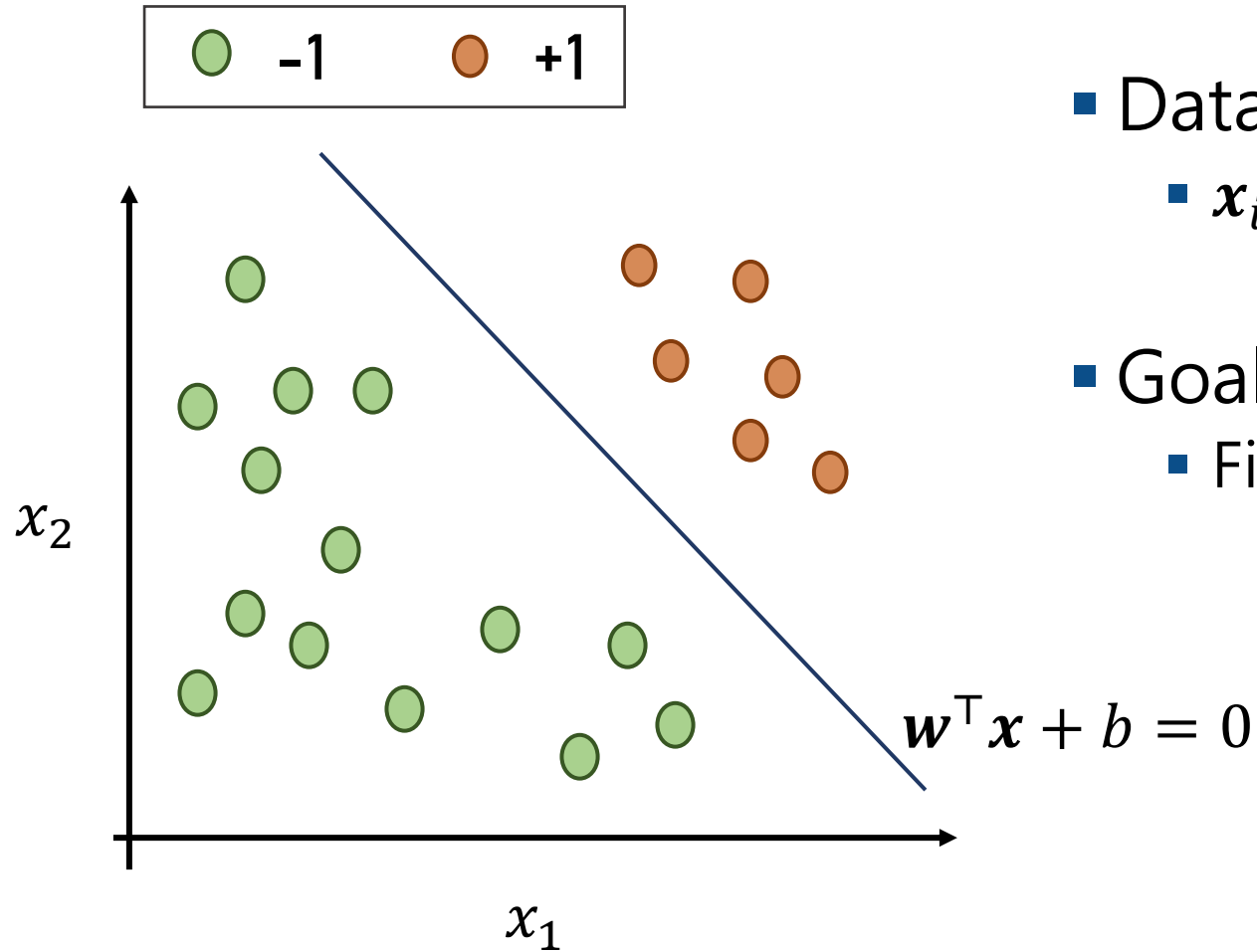


Non-linear SVM

Tennis example



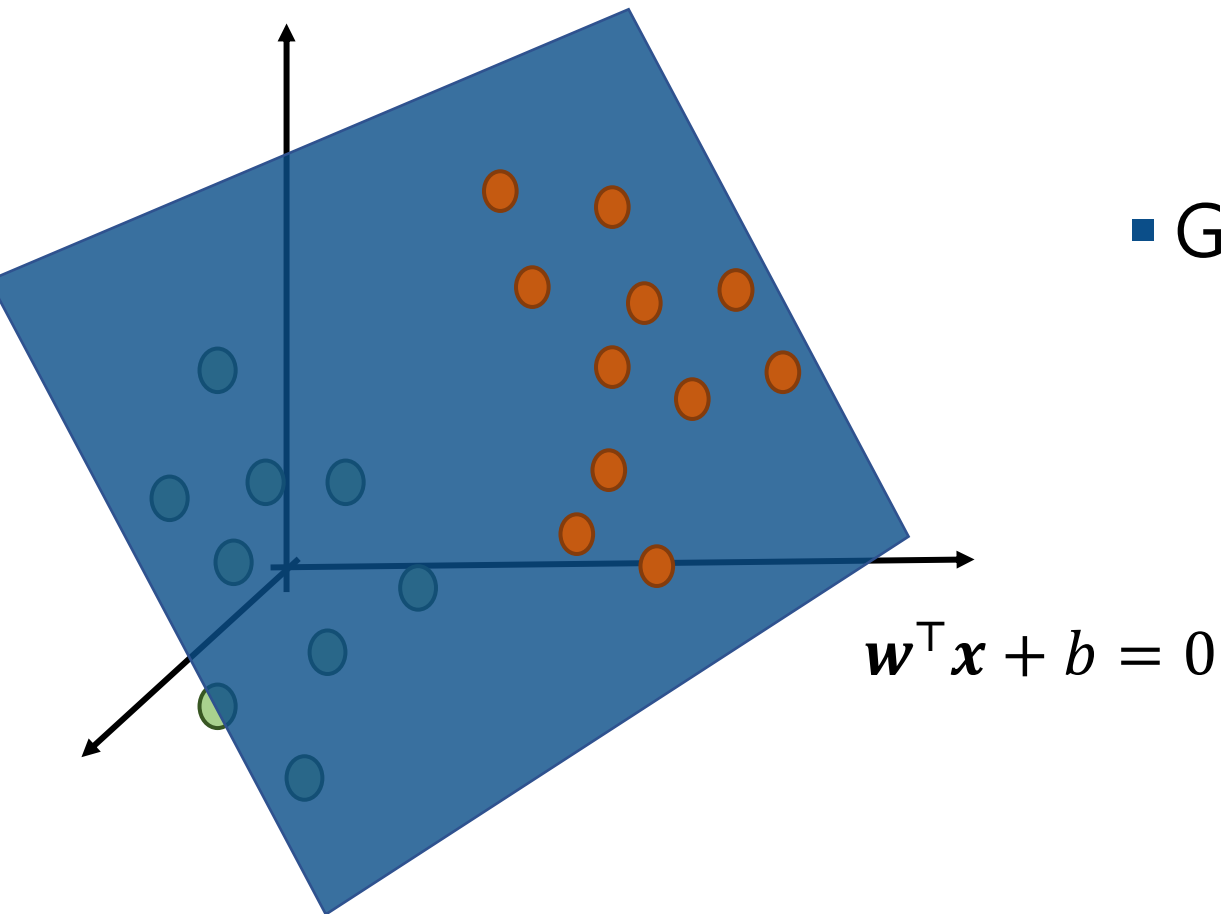
Linear SVM



- Data: $\langle x_i, y_i \rangle$ for $i = 1, \dots, n$
 - $x_i \in \mathbb{R}^d$, $y_i \in \{-1, +1\}$
- Goal
 - Finding a good separating hyperplane

$$w^T x + b = 0$$

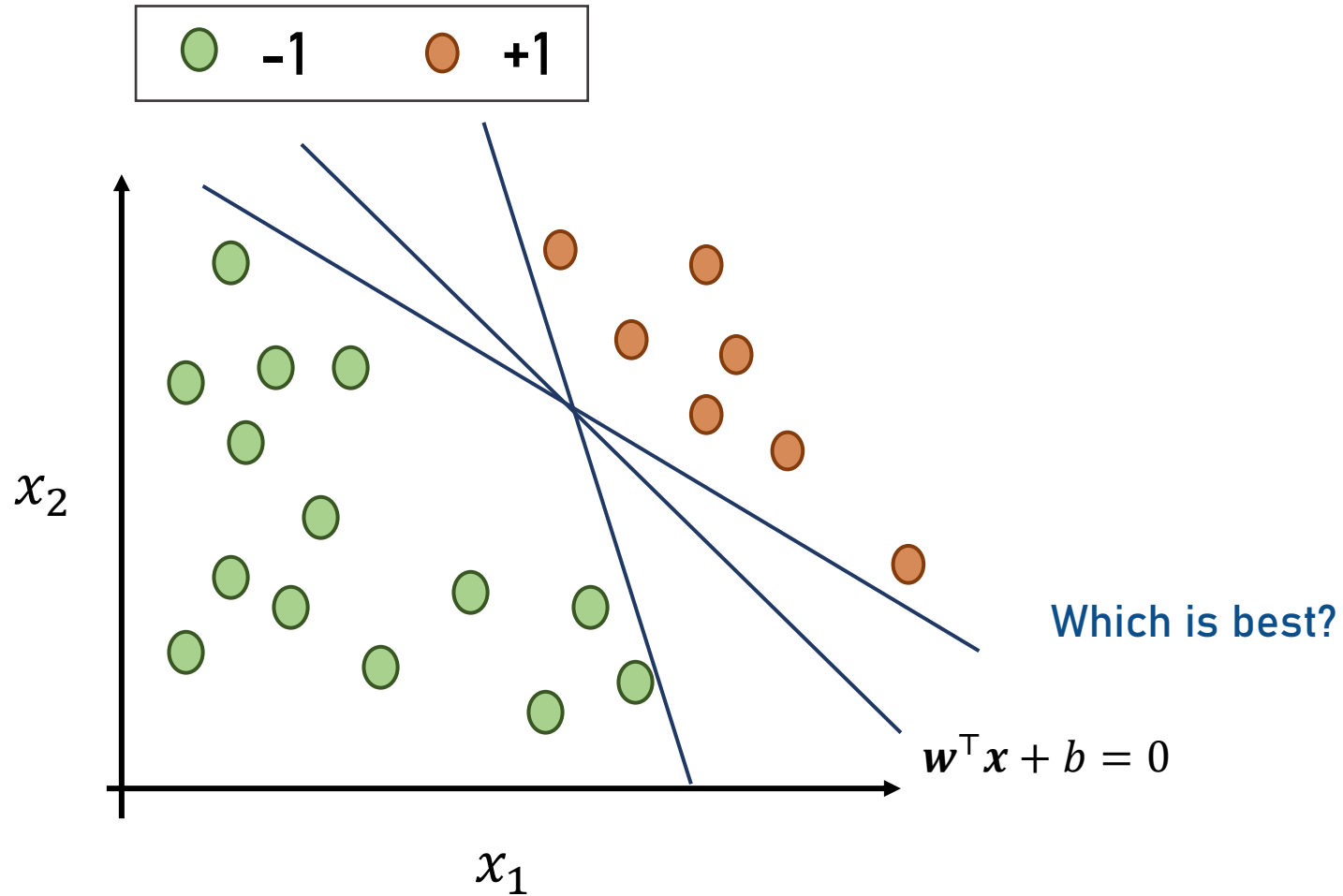
Linear SVM



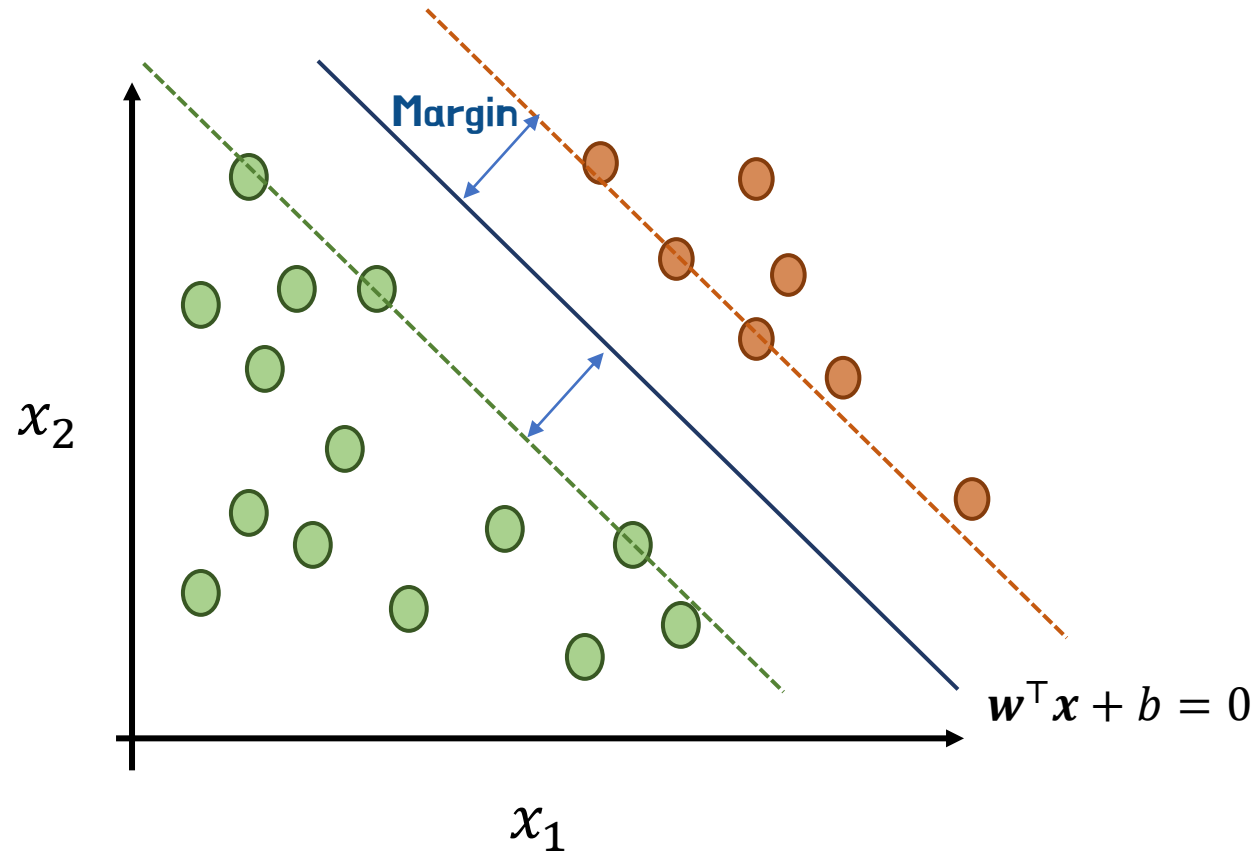
- Data: $\langle x_i, y_i \rangle$ for $i = 1, \dots, n$
 - $x_i \in \mathbb{R}^d$, $y_i \in \{-1, +1\}$
- Goal
 - Finding a good separating hyperplane

$$w^T x + b = 0$$

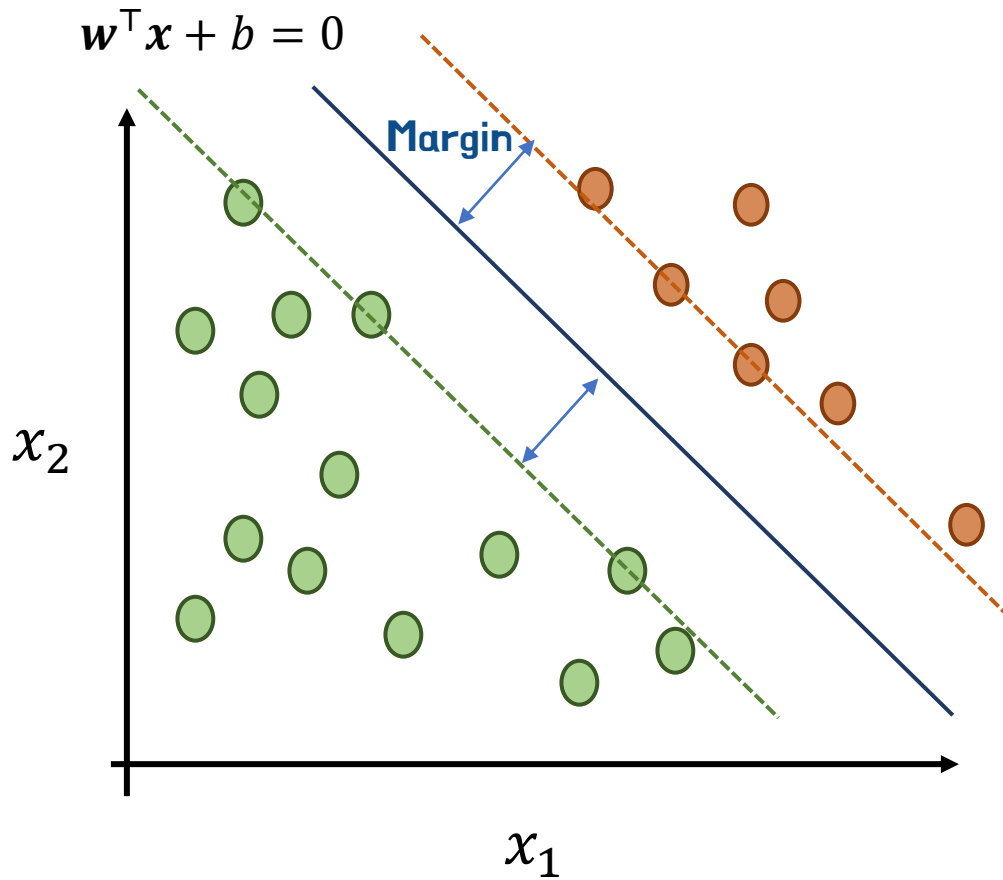
Separating Hyperplane



Separating Hyperplane

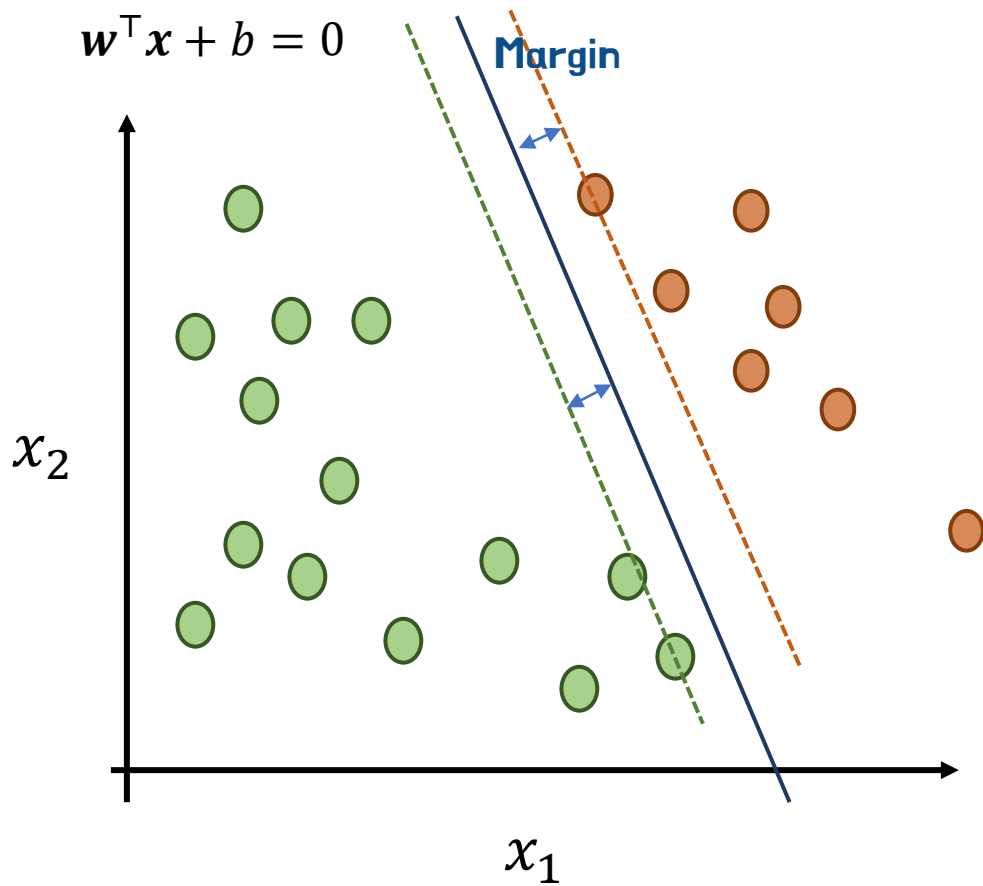


Margin



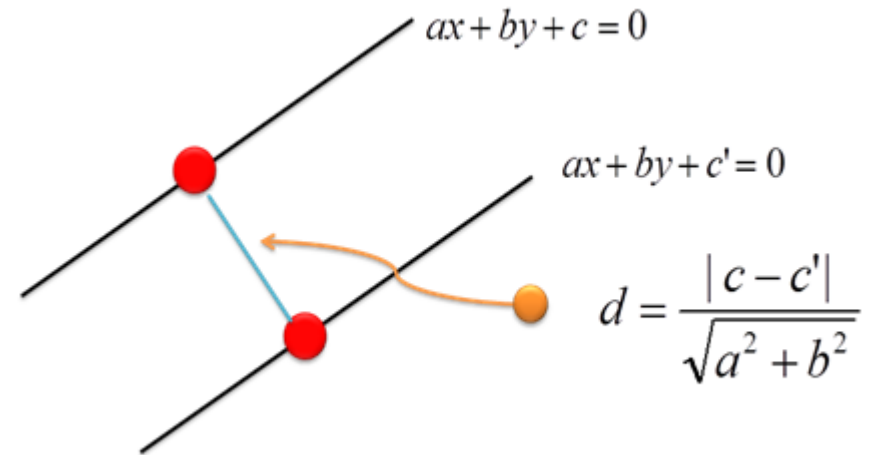
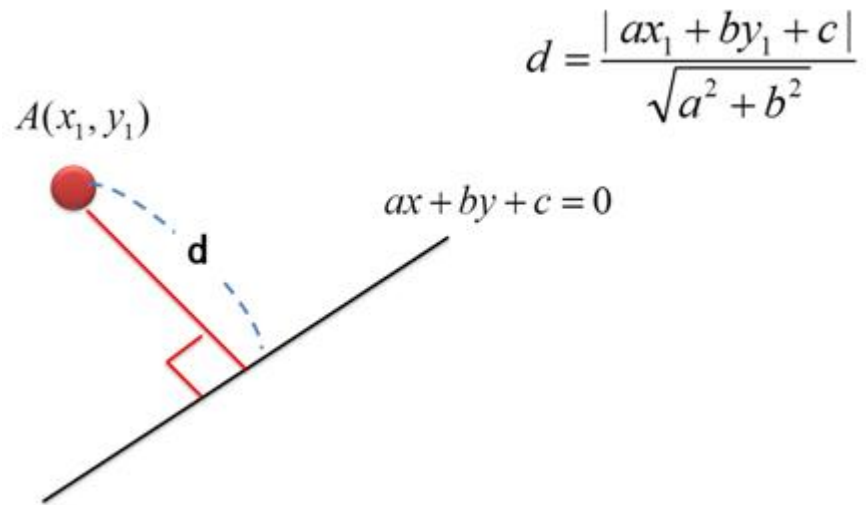
Maximizing **margin** over the training set
= Minimizing **generalization error**

Margin

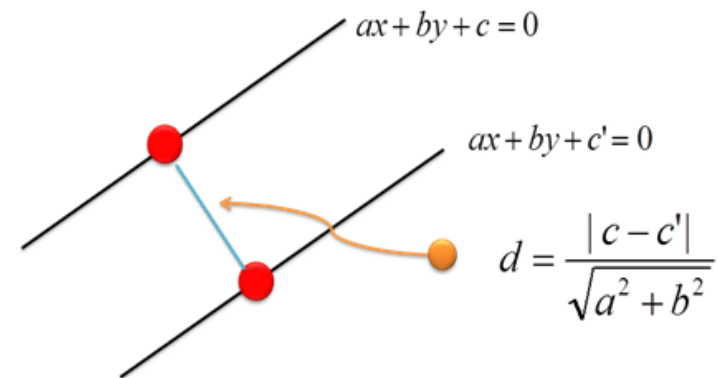
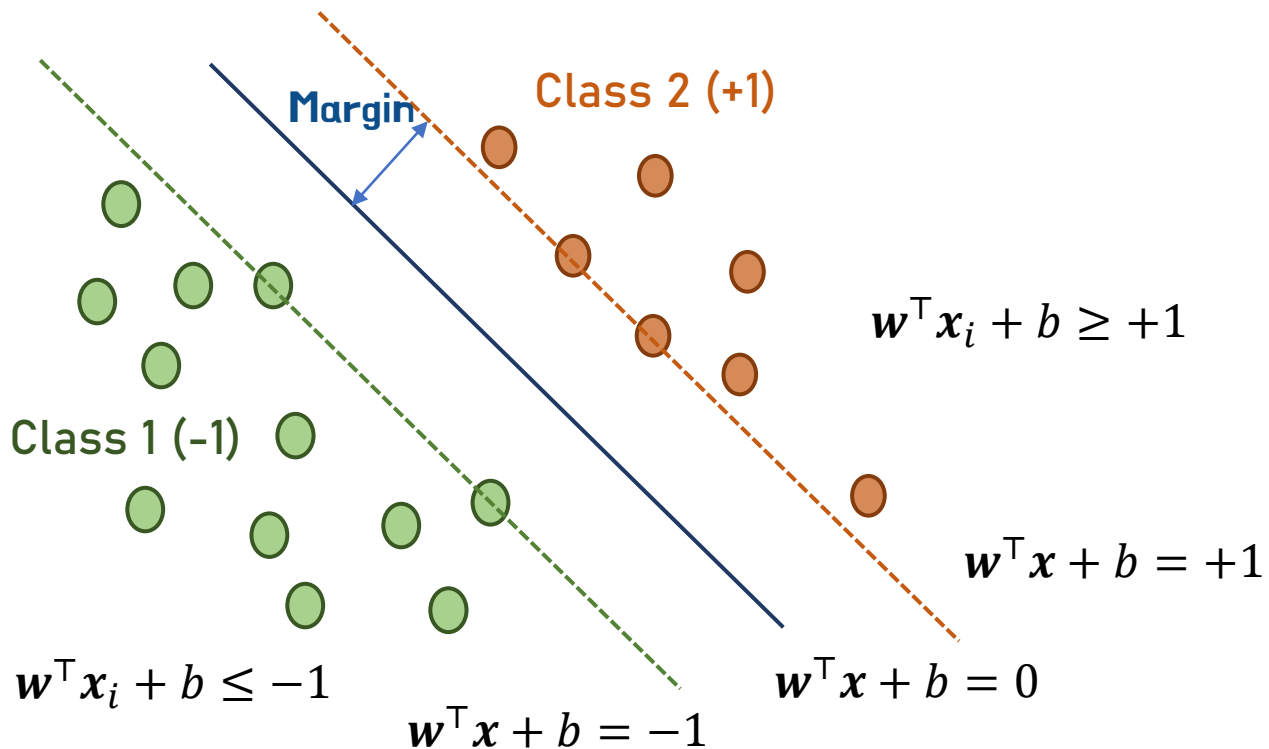


Maximizing **margin** over the training set
= Minimizing **generalization error**

Distance from a line



Margin



$$(Margin) = \frac{1}{\|w\|_2}$$

In order to maximize the margin,
we need to minimize $\|w\|$.
With the condition that there are
no datapoints between H1 and H2:
 $x_i \cdot w + b \geq +1$ when $y_i = +1$
 $x_i \cdot w + b \leq -1$ when $y_i = -1$
Can be combined into $y_i(x_i \cdot w) \geq 1$

Constrained Optimization Problem

$$\underset{\mathbf{w}, b}{\text{minimize}} \quad \frac{1}{2} \|\mathbf{w}\|_2^2$$

$$\text{subject to } y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1, i = 1, 2, \dots, n$$

- Learnable parameter는 \mathbf{w}, b
- Margin $\|\mathbf{w}\|_2$ 을 최소화
- Constraint 만족하면 training data를 완벽하게 separating하게 됨

Lagrangian Formulation

Original Problem

$$\text{minimize } \frac{1}{2} \|w\|_2^2$$

$$\text{subject to } y_i(w^T x_i + b) \geq 1, i = 1, 2, \dots, n$$

Lagrangian multiplier를 이용하여 Lagrangian primal문제로 변환

Lagrangian Primal

$$\max_{\alpha} \min_{w, b} \mathcal{L}(w, b, \alpha) = \frac{1}{2} \|w\|_2^2 - \sum_{i=1}^n \alpha_i (y_i(w^T x_i + b) - 1)$$

$$\text{subject to } \alpha_i \geq 0, i = 1, 2, \dots, n$$

Lagrangian Formulation

Lagrangian Primal

$$\max_{\alpha} \min_{w, b} \mathcal{L}(w, b, \alpha) = \frac{1}{2} \|w\|_2^2 - \sum_{i=1}^n \alpha_i (y_i (w^T x_i + b) - 1)$$

subject to $\alpha_i \geq 0, i = 1, 2, \dots, n$

Convex, continuous이기 때문에 미분 = 0에서 최소값을 가짐

$$\textcircled{1} \quad \frac{\partial \mathcal{L}(w, b, \alpha)}{\partial \underline{w}} = 0 \quad \Longrightarrow \quad \overset{w-}{w} = \sum_{i=1}^n \alpha_i y_i x_i$$

$$\textcircled{2} \quad \frac{\partial \mathcal{L}(w, b, \alpha)}{\partial \underline{b}} = 0 \quad \Longrightarrow \quad \sum_{i=1}^n \alpha_i y_i = 0$$

Lagrangian Formulation

$$\underbrace{\frac{1}{2} \|w\|_2^2}_{\textcircled{1}} - \underbrace{\sum_{i=1}^n \alpha_i (y_i (w^T x_i + b) - 1)}_{\textcircled{2}}$$

$$\textcircled{1} \quad \frac{1}{2} \|w\|_2^2 = \frac{1}{2} w^T w$$

$$= \frac{1}{2} w^T \sum_{j=1}^n \alpha_j y_j x_j$$

$$= \frac{1}{2} \sum_{j=1}^n \alpha_j y_j (w^T x_j)$$

$$= \frac{1}{2} \sum_{j=1}^n \alpha_j y_j \left(\sum_{i=1}^n \alpha_i y_i x_i^T x_j \right)$$

$$= \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j x_i^T x_j$$

RECALL

$$\frac{\partial \mathcal{L}(w, b, \alpha)}{\partial w} = 0 \implies w = \sum_{i=1}^n \alpha_i y_i x_i$$

Lagrangian Formulation

$$\underbrace{\frac{1}{2} \|w\|_2^2}_{\textcircled{1}} - \underbrace{\sum_{i=1}^n \alpha_i (y_i (w^T x_i + b) - 1)}_{\textcircled{2}}$$

$$\textcircled{2} \quad - \sum_{i=1}^n \alpha_i (y_i (w^T x_i + b) - 1)$$

$$= - \sum_{i=1}^n \alpha_i y_i (w^T x_i + b) + \sum_{i=1}^n \alpha_i$$

$$= - \sum_{i=1}^n \alpha_i y_i w^T x_i - b \sum_{i=1}^n \alpha_i y_i + \sum_{i=1}^n \alpha_i$$

$$= - \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j x_i^T x_j + \sum_{i=1}^n \alpha_i$$

RECALL

$$\begin{aligned} \frac{\partial \mathcal{L}(w, b, \alpha)}{\partial b} = 0 &\implies \sum_{i=1}^n \alpha_i y_i = 0 \\ \frac{\partial \mathcal{L}(w, b, \alpha)}{\partial w} = 0 &\implies w = \sum_{i=1}^n \alpha_i y_i x_i \end{aligned}$$

Lagrangian Dual

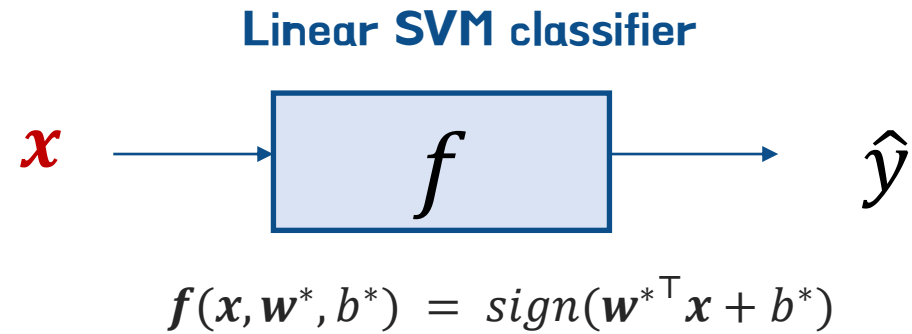
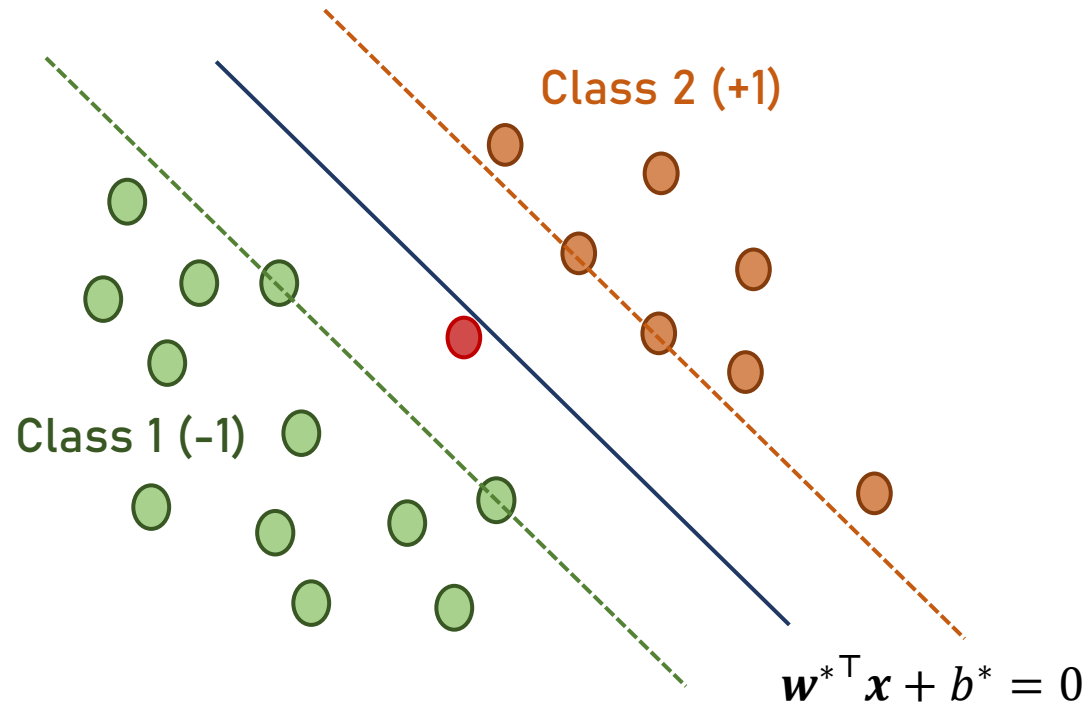
$$\min_{w,b} \frac{1}{2} \|w\|_2^2 - \sum_{i=1}^n \alpha_i (y_i (w^T x_i + b) - 1)$$

$$= \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j x_i^T x_j$$

$$\text{where } \sum_{i=1}^n \alpha_i y_i = 0$$

- Quadratic programming formulation
- Convex optimization을 통해 풀수 있음

Classifying a New Data Point



References

- Andrew W. Moore's slides:
 - <http://www.cs.cmu.edu/~awm/tutorials>
- Seoung Bum Kim's slides:
 - <https://youtu.be/qFg8cDnqYCI>
- Kyuseok Shim's slides