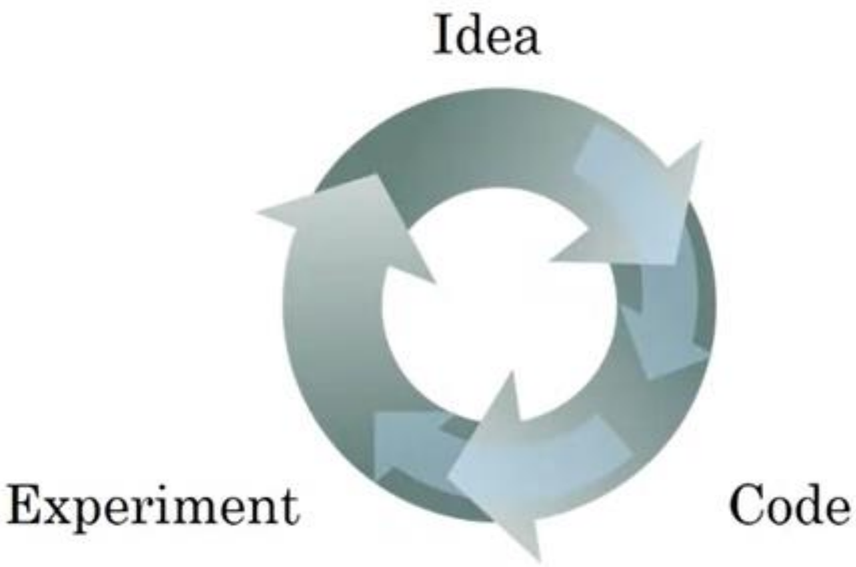
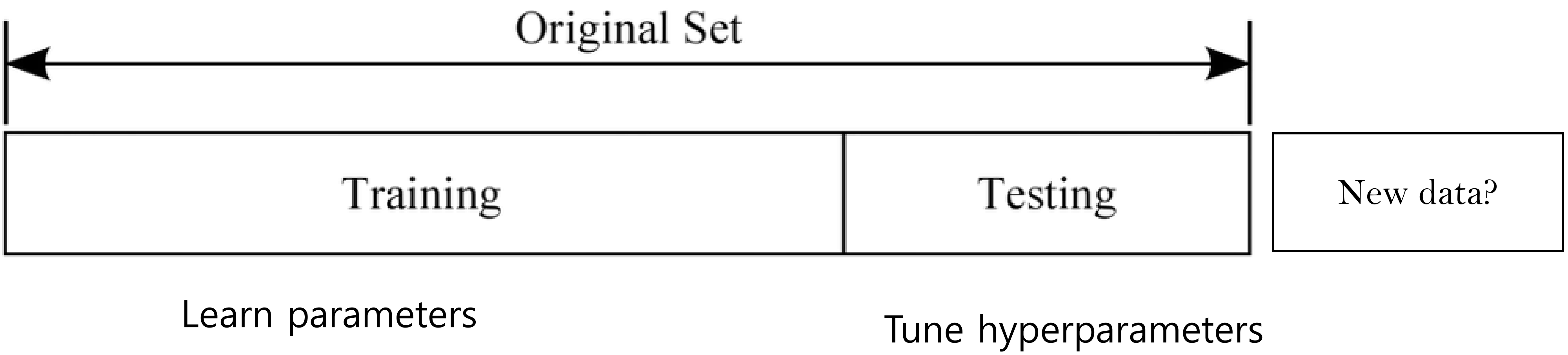
Train, Validation, Test datasets

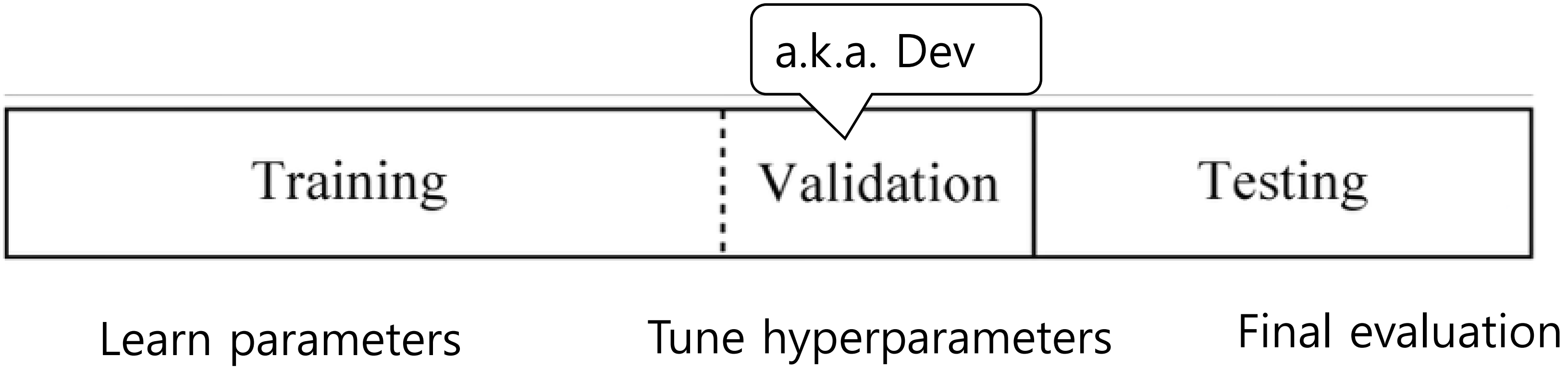
Applied ML is a highly iterative process

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



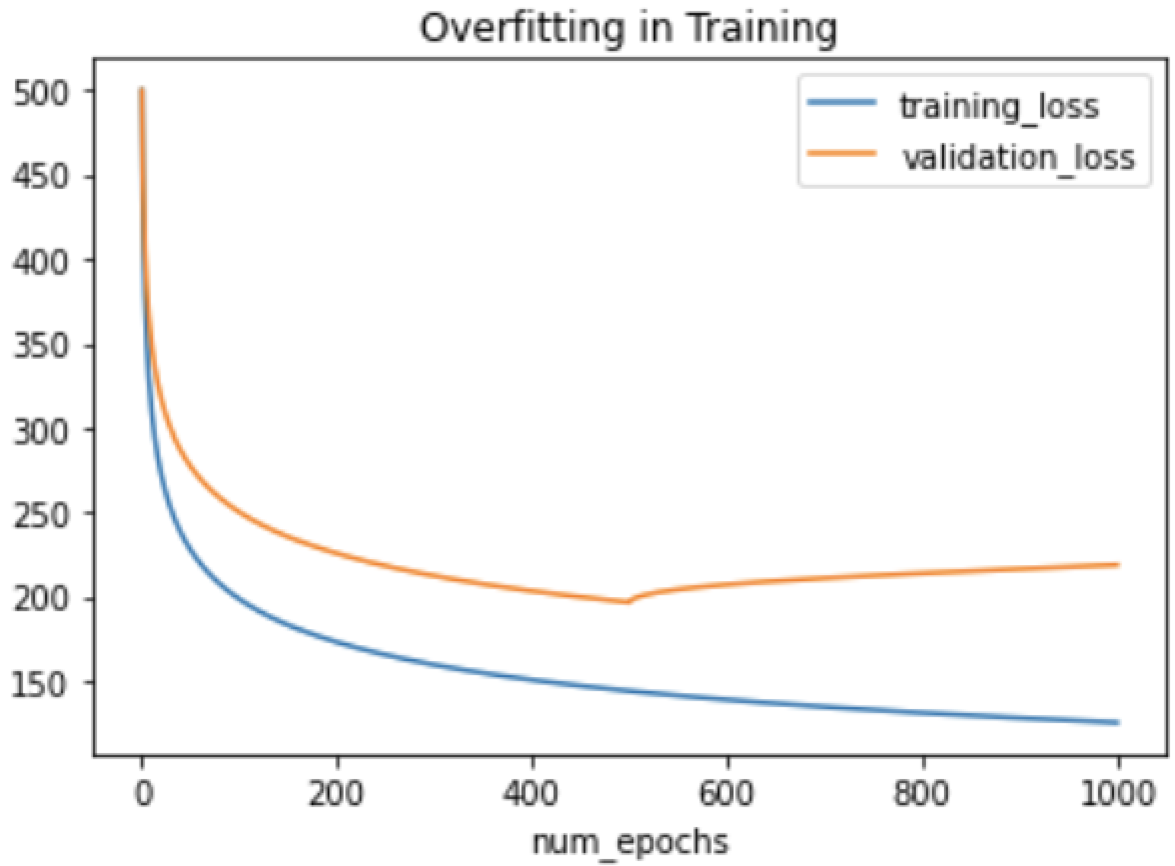
Train, Valid, Test datasets





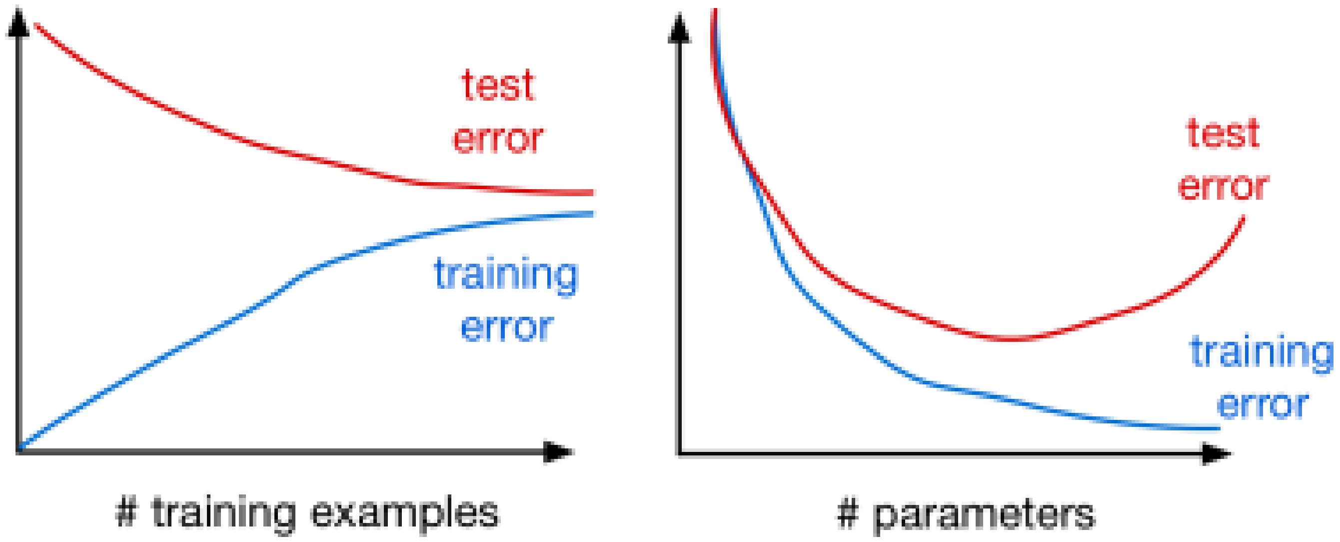
Early Stopping

"The number of epochs" is a hyperparameter!



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.

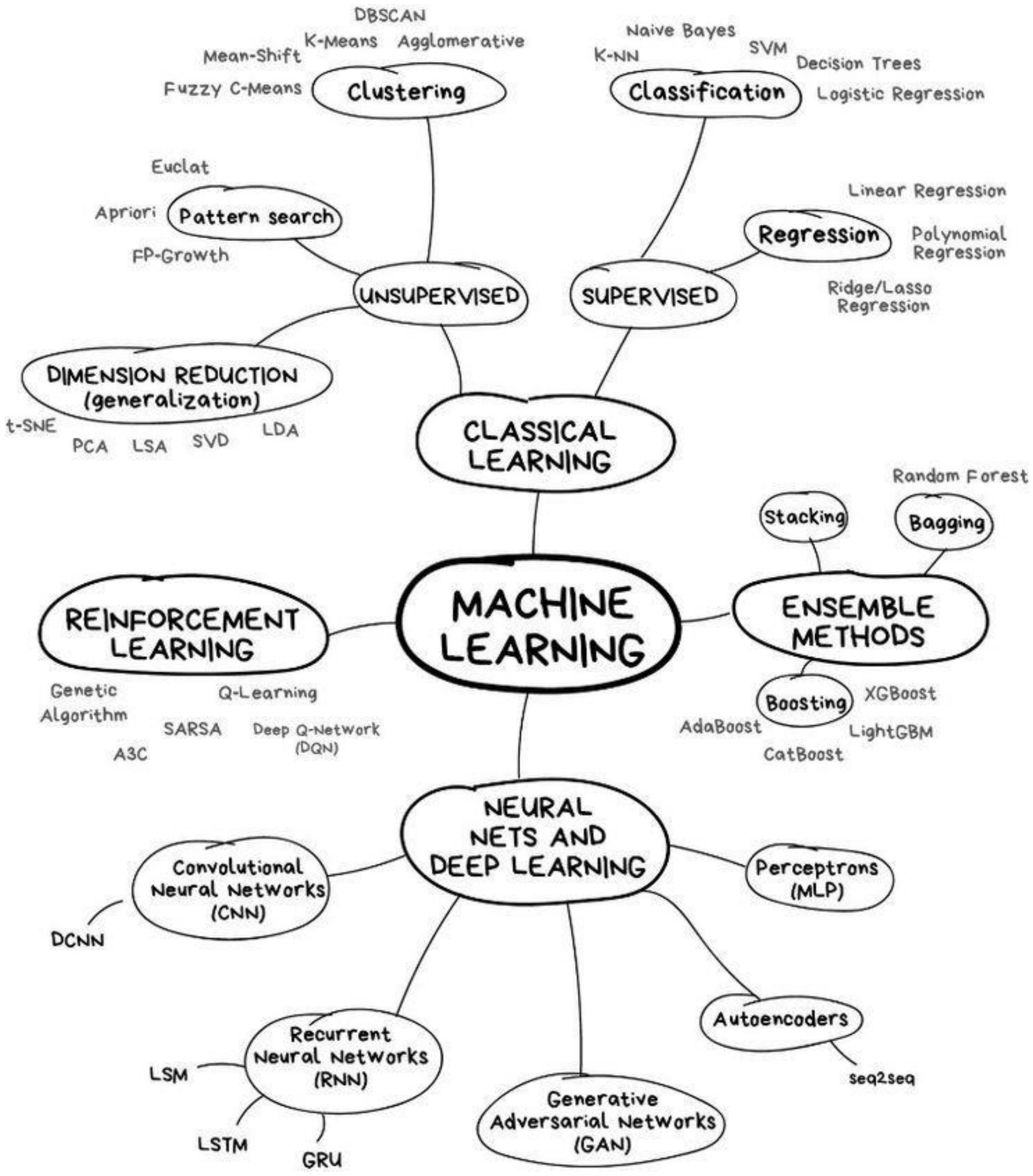


Improving Generalization Performance :Avoiding Overfitting

- 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)**



Support Vector Machine (SVM)

Basic Model

한양대학교 ERICA
소프트웨어융합대학
COLLEGE OF COMPUTING
인공지능학과
Department of Artificial Intelligence

정 우 환 (whjung@hanyang.ac.kr)

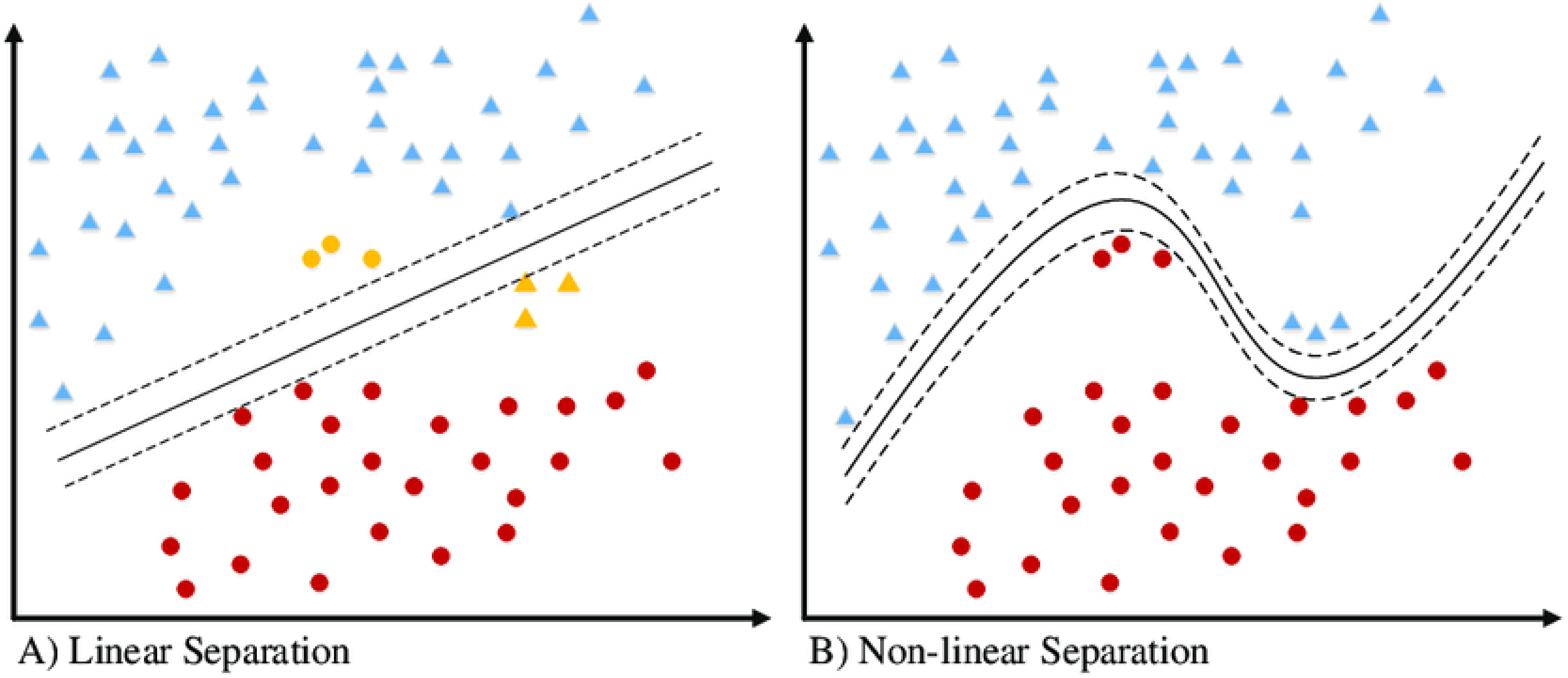
Fall 2022

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,…

http://scikit-learn.org/stable/modules/svm.html

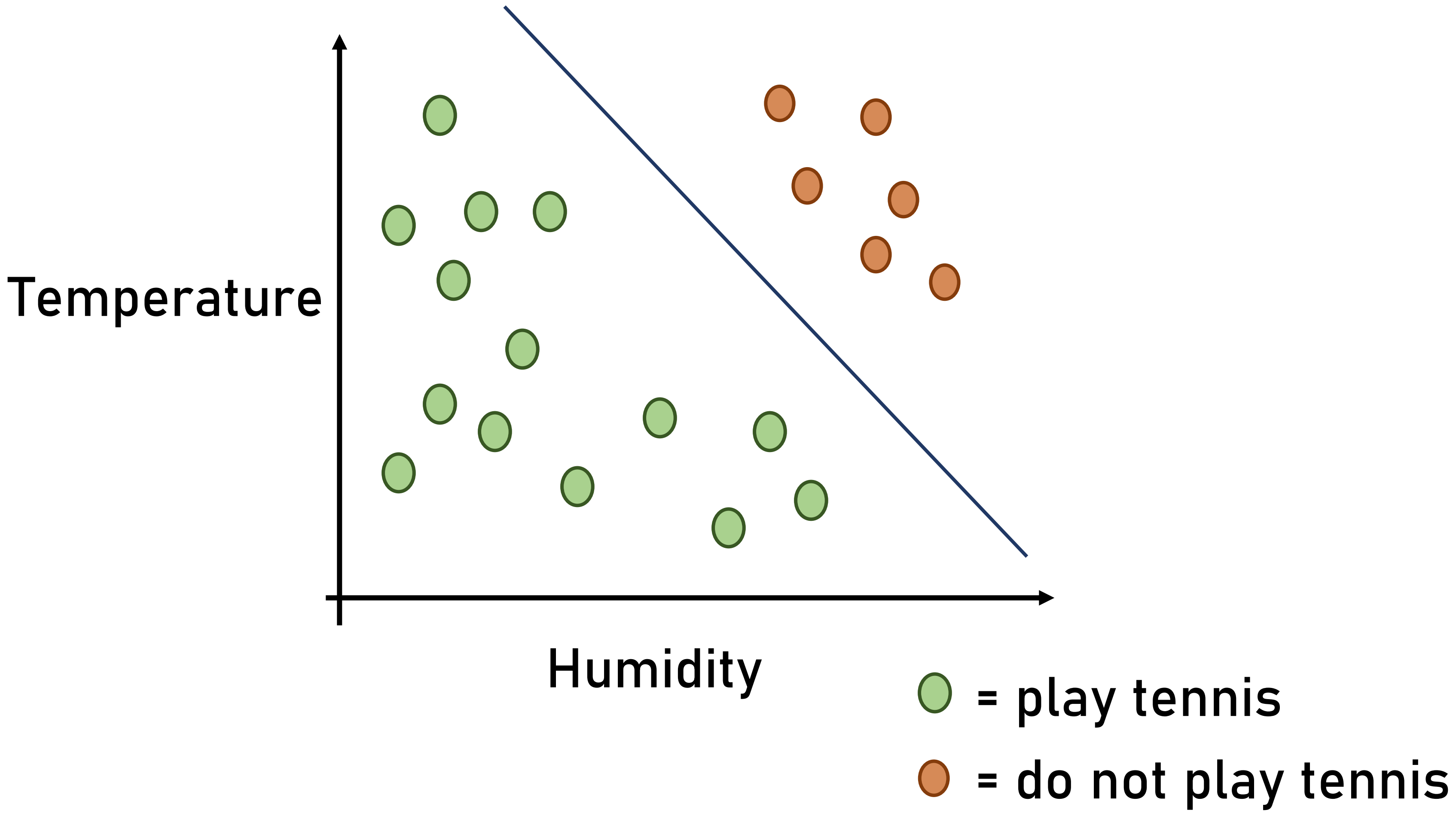
SVM: Support Vector Machines



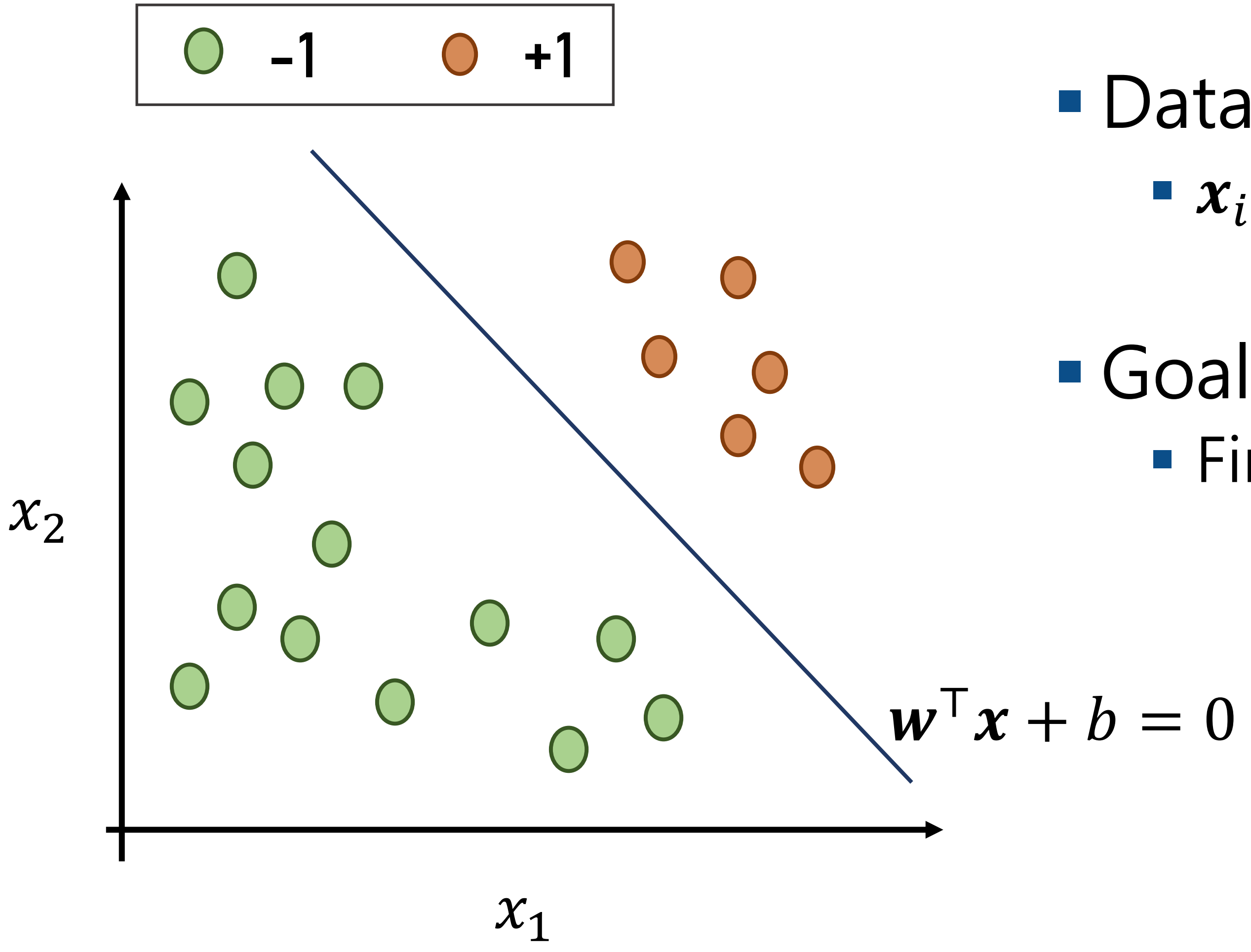
Linear SVM

Non-linear SVM

Tennis example



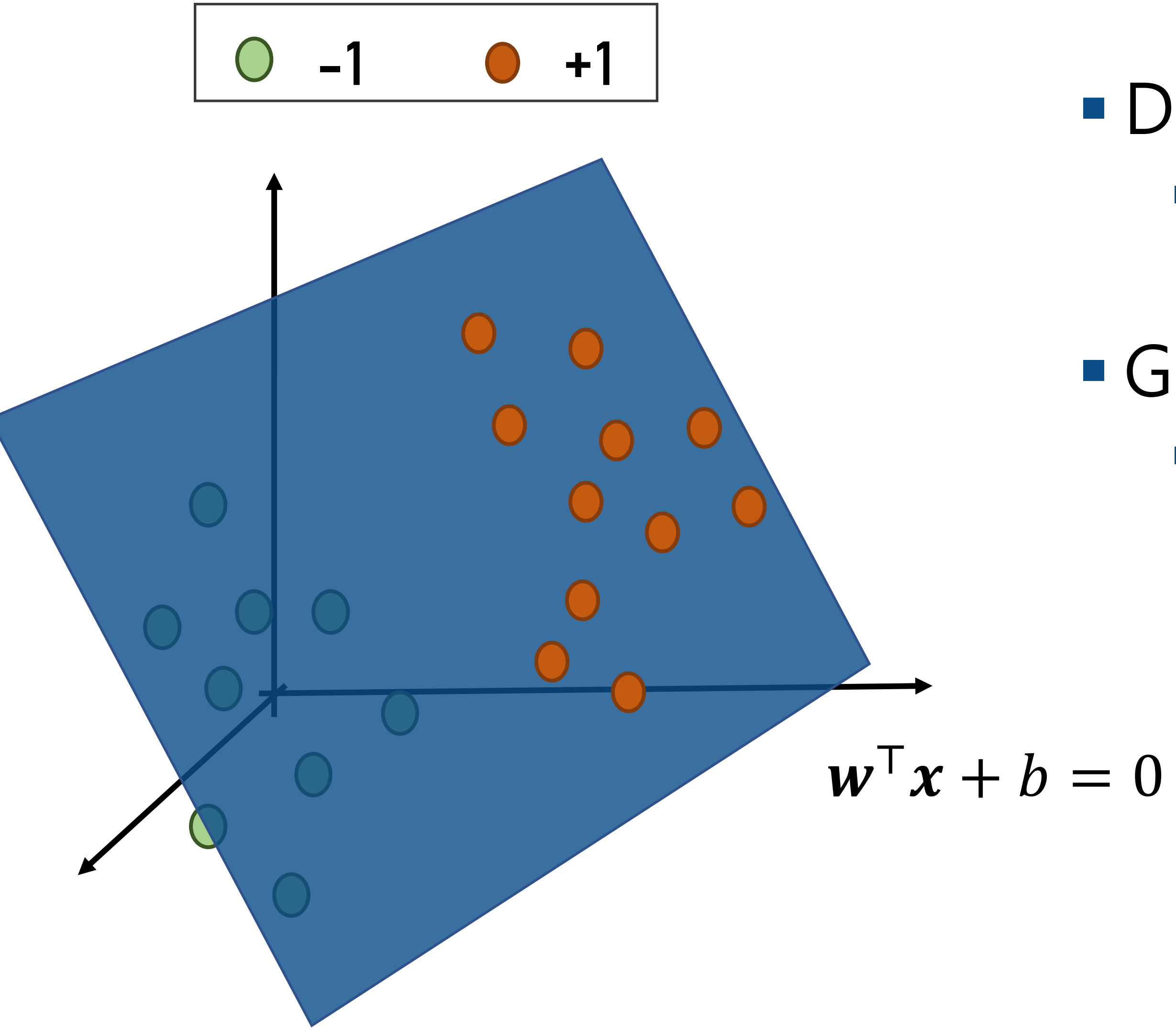
Linear SVM



- Data: $< \boldsymbol{x}_i, y_i >$ for $i = 1, \dots n$
  - $\boldsymbol{x}_i \in \mathbb{R}^d$, $y_i \in \{-1, +1\}$
- Goal
  - Finding a good separating hyperplane

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

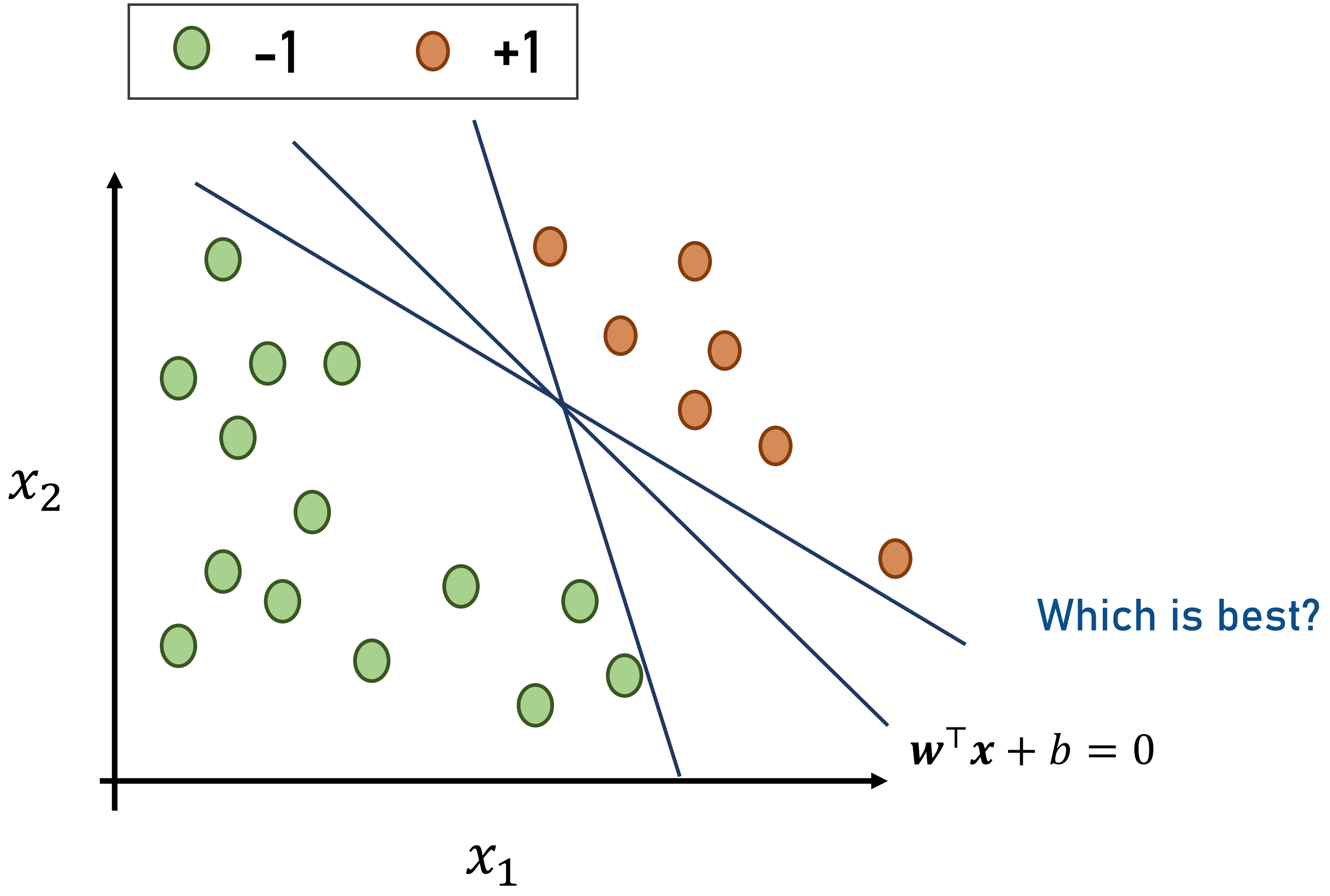
Linear SVM



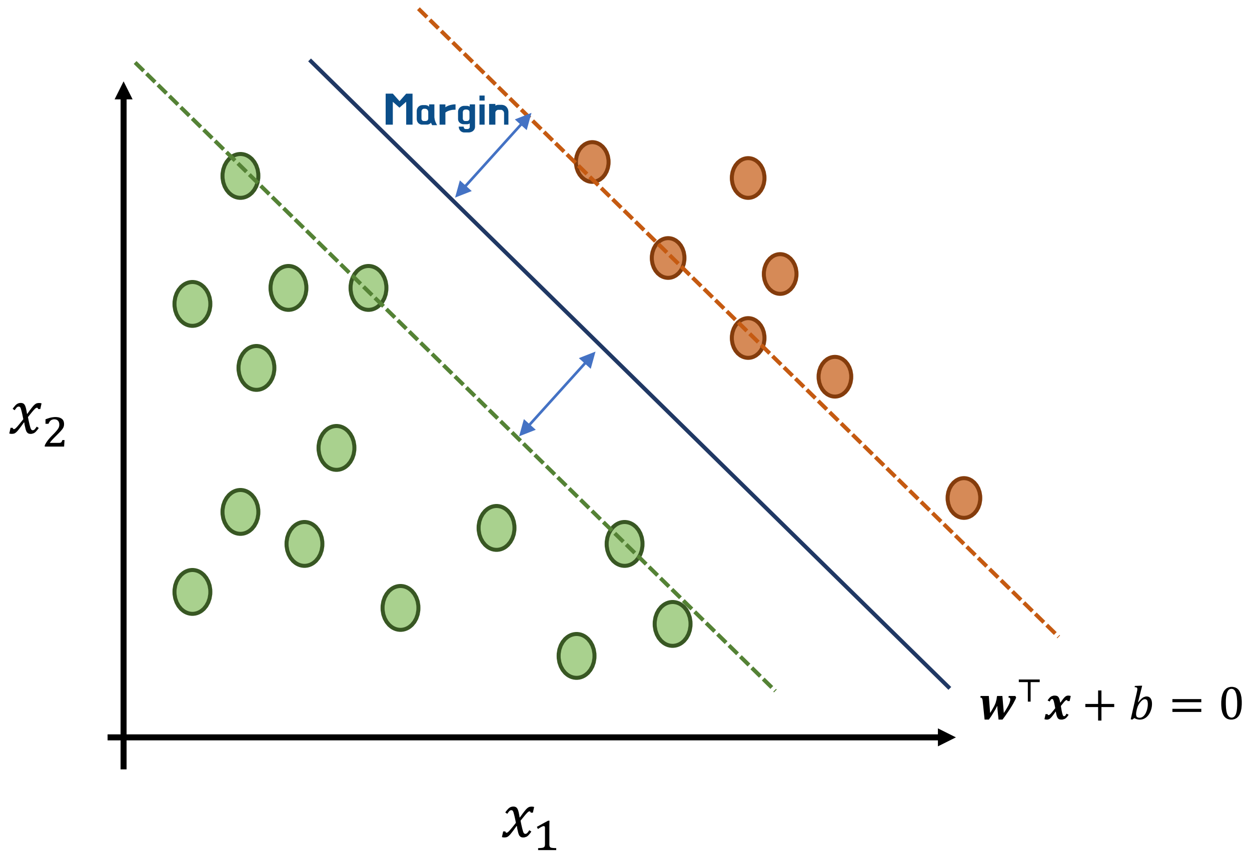
- Data: $< \boldsymbol{x}_i, y_i >$ for $i = 1, \ldots n$
  - $\boldsymbol{x}_i \in \mathbb{R}^d$, $y_i \in \{-1, +1\}$
- Goal
  - Finding a good separating hyperplane

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

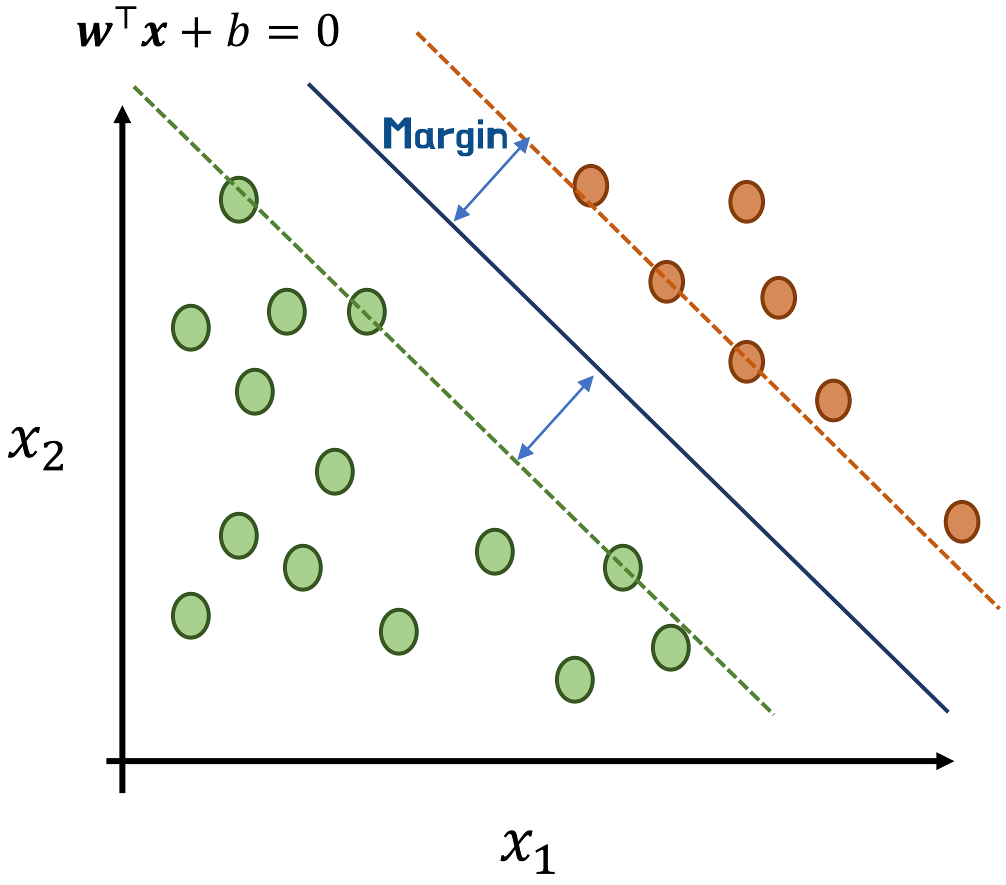
Separating Hyperplane



Separating Hyperplane

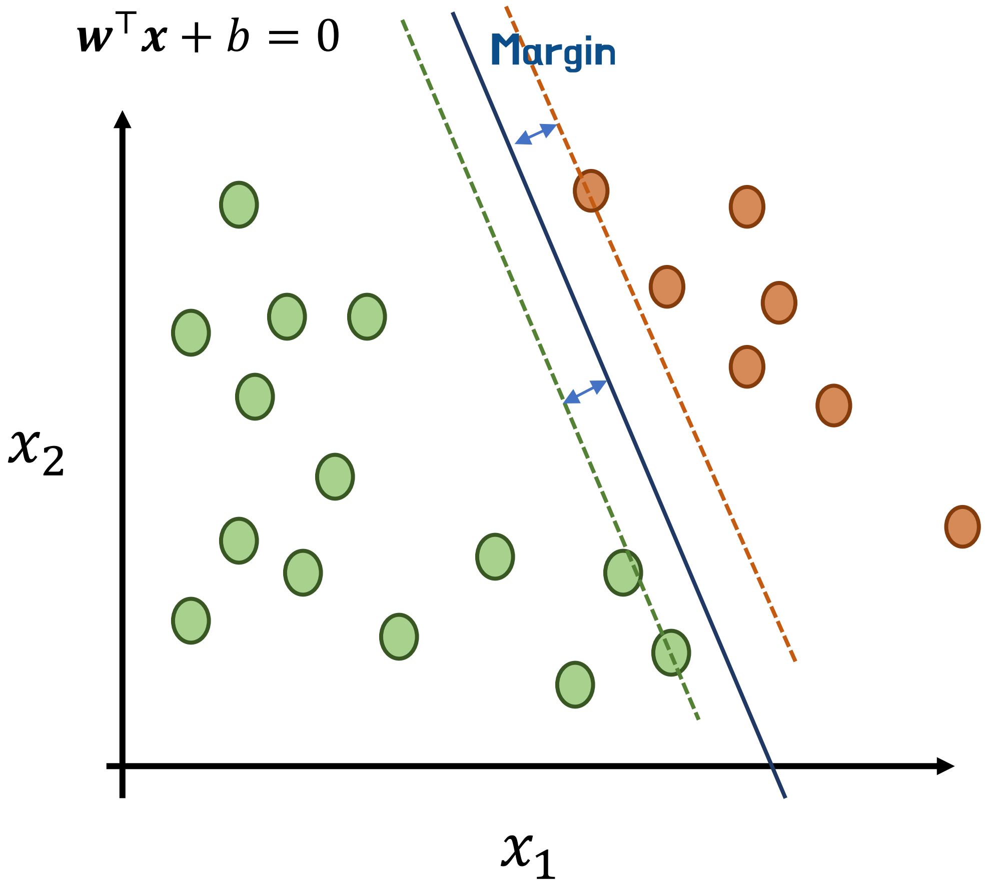


Margin



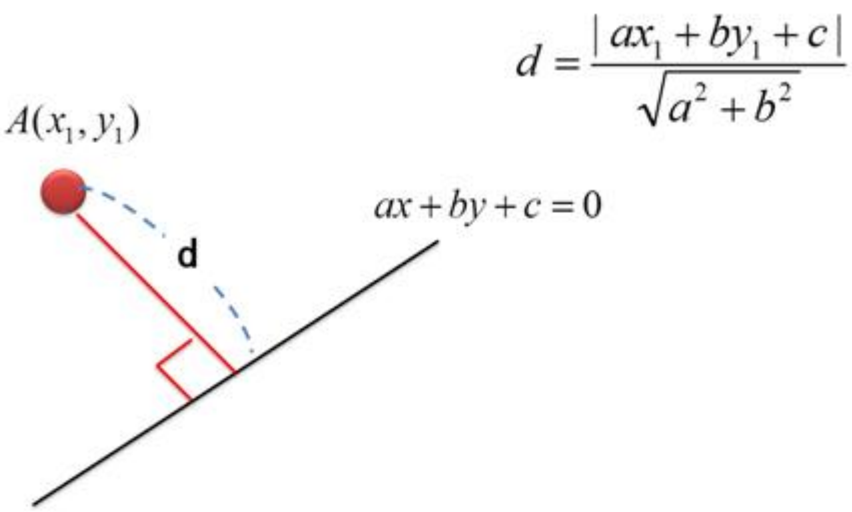
Maximizing margin over the training set
= Minimizing generalization error

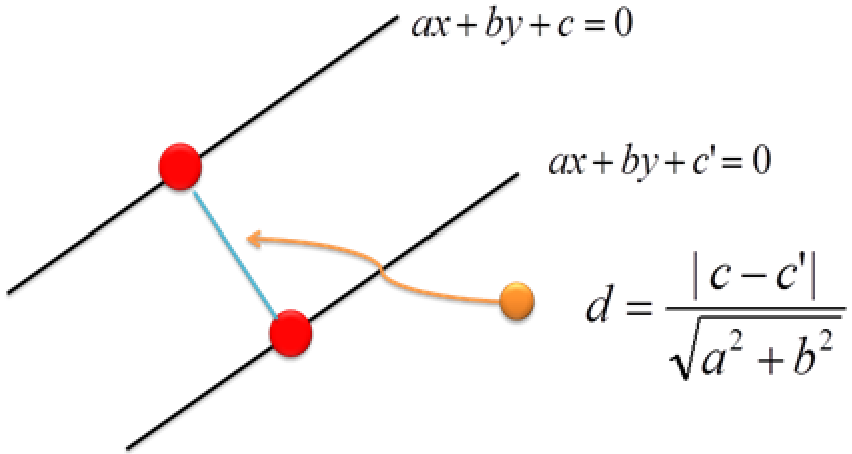
Margin



Maximizing margin over the training set
= Minimizing generalization error

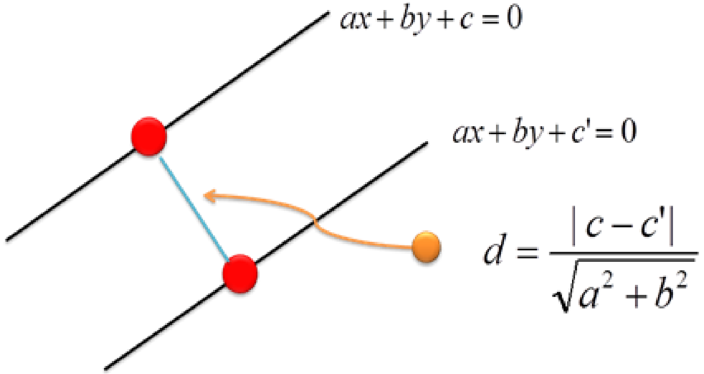
Distance from a line

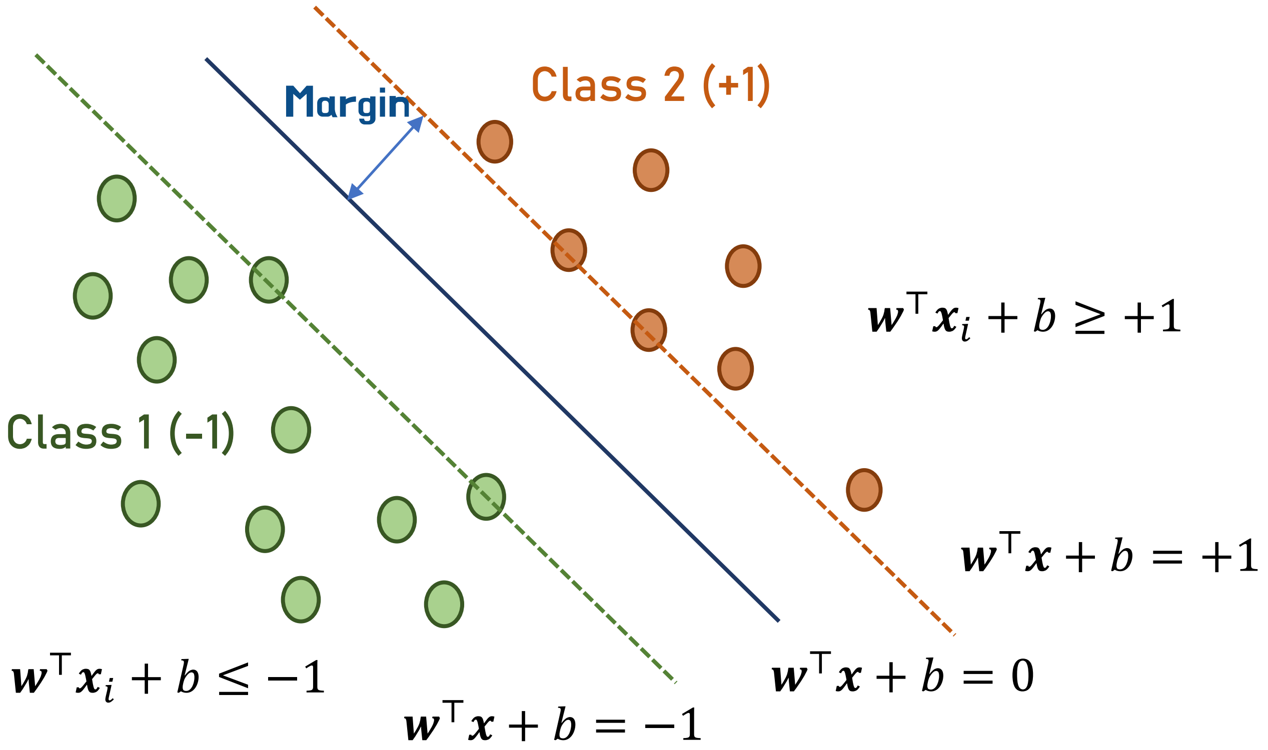




https://j1w2k3.tistory.com/491

Margin





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

In order to maximize the margin, we need to minimize ||w||.
With the condition that there are no datapoints between H1 and H2:

$\mathbf{x}_i \bullet \mathbf{w} + b \geq +1$ when $y_i = +1$

$\mathbf{x}_i \bullet \mathbf{w} + b \leq -1$ when $y_i = -1$

Can be combined into $\boldsymbol{y_i}(x_i \bullet w) \geq 1$

Constrained Optimization Problem

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

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

- Learnable parameter는 $\boldsymbol{w}, b$
- Loss function $\frac{1}{2}\|\boldsymbol{w}\|_2^2$
  - Margin $\frac{1}{\|\boldsymbol{w}\|_2}$ 을 최대화
- Constraint $y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b) \geq 1$
  - Training data를 완벽하게 separating
  - 두 boundary $(\boldsymbol{w}^\top \boldsymbol{x}_i + b = \pm 1)$ 사이에 데이터가 없음

Lagrangian Formulation

Original Problem

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

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

Lagrangian multiplier를 이용하여 Lagrangian primal문제로 변환

Lagrangian Primal

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

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

Seoung Bum Kim
https://youtu.be/qFg8cDnqYCI

Lagrangian Formulation

Lagrangian Primal

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

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

Convex, continuous이기 때문에 미분 = 0에서 최소값을 가짐

w −

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

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

Seoung Bum Kim
https://youtu.be/qFg8cDnqYCI

Lagrangian Formulation

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

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

RECALL

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

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

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

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

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

Seoung Bum Kim
https://youtu.be/qFg8cDnqYCI

Lagrangian Formulation

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

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

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

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

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

RECALL

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

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

Seoung Bum Kim
https://youtu.be/qFg8cDnqYCI

Lagrangian Dual

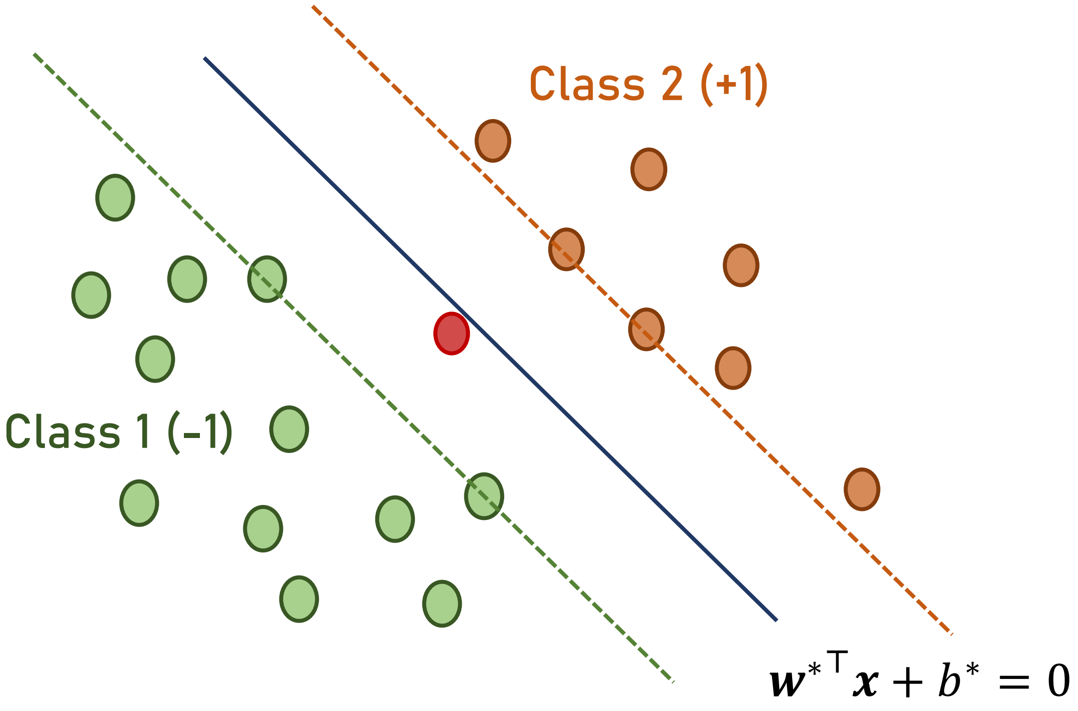
$$\min_{w,b} \frac{1}{2}\|w\|_2^2 - \sum_{i=1}^{n} \alpha_i (y_i (w^T x_i + b) - 1)$$

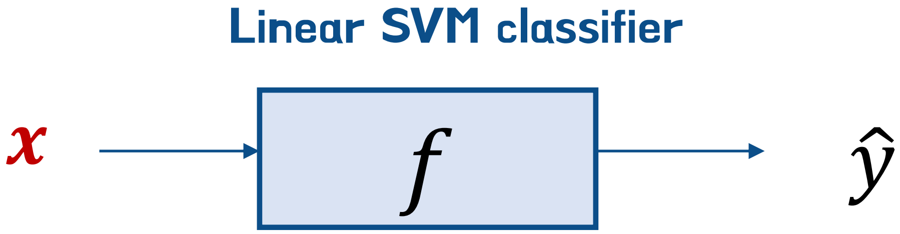
$$= \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^T x_j$$

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

- Quadratic programming formulation
- Convex optimization을 통해 풀수 있음

Classifying a New Data Point





$$\boldsymbol{f}(\boldsymbol{x}, \boldsymbol{w}^*, b^*) = sign(\boldsymbol{w^*}^\top \boldsymbol{x} + b^*)$$

Summary

1. Original formulation

$$\min_{\boldsymbol{w},b} \ \frac{1}{2}\|\boldsymbol{w}\|_2^2$$

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

2. Lagrangian primal

$$\max_{\alpha} \min_{\boldsymbol{w},b} \ \frac{1}{2}\|\boldsymbol{w}\|_2^2 - \sum_{i=1}^{n} \alpha_i(y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b) - 1)$$

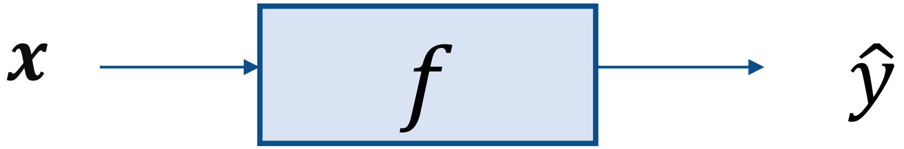
$$subject\ to\ \ \alpha_i \geq 1, i = 1,2,\dots,n$$

3. Lagrangian dual

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

$$subject\ to\ \alpha_i \geq 0, i = 1,2,\dots,n \qquad \sum_{i=1}^{n} \alpha_i y_i = 0$$

Linear SVM classifier



$$\boldsymbol{f}(\boldsymbol{x}, \boldsymbol{w}^*, b^*) \ = \ sign({\boldsymbol{w}^*}^\top \boldsymbol{x} + b^*)$$

where $\boldsymbol{w}^* = \sum_{i=1}^{n} \alpha_i^* y_i \boldsymbol{x}_i$

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