한국자료분석학회
The Korean Data Analysis Society

분류문제의 이해와 활용

2024 한국자료분석학회 동계학술대회 튜토리얼

고려대학교 통계학과 신승준 (sjshin@korea.ac.kr)

고려대학교

분류 모형

고려대학교

Kernel Logistic Regression

Support Vector Machine

Logistic Regression

Perceptron

Linear Discriminant Analysis

Neural Network

Tree

Naive Bayes Classifier

Gradient Boosting

Random Forest

k-Nearest Neighbor Classifier

한국자료분석학회
The Korean Data Analysis Society

분류 이해하기

고려대학교

분류문제를 해결하는 두가지 방법

고려대학교

데이터는 Random Variable

- 분석 – 데이터의 분포를 알아내는 것
- 이항 분류 (Binary Classification)
  - 데이터: $(y_i, \mathbf{x}_i) \in \{0,1\} \times \mathbb{R}^p, i = 1, \cdots, n$
  - 이항분류의 목표

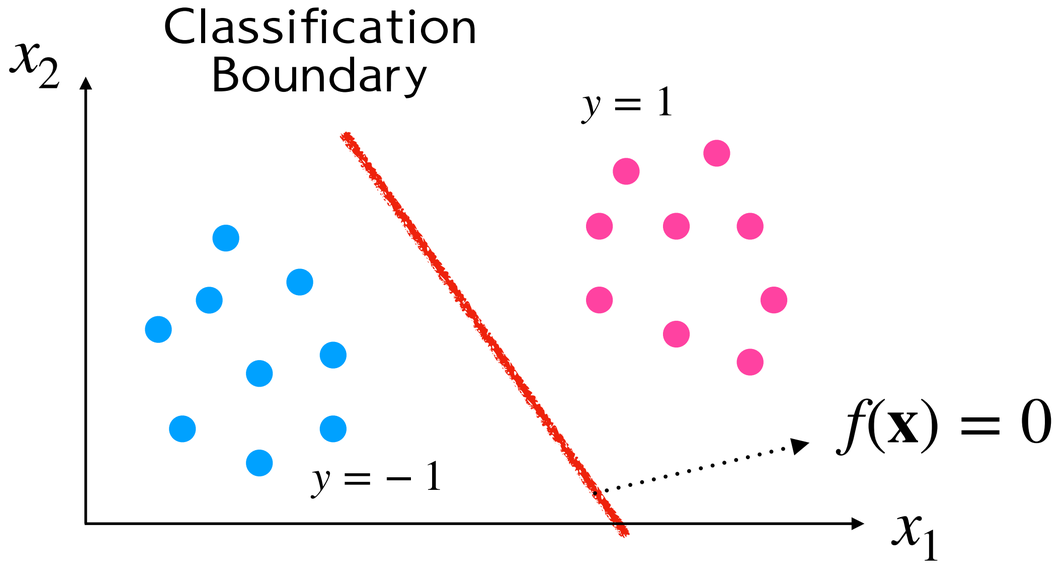
    → $p(\mathbf{x}) = P(y = 1 \mid \mathbf{x})$

    Class Probability
  - 분류 규칙: $p(\mathbf{x}) \geq 0.5 \rightarrow y = 1$

    $p(\mathbf{x}) < 0.5 \rightarrow y = 0$

데이터는 Numbers

- 데이터: $(y_i, \mathbf{x}_i) \in \{-1,1\} \times \mathbb{R}^p, i = 1, \cdots, n$



- 분류 규칙: $y = \text{sign}\{f(\mathbf{x})\}$

분포의 추정: Simple & Naive Approach

고려대학교

- Naive Bayes Classifier

$$p(\mathbf{x}) = P(Y = 1 \mid \mathbf{x})$$

$$= \frac{P(\mathbf{x} \mid Y = 1) \cdot P(Y = 1)}{P(\mathbf{x} \mid Y = 0) \cdot P(Y = 0) + P(\mathbf{x} \mid Y = 1) \cdot P(Y = 1)}$$

추정이 어려움 쉽게 추정 가능

- 독립을 가정하면!

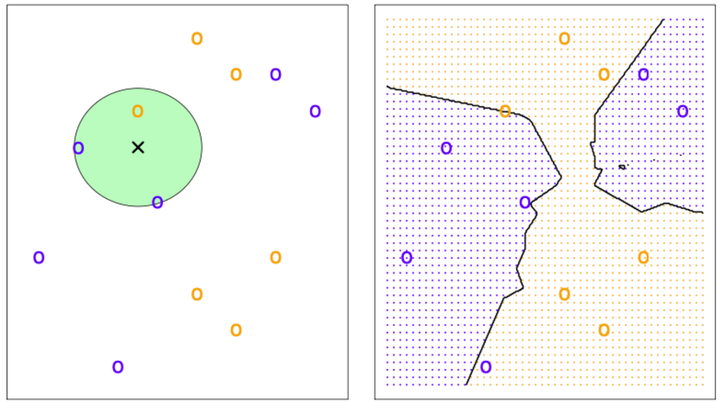
$$P(\mathbf{x} \mid Y = y) = \prod_{j=1}^{p} P(x_j \mid Y = y)$$

쉽게 추정 가능

- k-nearest Neighbor Classifier

$N_k(\mathbf{x}) = \{\mathbf{x}$와 가장 가까운 k개의 데이터 인덱스$\}$

$$p(\mathbf{x}) = P(Y = 1 \mid \mathbf{x}) \approx \sum_{i=N_k(\mathbf{x})} y_i/k$$



분포의 추정: LDA

고려대학교

- 선형판별분석 (Linear Discriminant Analysis)

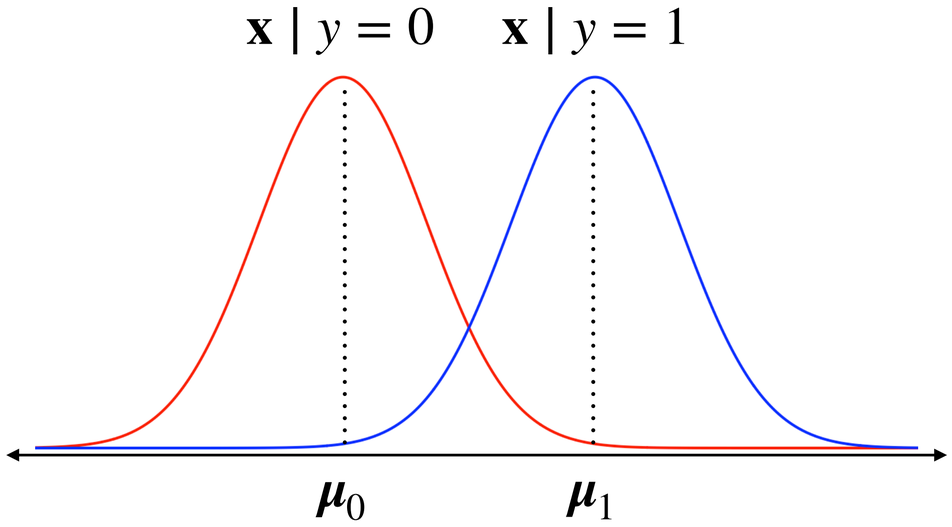
Model: $y \sim Bern(\pi), \ \mathbf{x} \mid y \sim N(\boldsymbol{\mu}_y, \boldsymbol{\Sigma})$

Classification Rule:

$$p(\mathbf{x}) > 0.5 \quad \longleftrightarrow \quad \frac{p(\mathbf{x})}{1-p(\mathbf{x})} > 1$$

$$\frac{p(\mathbf{x})}{1-p(\mathbf{x})} = \frac{\pi \cdot \phi(\mathbf{x}; \boldsymbol{\mu}_1, \boldsymbol{\Sigma})}{(1-\pi) \cdot \phi(\mathbf{x}; \boldsymbol{\mu}_0, \boldsymbol{\Sigma})}$$

손 쉽게 추정 가능



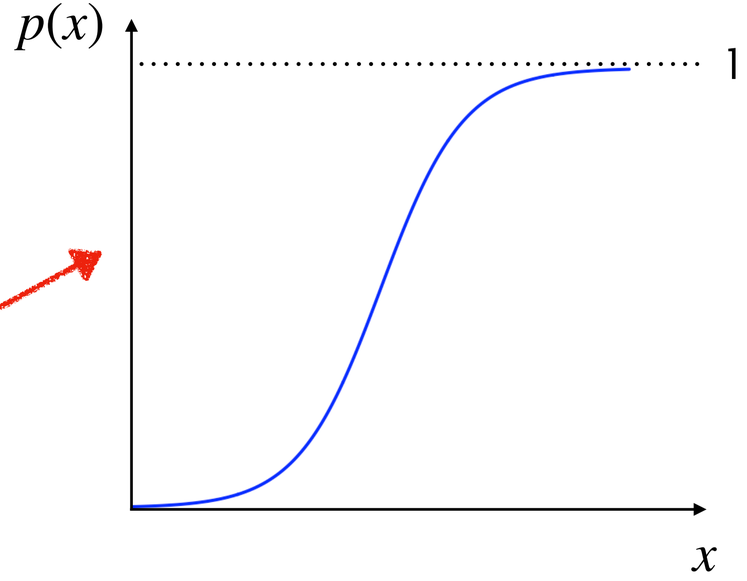
분포의 추정: Logistic Regression

고려대학교

- 로지스틱 (Logistic) 회귀 모형

$$y \mid \mathbf{x} \sim \text{Bern}(p(\mathbf{x}))$$

$$\log \frac{p(\mathbf{x})}{1-p(\mathbf{x})} = \beta_0 + \boldsymbol{\beta}^T \mathbf{x} \quad \longleftrightarrow \quad p(\mathbf{x}) = \frac{\exp(\beta_0 + \boldsymbol{\beta}^T \mathbf{x})}{1 + \exp(\beta_0 + \boldsymbol{\beta}^T \mathbf{x})}$$



- Maximum Likelihood Estimation

$$\arg\min_{\beta_0, \boldsymbol{\beta}} \ell(\beta_0, \boldsymbol{\beta}) = \sum_{i=1}^{n} \left[y_i \log p(\mathbf{x}_i) + (1-y_i) \log\{1 - p(\mathbf{x}_i)\} \right]$$

분포의 추정: Connection

고려대학교

- LDA

$$\frac{p(\mathbf{x})}{1-p(\mathbf{x})} = \frac{\pi \cdot \phi(\mathbf{x};\boldsymbol{\mu}_1,\boldsymbol{\Sigma})}{(1-\pi)\cdot\phi(\mathbf{x};\boldsymbol{\mu}_0,\boldsymbol{\Sigma})} = \frac{\pi}{1-\pi}\cdot\frac{\exp\left\{(\mathbf{x}-\boldsymbol{\mu}_1)^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_1)\right\}}{\exp\left\{(\mathbf{x}-\boldsymbol{\mu}_0)^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_0)\right\}}$$

$$\log\frac{p(\mathbf{x})}{1-p(\mathbf{x})} = \log\frac{\pi}{1-\pi} + (\boldsymbol{\mu}_0^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1)/2 + (\boldsymbol{\mu}_1-\boldsymbol{\mu}_0)\boldsymbol{\Sigma}^{-1}\cdot\mathbf{x}$$

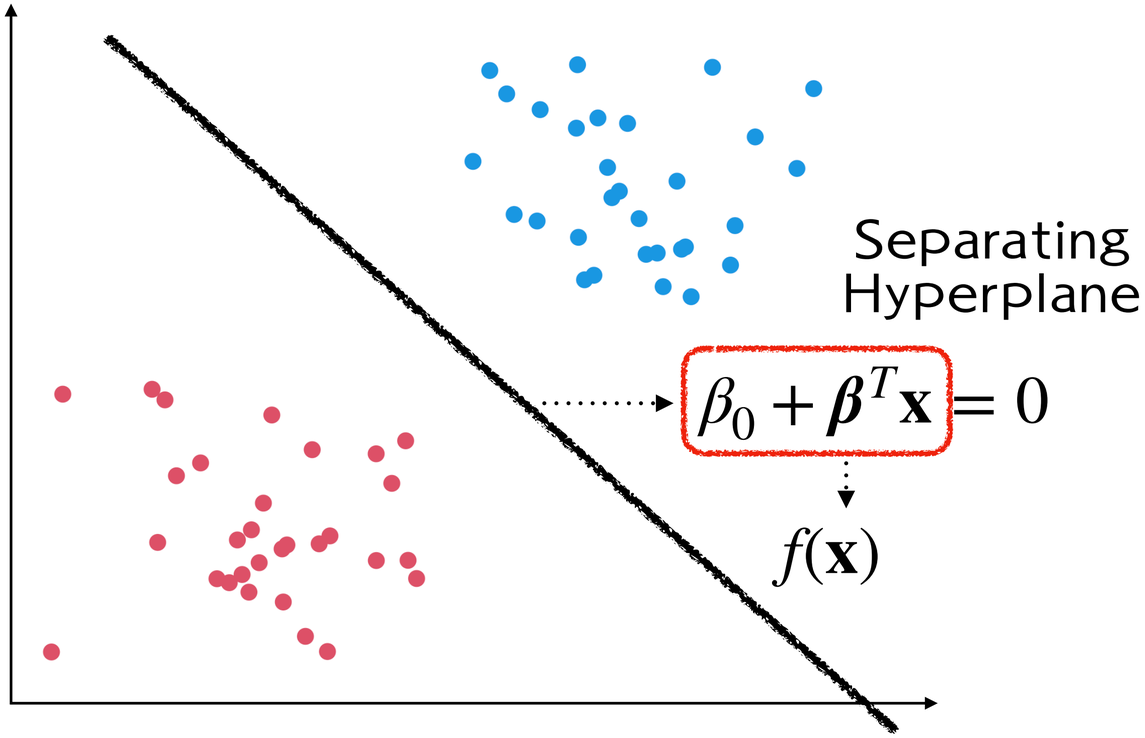
$$\beta_0 + \boldsymbol{\beta}^T\mathbf{x}$$

⇒ Logistic Regression

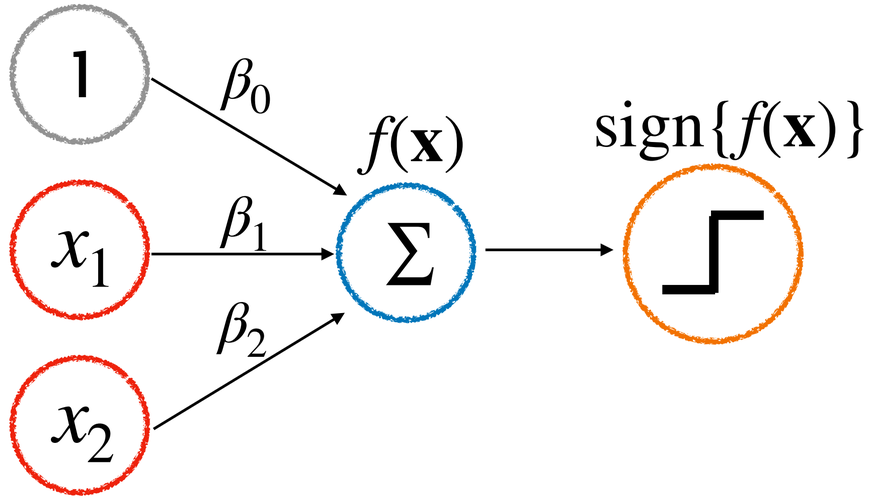
집단 나누기: Perceptron

고려대학교

- 데이터: $(y_i, \mathbf{x}_i) \in \{-1,1\} \times \mathbb{R}^p, i = 1, \cdots, n$



Perceptron
(1957, Frank Rosenblatt)



두 집단을 나누는 직선을 찾는
최초의 알고리즘

집단 나누기: Maximal Margin Classifier

고려대학교

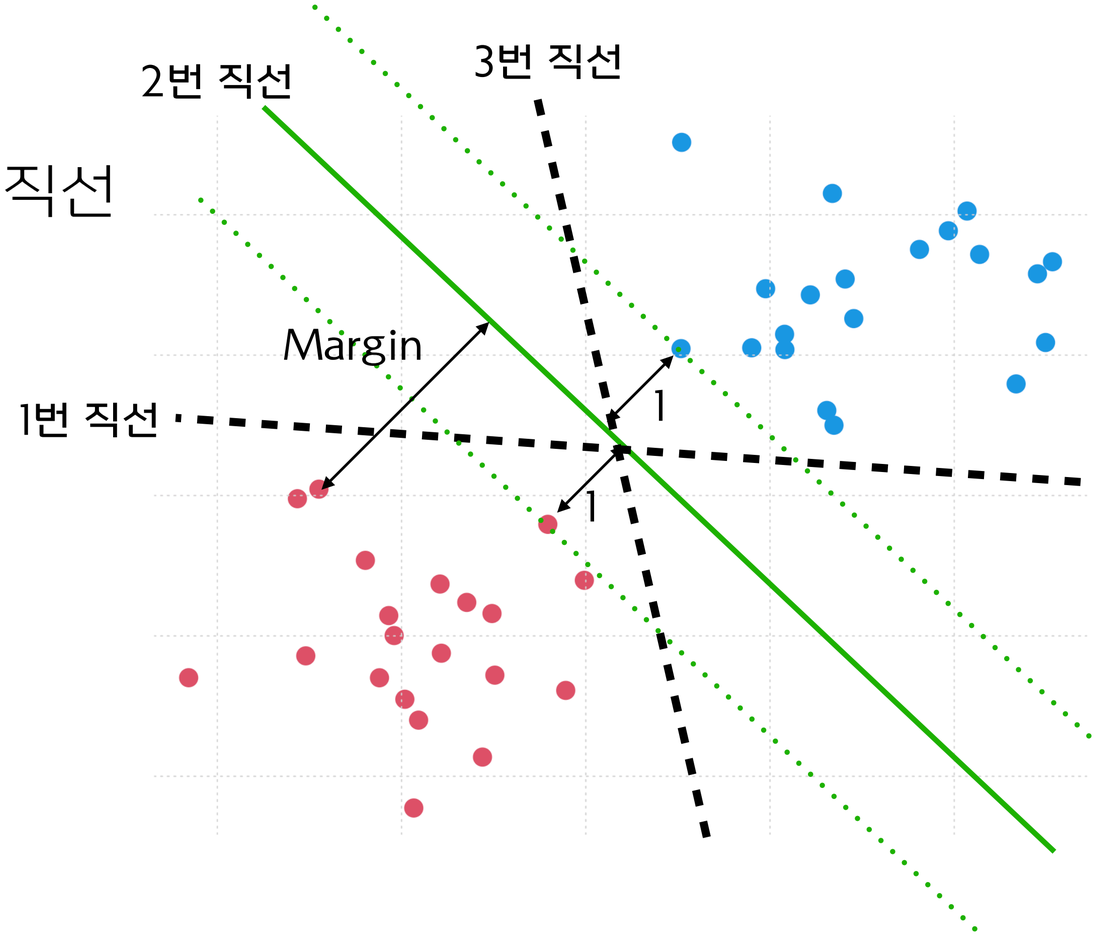
- 가장 좋은 분류 직선은?
  - 분류 경계와 집단간의 거리를 최대화 하는 직선을 찾자!

$$\min_{\beta_0,\boldsymbol{\beta}} \|\boldsymbol{\beta}\|^2$$

$$\text{s.t } y_i(\beta_0 + \boldsymbol{\beta}^T \mathbf{x}_i) \geq 1, \quad i = 1,2,\cdots,n$$

$y_i(\beta_0 + \boldsymbol{\beta}^T \mathbf{x}_i)$: $\mathbf{x}_i$와 분류경계 간의 거리 Margin

≥ 1: 분류경계와 가장 가까운 개체간의 거리는 항상 1로 표준화



집단 나누기: Make it Soft!

고려대학교

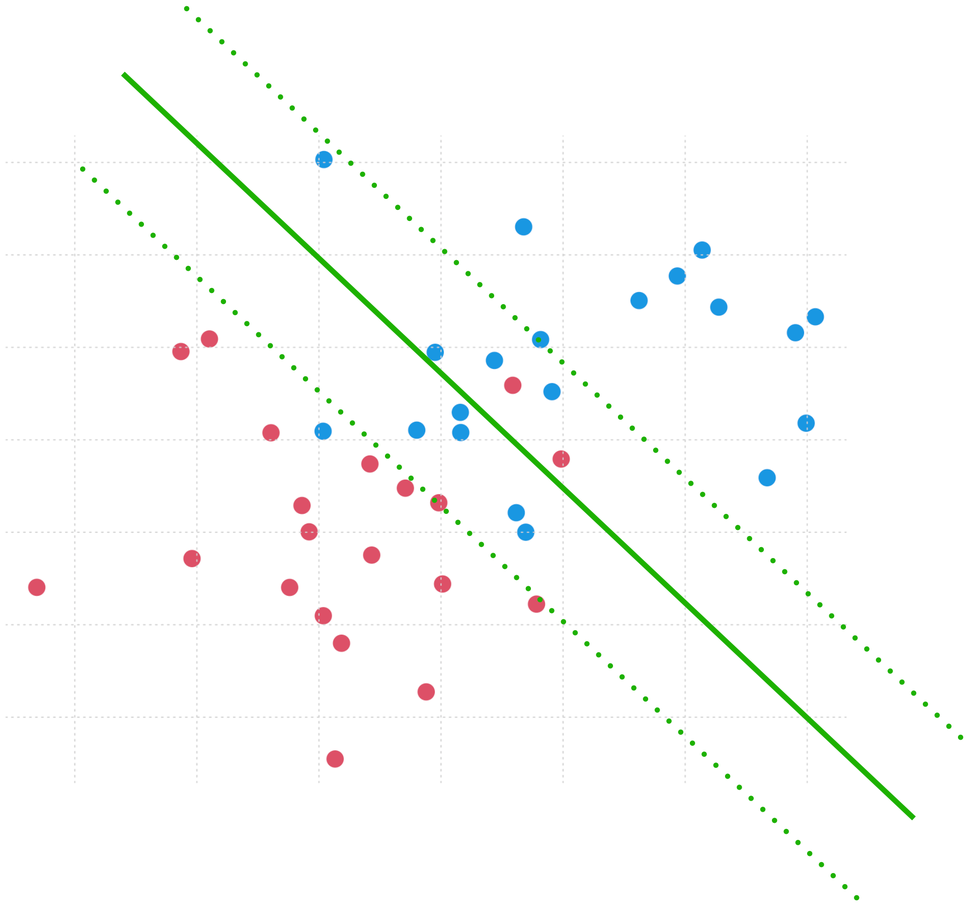
- 두 집단이 선형적으로 분리되지 않는다면?
  - 제약 조건을 완화하자!

$$\min_{\beta_0,\boldsymbol{\beta}} \|\boldsymbol{\beta}\|^2 + C\sum_{i=1}^{n} \xi_i$$

$$\text{s.t } y_i(\beta_0 + \boldsymbol{\beta}^T \mathbf{x}_i) \geq 1 - \xi_i, \quad i = 1,2,\cdots,n$$

$$\xi_i \geq 0, \quad i = 1,2,\cdots,n$$

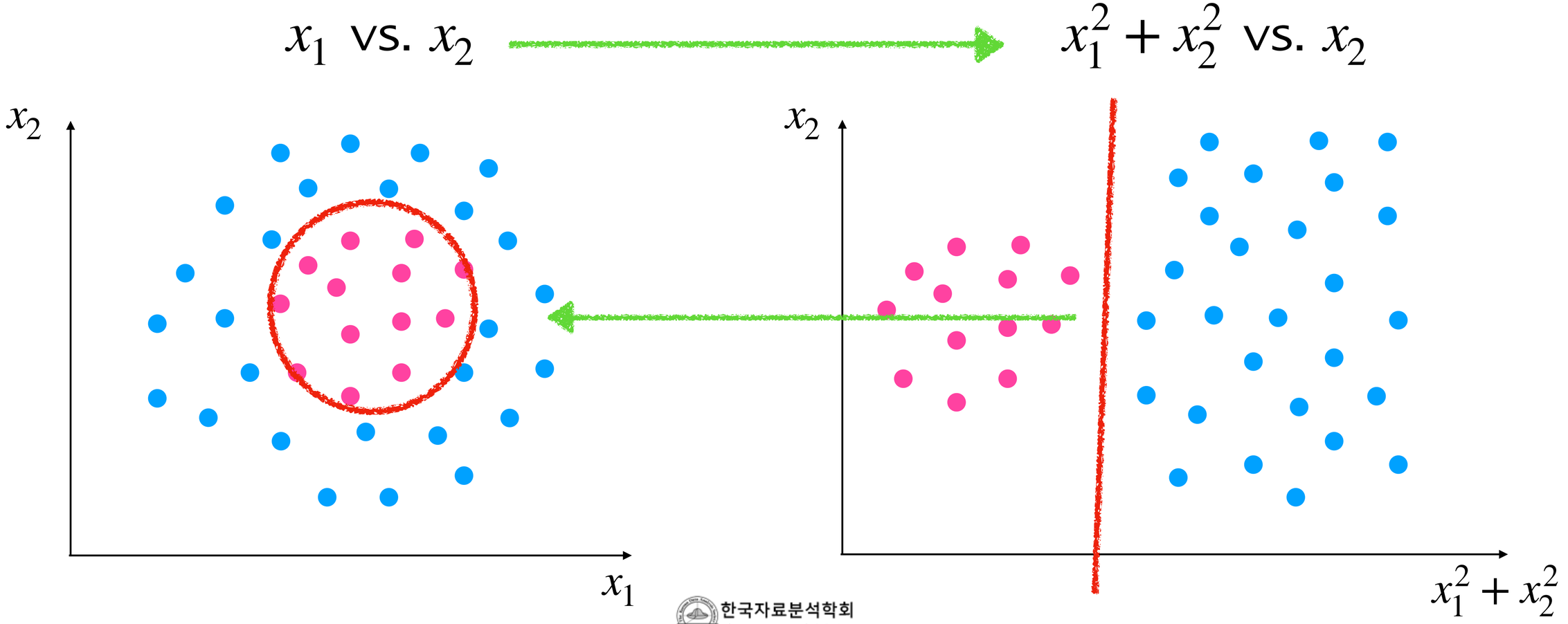
Soft Margin Classifier
Linear SVM



비선형으로 집단 나누기

고려대학교

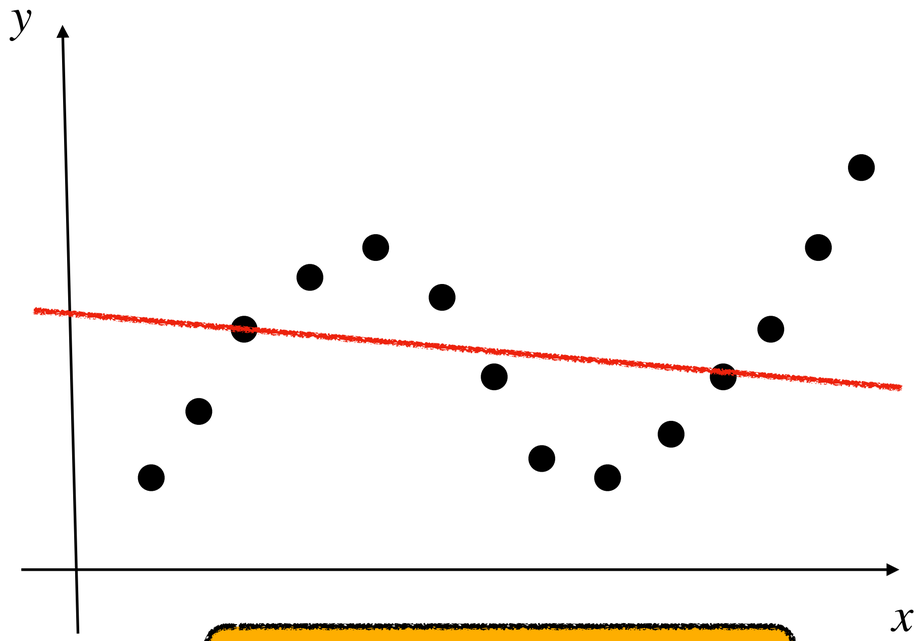
- 비선형으로 나누어진 경우는 어떻게 할까?



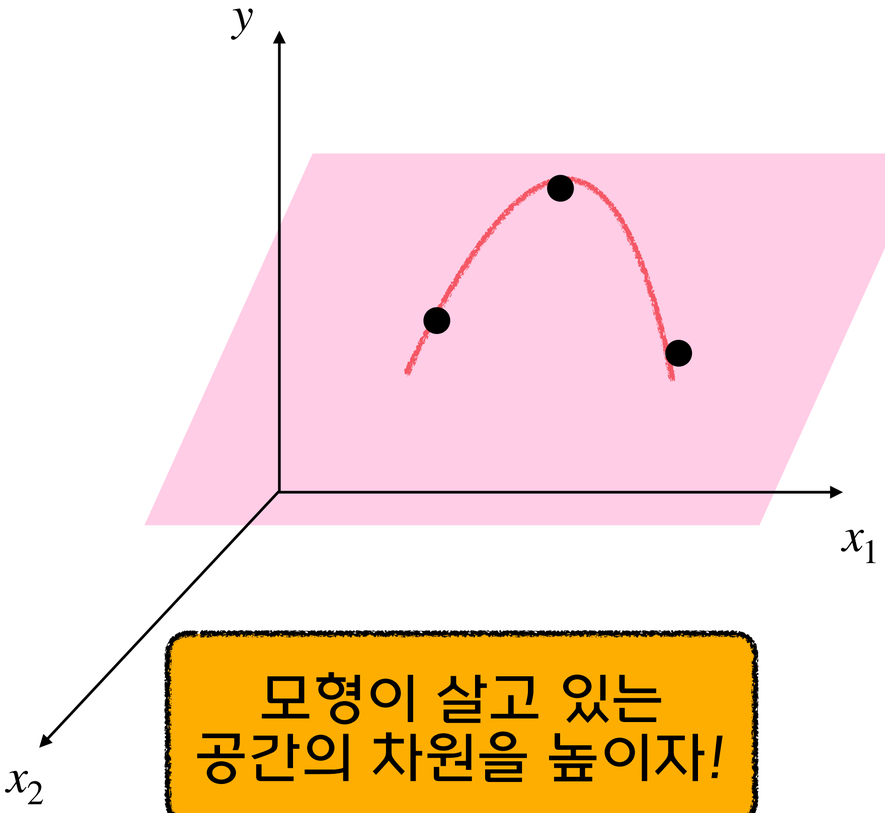
비선형으로 집단 나누기: 변환을 찾아서

고려대학교

- 어떻게 하면 변환된 공간에서 선형이 되도록 할 수 있을까?



변환의 복잡도를
올리자!

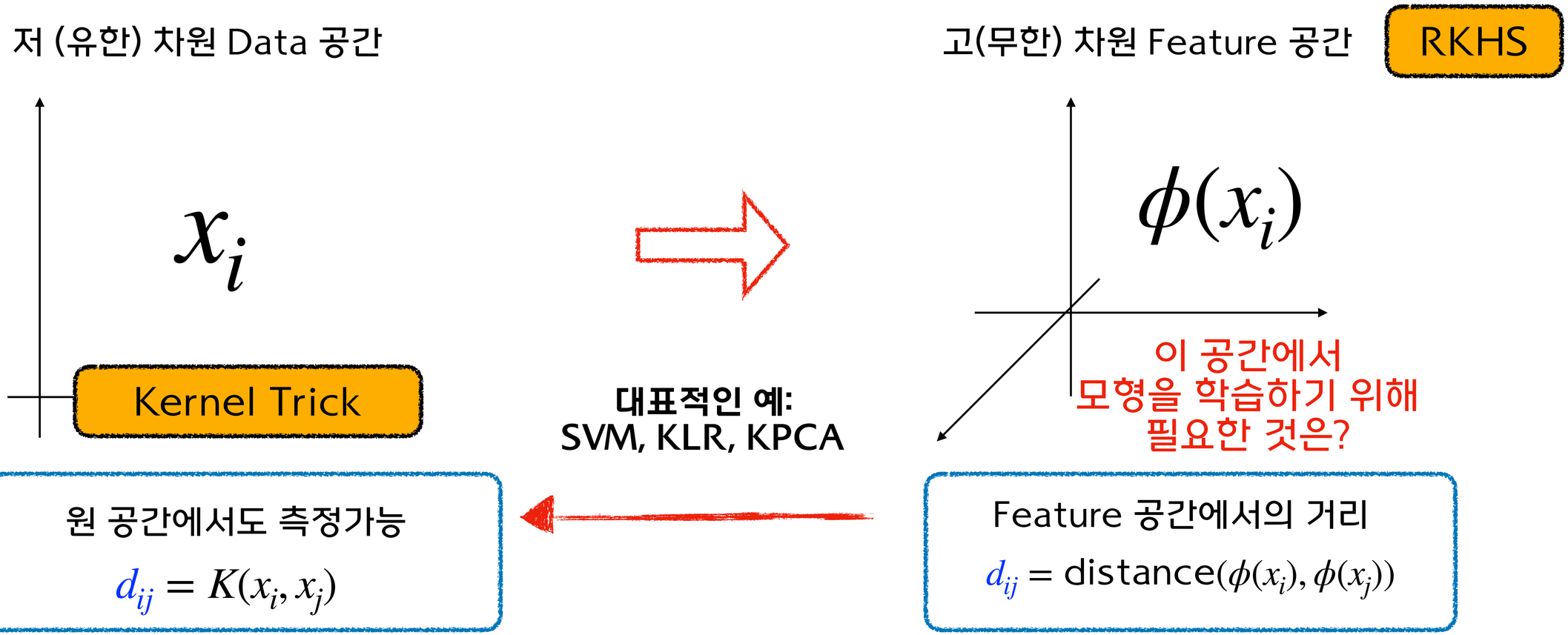


모형이 살고 있는
공간의 차원을 높이자!

비선형으로 집단 나누기 – Kernel Trick

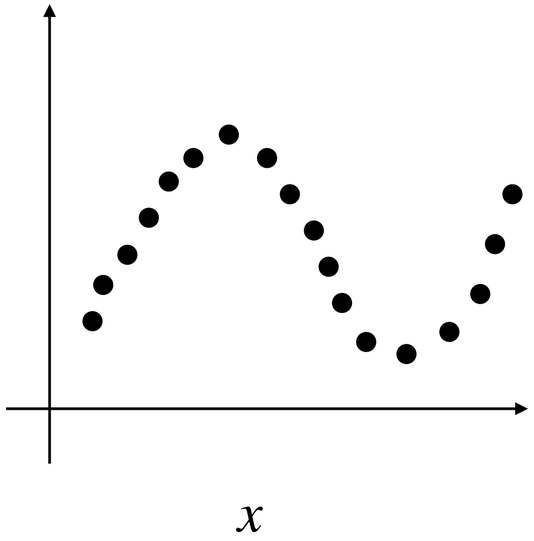
고려대학교

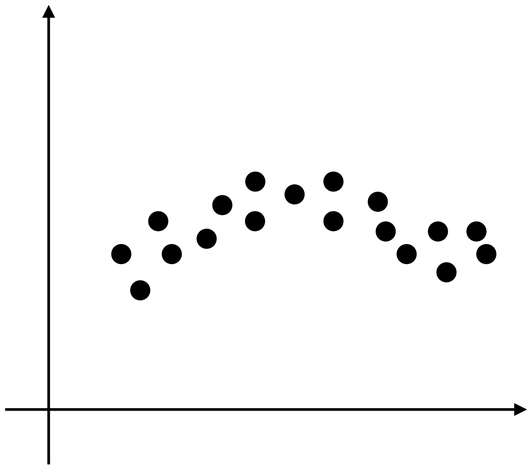
- 데이터를 고차원 공간으로 한번에 Mapping시켜 학습



비선형으로 집단 나누기 – Neural Net

고려대학교

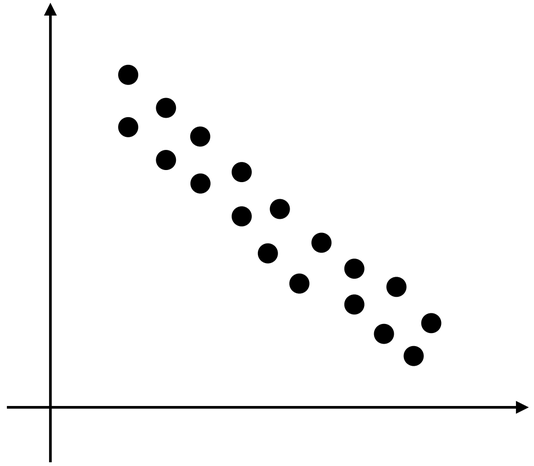




Neural Network

계속하면!

… …



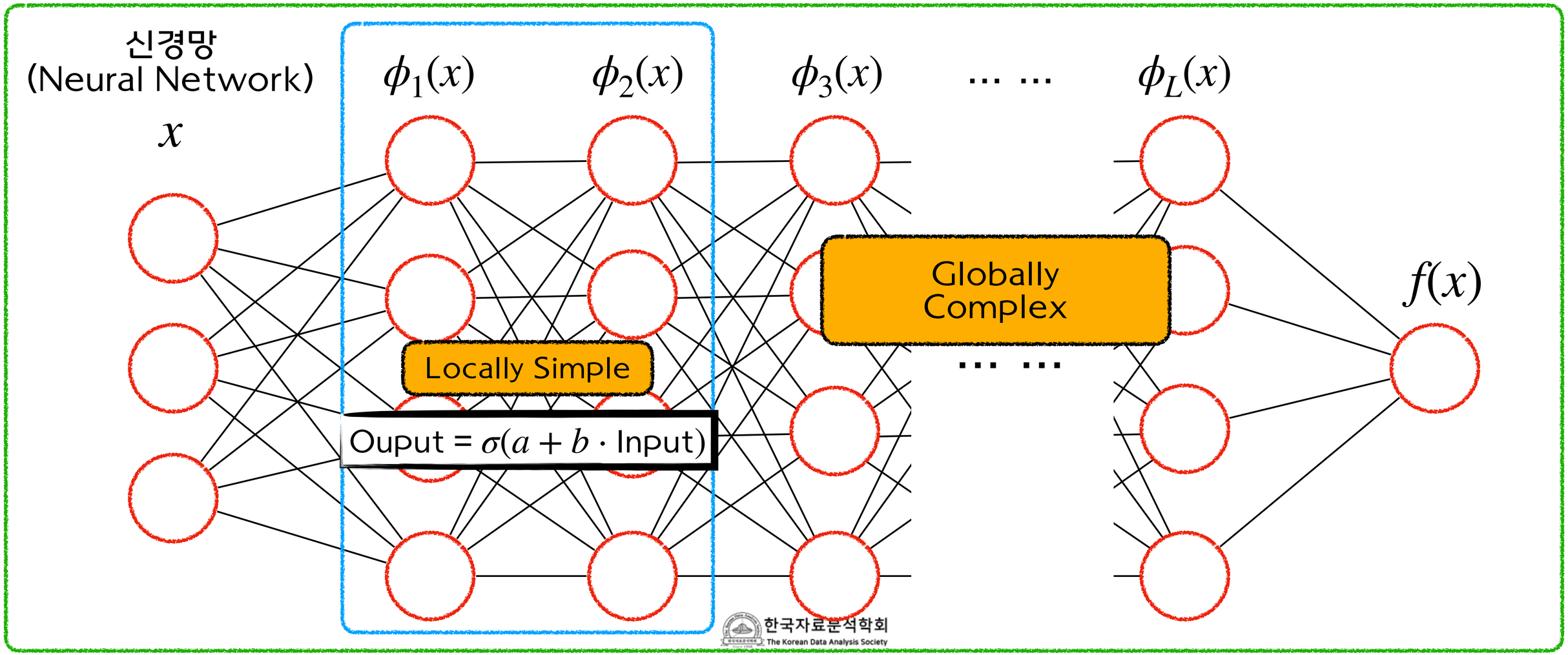
변환! $\phi_1(x) = \sigma(a_1 + b_1 x)$

한번 더! $\phi_2(x) = \sigma(a_2 + b_2 \phi_1(x))$

또 한번 더! $\phi_3(x) = \sigma(a_3 + b_3 \phi_2(x))$ … … $\phi_L(x) = \sigma(a_L + b_L \phi_{L-1}(x))$

비선형으로 집단 나누기 – Neural Net

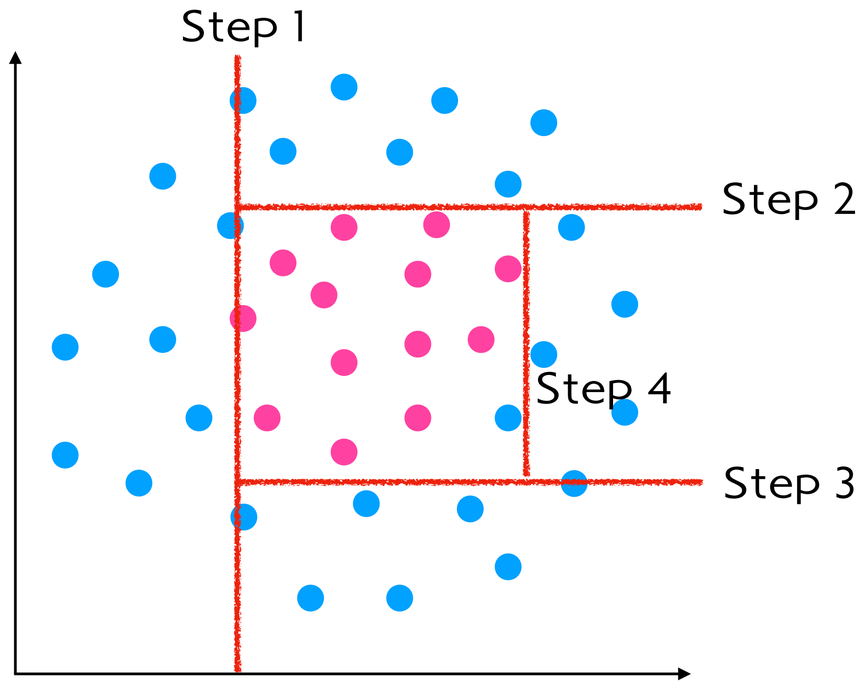
고려대학교



비선형으로 집단 나누기 – Tree

고려대학교

- Tree 모형의 아이디어
  - Binary Recursive Splitting
- **간단하지만 성능이 좋지 않음**
  - 모형의 분산이 크기 때문
- **앙상블: 주어진 데이터에 대해 여러번 Tree 모형을 적합해서 합치자!**
  - Random Forest (Over-fitted Tree)
  - Boosting (Under-fitted Tree)



한국자료분석학회
The Korean Data Analysis Society

분류 모형 좀 더 들어다 보기

고려대학교

Population Level

고려대학교

- 좋은 분류기는 무엇인가?
  - 오분류율 (misclassification rate)을 최소화 하는 분류함수!

$$P(Y \neq \hat{Y} \mid \mathbf{x}) = P\left(Y \neq \text{sign}\{f(\mathbf{x})\} \mid \mathbf{x}\right) = P\left(\underbrace{Y \cdot f(\mathbf{x})}_{\text{Margin}} < 0 \mid \mathbf{x}\right)$$

- Bayes Classifier (Optimal classification function)

$$f^{\text{Bayes}}(\mathbf{x}) = \arg\min_{f(\mathbf{x}) \in \mathbb{R}} P\left(\underbrace{Y \cdot f(\mathbf{x})}_{\substack{\text{Classification} \\ \text{function}}} < 0 \mid \mathbf{x}\right) \quad \longleftrightarrow \quad \text{sign}\{f^{\text{Bayes}}(\mathbf{x})\} = \text{sign}\{\underbrace{p(\mathbf{x})}_{\substack{\text{Class Probability}}} - 0.5\}$$

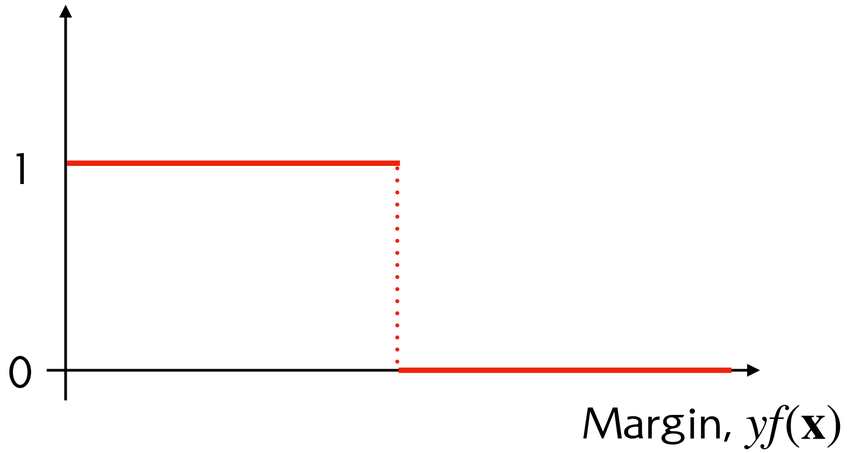
Sample Level

고려대학교

- 데이터: $(y_i, \mathbf{x}_i) \in \{-1,1\} \times \mathbb{R}^p, i = 1, \cdots, n$

$$\arg\min_f \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\{y_i \cdot f(\mathbf{x}_i) < 0\} \longrightarrow E\left[\mathbb{I}\{Y \cdot f(\mathbf{x}) < 0 \mid \mathbf{x}\}\right] = P\left(Y \cdot f(\mathbf{x}) < 0\right)$$

0-1 손실함수



최적화가 쉽지 않음!

Surrogate Loss

고려대학교

- 최적화가 까다로운 0-1 손실함수를 좀더 다루기 쉬운 손실함수로 대체
  - Logistic Loss

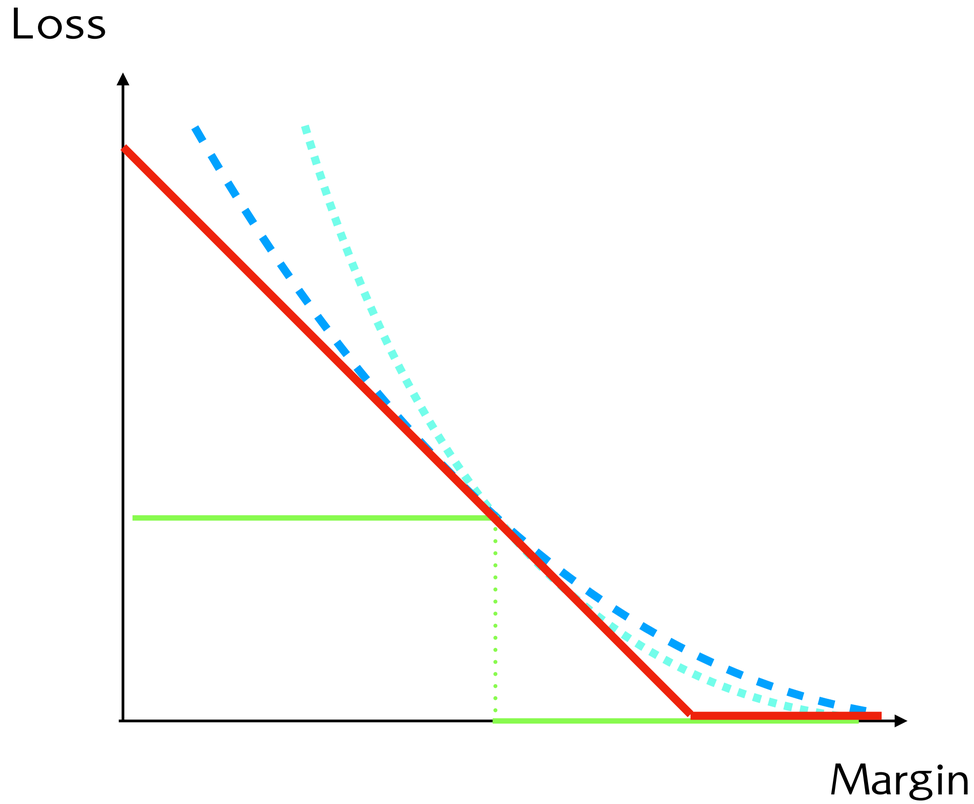
$$L(y,f) = \log\{1 + \exp(-yf)\}$$

  - Hinge Loss (SVM)

$$L(y,f) = [1 - yf]_+$$

  - Exponential Loss (Boosting)

$$L(y,f) = \exp(-yf)$$



Surrogate Loss: Examples

고려대학교

Logistic Regression

$$\min_{\beta_0,\boldsymbol{\beta}} -\sum_{i=1}^{n}\left[y_i \log \frac{e^{\beta_0+\boldsymbol{\beta}^T\mathbf{x}}}{1+e^{\beta_0+\boldsymbol{\beta}^T\mathbf{x}}} + (1-y_i)\log\left\{1-\frac{e^{\beta_0+\boldsymbol{\beta}^T\mathbf{x}}}{1+e^{\beta_0+\boldsymbol{\beta}^T\mathbf{x}}}\right\}\right]$$

- $y_i \in \{-1,1\}$ 코딩으로 바꾸면

$$\min_{\beta_0,\boldsymbol{\beta}} \sum_{i=1}^{n} \log[1+\exp\{y(\beta_0+\boldsymbol{\beta}^T\mathbf{x})\}]$$

Support Vector Machine

$$\min_{\beta_0,\boldsymbol{\beta}} \|\boldsymbol{\beta}\|^2 + C\sum_{i=1}^{n}\xi_i$$

$$y_i(\beta_0+\boldsymbol{\beta}^T\mathbf{x}_i) \geq 1-\xi_i,\ \forall i$$

$$\xi_i \geq 0,\ \forall i$$

- Margin이 1보다 크면 ξ_i는 0
- Margin이 1보다 작으면 ξ_i는 딱 그 작은 만큼

$$\min_{\beta_0,\boldsymbol{\beta}} \frac{1}{C}\|\boldsymbol{\beta}\|^2 + \sum_{i=1}^{n}[1-y_i(\beta_0+\boldsymbol{\beta}^T\mathbf{x}_i)]_+$$

Surrogate Loss: Fisher Consistency

고려대학교

- 0-1손실함수를 대체하기 위한 Surrogate Loss의 최소한의 조건은?
- 대체된 손실함수의 Risk를 최소화하면 0-1 Risk를 최소화 한 것과 동일해야함!

Definition (Fisher Consistency/Classification Calibrated)

A loss function L is Fisher consistent (or classification calibrated) if its population risk minimizer leads the Bayes classification rule.

- Fisher Consistency에 대한 충분조건은?

Theorem (Bartlett et al., 2006)

Let L is convex. If L is differentiable at $m = 0$ and $L'(0) < 0$, then the convex loss L is Fisher consistent.

Empirical Risk Minimization

고려대학교

- ERM Formulation

결국 오분류율을 최소화하는 것이 목표!

$$\min_f \frac{1}{n}\sum_{i=1}^{n} L(y_i, f(\mathbf{x}_i)) + J_\lambda(f) \longrightarrow P(Y \cdot f(\mathbf{x}) < 1 \mid \mathbf{x})$$

- 손실함수 L의 선택에 따른 분류

  로지스틱 회귀, SVM, Boosting 등

- 분류함수 f의 선택에 따른 분류

  선형, Tree, Neural Network, Generalized Additive Model 등

- 패널티함수 J의 선택에 따른 분류

  Lasso, Ridge 등

분류모형은
(손실함수, 분류함수, 패널티함수)
이 세가지 Component를 어떻게
선택하는가의 차이

분균형 분류

고려대학교

- 때로는 두 집단의 중요도가 다르거나, 혹은 오분류로 인한 비용이 비대칭적인 경우 발생
  - 예) 암진단, 대출심사 등
- 집단의 중요도 / 오분류 비용을 반영

$$f^{\text{Bayes}}(\mathbf{x}) = \arg\min_{f(\mathbf{x})\in\mathbb{R}} E\left(\pi(Y)\mathbb{I}(Y\cdot f(\mathbf{x})<0)\mid \mathbf{x}\right) \qquad \text{where } \pi(Y)=\begin{cases}1-\pi & \text{if } Y=1\\ \pi & \text{if } Y=-1\end{cases}$$

for a given $0<\pi<1$

$$\longleftrightarrow \quad \text{sign}\{f^{\text{Bayes}}(\mathbf{x})\} = \text{sign}\{p(\mathbf{x})-\pi\}$$

분균형 분류

고려대학교

- Weighted 로지스틱 회귀

$$\min_{\beta_0,\boldsymbol{\beta}} \sum_{i=1}^{n} \pi(y_i) \cdot \log[1 + \exp\{y(\beta_0 + \boldsymbol{\beta}^T \mathbf{x})\}]$$

- Weighted SVM

$$\min_{\beta_0,\boldsymbol{\beta}} \sum_{i=1}^{n} \pi(y_i) \cdot [1 - y_i(\beta_0 + \boldsymbol{\beta}^T \mathbf{x}_i)]_+ + \lambda \|\boldsymbol{\beta}\|^2$$

⋮

→ 비대칭성을 고려한
오분류 비용을 최소화하는 것이 목표!

분균형 분류

고려대학교

- $p(\mathbf{x})$가 주어지는 경우, 간단히 불균형 분류를 해결할 수 있음

$$\begin{cases} p(\mathbf{x}) \geq \pi \rightarrow y = 1 \\ p(\mathbf{x}) < \pi \rightarrow y = 0 \end{cases}$$

- $p(\mathbf{x})$를 추정하는 방법의 경우, 가중치를 이용하기 보다는 상기의 방법을 활용
- 실제 분석에서 π를 선택하는 문제는 생각보다 쉽지 않음
  - 정확한 오분류 비용을 모르기 때문

불균형 분류: 성능 평가

고려대학교

- 불균형 분류에서 π의 값에 따라 분류결과가 달라짐

y
1
0
1
1
0
0
0
0
1
1

Learning →

$p(\mathbf{x})$
0.75
0.12
0.93
0.53
0.15
0.31
0.12
0.30
0.41
0.75

Prediction →

$\pi = .5$
1
0
1
1
0
0
0
0
0
1

Pred / True	−	+
−	5	0
+	1	4

Prediction →

Pred / True	−	+
−	3	2
+	0	5

$\pi = .25$
1
0
1
1
0
1
0
1
1
1

Error Rate

90 % vs. 80 %

TPR / Sensitivity

80 % vs. 100 %

TNR / Specificity

100 % vs. 60 %

불균형 분류: ROC Curve

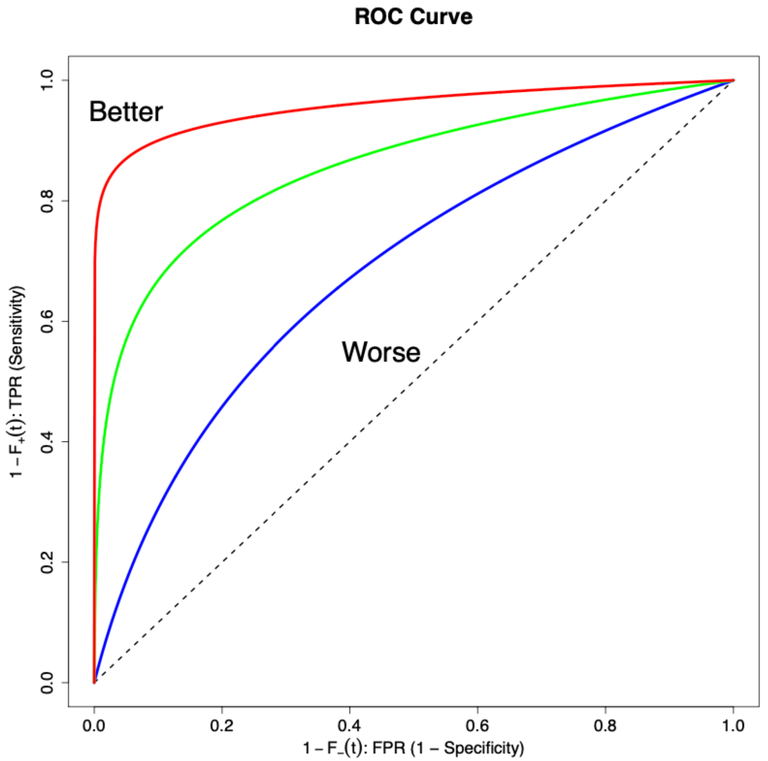
고려대학교

- Receiver Operation Characteristic (ROC) 곡선
  - 가능한 모든 분류 기준값에 대하여 예측의 민감도와 특이도를 그림으로 표현
- Area Under the ROC Curve (AUC)
  - 1에 가까울 수록 좋은 성능을 의미

$$AUC(f) = P(f(\mathbf{x}_+) \geq f(\mathbf{x}\_))$$

$$= E(\mathbb{I}\{\underbrace{f(\mathbf{x}_+) - f(\mathbf{x}\_)}_{\text{ROC Margin}} \geq 0\})$$

0-1 손실함수



AUC-Opimizing Classifier

다범주 분류

고려대학교

- $p(\mathbf{x})$를 추정하는 방법은 손쉽게 확장 가능
  - Logistic Regression
  - LDA
  - kNN
  - Naive Bayes

이항분포 대신 다항분포를 활용!

다범주 분류

고려대학교

- k개의 분류경계 $\{f_1(\mathbf{x}), \cdots, f_k(\mathbf{x})\}$를 찾는 방법은 상대적으로 쉽지 않음
  - 이항분류기를 반복적으로 적용 (One vs. One / One vs. Rest)
  - 다범주 SVM

a.	(Lee et al., 2004)	$\sum_{k \neq y}[1 + f_k(\mathbf{x})]_+$
b.	(Naive Hinge)	$[1 - f_y(\mathbf{x})]_+$
c.	(Vapnik, 1998)	$\sum_{k \neq y}[1 - (f_y(\mathbf{x}) - f_j(\mathbf{x})]_+$
d.	(Crammer and Singer, 2001)	$[1 - \min_j(f_y(\mathbf{x}) - f_j(\mathbf{x}))]_+$

  - Neural Net / Tree 기반의 방법은 매우 간단하게 확장 가능

한국자료분석학회
The Korean Data Analysis Society

정리

고려대학교

요약 / 정리

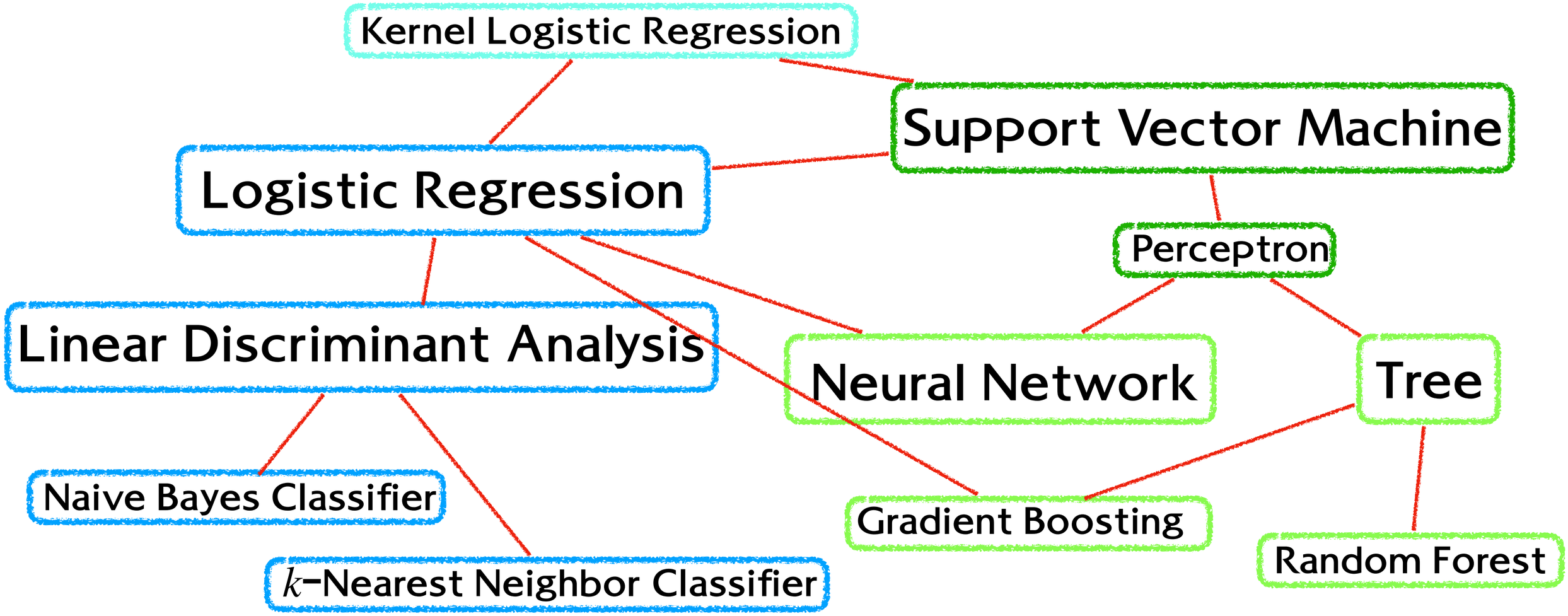
고려대학교

- **분류를 해결하는 원리는 두가지**
  - 분포를 추정하나거나, 분류경계를 학습하거나
- **두가지 방법은 오분류율을 최소화한다는 관점에서 연결**
- **이를 바탕으로 ERM 문제로 일반화 가능**
  - 다양한 방식으로 확장가능

    불균형 분류, 다범주 분류, … …

분류 모형

고려대학교


Kernel Logistic Regression
Support Vector Machine
Logistic Regression
Perceptron
Linear Discriminant Analysis
Neural Network
Tree
Naive Bayes Classifier
Gradient Boosting
Random Forest
k-Nearest Neighbor Classifier