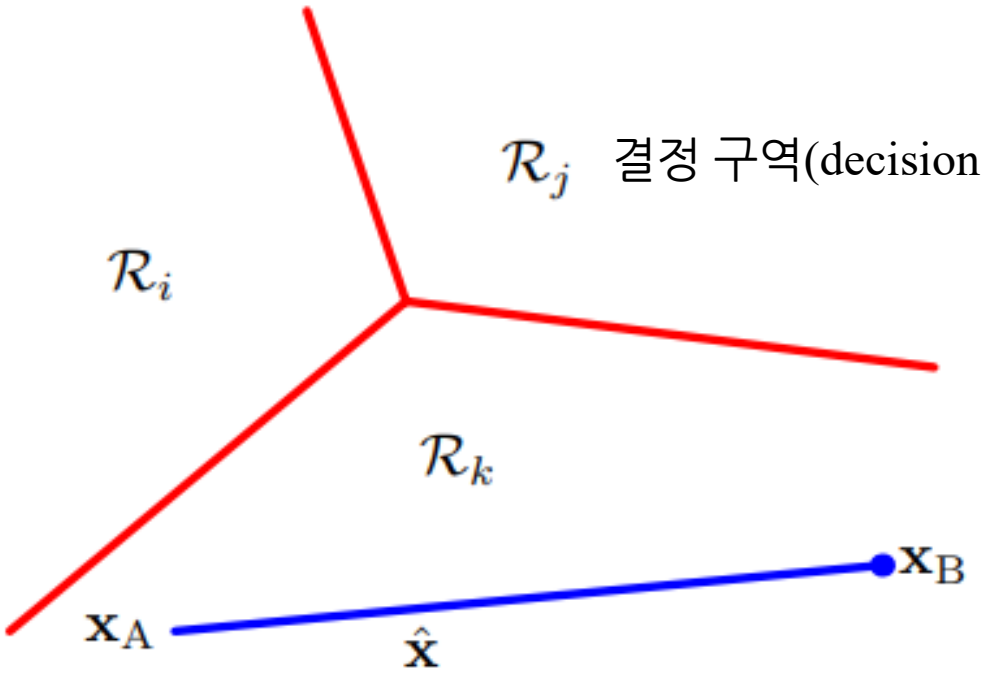
Ch. 4
Linear Models for Classification

조윤정

Contents

4 선형 분류 모델

입력
X – (D)



결정 구역(decision region)

결정 경계(decision boundary)
결정 표면(decision surface) – (D-1)

데이터 집합 X : 선형 분리 가능(linearly separable)한 집합

4 선형 분류 모델

One-hot-encoding

$$\mathbf{t} = (0, 1, 0, 0, 0)^{\mathrm{T}}$$

t_k :클래스 C_k일 확률

타깃변수 t
해석

확률적 모델

1) $t \in \{0, 1\}$
   $t=1$ C_1
   $t=0$ C_2

2) t : C_1에 속할 확률

4 선형 분류 모델

1장 - 분류 문제 푸는 3가지 방법

1) 판별 함수 (discriminant function) 만들어 활용
2) 추론 단계에 조건부 확률 분포 모델 만들고 이를 활용(추론, 결정 분리)
   1) 직접 모델 – 조건부 확률을 매개변수 모델로 표현
   2) 생성적인 방법 – 베이지안 정리 이용

일반화된 선형 모델(generalized linear model)

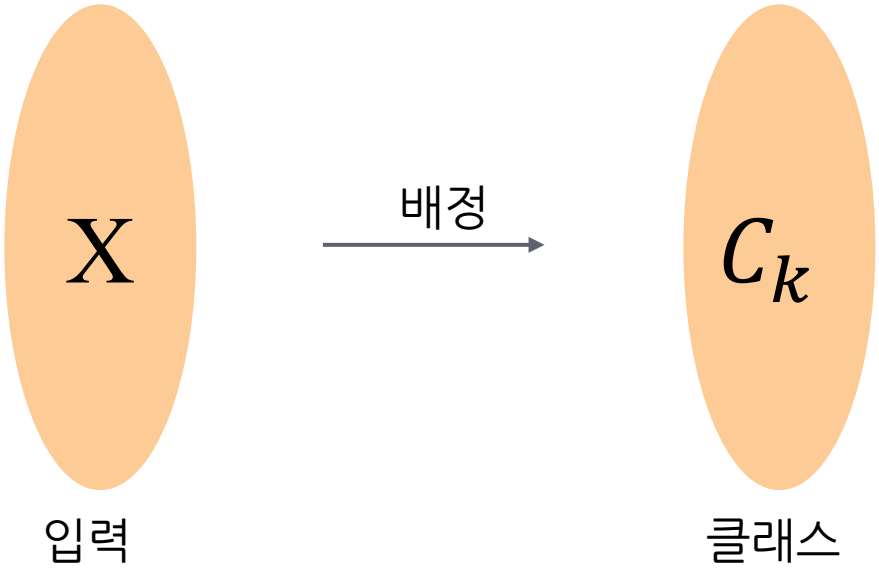
f 는 비선형이더라도 결정 경계면은 x에 대한 선형 함수
→ 매개변수에 대해 선형적 X

$$y(\mathbf{x}) = f\left(\mathbf{w}^{\mathrm{T}}\mathbf{x} + w_0\right).$$

활성화 함수(activation function) ⟶ 역함수 ⟶ 연결 함수(link function)

결정 경계면 : $y(\mathbf{x}) = \text{constant} \longrightarrow \mathbf{w}^{\mathrm{T}}\mathbf{x} + w_0 = \text{constant}$

4.1 판별함수



선형 판별(linear discriminant) : 결정표면이 초평면

4.1.1 두 개의 클래스

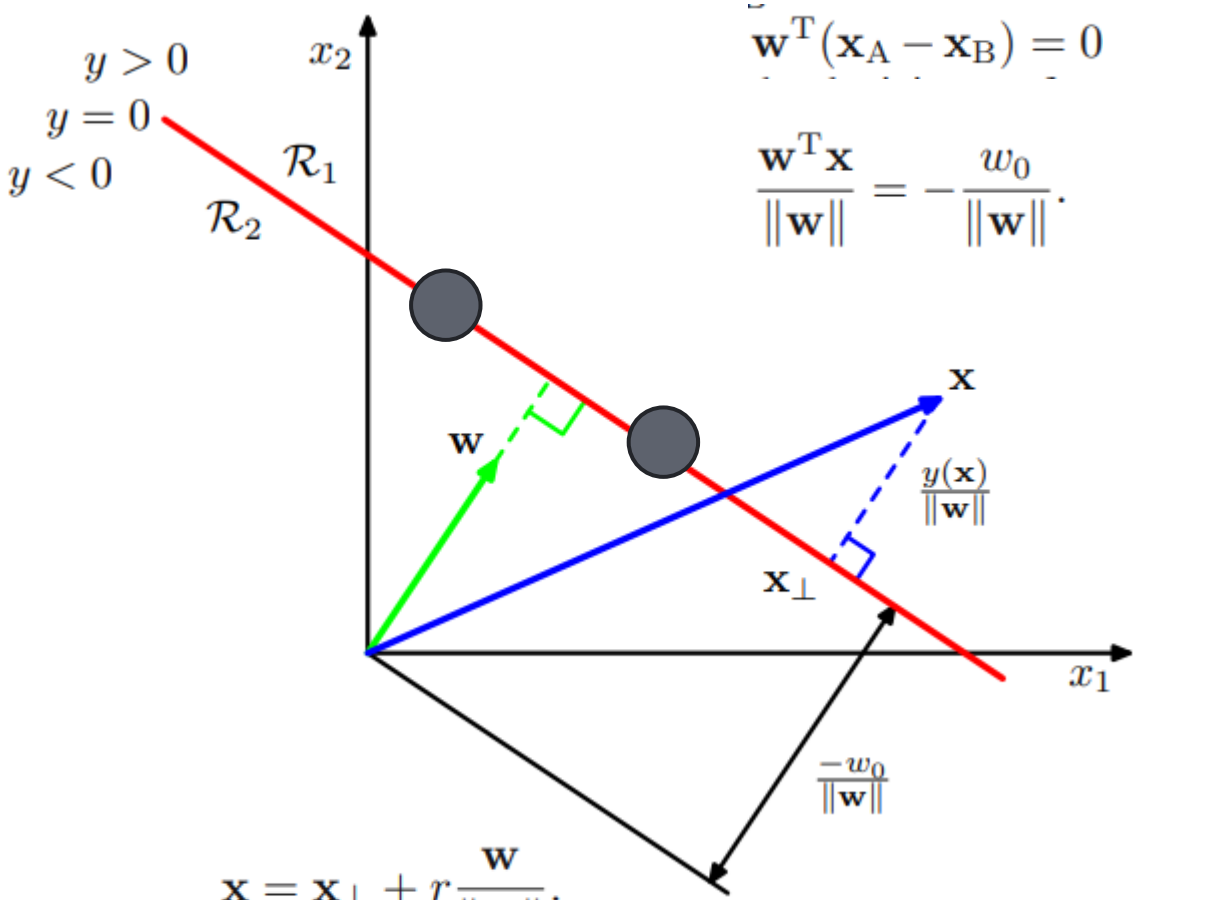
선형 판별 함수 가장 단순하게 표현

$$y(\mathbf{x}) = \mathbf{w}^{\mathrm{T}}\mathbf{x} + w_0$$

w : 가중 벡터
w_0 : 편향
threshold(임계값) : 편향의 음의 값

$y(\mathbf{x}) = 0$, 결정경계

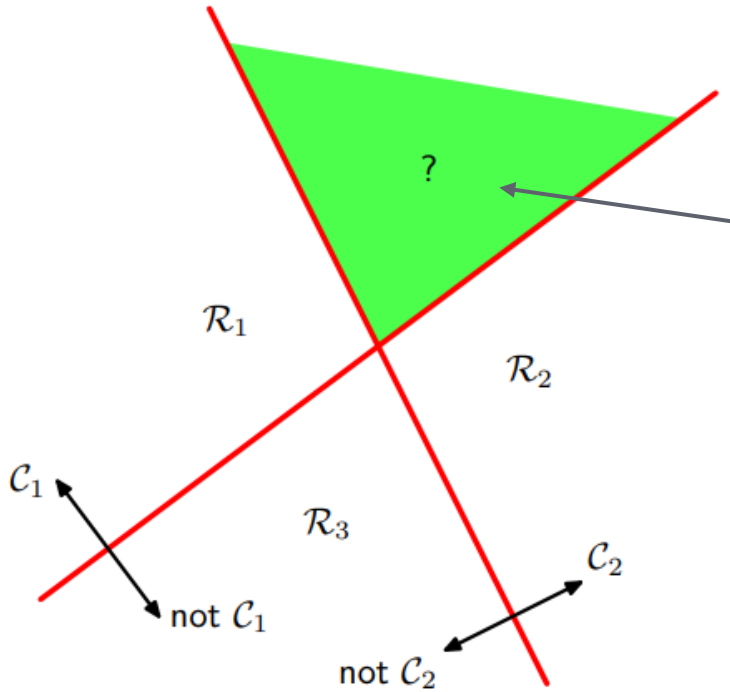
입력 차원 : D
초평면 : D차원 입력 공간 상의 (D-1)차원



4.1.2 다중 클래스

일대다 분류기

2클래스 판별 함수들 —(K -1)개→ K클래스 판별 함수 (K > 2)

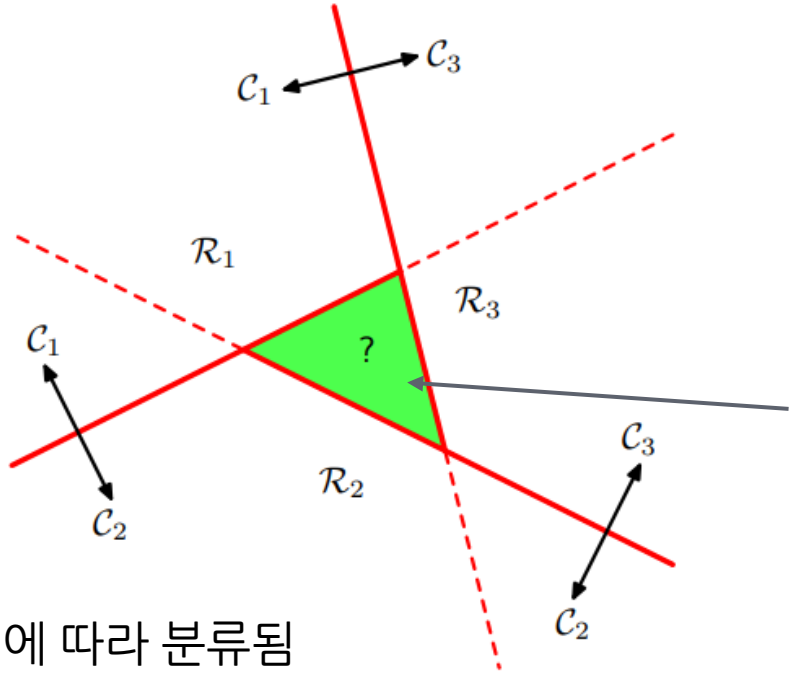


C_1이자 C_2 인 영역이 생김,
분류해야하는데 중복,
즉 불확실한 영역이 발생

4.1.2 다중 클래스

일대일 분류기

2클래스 판별 함수들 —K(K -1)/2 개→ K클래스 판별 함수 (K > 2)



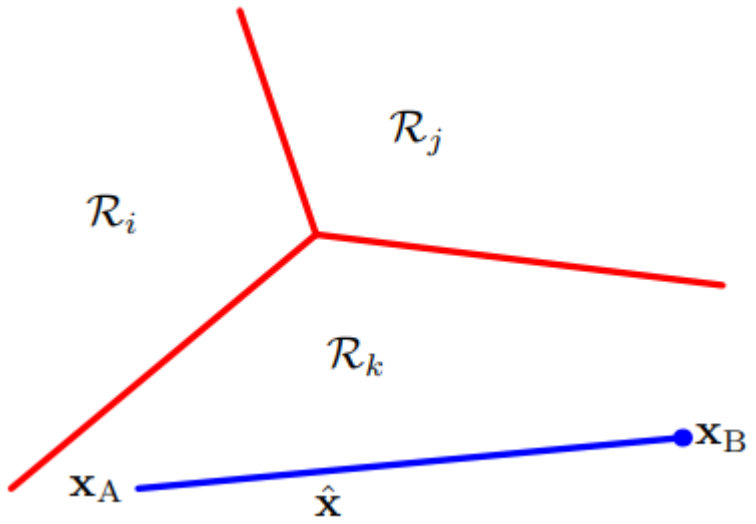
C_1이자 C_2 이자 C_3 인 영역이 생김,
분류해야하는데 중복,
즉 불확실한 영역이 발생

각 점은 판별 함수들의 다수결에 따라 분류됨

4.1.2 다중 클래스

문제 해결

선형 함수들 —K개→ 하나의 K클래스 판별 함수 (K > 2)



$$y_k(\mathbf{x}) = \mathbf{w}_k^{\mathrm{T}}\mathbf{x} + w_{k0}$$

if $y_k(\mathbf{x}) > y_j(\mathbf{x})$ for all $j \neq k$.

포인트 x를 C_k 에 배정

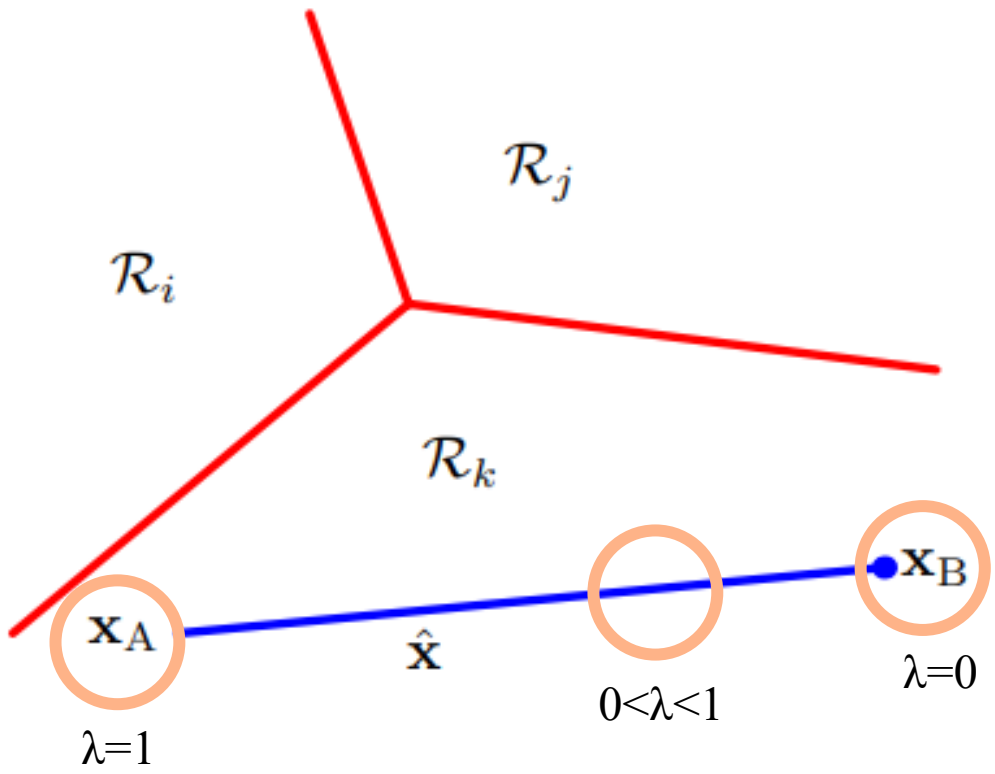
클래스 C_k 와 클래스 C_j 의 결정경계 : $y_k(\mathbf{x}) = y_j(\mathbf{x})$ ⟶ 초평면 : $(\mathbf{w}_k - \mathbf{w}_j)^{\mathrm{T}}\mathbf{x} + (w_{k0} - w_{j0}) = 0.$

4.1.2 다중 클래스

$$y_k(\hat{\mathbf{x}}) = \lambda y_k(\mathbf{x}_\mathrm{A}) + (1-\lambda) y_k(\mathbf{x}_\mathrm{B}).$$

볼록 집합 (Convex Set):

정의: 집합이 볼록하다는 것은,
그 집합 내의 임의의 두 점을 선택했을 때,
그 두 점을 잇는 선분이 항상 그 집합 내에 존재하는 것을 의미함



$$\hat{\mathbf{x}} = \lambda \mathbf{x}_\mathrm{A} + (1-\lambda)\mathbf{x}_\mathrm{B}$$

$$y_k(\mathbf{x}_\mathrm{A}) > y_j(\mathbf{x}_\mathrm{A}), \quad y_k(\mathbf{x}_\mathrm{B}) > y_j(\mathbf{x}_\mathrm{B}) \longrightarrow y_k(\hat{\mathbf{x}}) > y_j(\hat{\mathbf{x}})$$

R_k는 Convex

선형 보간(linear interpolation) :
두 점 사이 선분 위의 점을 찾는 법

4.1.3 ~ 4.1.7

선형 판별 함수의 '매개변수' 학습

4.1.3 분류를 위한 최소 제곱법

4.1.4 피셔의 선형 판별

(4.1.5 두 클래스 문제에서 피셔는 최소 제곱법의 특별 케이스)

4.1.6 다중 클래스에 대한 피셔 판별식

4.1.7 퍼셉트론 알고리즘

4.1.3 분류를 위한 최소 제곱법

사용하는 이유 : 입력 벡터가 주어졌을 때 표적 벡터의 조건부 기댓값($E[t|x]$)의 근삿값을 구하는 방법이기에

각각의 클래스 C_k 들을 선형 모델로 표현

$$y_k(\mathbf{x}) = \mathbf{w}_k^{\mathrm{T}}\mathbf{x} + w_{k0}$$

↓ 벡터 표기 이용

$$\mathbf{y}(\mathbf{x}) = \widetilde{\mathbf{W}}^{\mathrm{T}}\widetilde{\mathbf{x}}$$

이진 부호화 : $E[t|x]) \rightarrow$ 사후 글래스 확률의 벡터

성능 ↓ (선형 모델의 제한적인 유연성 ← (0,1) 범위 밖 가질 수 있음)

제곱합의 오류 함수

$$E_D(\widetilde{\mathbf{W}}) = \frac{1}{2}\mathrm{Tr}\left\{(\widetilde{\mathbf{X}}\widetilde{\mathbf{W}} - \mathbf{T})^{\mathrm{T}}(\widetilde{\mathbf{X}}\widetilde{\mathbf{W}} - \mathbf{T})\right\}.$$

미분 ↓

$$\widetilde{\mathbf{W}} = (\widetilde{\mathbf{X}}^{\mathrm{T}}\widetilde{\mathbf{X}})^{-1}\widetilde{\mathbf{X}}^{\mathrm{T}}\mathbf{T} = \widetilde{\mathbf{X}}^{\dagger}\mathbf{T}$$

최종 판별 함수

$$\mathbf{y}(\mathbf{x}) = \widetilde{\mathbf{W}}^{\mathrm{T}}\widetilde{\mathbf{x}} = \mathbf{T}^{\mathrm{T}}\left(\widetilde{\mathbf{X}}^{\dagger}\right)^{\mathrm{T}}\widetilde{\mathbf{x}}.$$

4.1.3 분류를 위한 최소 제곱법

$$\mathbf{a}^{\mathrm{T}}\mathbf{t}_n + b = 0$$

어떤 훈련 집합의 표적 벡터들이 전부 아래 식을 만족한다면

$$\mathbf{a}^{\mathrm{T}}\mathbf{y}(\mathbf{x}) + b = 0.$$

어떤 x값에 대한 모델 예측값 등 같은 제약 조건 (왼쪽 식)을 만족

최종 판별 함수

$$\mathbf{y}(\mathbf{x}) = \widetilde{\mathbf{W}}^{\mathrm{T}}\widetilde{\mathbf{x}} = \mathbf{T}^{\mathrm{T}}\left(\widetilde{\mathbf{X}}^{\dagger}\right)^{\mathrm{T}}\widetilde{\mathbf{x}}.$$

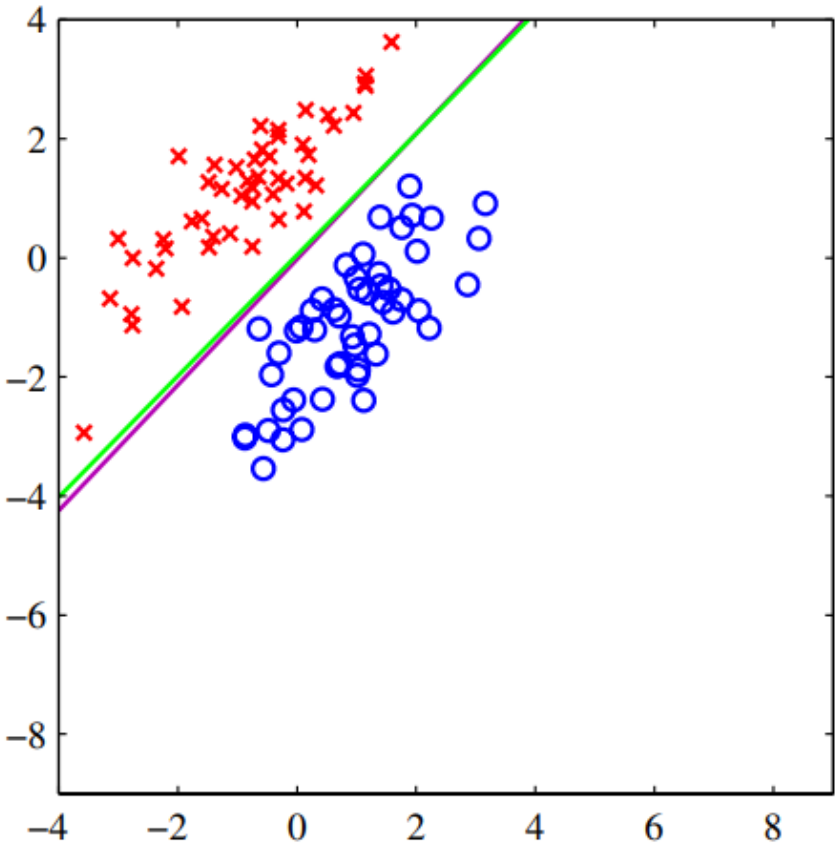
원 핫 인코딩을 한 K개의 클래스 y(x)의 원소들을 전부 합하면 1

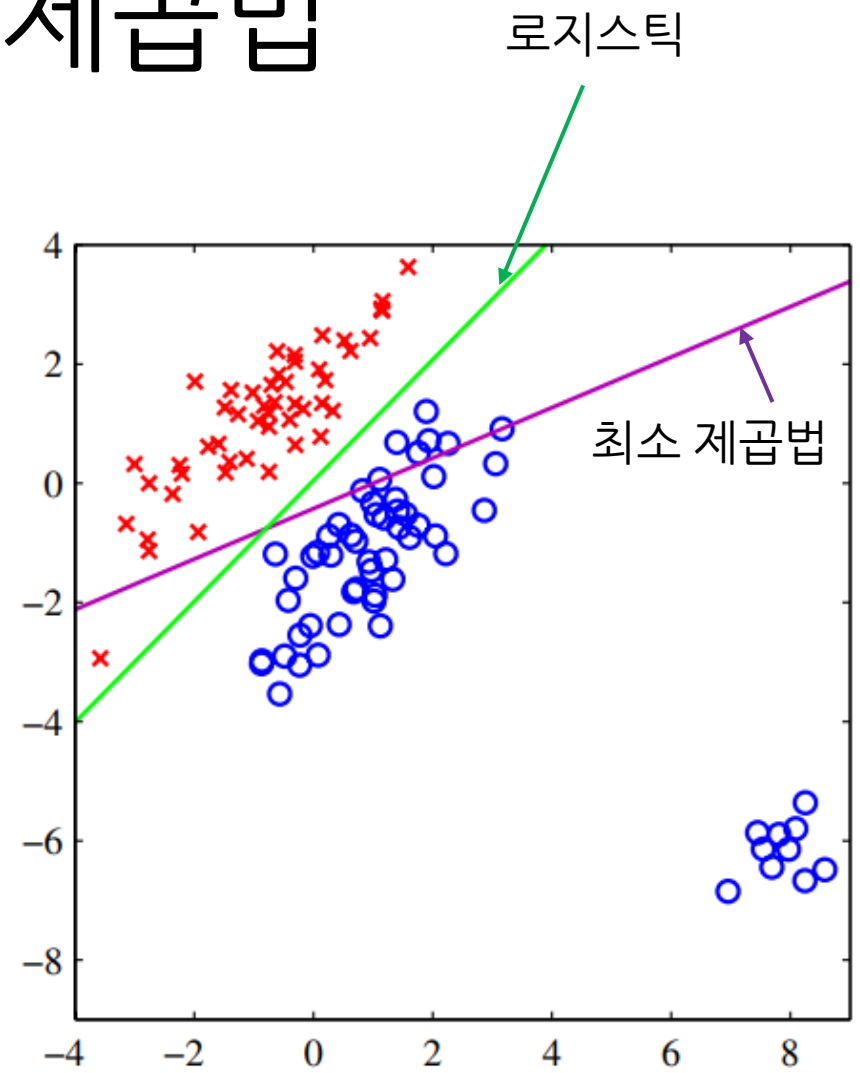
Ex) (0,1,0,0,0), K=5

Why? (0,1) 구간 사이에 값이 존재해야한다는 '제약 조건 ' 이 없음

모델의 출력값을 확률로 해석

4.1.3 분류를 위한 최소 제곱법

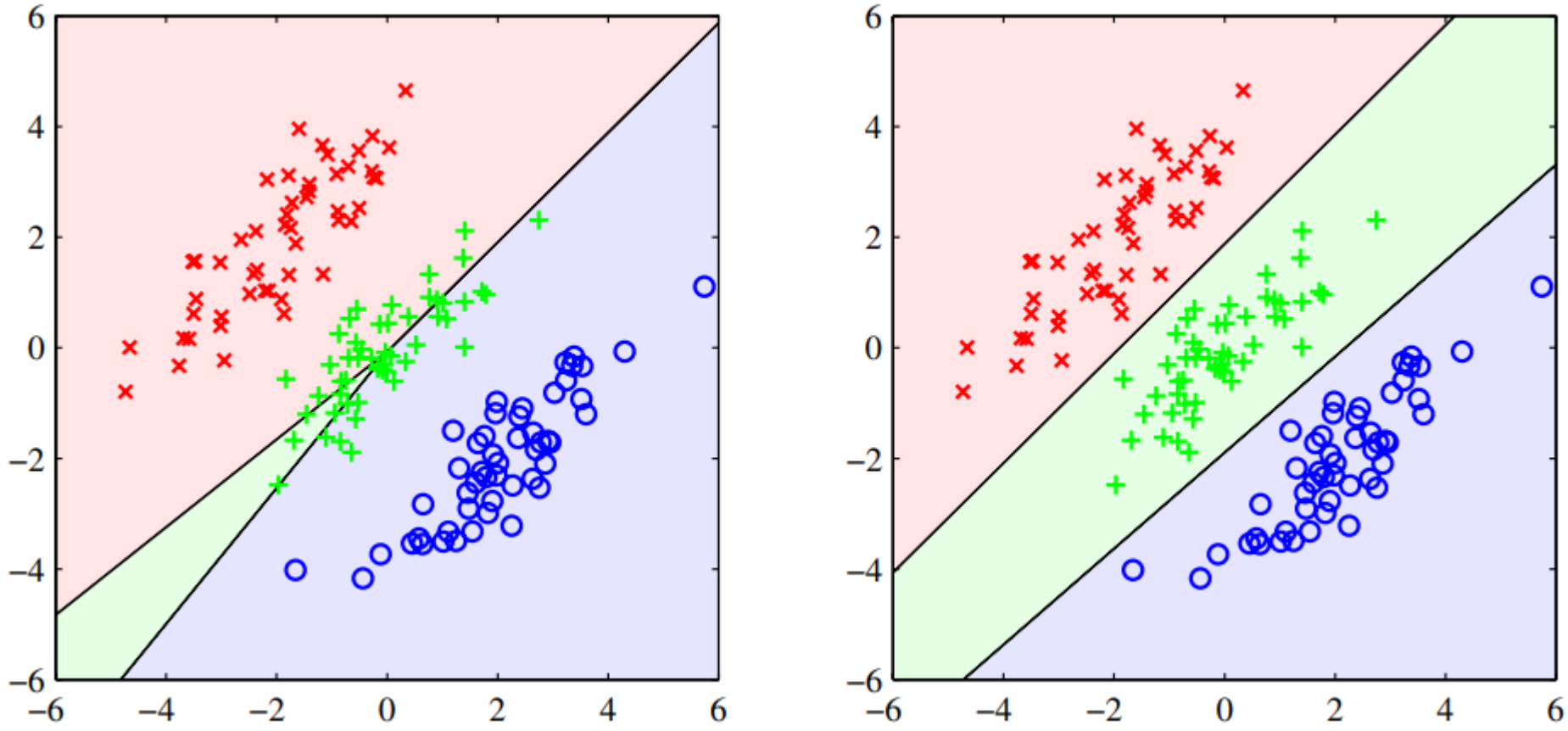




이상값에 민감하다! → 강건성이 부족

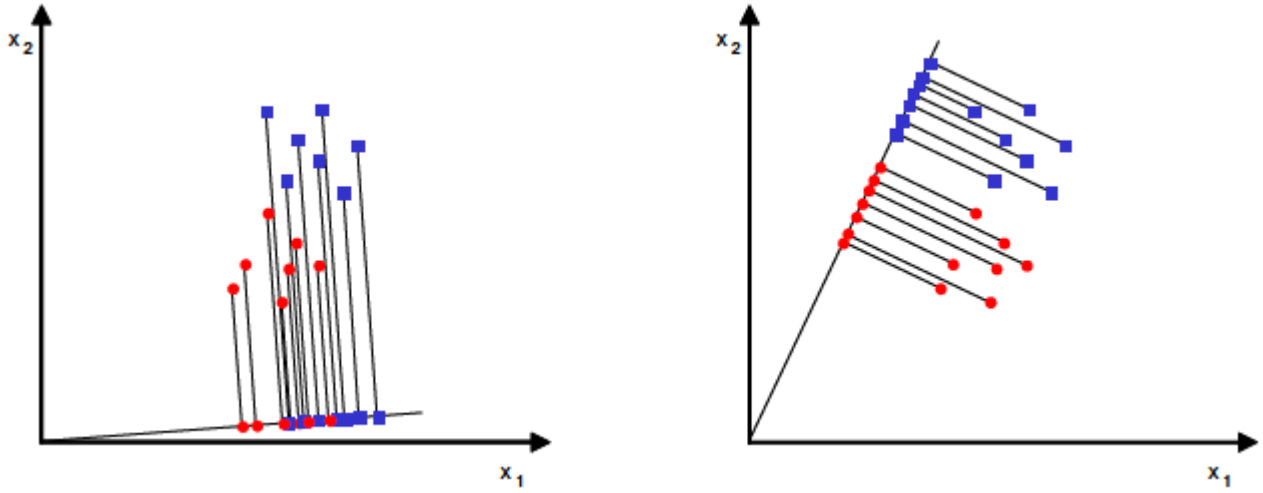
4.1.3 분류를 위한 최소 제곱법

각 클래스마다 결정 경계 존재



최소 제곱법은 가우시안 조건부 분포를 가정하기에 이런 일이 발생함

4.1.4 피셔의 선형 판별



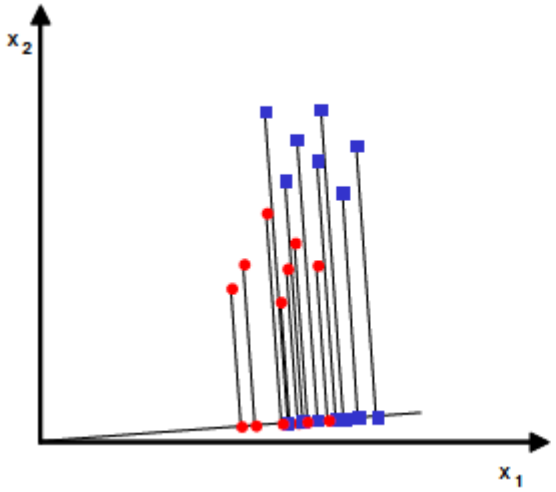
표본의 두 집단을 가장 잘 분리시키는 선에 정사영 하는 방법

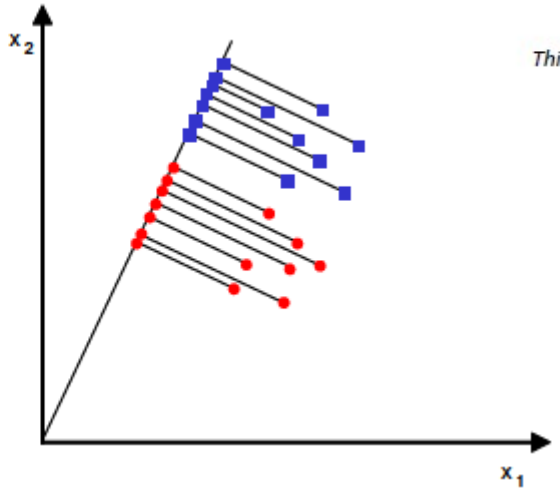
1차원 투영 시 상당한 양의 정보 손실 &
원래 잘 분리되었던 분리가 잘 이루어지지 X

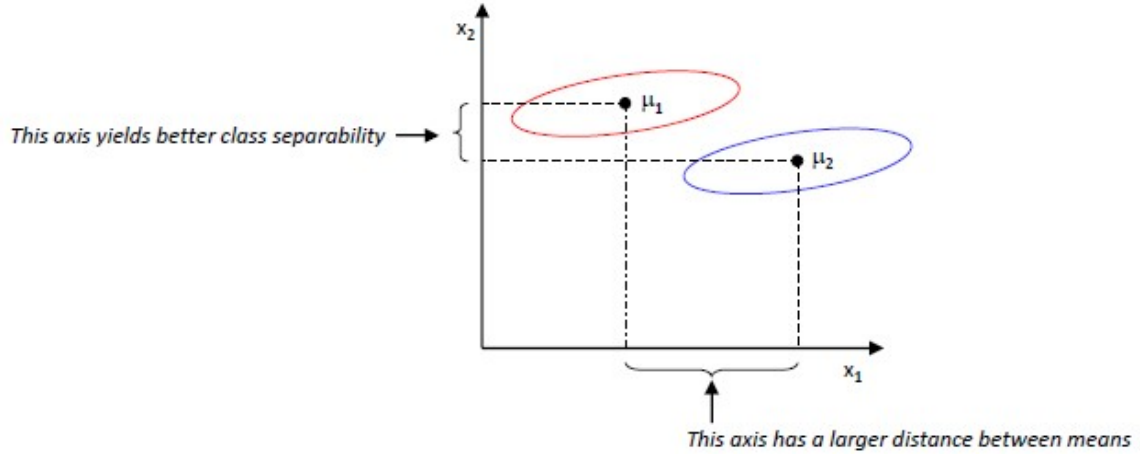
→

가중 벡터 w의 성분들을 잘 조절하여
클래스 간의 분리를 최대화 해야함.

4.1.4 피셔의 선형 판별







분리 측도 (measure of separation)

1. 정사영된 곳에서 각 집합의 평균 차이 : 좋지X

   각각 집단의 표본 분산을 고려하지 않았서 안 좋음

x,y 공간들의 평균 벡터

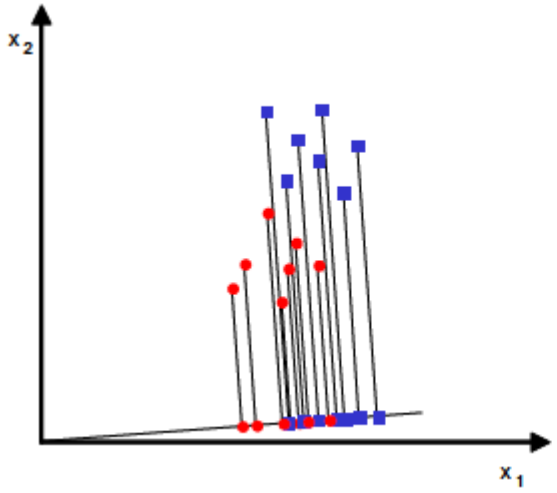
$$\mu_i = \frac{1}{N_i}\sum_{x\in\omega_i} x \text{ and } \tilde{\mu}_i = \frac{1}{N_i}\sum_{y\in\omega_i} y = \frac{1}{N_i}\sum_{x\in\omega_i} w^T x = w^T\mu_i$$

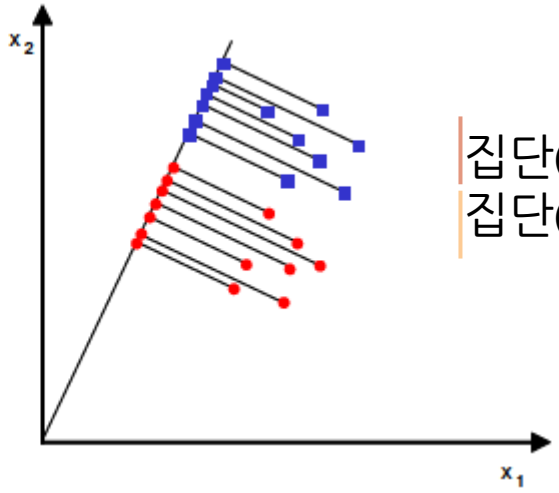
정사영된 평균의 차

$$J(w) = |\tilde{\mu}_1 - \tilde{\mu}_2| = |w^T(\mu_1 - \mu_2)|$$

4.1.4 피셔의 선형 판별

클래스 내 분산 $s_k^2 = \sum_{n \in C_k} (y_n - m_k)^2$





집단(클래스) 간 분산 값 ↑
집단(클래스) 내 분산 값 ↓

분리 측도 (measure of separation)

2. 피셔의 해법

$$J(\mathbf{w}) = \frac{(m_2 - m_1)^2}{s_1^2 + s_2^2}.$$

투영된 클래스 평균 사이의 분리 정도 ↑
각 클래스 내의 분산 ↓
→ 클래스 간 중복 최소화

4.1.4 피셔의 선형 판별

$$(\mathbf{w}^{\mathrm{T}}\mathbf{S}_{\mathrm{B}}\mathbf{w})\mathbf{S}_{\mathrm{W}}\mathbf{w} = (\mathbf{w}^{\mathrm{T}}\mathbf{S}_{\mathrm{W}}\mathbf{w})\mathbf{S}_{\mathrm{B}}\mathbf{w}.$$

J(w)는 위의 경우에서 극대화됨

$y = \mathbf{w}^{\mathrm{T}}\mathbf{x}.$ 선형 함수 정의

$m_k = \mathbf{w}^{\mathrm{T}}\mathbf{m}_k$

$$s_k^2 = \sum_{n \in \mathcal{C}_k} (y_n - m_k)^2$$

$\longrightarrow$ $J(\mathbf{w}) = \dfrac{\mathbf{w}^{\mathrm{T}}\mathbf{S}_{\mathrm{B}}\mathbf{w}}{\mathbf{w}^{\mathrm{T}}\mathbf{S}_{\mathrm{W}}\mathbf{w}}$

w에 대해 미분 ↑

스칼라만 보기 ↓

$$\mathbf{w} \propto \mathbf{S}_{\mathrm{W}}^{-1}(\mathbf{m}_2 - \mathbf{m}_1).$$

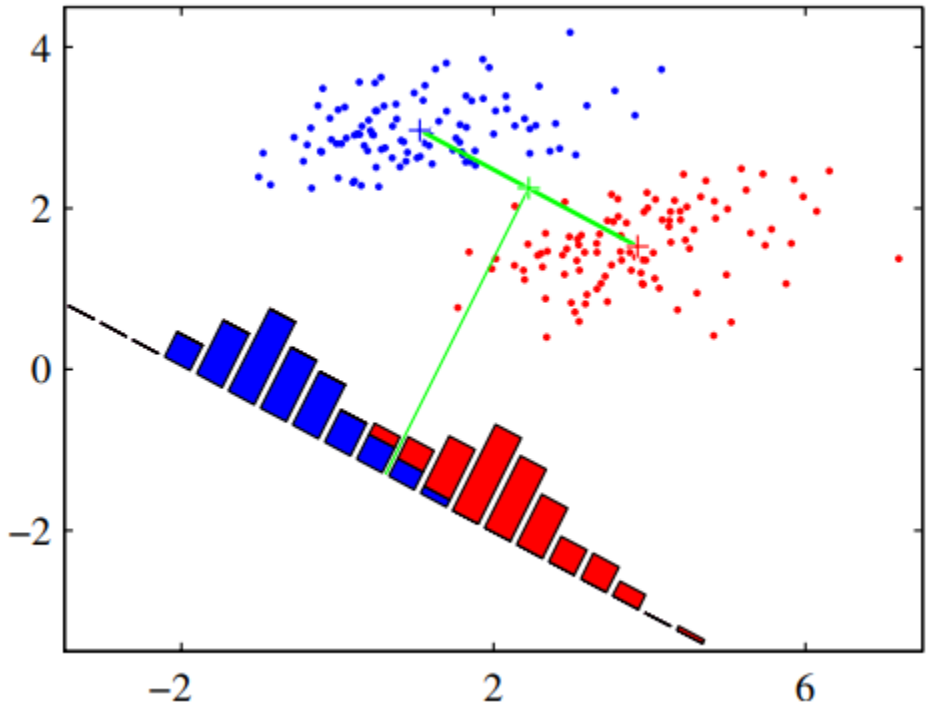
클래스 간 공분산 행렬(between class)

$$\mathbf{S}_{\mathrm{B}} = (\mathbf{m}_2 - \mathbf{m}_1)(\mathbf{m}_2 - \mathbf{m}_1)^{\mathrm{T}}$$

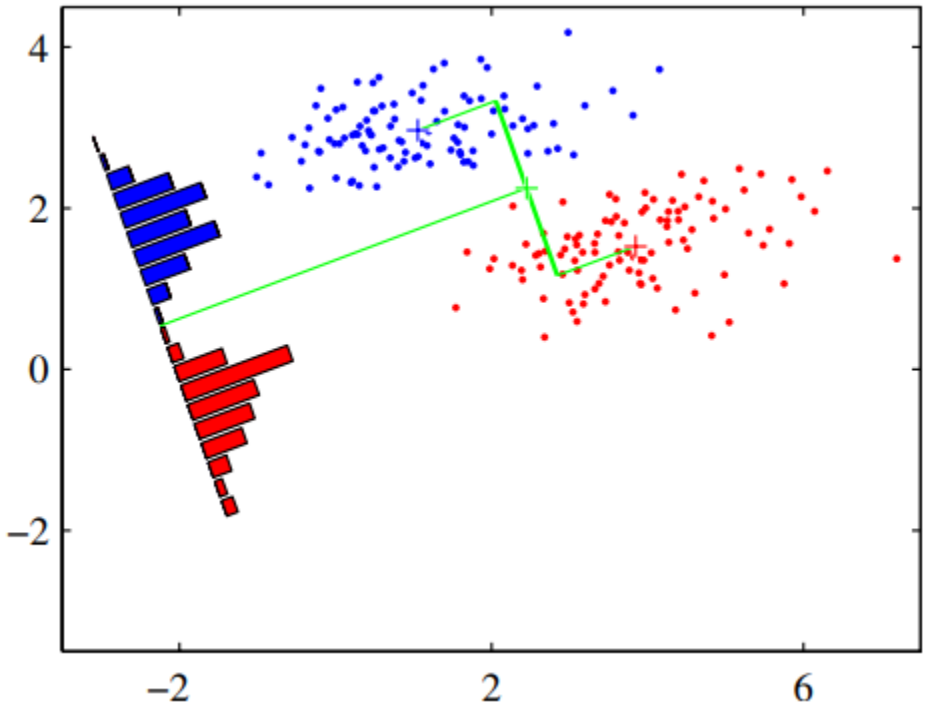
클래스 내 공분산 행렬(within class)

$$\mathbf{S}_{\mathrm{W}} = \sum_{n \in \mathcal{C}_1} (\mathbf{x}_n - \mathbf{m}_1)(\mathbf{x}_n - \mathbf{m}_1)^{\mathrm{T}} + \sum_{n \in \mathcal{C}_2} (\mathbf{x}_n - \mathbf{m}_2)(\mathbf{x}_n - \mathbf{m}_2)^{\mathrm{T}}.$$

4.1.4 피셔의 선형 판별

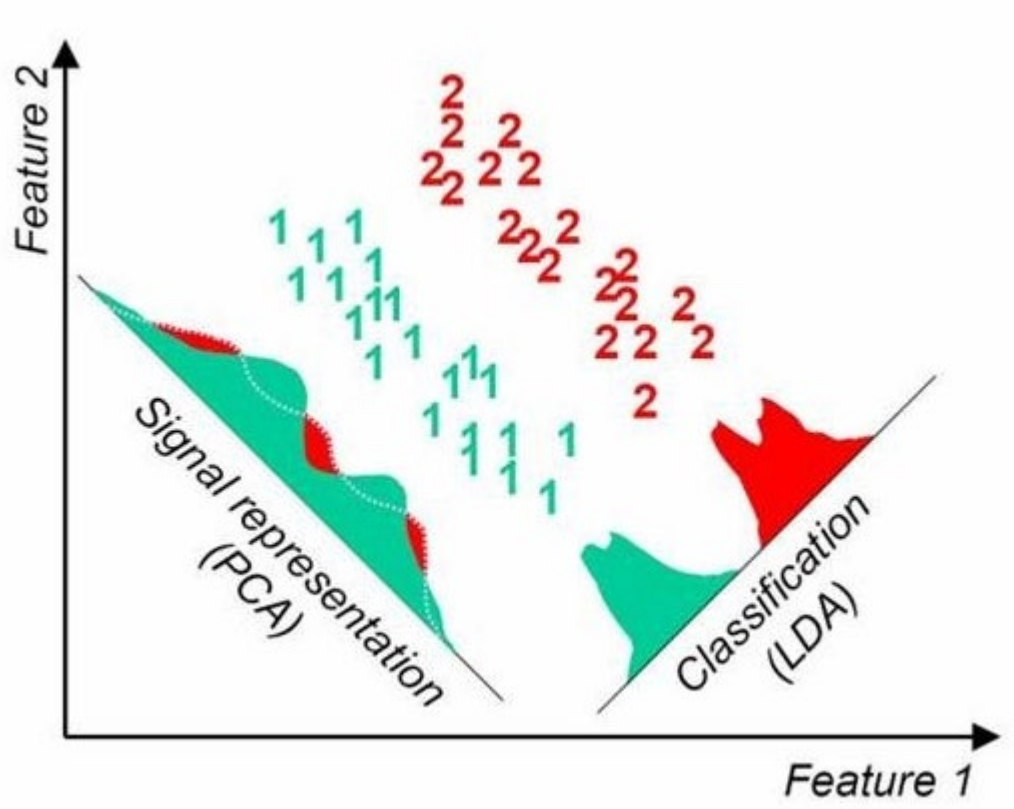


분산 고려 X



클래스 내 분산 고려

4.1.4 피셔의 선형 판별


Feature 2
Feature 1
Signal representation
(PCA)
Classification
(LDA)

4.1.5 최소 제곱법과의 관계

선형 판별 함수의 '**매개변수**' 학습

출력 공간 상에서 클래스 간의 분리를 최대화 하기 위해

모델의 예측값을 가능한 한 표적값에 가깝게 하기 위해

4.1.3 분류를 위한 **최소 제곱법**

4.1.4 **피셔의 선형 판별**

(4.1.5 두 클래스 문제에서 피셔는 최소 제곱법의 특별 케이스)

4.1.6 다중 클래스에 대한 피셔 판별식

4.1.7 **퍼셉트론 알고리즘**

피셔 기준은 최소 제곱법의 특별 케이스

4.1.5 최소 제곱법과의 관계

원 핫 인코딩을 표적값에 적용

↓

다른 부호화 적용

→ 최소 제곱 해는 곧 피셔 해

C_1 의 표적값 : N / N_1

C_2 의 표적값 : - N / N_2

N : 전체 패턴들의 숫자

N_1 : 클래스 C_1 에 있는 패턴들의 숫자

N_2 : 클래스 C_2 에 있는 패턴들의 숫자

4.1.5 최소 제곱법과의 관계

$$\mathbf{w} \propto \mathbf{S}_{\mathrm{W}}^{-1}(\mathbf{m}_2 - \mathbf{m}_1)$$

$\mathcal{C}_1$ if $y(\mathbf{x}) = \mathbf{w}^{\mathrm{T}}(\mathbf{x}-\mathbf{m}) > 0$ and class $\mathcal{C}_2$ otherwise.

$$E = \frac{1}{2}\sum_{n=1}^{N}\left(\mathbf{w}^{\mathrm{T}}\mathbf{x}_n + w_0 - t_n\right)^2.$$

↓ w_0 와 w에 대해 미분, 각 미분값을 0으로 설정

$$\sum_{n=1}^{N}\left(\mathbf{w}^{\mathrm{T}}\mathbf{x}_n + w_0 - t_n\right) = 0$$

$$\sum_{n=1}^{N}\left(\mathbf{w}^{\mathrm{T}}\mathbf{x}_n + w_0 - t_n\right)\mathbf{x}_n = 0.$$

표적값 부호화를 t_n에 대해 적용

$$w_0 = -\mathbf{w}^{\mathrm{T}}\mathbf{m}$$

t_n 식

$$\sum_{n=1}^{N} t_n = N_1\frac{N}{N_1} - N_2\frac{N}{N_2} = 0$$

m은 전체 데이터 집합의 평균

$$\mathbf{m} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n = \frac{1}{N}(N_1\mathbf{m}_1 + N_2\mathbf{m}_2).$$

정리한 후 t_n에 대해 부호화 적용

$$\left(\mathbf{S}_{\mathrm{W}} + \frac{N_1 N_2}{N}\mathbf{S}_{\mathrm{B}}\right)\mathbf{w} = N(\mathbf{m}_1 - \mathbf{m}_2)$$

$$\mathbf{S}_{\mathrm{B}} = (\mathbf{m}_2 - \mathbf{m}_1)(\mathbf{m}_2 - \mathbf{m}_1)^{\mathrm{T}}$$

$$\mathbf{S}_{\mathrm{W}} = \sum_{n\in\mathcal{C}_1}(\mathbf{x}_n - \mathbf{m}_1)(\mathbf{x}_n - \mathbf{m}_1)^{\mathrm{T}} + \sum_{n\in\mathcal{C}_2}(\mathbf{x}_n - \mathbf{m}_2)(\mathbf{x}_n - \mathbf{m}_2)^{\mathrm{T}}.$$

4.1.6 다중 클래스에 대한 피셔 판별식

D > K

클래스 K개 (K > 2)

$$\mathbf{y} = \mathbf{W}^{\mathrm{T}}\mathbf{x}.$$ 특징값 벡터 y

$$\mathbf{S}_{\mathrm{W}} = \sum_{k=1}^{K} \mathbf{S}_k$$ 클래스 내 공분산 행렬

$$\begin{aligned} \mathbf{S}_k &= \sum_{n \in \mathcal{C}_k} (\mathbf{x}_n - \mathbf{m}_k)(\mathbf{x}_n - \mathbf{m}_k)^{\mathrm{T}} \\ \mathbf{m}_k &= \frac{1}{N_k} \sum_{n \in \mathcal{C}_k} \mathbf{x}_n \end{aligned}$$

전체 공분산 행렬

$$\mathbf{S}_{\mathrm{T}} = \sum_{n=1}^{N} (\mathbf{x}_n - \mathbf{m})(\mathbf{x}_n - \mathbf{m})^{\mathrm{T}}$$

m은 데이터 집합 평균

$$\mathbf{m} = \frac{1}{N} \sum_{n=1}^{N} \mathbf{x}_n = \frac{1}{N} \sum_{k=1}^{K} N_k \mathbf{m}_k$$

4.1.6 다중 클래스에 대한 피셔 판별식

D > K

$N = \sum_k N_k$ 전체 데이터 포인트들의 개수

클래스 K개 (K > 2)

클래스 내 공분산 행렬

$$\mathbf{S}_{\mathrm{T}} = \mathbf{S}_{\mathrm{W}} + \mathbf{S}_{\mathrm{B}}$$

추가적인 행렬

$$\mathbf{S}_{\mathrm{B}} = \sum_{k=1}^{K} N_k(\mathbf{m}_k - \mathbf{m})(\mathbf{m}_k - \mathbf{m})^{\mathrm{T}}.$$

클래스 간 공분산

$$\mathbf{s}_{\mathrm{W}} = \sum_{k=1}^{K} \sum_{n \in \mathcal{C}_k} (\mathbf{y}_n - \boldsymbol{\mu}_k)(\mathbf{y}_n - \boldsymbol{\mu}_k)^{\mathrm{T}}$$

$$\mathbf{s}_{\mathrm{B}} = \sum_{k=1}^{K} N_k(\boldsymbol{\mu}_k - \boldsymbol{\mu})(\boldsymbol{\mu}_k - \boldsymbol{\mu})^{\mathrm{T}}$$

원 x 공간상에 정의

$$\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{n \in \mathcal{C}_k} \mathbf{y}_n, \qquad \boldsymbol{\mu} = \frac{1}{N} \sum_{k=1}^{K} N_k \boldsymbol{\mu}_k.$$

비슷한 행렬을 투영된 D'차원의 y 공간상에 정의

4.1.6 다중 클래스에 대한 피셔 판별식

D > K

클래스 K개 (K > 2)

클래스 간 공분산 ↑
클래스 내 공분산 ↓

$$J(\mathbf{W}) = \mathrm{Tr}\left\{\mathbf{s}_{\mathrm{W}}^{-1}\mathbf{s}_{\mathrm{B}}\right\}.$$

↓ 투영 행렬 W에 대한 명시적인 함수

$$J(\mathbf{w}) = \mathrm{Tr}\left\{(\mathbf{W}\mathbf{S}_{\mathrm{W}}\mathbf{W}^{\mathrm{T}})^{-1}(\mathbf{W}\mathbf{S}_{\mathrm{B}}\mathbf{W}^{\mathrm{T}})\right\}.$$

가중치 값은 $\mathbf{S}_{\mathrm{W}}^{-1}\mathbf{S}_{\mathrm{B}}$ 의 D'개의 고윳값
에 해당하는 고유 벡터들에 의해 결정

4.1.6 다중 클래스에 대한 피셔 판별식

$$\mathbf{S}_\mathrm{T} = \mathbf{S}_\mathrm{W} + \mathbf{S}_\mathrm{B}$$

$$\mathbf{S}_\mathrm{B} = \sum_{k=1}^{K} N_k (\mathbf{m}_k - \mathbf{m})(\mathbf{m}_k - \mathbf{m})^\mathrm{T}.$$

K개의 행렬
각각의 행렬은 두 벡터의 외적으로 계수(rank)가 1

→ (K-1)개의 행렬들만 독립적
최대 (K-1)개의 행렬 계수(=0이 아닌 고윳값) 가짐

여기서 고유 벡터들에 의해 (K-1) 차원의 부분 공간에 투영 → J(W)의 값을 바꾸지 않음, 이를 이용해서 (K-1)개보다 많은 선형 '특징' 찾는 것은 불가능

4.1.7 퍼셉트론 알고리즘

$K = 2$

$t \in \{0, 1\}$

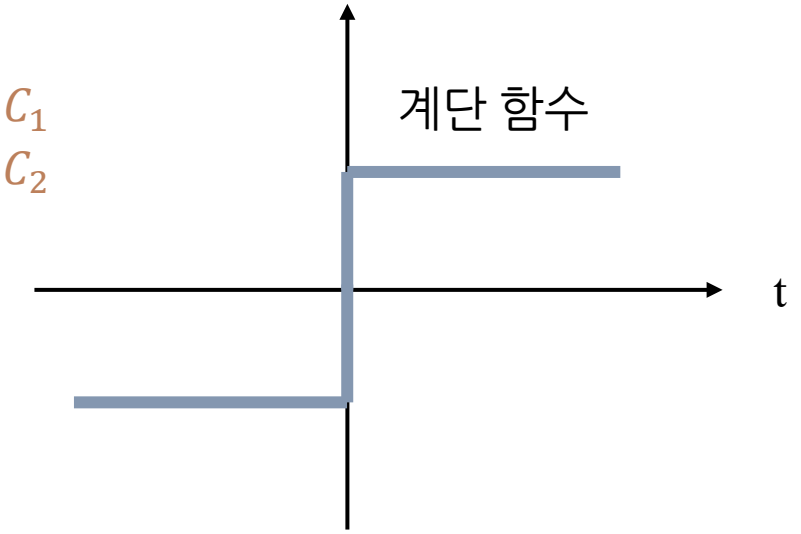
→ 비선형 변환

$$y(\mathbf{x}) = f\left(\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x})\right)$$

입력 벡터 x　　특징 벡터

$$f(a) = \begin{cases} +1, & a \geqslant 0 \quad C_1 \\ -1, & a < 0. \quad C_2 \end{cases}$$

비선형 활성화 함수



매개변수 w를 구하자

→ 오류 함수를 최소화하자

~~오류 함수 : 오분류된 패턴의 총 숫자~~

학습 알고리즘 복잡(w에 대해 조각별 상수 함수)
→ 불연속성, 기울기 이용 못 함.

→ 오류 함수 : 퍼셉트론 기준(perceptron criterion)

4.1.7 퍼셉트론 알고리즘

$$y(\mathbf{x}) = f\left(\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x})\right)$$

$$f(a) = \begin{cases} +1, & a \geqslant 0 \\ -1, & a < 0. \end{cases}$$

→ 이 식을 만족하는 가중치 벡터 **w** 찾기

$$t \in \{-1, +1\}$$

$$\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n)t_n > 0.$$

오류 함수 : 퍼셉트론 기준(perceptron criterion)

$$E_{\mathrm{P}}(\mathbf{w}) = -\sum_{n \in \mathcal{M}} \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}_n t_n$$

$\boldsymbol{\phi}(\mathbf{x}_n)$

오분류된 패턴들의 전체 집합

올바르게 분류된 패턴 – 0 배정

오분류된 패턴 x_n에 대해서 $\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n)t_n$ 최소화

4.1.6 다중 클래스에 대한 피셔 판별식

알고리즘의 단계에 대한 지표(정수)

확률적 경사 하강법 적용 $\mathbf{w}^{(\tau+1)} = \mathbf{w}^{(\tau)} - \eta \nabla E_{\mathrm{P}}(\mathbf{w}) = \mathbf{w}^{(\tau)} + \eta \boldsymbol{\phi}_n t_n$

학습률

참고 : 훈련 중 w가 변함 → 오분류가 되는 패턴 집합도 변함.

올바르게 분류 → w 값 안 변함.

틀리게 분류 → C_1 벡터 $\boldsymbol{\phi}(\mathbf{x}_n)$ + w의 예측치

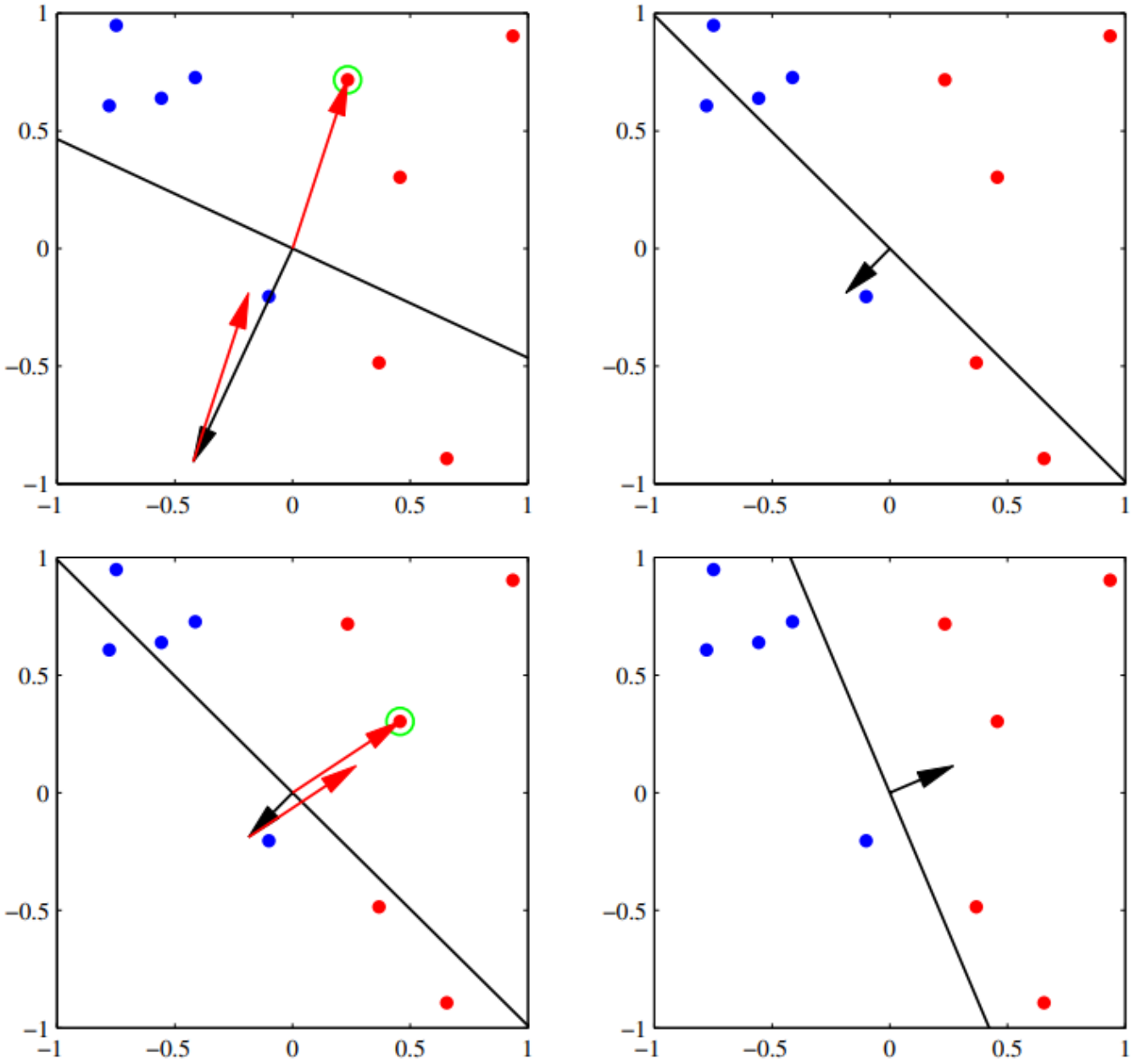
C_2 벡터 $\boldsymbol{\phi}(\mathbf{x}_n)$ - w의 예측치

$\eta = 1$

$(\varphi_n\ t_n)^2 > 0$

$$-\mathbf{w}^{(\tau+1)\mathrm{T}} \boldsymbol{\phi}_n t_n = -\mathbf{w}^{(\tau)\mathrm{T}} \boldsymbol{\phi}_n t_n - (\boldsymbol{\phi}_n t_n)^{\mathrm{T}} \boldsymbol{\phi}_n t_n < -\mathbf{w}^{(\tau)\mathrm{T}} \boldsymbol{\phi}_n t_n$$

각 단계가 전체 오류 함수의 값을 줄인다고 보장 X



4.1.7 퍼셉트론 알고리즘

4.1.7 퍼셉트론 알고리즘

퍼셉트론 수렴 정리(perceptron convergence theorem)

If 문제가 정확한 해를 가지고 있기만 한다면
(=훈련 집합이 선형적으로 분리가 가능하다면)

분리 가능 여부 모름
서로 다른 여러 해 존재 가능성

정확한 해를 유한한 단계 안에 확실히 구할 수 있다.
(단, 수렴을 위해 필요한 단계의 수는 알 수 없음)

시간 문제

매개변수 초기화 방법
데이터 포인트들 입력 순서

한계점

학습 알고리즘 자체가 어려움
확률적인 출력값 내지 않음
K>2 클래스 문제에 대한 일반화X
고정된 기저 함수들의 선형 결합 → 가장 큰 한계점 발생

4.2 확률적 생성 모델

$$\sigma(-a) = 1 - \sigma(a) \qquad \text{역} : a = \ln\left(\frac{\sigma}{1-\sigma}\right)$$

Sigmoid 성질 로짓 함수

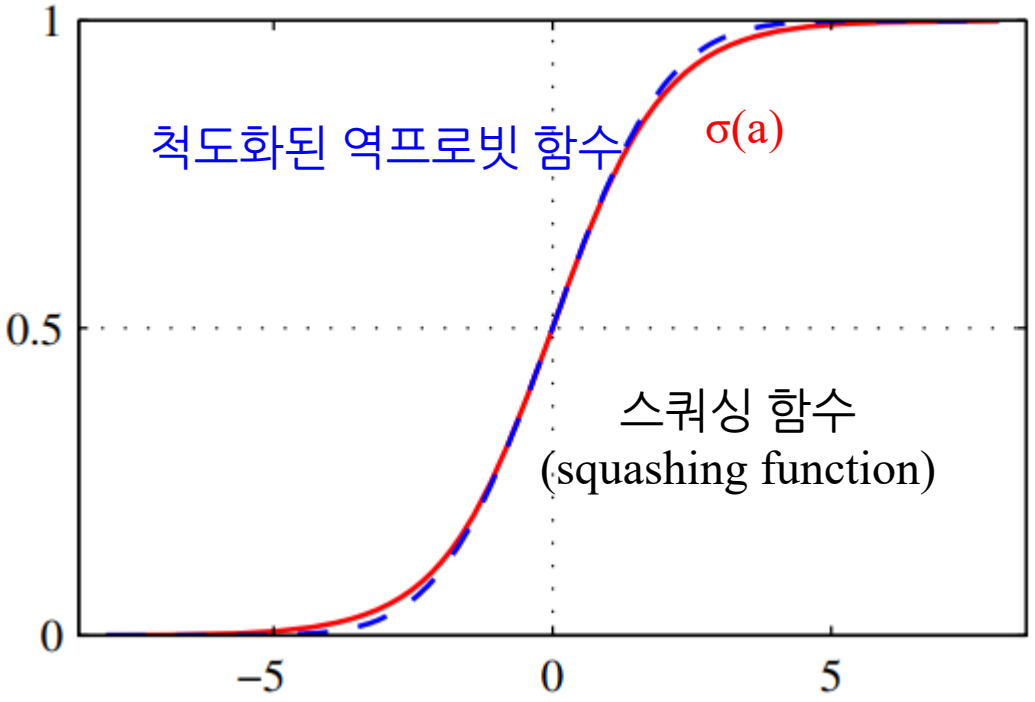
$\ln [p(\mathcal{C}_1|\mathbf{x})/p(\mathcal{C}_2|\mathbf{x})]$ log odds

클래스 2개

$p(\mathbf{x}|\boldsymbol{C_k})$: 클래스별 조건부 밀도

$p(\boldsymbol{C_k})$: 클래스 사전 분포

⟶ $p(\boldsymbol{C_k}|\mathbf{x})$: 사후 확률



$$p(\mathcal{C}_1|\mathbf{x}) = \frac{p(\mathbf{x}|\mathcal{C}_1)p(\mathcal{C}_1)}{p(\mathbf{x}|\mathcal{C}_1)p(\mathcal{C}_1) + p(\mathbf{x}|\mathcal{C}_2)p(\mathcal{C}_2)}$$

$$= \frac{1}{1+\exp(-a)} = \sigma(a)$$

$$a = \ln \frac{p(\mathbf{x}|\mathcal{C}_1)p(\mathcal{C}_1)}{p(\mathbf{x}|\mathcal{C}_2)p(\mathcal{C}_2)}$$

$$\sigma(a) = \frac{1}{1+\exp(-a)}$$

σ(a) : Logistic Sigmoid

4.2 확률적 생성 모델

클래스 K개 (K > 2)

$$p(\mathcal{C}_k|\mathbf{x}) = \frac{p(\mathbf{x}|\mathcal{C}_k)p(\mathcal{C}_k)}{\sum_j p(\mathbf{x}|\mathcal{C}_j)p(\mathcal{C}_j)} = \frac{\exp(a_k)}{\sum_j \exp(a_j)}$$

⟶ 정규화된 지수 함수(normalized exponential function) = softmax function

logistic sigmoid를 여러 클래스에 대해 일반화한 형태

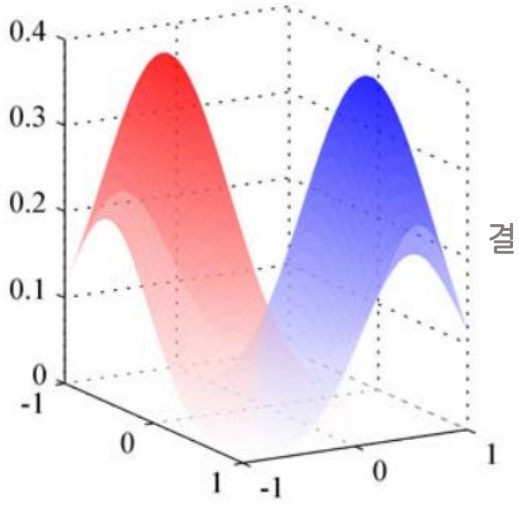
$$a_k = \ln p(\mathbf{x}|\mathcal{C}_k)p(\mathcal{C}_k)$$

$j \neq k, \quad a_k \gg a_j$

$$p(\mathcal{C}_k|\mathbf{x}) \simeq 1, \text{ and } p(\mathcal{C}_j|\mathbf{x}) \simeq 0.$$

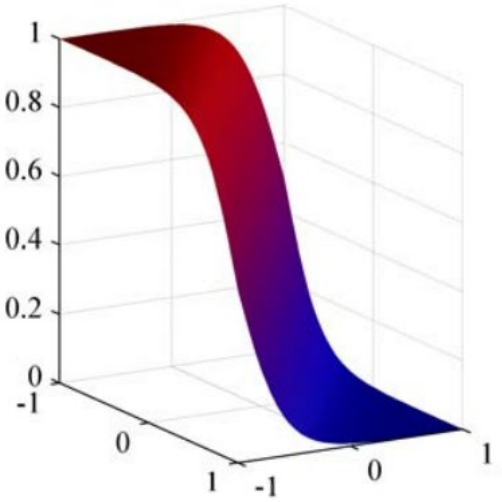
4.2.1 연속 입력

클래스별 조건부 밀도 → 가우시안
모든 클래스가 같은 공분산 행렬 공유



두 클래스들의 클래스 조건부 밀도

결정경계는 사후 확률
$p(C_k|\mathbf{x})$이 상수
$\mathbf{x}$의 선형 함수



$$p(\mathbf{x}|\mathcal{C}_k) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_k)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_k)\right\}.$$

$p(C_1|\mathbf{x})$: 사후 확률

$p(C_2|\mathbf{x})$: 사후 확률
$p(C_2|\mathbf{x}) = 1 - p(C_1|\mathbf{x})$

$p(C_k)$는 w_0 통해서만 연관 → 결정 경계 평행 이동

클래스 2개

$$p(\mathcal{C}_1|\mathbf{x}) = \sigma(\mathbf{w}^{\mathrm{T}}\mathbf{x} + w_0)$$

정의

$$\begin{aligned} \mathbf{w} &= \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \\ w_0 &= -\frac{1}{2}\boldsymbol{\mu}_1^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1 + \frac{1}{2}\boldsymbol{\mu}_2^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2 + \ln\frac{p(\mathcal{C}_1)}{p(\mathcal{C}_2)} \end{aligned}$$

4.2.1 연속 입력

클래스 K개 (K > 2)

$$a_k(\mathbf{x}) = \mathbf{w}_k^{\mathrm{T}}\mathbf{x} + w_{k0}$$

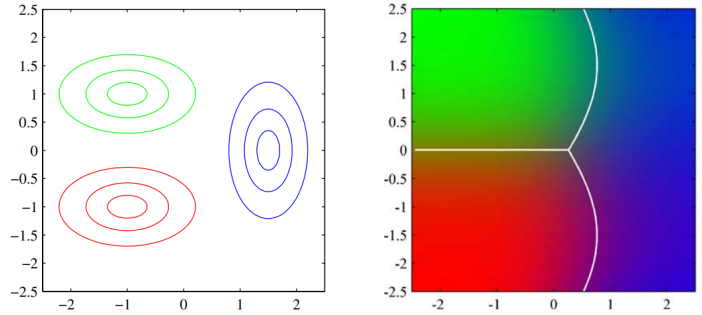
정의

$$\begin{aligned} \mathbf{w}_k &= \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k \\ w_{k0} &= -\frac{1}{2}\boldsymbol{\mu}_k^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k + \ln p(\mathcal{C}_k). \end{aligned}$$

클래스별 조건부 밀도 → 가우시안
~~모든 클래스가 같은 공분산 행렬 공유~~

$p(\mathbf{x}|\boldsymbol{C_k})$: 클래스별 조건부 밀도 ⟶ 각각 공분산 행렬 $\boldsymbol{\Sigma_k}$ 가진다

이차항 안 사라짐 → x의 이차 함수 : 이차 판별식(quadratic discriminant)



4.2.2 최대 가능도 해

"특정 분포인 경우"

$p(\mathbf{x}|\boldsymbol{C_k})$: 클래스별 조건부 밀도 → 매개변수적 함수

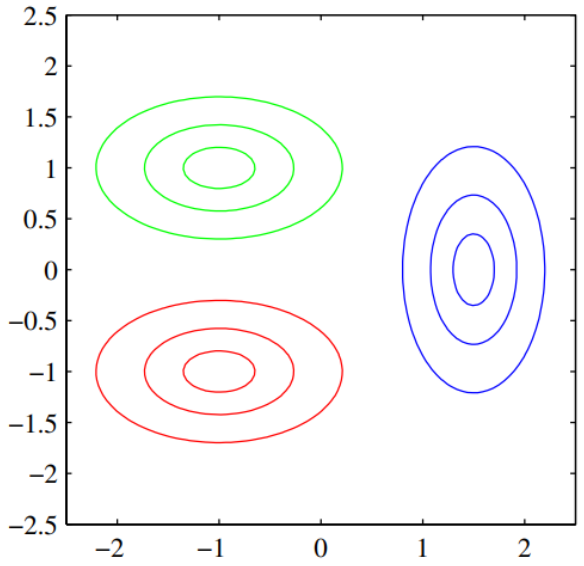
↓ 최대 가능도 방법 이용 ← 필요 — 관측값 x, 라벨링 데이터 집합

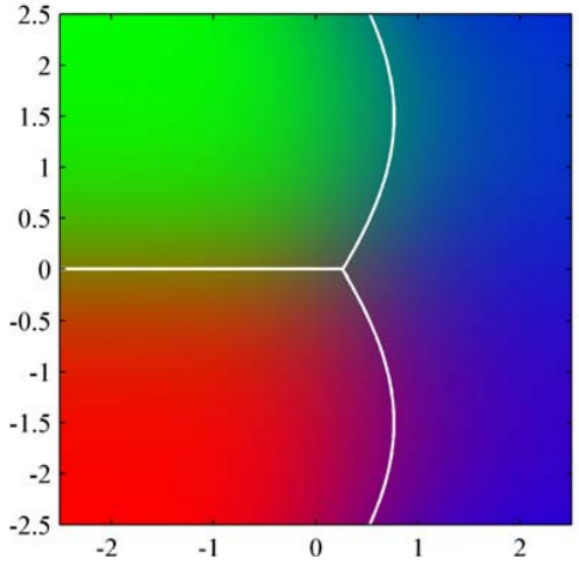
매개변수 값

$p(\boldsymbol{C_k})$: 사전 클래스 확률

4.2.2 최대 가능도 해

클래스별 조건부 밀도 → 가우시안
모든 클래스가 같은 공분산 행렬 공유





클래스 2개 $p(\boldsymbol{C_1})$: π , $p(\boldsymbol{C_2})$: 1 - π

C_1 $\boldsymbol{t_n} = 0$ $p(\mathbf{x}_n, \mathcal{C}_1) = p(\mathcal{C}_1)p(\mathbf{x}_n|\mathcal{C}_1) = \pi\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_1, \boldsymbol{\Sigma}).$

C_2 $\boldsymbol{t_n} = 1$ $p(\mathbf{x}_n, \mathcal{C}_2) = p(\mathcal{C}_2)p(\mathbf{x}_n|\mathcal{C}_2) = (1-\pi)\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_2, \boldsymbol{\Sigma}).$

가능도 함수 $$p(\mathbf{t}|\pi, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}) = \prod_{n=1}^{N} [\pi\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_1, \boldsymbol{\Sigma})]^{t_n} [(1-\pi)\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_2, \boldsymbol{\Sigma})]^{1-t_n}$$

$$\mathbf{t} = (t_1, \ldots, t_N)^{\mathrm{T}}$$

4.2.2 최대 가능도 해

C_1 데이터 포인트 수 : N_1
C_2 데이터 포인트 수 : N_2

클래스 2개

π에 대해 종속적인 로그 가능도 함수 항

$$\sum_{n=1}^{N} \{t_n \ln \pi + (1 - t_n) \ln(1 - \pi)\}$$

π에 대해 미분, 각 미분값을 0으로 설정

$$\pi = \frac{1}{N} \sum_{n=1}^{N} t_n = \frac{N_1}{N} = \frac{N_1}{N_1 + N_2}$$

μ들에 대해 같은 방식으로 해서 구하기

$$\boldsymbol{\mu}_1 = \frac{1}{N_1} \sum_{n=1}^{N} t_n \mathbf{x}_n$$

$$\boldsymbol{\mu}_2 = \frac{1}{N_2} \sum_{n=1}^{N} (1 - t_n) \mathbf{x}_n$$

μ 에 대해 종속적인 로그 가능도 함수 항

$$\sum_{n=1}^{N} t_n \ln \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_1, \boldsymbol{\Sigma}) = -\frac{1}{2} \sum_{n=1}^{N} t_n (\mathbf{x}_n - \boldsymbol{\mu}_1)^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{x}_n - \boldsymbol{\mu}_1) + \text{const.}$$

4.2.2 최대 가능도 해

공유된 공분산 행렬Σ에 대한 최대 가능도 해 고려

$$-\frac{1}{2}\sum_{n=1}^{N} t_n \ln|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{n=1}^{N} t_n (\mathbf{x}_n - \boldsymbol{\mu}_1)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}_n - \boldsymbol{\mu}_1)$$

$$-\frac{1}{2}\sum_{n=1}^{N} (1-t_n) \ln|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{n=1}^{N} (1-t_n)(\mathbf{x}_n - \boldsymbol{\mu}_2)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}_n - \boldsymbol{\mu}_2)$$

$$= -\frac{N}{2}\ln|\boldsymbol{\Sigma}| - \frac{N}{2}\mathrm{Tr}\left\{\boldsymbol{\Sigma}^{-1}\mathbf{S}\right\}$$

$$\begin{aligned}
\mathbf{S} &= \frac{N_1}{N}\mathbf{S}_1 + \frac{N_2}{N}\mathbf{S}_2 \\
\mathbf{S}_1 &= \frac{1}{N_1}\sum_{n\in\mathcal{C}_1}(\mathbf{x}_n - \boldsymbol{\mu}_1)(\mathbf{x}_n - \boldsymbol{\mu}_1)^{\mathrm{T}} \\
\mathbf{S}_2 &= \frac{1}{N_2}\sum_{n\in\mathcal{C}_2}(\mathbf{x}_n - \boldsymbol{\mu}_2)(\mathbf{x}_n - \boldsymbol{\mu}_2)^{\mathrm{T}}.
\end{aligned}$$

정의 이용

$$\boldsymbol{\Sigma} = \mathbf{S}$$

두 클래스 쌍의 해당하는 공분산 행렬들의 가중 평균

클래스 K개 (K > 2)

확장 가능, 하지만 이상점이 포함되어 있는 경우는 강건X

4.2.3 이산 특징

입력 : D
일반적인 분포 : 2^D
(D차원의 이진 입력을 고려할 때
가능한 모든 조합에 대한 클래스 확률)

합산 제약 조건 → 이 분포는 2^D - 1개의 독립 변수 가짐

$x_i \in \{0, 1\}$

특징의 수 제한을 위한
제한된 표현 방법 필요

나이브 베이즈 해 가정

각각의 값들이 클래스 C_k 에 대해 조
건부일 때 서로 독립적으로 취급

$$p(\mathbf{x}|\mathcal{C}_k) = \prod_{i=1}^{D} \mu_{ki}^{x_i}(1-\mu_{ki})^{1-x_i}$$ ← D개의 독립변수 가짐

$$a_k = \ln p(\mathbf{x}|\mathcal{C}_k)p(\mathcal{C}_k)$$

클래스 2개

로지스틱 시그모이드
함수 고려 가능

$$a_k(\mathbf{x}) = \sum_{i=1}^{D} \{x_i \ln \mu_{ki} + (1-x_i)\ln(1-\mu_{ki})\} + \ln p(\mathcal{C}_k)$$ ⟶ x_i 에 대한 선형 함수

4.2.4 지수족

각각의 클래스들이 각자의 매개변수 벡터 λ_k 를 가지지만, 척도 매개변수 s는 공유한다고 가정

모든 지수족 분포엔 고유 파라미터가 있음, 그래서 클래스별로 달라야 해서 λ_k 이용, 모든 클래스가 동일한 스케일 파라미터 가정

클래스 K = 2 로지스틱 시그모이드 함수

클래스 K개 (K >= 2) 소프트맥스 함수

→ 더 일반화

$$p(\mathbf{x}|\boldsymbol{\lambda}_k) = h(\mathbf{x})g(\boldsymbol{\lambda}_k)\exp\left\{\boldsymbol{\lambda}_k^{\mathrm{T}}\mathbf{u}(\mathbf{x})\right\}$$ 지수족 성질을 이용하여 x에 대한 분포 작성

u(x) = x인 부분 집합들에 대해서만 고려

$$p(\mathbf{x}|\boldsymbol{\lambda}_k, s) = \frac{1}{s}h\left(\frac{1}{s}\mathbf{x}\right)g(\boldsymbol{\lambda}_k)\exp\left\{\frac{1}{s}\boldsymbol{\lambda}_k^{\mathrm{T}}\mathbf{x}\right\}$$

←

$$p(x|\sigma) = \frac{1}{\sigma}f\left(\frac{x}{\sigma}\right)$$

척도 매개변수 s 도입

4.2.4 지수족

클래스 K = 2 ⟶ $a = \ln \frac{p(\mathbf{x}|\mathcal{C}_1)p(\mathcal{C}_1)}{p(\mathbf{x}|\mathcal{C}_2)p(\mathcal{C}_2)}$ 식 4.58 대입

$$a(\mathbf{x}) = (\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2)^{\mathrm{T}}\mathbf{x} + \ln g(\boldsymbol{\lambda}_1) - \ln g(\boldsymbol{\lambda}_2) + \ln p(\mathcal{C}_1) - \ln p(\mathcal{C}_2).$$

사후 클래스 확률이 선형 함수 a(x)에 대한 로지스틱 시그모이드 함수로 주어짐

클래스 K개 (K >= 2) ⟶ $a_k = \ln p(\mathbf{x}|\mathcal{C}_k)p(\mathcal{C}_k).$ 식 4.63 대입

$$a_k(\mathbf{x}) = \boldsymbol{\lambda}_k^{\mathrm{T}}\mathbf{x} + \ln g(\boldsymbol{\lambda}_k) + \ln p(\mathcal{C}_k)$$

x에 대한 선형 함수

4.3 확률적 판별 모델

생성적 모델링 4.2 내용

$p(\mathbf{x}|C_k)$: 클래스별 조건부 밀도

$p(C_k)$: 사전 클래스 확률

→ $p(C_k|\mathbf{x})$: 사후 확률 결정(간접적으로 파라미터 결정)

판별적 모델링 <반복 재가중 최소 제곱법 = IRLS 알고리즘>
<iterative reweighted least squares>

일반화된 선형 모델 함수 형태를 명시적으로 사용

최대 가능도 ↓

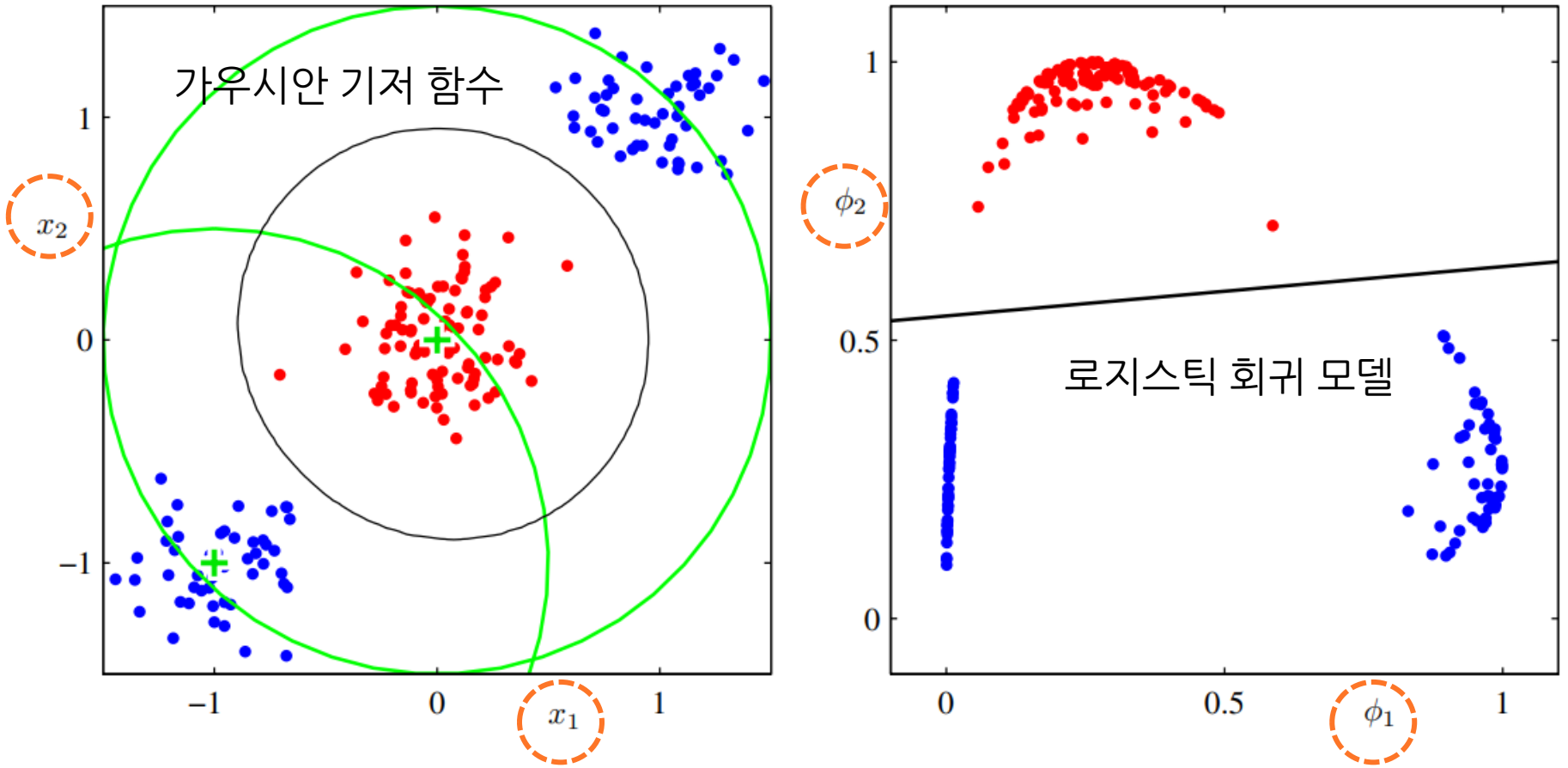
직접적으로 매개변수 값 얻기
(생성적 모델보에 비해 구해야 할 적응
매개변수 숫자가 적음.)

← $p(C_k|\mathbf{x})$ 정의

MLE 활용

생성적 모델링이 잘 안 맞을 때 쓰면 좋은 결과 얻기 가능

4.3.1 고정된 기저 함수



입력 벡터를 φ_x 로 변환하는 상태 이용할 것

4.3.1 고정된 기저 함수

실제 데이터

$p(\mathbf{x}|\boldsymbol{C_k})$: 클래스별 조건부 밀도 ← 각각의 클래스 사이에서 중첩되기 쉬움

↓

각각의 값이 0, 1과 같이 명확하게 떨어지지 않음

↓

1장) 표준 결정 이론 적용
→ 최적의 해 구하기 가능

한계점

비선형 변환 φ(x)는 클래스 중첩 제거 잘 못 함.

↓

클래스 중첩 증가
원래 관측 공간에서 없던 새로운 중첩 유발

그럼에도 불구하고 적절한 비선형성 선택 시 사후 확률 모델링 과정 쉬워짐

4.3.1 고정된 기저 함수

한계점

기저 함수 자체가 데이터에 대해 적응되도록 만들면 해결 가능

1. 비선형성과 복잡한 패턴의 제한 (데이터가 복잡한 비선형 패턴 잘 포착X)
2. 기저 함수 선택의 어려움 (데이터에 대한 깊은 이해도 필요)
3. 고차원 데이터에서의 한계 (차원의 저주, 효율성 저하)
4. 유연성의 부족 (적응성 결여, 기저 함수 동적 조절 불가)
5. 과적합 문제 (기저 함수 너무 복잡하다면)

4.3.2 로지스틱 회귀

4.2 내용

클래스 K = 2 C_1 에 대한 사후 확률 :
특정 벡터의 선형 함수에 대한
로지스틱 시그모이드 함수

$$\sigma(a) = \frac{1}{1 + \exp(-a)}$$

클래스 K = 2 일반화된 선형 모델

M차원 특정 공간 φ

$$p(\mathcal{C}_1|\boldsymbol{\phi}) = y(\boldsymbol{\phi}) = \sigma\left(\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}\right)$$ "로지스틱 회귀"

$$p(\mathcal{C}_2|\boldsymbol{\phi}) = 1 - p(\mathcal{C}_1|\boldsymbol{\phi})$$

매개변수 : M개

"가우시안 클래스 조건부 밀도"

매개변수(평균값) : 2M개
매개변수(공분산 행렬) : M(M+5)/2 +1개

매개변수 개수 M이 큰 경우에는 로지스틱 회귀 모델을 다루는 것 이 더 유리할 수 있음

4.3.2 로지스틱 회귀

로지스틱 회귀 모델 매개변수 구하기 ← 최대 가능도 이용

$$p(\mathcal{C}_1|\boldsymbol{\phi}) = y(\boldsymbol{\phi}) = \sigma\left(\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}\right)$$

$$\frac{d\sigma}{da} = \sigma(1-\sigma).$$

a data set $\{\boldsymbol{\phi}_n, t_n\}$

$\boldsymbol{\phi}_n = \boldsymbol{\phi}(\mathbf{x}_n)$

$t_n \in \{0, 1\}$, n = 1,...,N

가능도 함수

$$p(\mathbf{t}|\mathbf{w}) = \prod_{n=1}^{N} y_n^{t_n}\{1-y_n\}^{1-t_n}$$

$\mathbf{t} = (t_1, \ldots, t_N)^{\mathrm{T}}$

$y_n = p(\mathcal{C}_1|\boldsymbol{\phi}_n)$

음의 로그값 ↓

교차 엔트로피 (cross entropy)

$$E(\mathbf{w}) = -\ln p(\mathbf{t}|\mathbf{w}) = -\sum_{n=1}^{N}\{t_n \ln y_n + (1-t_n)\ln(1-y_n)\}$$

4.3.2 로지스틱 회귀

교차 엔트로피
(cross entropy)

$$E(\mathbf{w}) = -\ln p(\mathbf{t}|\mathbf{w}) = -\sum_{n=1}^{N}\{t_n \ln y_n + (1-t_n)\ln(1-y_n)\}$$

$$\frac{d\sigma}{da} = \sigma(1-\sigma).$$

이용

$$y_n = \sigma(a_n)$$

$$a_n = \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}_n$$

w에 대하여 오류 함수 계산

$$\nabla E(\mathbf{w}) = \sum_{n=1}^{N}(y_n - t_n)\boldsymbol{\phi}_n$$

오류

$$\nabla \ln p(\mathbf{t}|\mathbf{w},\beta) = \sum_{n=1}^{N}\{t_n - \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n)\}\boldsymbol{\phi}(\mathbf{x}_n)^{\mathrm{T}}.$$

선형 회귀 모델의 제곱합 오류 함수의 기울기

$$\mathbf{w}^{(\tau+1)} = \mathbf{w}^{(\tau)} - \eta\nabla E_n$$

가중치 벡터 업데이트 식

4.3.2 로지스틱 회귀

선형 분리 가능한 데이터 집합에 최대 가능도 방법 사용
→ 심각한 과적합 문제

헤비사이드 계단 함수, 가각의 클래스 K에서 온 모든 훈련 포인트들이 사후 확률 1을 가지게 됨

최대 가능도 방법을 통해서 어떤 하나의 해를 다른 해보다 선호하게 할 수 없음. 해를 찾는 건 최적화 알고리즘과 매개변수 초기화에 달림

→ 사전 확률 포함, w에 대해 MAP해를 찾는 방식, 오류함수에 정규화항 추가하는 방식으로 해결 가능

4.3.3 반복 재가중 최소 제곱법

가우시안 노이즈 모델 – 최대 가능도 해 : 닫힌 형태
why? 로그 가능도 함수가 매개변수 벡터 **w**에 대해 이차 종속성 가짐

로지스틱 회귀 모델

닫힌 해X

로지스틱 시그모이드 함수의 비선형성

이차식 형태 → 오류 함수 Convex → 유일한 최솟값

<뉴턴 라프슨(Newton-Raphson) >

로그 가능도 함수에 대한 지역적인 이차식 근삿값을 구하는 방식

$$\mathbf{w}^{(\text{new})} = \mathbf{w}^{(\text{old})} - \mathbf{H}^{-1}\nabla E(\mathbf{w})$$

헤시안 행렬 최소화 대상

$\mathbf{H} = \nabla\nabla E(\mathbf{w})$ 각 원소 : $E(\mathbf{w})$를 **w**의 각 성분으로 이차 미분한 값

4.3.3 반복 재가중 최소 제곱법

$$y(\mathbf{x}, \mathbf{w}) = \sum_{j=0}^{M-1} w_j \phi_j(\mathbf{x}) = \mathbf{w}^T \phi(\mathbf{x}) \qquad (3.3)$$

$$E_D(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^{N} \{t_n - \mathbf{w}^T \phi(\mathbf{x}_n)\}^2 \qquad (3.12)$$

<뉴턴 라프슨 적용>

$$\nabla E(\mathbf{w}) = \sum_{n=1}^{N} (\mathbf{w}^{\mathrm{T}} \boldsymbol{\phi}_n - t_n) \boldsymbol{\phi}_n = \boldsymbol{\Phi}^{\mathrm{T}} \boldsymbol{\Phi} \mathbf{w} - \boldsymbol{\Phi}^{\mathrm{T}} \mathbf{t}$$

$$\mathbf{H} = \nabla \nabla E(\mathbf{w}) = \sum_{n=1}^{N} \boldsymbol{\phi}_n \boldsymbol{\phi}_n^{\mathrm{T}} = \boldsymbol{\Phi}^{\mathrm{T}} \boldsymbol{\Phi}$$

Φ is the N × M design matrix

$$\mathbf{w}^{(\mathrm{new})} = \mathbf{w}^{(\mathrm{old})} - (\boldsymbol{\Phi}^{\mathrm{T}} \boldsymbol{\Phi})^{-1} \left\{ \boldsymbol{\Phi}^{\mathrm{T}} \boldsymbol{\Phi} \mathbf{w}^{(\mathrm{old})} - \boldsymbol{\Phi}^{\mathrm{T}} \mathbf{t} \right\}$$

$$= (\boldsymbol{\Phi}^{\mathrm{T}} \boldsymbol{\Phi})^{-1} \boldsymbol{\Phi}^{\mathrm{T}} \mathbf{t}$$

이 결괏값은 표준 최소 제곱 해
오류 함수 : 이차식
뉴턴 라프슨 공식 한 단계 만에 정확한 해 구해냄

4.3.3 반복 재가중 최소 제곱법

$$\frac{d\sigma}{da} = \sigma(1-\sigma).$$

이용

$$E(\mathbf{w}) = -\ln p(\mathbf{t}|\mathbf{w}) = -\sum_{n=1}^{N}\{t_n \ln y_n + (1-t_n)\ln(1-y_n)\}$$ 교차 엔트로피 함수

$$\nabla E(\mathbf{w}) = \sum_{n=1}^{N}(y_n - t_n)\boldsymbol{\phi}_n = \boldsymbol{\Phi}^{\mathrm{T}}(\mathbf{y}-\mathbf{t})$$

오류 함수 기울기

$$\mathbf{H} = \nabla\nabla E(\mathbf{w}) = \sum_{n=1}^{N} y_n(1-y_n)\boldsymbol{\phi}_n\boldsymbol{\phi}_n^{\mathrm{T}} = \boldsymbol{\Phi}^{\mathrm{T}}\mathbf{R}\boldsymbol{\Phi}$$

헤시안

$$R_{nn} = y_n(1-y_n).$$

$N \times N$: 대각행렬 $\mathbf{R}$

헤시안 – 상수 ×

가중 행렬 **R**을 통해 **w**에 종속성 가짐

→ 오류 함수가 quadratic 아님

시그모이드 출력 범위 $0 < y_n < 1$

$$\mathbf{u}^{\mathrm{T}}\mathbf{H}\mathbf{u} > 0$$

H는 양의 정부호 행렬

→ 오류 함수 w에 대한 convex, 유일한 최솟값 가짐

4.3.3 반복 재가중 최소 제곱법

로지스틱 회귀 모델에 대한 뉴턴 라프슨 업데이트

$$\begin{aligned}\mathbf{w}^{(\text{new})} &= \mathbf{w}^{(\text{old})} - (\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{R}\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^{\mathrm{T}}(\mathbf{y}-\mathbf{t}) \\ &= (\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{R}\boldsymbol{\Phi})^{-1}\left\{\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{R}\boldsymbol{\Phi}\mathbf{w}^{(\text{old})} - \boldsymbol{\Phi}^{\mathrm{T}}(\mathbf{y}-\mathbf{t})\right\} \\ &= (\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{R}\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{R}\mathbf{z}\end{aligned}$$

$$\mathbf{z} = \boldsymbol{\Phi}\mathbf{w}^{(\text{old})} - \mathbf{R}^{-1}(\mathbf{y}-\mathbf{t}).$$

Z는 N차원 벡터

→ 가중된 최소 제곱 문제의 정규 방정식의 집합 형태

행렬 R이 상수가 아니고 벡터 w에 종속적 이기에
정규 방정식들을 '반복적'으로 적용해야 함.
매 반복마다 새 가중 벡터 w를 이용하여 수정된 가중 행렬 R을 구해야 함.

4.3.3 반복 재가중 최소 제곱법 = IRLS 알고리즘

대각 가중 행렬 R의 원소 → 분산으로 해석

↓(로지스틱 회귀 모델에서의 t의 평균과 분산)

$$\begin{aligned}\mathbb{E}[t] &= \sigma(\mathbf{x}) = y \\ \mathrm{var}[t] &= \mathbb{E}[t^2] - \mathbb{E}[t]^2 = \sigma(\mathbf{x}) - \sigma(\mathbf{x})^2 = y(1-y)\end{aligned}$$

$t \in \{0, 1\}$

$t = t^2$

IRLS를 $a = \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}.$

공간상에서 선형된 문제의 해로 해석

$$\begin{aligned}a_n(\mathbf{w}) &\simeq a_n(\mathbf{w}^{(\mathrm{old})}) + \left.\frac{\mathrm{d}a_n}{\mathrm{d}y_n}\right|_{\mathbf{w}^{(\mathrm{old})}} (t_n - y_n) \\ &= \boldsymbol{\phi}_n^{\mathrm{T}}\mathbf{w}^{(\mathrm{old})} - \frac{(y_n - t_n)}{y_n(1-y_n)} = z_n.\end{aligned}$$

4.3.4 다중 클래스 로지스틱 회귀

생성적 모델링 4.2 내용

$p(\mathbf{x}|\boldsymbol{C_k})$: 클래스별 조건부 밀도

$p(\boldsymbol{C_k})$: 사전 클래스 확률

→ $p(\boldsymbol{C_k}|\mathbf{x})$: 사후 확률 결정(간접적으로 파라미터 결정)

$$p(\mathcal{C}_k|\boldsymbol{\phi}) = y_k(\boldsymbol{\phi}) = \frac{\exp(a_k)}{\sum_j \exp(a_j)} \quad \Big| \quad a_k = \mathbf{w}_k^{\mathrm{T}}\boldsymbol{\phi}.$$

$\boldsymbol{w_k}$ ← 최대 가능도 방법 활용하여 직업 구하기

$$\frac{\partial y_k}{\partial a_j} = y_k(I_{kj} - y_j)$$

항등 행렬의 원소

4.3.4 다중 클래스 로지스틱 회귀

가능도 함수(원 핫 인코딩 사용해서 표현)

$$p(\mathbf{T}|\mathbf{w}_1,\ldots,\mathbf{w}_K) = \prod_{n=1}^{N}\prod_{k=1}^{K} p(\mathcal{C}_k|\boldsymbol{\phi}_n)^{t_{nk}} = \prod_{n=1}^{N}\prod_{k=1}^{K} y_{nk}^{t_{nk}} \quad \Big| \quad y_{nk} = y_k(\boldsymbol{\phi}_n)$$

음의 로그값

$$E(\mathbf{w}_1,\ldots,\mathbf{w}_K) = -\ln p(\mathbf{T}|\mathbf{w}_1,\ldots,\mathbf{w}_K) = -\sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk}\ln y_{nk}$$

N × K 행렬, 원소 t_{nk}

다중 클래스 분류 문제상에서의 교차 엔트로피(cross entropy) 오류 함수

4.3.4 다중 클래스 로지스틱 회귀

$$\nabla_{\mathbf{w}_j} E(\mathbf{w}_1, \ldots, \mathbf{w}_K) = \sum_{n=1}^{N} (y_{nj} - t_{nj})\, \boldsymbol{\phi}_n$$

기저함수 (기저함수 $\boldsymbol{\phi}_n$)

오룻값 ($y_{nj} - t_{nj}$)

이용

$$\frac{\partial y_k}{\partial a_j} = y_k(I_{kj} - y_j)$$

$$\sum_k t_{nk} = 1$$

소프트맥스의 특징

선형 모델에 제곱합 오류 함수를 사용했을 경우의 기울기 함수 형태가 같음.

→ 순차적 알고리즘 만들 수 있음.

교차 엔트로피 오류 함수 $\nabla E(\mathbf{w}) = \sum_{n=1}^{N} (y_n - t_n)\boldsymbol{\phi}_n$

소프트맥스 활성화 함수, 다중 클래스 식 모두 같은 형태

$$\nabla \ln p(\mathbf{t}|\mathbf{w}, \beta) = \sum_{n=1}^{N} \{t_n - \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n)\}\, \boldsymbol{\phi}(\mathbf{x}_n)^{\mathrm{T}}.$$

선형 회귀 모델의 제곱합 오류 함수의 기울기

$$\mathbf{w}^{(\tau+1)} = \mathbf{w}^{(\tau)} - \eta \nabla E_n$$

가중치 벡터 업데이트 식

4.3.4 다중 클래스 로지스틱 회귀

Batch 알고리즘 찾기

→ 뉴턴 라프슨 업데이트 적용
다중 클래스상에서 IRLS 알고리즘 구하기

계산 $$\nabla_{\mathbf{w}_k}\nabla_{\mathbf{w}_j}E(\mathbf{w}_1,\ldots,\mathbf{w}_K) = -\sum_{n=1}^{N} y_{nk}(I_{kj} - y_{nj})\boldsymbol{\phi}_n\boldsymbol{\phi}_n^{\mathrm{T}}.$$

다중 클래스 로지스틱 회귀 모델(이중 클래스와 마찬가지로)

H는 양의 정부호 행렬

→ 오류 함수 w에 대한 convex, 유일한 최솟값 가짐

4.3.5 프로빗 회귀

4.2.4 지수족

클래스 K = 2 ⟶ $a = \ln \frac{p(\mathbf{x}|\mathcal{C}_1)p(\mathcal{C}_1)}{p(\mathbf{x}|\mathcal{C}_2)p(\mathcal{C}_2)}$ 식 4.58 대입

$$a(\mathbf{x}) = (\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2)^{\mathrm{T}}\mathbf{x} + \ln g(\boldsymbol{\lambda}_1) - \ln g(\boldsymbol{\lambda}_2) + \ln p(\mathcal{C}_1) - \ln p(\mathcal{C}_2).$$

사후 클래스 확률이 선형 함수 a(x)에 대한 로지스틱 시그모이드 함수로 주어짐

클래스 K개 (K >= 2) ⟶ $a_k = \ln p(\mathbf{x}|\mathcal{C}_k)p(\mathcal{C}_k)$. 식 4.63 대입

$$a_k(\mathbf{x}) = \boldsymbol{\lambda}_k^{\mathrm{T}}\mathbf{x} + \ln g(\boldsymbol{\lambda}_k) + \ln p(\mathcal{C}_k)$$

x에 대한 선형 함수

이런 단순한 형태의 사후 확률이 모든 종류의 클래스 조건부 밀도 분포에 대해서 결과로 나오는 것은 아님.

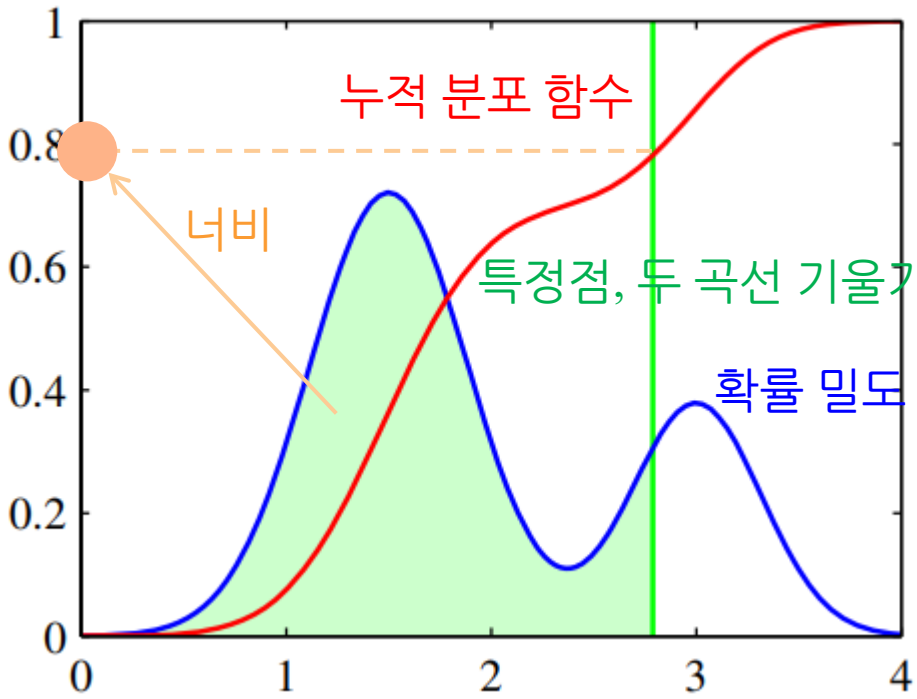
4.3.5 프로빗 회귀

$p(t=1|a) = f(a)$ 일반화된 선형 모델의 틀

$a = \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}$ f는 활성 함수

대안적인 연결 함수 :노이즈 임계값 모델

$$\begin{cases} t_n = 1 & \text{if } a_n \geqslant \theta \\ t_n = 0 & \text{otherwise.} \end{cases}$$



$$f(a) = \int_{-\infty}^{a} p(\theta)\,\mathrm{d}\theta$$

확률 밀도

대안적인 연결 함수

일반화 선형 모델에서 사용되며, 선형 예측값을 종속 변수의 기대값과 연결합니다. 주로 확률적 성격을 띠는 모델에서 사용되며, 특정 확률 분포와 관련이 있습니다.

4.3.5 프로빗 회귀

$$f(a) = \int_{-\infty}^{a} p(\theta)\,\mathrm{d}\theta$$

확률 밀도

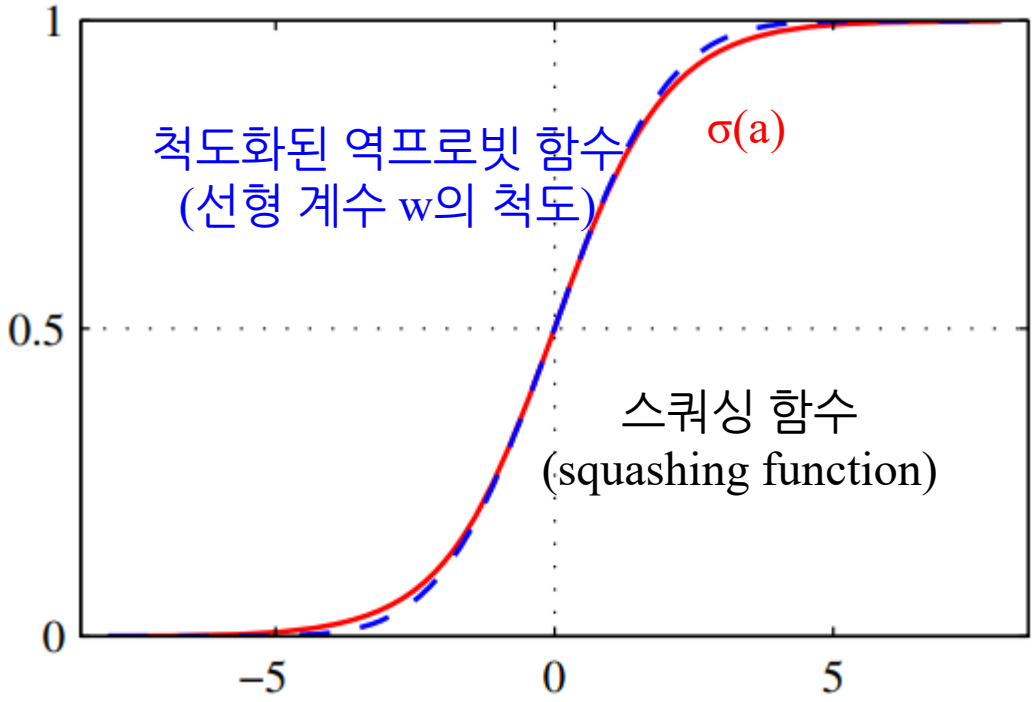
가우시안 분포 가정:평균 0, 분산 1

$$\Phi(a) = \int_{-\infty}^{a} \mathcal{N}(\theta|0,1)\,\mathrm{d}\theta$$

역프로빗(inverse probit)

오차 함수(오류 함수 아님)

$$\mathrm{erf}(a) = \frac{2}{\sqrt{\pi}} \int_{0}^{a} \exp(-\theta^2/2)\,\mathrm{d}\theta$$



역프로빗과 오차 함수 관계

$$\mathbf{\Phi}(a) = \frac{1}{2}\left\{1 + \frac{1}{\sqrt{2}}\mathrm{erf}(a)\right\}$$

역프로빗 활성화 함수를 바탕으로 한 일반화된 선형 모델 : 프로빗 회귀

4.3.5 프로빗 회귀

실제 응용 사례 이슈

“이상값(outlier)”

입력 벡터 x 측정 시 오류
표적 벡터 t 잘못 라벨링

이상적인 결정 경계면을 기준으로 잘못된 방향에 놓임
→ 분류기를 심각하게 왜곡할 수 있음.

$x \rightarrow \infty$

반응

로지스틱 회귀 모델 — 점근적으로 $\exp(-x)$와 같이 감소

프로빗 회귀 모델 — 점근적으로 $\exp(-x^2)$와 같이 감소

더 예민하게 반응

4.3.5 프로빗 회귀

데이터가 올바르게 라벨링 되었다고 가정

표적값 t가 잘못된 값을 부여받았을 확률 ϵ

$$
\begin{aligned}
p(t|\mathbf{x}) &= (1-\epsilon)\sigma(\mathbf{x}) + \epsilon(1-\sigma(\mathbf{x})) \\
&= \epsilon + (1-2\epsilon)\sigma(\mathbf{x})
\end{aligned}
$$

← 잘못된 라벨링의 효과를 확률적 모델에 적용

$\sigma(\mathbf{x})$ 활성화 함수

ϵ : 미리 정해두거나 데이터로부터 유추할 수 있는 초매개변수처럼 다루기 가능

4.3.6 정준 연결 함수

$$\nabla_{\mathbf{w}_j} E(\mathbf{w}_1, \ldots, \mathbf{w}_K) = \sum_{n=1}^{N} (y_{nj} - t_{nj})\, \phi_n$$

기저함수 / 오릇값

이용

$$\frac{\partial y_k}{\partial a_j} = y_k(I_{kj} - y_j)$$

$$\sum_k t_{nk} = 1$$

소프트맥스의 특징

선형 모델에 제곱합 오류 함수를 사용했을 경우의 기울기 함수 형태가 같음.

→ 순차적 알고리즘 만들 수 있음.

교차 엔트로피 오류 함수 $\nabla E(\mathbf{w}) = \sum_{n=1}^{N} (y_n - t_n)\phi_n$

소프트맥스 활성화 함수, 다중 클래스 식 모두 같은 형태

$$\nabla \ln p(\mathbf{t}|\mathbf{w}, \beta) = \sum_{n=1}^{N} \{t_n - \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n)\}\, \boldsymbol{\phi}(\mathbf{x}_n)^{\mathrm{T}}.$$

선형 회귀 모델의 제곱합 오류 함수의 기울기

$$\mathbf{w}^{(\tau+1)} = \mathbf{w}^{(\tau)} - \eta \nabla E_n$$

가중치 벡터 업데이트 식

타깃 변수의 조건부 분포가 지수족에 포함되는 경우에 일반적으로 얻을 수 있다는 증명 부분

이때의 활성화 함수

정준 연결 함수 (canonical link function)

4.3.6 정준 연결 함수

$$p(\mathbf{x}|\boldsymbol{\lambda}_k) = h(\mathbf{x})g(\boldsymbol{\lambda}_k)\exp\left\{\boldsymbol{\lambda}_k^{\mathrm{T}}\mathbf{u}(\mathbf{x})\right\}$$ 지수족 성질을 이용하여 x에 대한 분포 작성

u(x) = x인 부분 집합들에 대해서만 고려

$$p(\mathbf{x}|\boldsymbol{\lambda}_k, s) = \frac{1}{s}h\left(\frac{1}{s}\mathbf{x}\right)g(\boldsymbol{\lambda}_k)\exp\left\{\frac{1}{s}\boldsymbol{\lambda}_k^{\mathrm{T}}\mathbf{x}\right\}$$ ← $$p(x|\sigma) = \frac{1}{\sigma}f\left(\frac{x}{\sigma}\right)$$

척도 매개변수 s 도입

이번에는 타깃 변수 t를 지수족 분포로 가정

$$p(t|\eta, s) = \frac{1}{s}h\left(\frac{t}{s}\right)g(\eta)\exp\left\{\frac{\eta t}{s}\right\}.$$

y로 표현되는 t의 조건부 평균

$$y \equiv \mathbb{E}[t|\eta] = -s\frac{d}{d\eta}\ln g(\eta).$$

y와 η는 반드시 연관성이 있어야 함

$$\eta = \psi(y).$$

4.3.6 정준 연결 함수

머신 러닝 문헌 : f는 활성화 함수
통계학 문헌 : f의 역함수는 연결 함수

가정 : 모든 관측값은 동일한 척도 매개변수 공유
(가우시안 분포는 노이즈의 분산) → s는 n에 대해 독립적

Nelder and Wedderburn에 따르면
일반화된 선형 모델은 y가 입력(또는 특징)의 선형 결합을 비선형 함수에 넣은 결과로 표현되는 모델이라 정의

$$y = f(\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi})$$

η의 함수로 표현되는 이 모델의 로그 기능도 함수

$$\ln p(\mathbf{t}|\eta, s) = \sum_{n=1}^{N} \ln p(t_n|\eta, s) = \sum_{n=1}^{N} \left\{ \ln g(\eta_n) + \frac{\eta_n t_n}{s} \right\} + \text{const}$$

↓ 모델 매개변수 w에 대해 미분

$$\begin{aligned} \nabla_{\mathbf{w}} \ln p(\mathbf{t}|\eta, s) &= \sum_{n=1}^{N} \left\{ \frac{d}{d\eta_n} \ln g(\eta_n) + \frac{t_n}{s} \right\} \frac{d\eta_n}{dy_n} \frac{dy_n}{da_n} \nabla a_n \\ &= \sum_{n=1}^{N} \frac{1}{s} \{t_n - y_n\} \psi'(y_n) f'(a_n) \boldsymbol{\phi}_n \end{aligned}$$

$$a_n = \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}_n$$

$$y \equiv \mathbb{E}[t|\eta] = -s\frac{d}{d\eta} \ln g(\eta).$$

$$y_n = f(a_n)$$

$$f^{-1}(y) = \psi(y)$$

$$f(\psi(y)) = y$$

$$f'(\psi)\psi'(y) = 1$$

$$\nabla \ln E(\mathbf{w}) = \frac{1}{s} \sum_{n=1}^{N} \{y_n - t_n\} \boldsymbol{\phi}_n.$$

오류 함수의 기울기

4.4 라플라스 근사

목적 : 연속 변수의 집합에 대해 정의된 확률 밀도의 가우시안 근사치 찾기

$$p(z) = \frac{1}{Z} f(z)$$ z는 단일 연속 변수

정규화 계수 $Z = \int f(z)\,\mathrm{d}z$

"분포 p(z)의 최빈값을 중심으로 한 가우시안 근사 q(z) 찾기"

1) p(z)의 최빈값 구하기

$\mathbf{p'}(z_0) = 0$이 되는 z_0값 찾기

$$\left.\frac{df(z)}{dz}\right|_{z=z_0} = 0.$$

4.4 라플라스 근사

$$\ln f(z) \simeq \ln f(z_0) - \frac{1}{2}A(z - z_0)^2$$

$$A = -\left.\frac{d^2}{dz^2}\ln f(z)\right|_{z=z_0}.$$

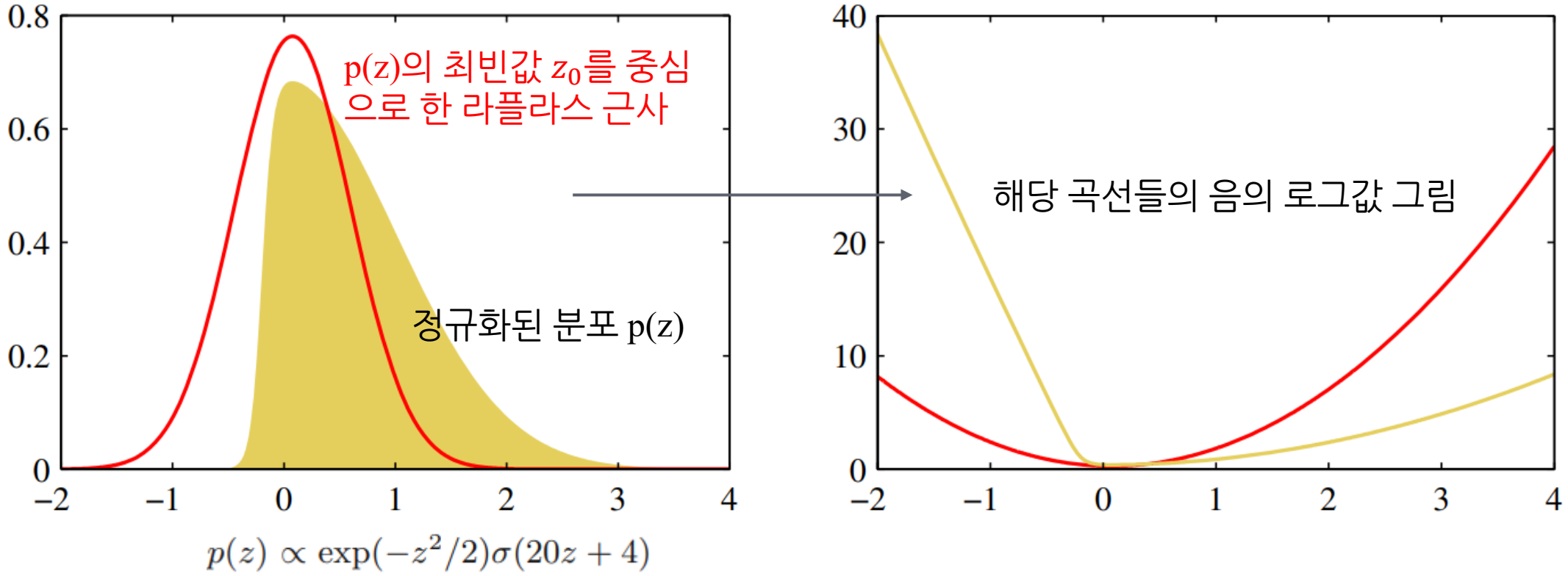
z_0가 분포의 최댓값 → 테일러 전개 일차항 안 나타남

$$f(z) \simeq f(z_0)\exp\left\{-\frac{A}{2}(z - z_0)^2\right\}$$

정규화된 가우시안 분포에 대한 표준 방법 이용
→ 정규화된 분포 q(z)를 구할 수 있음.

$$q(z) = \left(\frac{A}{2\pi}\right)^{1/2}\exp\left\{-\frac{A}{2}(z - z_0)^2\right\}.$$

4.4 라플라스 근사



분포의 정밀도 A가 0보다 큰 경우에만 잘 정의
(임계점 z_0가 지역적 최댓값, f(z)의 z_0에서의 2차 미분값이 음수여야함

4.4 라플라스 근사

m차원 공간 z에 대해 정의된 분포
p(z) = f(z) / Z에 대해서 라플라스 근사 확장

$$\nabla f(\mathbf{z})$$

임계점에서 기울기 0
→ 임계점 근처에 대해서 전개 시행

$$\ln f(\mathbf{z}) \simeq \ln f(\mathbf{z}_0) - \frac{1}{2}(\mathbf{z}-\mathbf{z}_0)^{\mathrm{T}}\mathbf{A}(\mathbf{z}-\mathbf{z}_0)$$

$$\mathbf{A} = -\nabla\nabla \ln f(\mathbf{z})|_{\mathbf{z}=\mathbf{z}_0}$$

M × M 헤시안 행렬 A의 정의

∇는 기울기 연산자, 양쪽 변에 대해 지수 함수 취하기

$$f(\mathbf{z}) \simeq f(\mathbf{z}_0)\exp\left\{-\frac{1}{2}(\mathbf{z}-\mathbf{z}_0)^{\mathrm{T}}\mathbf{A}(\mathbf{z}-\mathbf{z}_0)\right\}.$$

분포p(z)는 f(z)에 비례, 적절한 정규화 계수는 다변량 가우시안에 대한 표준 결과를 이용해서 구하기 가능

$$q(\mathbf{z}) = \frac{|\mathbf{A}|^{1/2}}{(2\pi)^{M/2}}\exp\left\{-\frac{1}{2}(\mathbf{z}-\mathbf{z}_0)^{\mathrm{T}}\mathbf{A}(\mathbf{z}-\mathbf{z}_0)\right\} = \mathcal{N}(\mathbf{z}|\mathbf{z}_0, \mathbf{A}^{-1})$$

4.4.1 모델 비교와 베이지안 정보 기준

P(z) 말고 Z도 근사치 구할 수 있음

$$
\begin{aligned}
Z &= \int f(\mathbf{z})\,\mathrm{d}\mathbf{z} \\
&\simeq f(\mathbf{z}_0)\int \exp\left\{-\frac{1}{2}(\mathbf{z}-\mathbf{z}_0)^{\mathrm{T}}\mathbf{A}(\mathbf{z}-\mathbf{z}_0)\right\}\mathrm{d}\mathbf{z} \\
&= f(\mathbf{z}_0)\frac{(2\pi)^{M/2}}{|\mathbf{A}|^{1/2}}
\end{aligned}
$$

$$
p(\mathcal{D}) = \int p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta}.
$$

$$
f(\boldsymbol{\theta}) = p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta}) \text{ and } Z = p(\mathcal{D})
$$

$$
\ln p(\mathcal{D}) \simeq \ln p(\mathcal{D}|\boldsymbol{\theta}_{\mathrm{MAP}}) + \underbrace{\ln p(\boldsymbol{\theta}_{\mathrm{MAP}}) + \frac{M}{2}\ln(2\pi) - \frac{1}{2}\ln|\mathbf{A}|}_{\text{Occam factor}}
$$

$$
\mathbf{A} = -\nabla\nabla \ln p(\mathcal{D}|\boldsymbol{\theta}_{\mathrm{MAP}})p(\boldsymbol{\theta}_{\mathrm{MAP}}) = -\nabla\nabla \ln p(\boldsymbol{\theta}_{\mathrm{MAP}}|\mathcal{D}).
$$

4.4.1 모델 비교와 베이지안 정보 기준

$$\ln p(\mathcal{D}) \simeq \ln p(\mathcal{D}|\boldsymbol{\theta}_{\mathrm{MAP}}) + \underbrace{\ln p(\boldsymbol{\theta}_{\mathrm{MAP}}) + \frac{M}{2}\ln(2\pi) - \frac{1}{2}\ln|\mathbf{A}|}_{\text{Occam factor}}$$

매개변수에 대한 가우시안 사전 분포가 넓게 퍼졌고,
헤시안 행렬이 최대 계수를 가진다 가정

$$\ln p(\mathcal{D}) \simeq \ln p(\mathcal{D}|\boldsymbol{\theta}_{\mathrm{MAP}}) - \frac{1}{2}M\ln N$$

베이지안 정보 기준(BIC) = 슈바르츠 기준

아카이케 정보량 기준(AIC)과 비교했을 때
베이지안 정보 기준은 모델의 복잡도에 대해 더 큰 불이익을 줌.

4.4.1 모델 비교와 베이지안 정보 기준

복잡도를 측정하는 단위는 측정이 쉽다는 장점이 있지만

잘못된 결과도 끌어낼 수 있음.

Ex) 헤시안 행렬이 완전 행렬 계수를 가진다는 가정은
대부분 사실이 아님 ← 매개변수들이 잘 확정되지 않기 때문

4.5 베이지안 로지스틱 회귀

로지스틱 회귀의 정확한 베이지안 추론은 다루기 아주 어려움

사후 분포 계산 ← 정규화 (사전 분포 * 가능도 함수)

각 데이터 포인트마다
로지스틱 시그모이드
함수를 모두 곱한 값

예측 분포 계산 ← 사후 분포와 비슷한 이유로 다루기 어려움.

→ 베이지안 로지스틱 회귀 문제에 라플라스 근사를 적용할 것

4.5.1 라플라스 근사

사후 분포의 가우시안 표현 찾기

$p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_0, \mathbf{S}_0)$ 가우시안 사전 분포

고정된 초매개변수

$p(\mathbf{w}|\mathbf{t}) \propto p(\mathbf{w})p(\mathbf{t}|\mathbf{w})$ w에 대한 사후 분포

$\mathbf{t} = (t_1, \ldots, t_N)^{\mathrm{T}}$

양변에 로그
사전 분포 식
기능도 함수 →

$$p(\mathbf{t}|\mathbf{w}) = \prod_{n=1}^{N} y_n^{t_n} \{1 - y_n\}^{1-t_n}$$

$$\ln p(\mathbf{w}|\mathbf{t}) = -\frac{1}{2}(\mathbf{w} - \mathbf{m}_0)^{\mathrm{T}}\mathbf{S}_0^{-1}(\mathbf{w} - \mathbf{m}_0) + \sum_{n=1}^{N} \{t_n \ln y_n + (1 - t_n) \ln(1 - y_n)\} + \text{const}$$

$$y_n = \sigma(\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}_n).$$

4.5.1 라플라스 근사

사후 분포의 가우시안 표현 찾기

사후 분포를 최대화 → 최대 사후 분포 해인 $\boldsymbol{w_{map}}$ 을 구해야 함.

가우시안 분포의 평균값 정의 가능

공분산 : 음의 로그 가능도 함수의 이차 미분값 행렬의 역

$$\mathbf{S}_N = -\nabla\nabla \ln p(\mathbf{w}|\mathbf{t}) = \mathbf{S}_0^{-1} + \sum_{n=1}^{N} y_n(1-y_n)\boldsymbol{\phi}_n\boldsymbol{\phi}_n^{\mathrm{T}}.$$

$$q(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{w}_{\mathrm{MAP}}, \mathbf{S}_N).$$

사후 분포의 가우시안 근삿값

4.5.2 예측 분포

실제 예측을 위해 주변화 시행

새 특징 벡터 : φ(x), 클래스 C_1에 대한 예측 분포

↓ 사후 분포 p(w|t)에 대해 주변화 시행

가우시안 분포 q(w)로 근사 가능

$$p(\mathcal{C}_1|\boldsymbol{\phi},\mathbf{t}) = \int p(\mathcal{C}_1|\boldsymbol{\phi},\mathbf{w})p(\mathbf{w}|\mathbf{t})\,\mathrm{d}\mathbf{w} \simeq \int \sigma(\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi})q(\mathbf{w})\,\mathrm{d}\mathbf{w}$$

$$\sigma(\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}) = \int \delta(a - \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi})\sigma(a)\,\mathrm{d}a$$

← w에 대해 종속적인 부분 : w 의 φ에 대한 투영

$$a = \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}$$

클래스 C_2에 대한 확률 $p(\mathcal{C}_2|\boldsymbol{\phi},\mathbf{t}) = 1 - p(\mathcal{C}_1|\boldsymbol{\phi},\mathbf{t})$

4.5.2 예측 분포

실제 예측을 위해 주변화 시행

$$\sigma(\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}) = \int \delta(a - \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi})\sigma(a)\,\mathrm{d}a$$

δ는 디락(Dirac) 델타 함수이기에 다음을 구할 수 있음.

$$\int \sigma(\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi})q(\mathbf{w})\,\mathrm{d}\mathbf{w} = \int \sigma(a)p(a)\,\mathrm{d}a$$

$$p(a) = \int \delta(a - \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi})q(\mathbf{w})\,\mathrm{d}\mathbf{w}.$$

델타 함수로 인해 w에 대해 선향 제약 조건이 걸림

4.5.2 예측 분포

$$p(a) = \int \delta(a - \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}) q(\mathbf{w})\, \mathrm{d}\mathbf{w}.$$

결합 분포 q(w)를 φ에 대해 직교하는
모든 방향으로 적분하여 주변 분포 만들기

q(w) : 가우시안 분포 → 주변 분포도 가우시안

평균 $\mu_a = \mathbb{E}[a] = \int p(a) a \, \mathrm{d}a = \int q(\mathbf{w}) \mathbf{w}^{\mathrm{T}} \boldsymbol{\phi} \, \mathrm{d}\mathbf{w} = \mathbf{w}_{\mathrm{MAP}}^{\mathrm{T}} \boldsymbol{\phi}$

공분산 $\sigma_a^2 = \mathrm{var}[a] = \int p(a) \left\{ a^2 - \mathbb{E}[a]^2 \right\} \mathrm{d}a$

$$= \int q(\mathbf{w}) \left\{ (\mathbf{w}^{\mathrm{T}} \boldsymbol{\phi})^2 - (\mathbf{m}_N^{\mathrm{T}} \boldsymbol{\phi})^2 \right\} \mathrm{d}\mathbf{w} = \boldsymbol{\phi}^{\mathrm{T}} \mathbf{S}_N \boldsymbol{\phi}.$$

4.5.2 예측 분포

$$p(\mathcal{C}_1|\mathbf{t}) = \int \sigma(a)p(a)\,\mathrm{d}a = \int \sigma(a)\mathcal{N}(a|\mu_a, \sigma_a^2)\,\mathrm{d}a.$$

a의 분포가 아래의 선형 회귀 모델에서의 예측 분포가 노이즈 분산이 0일 때와 같은 형태를 띤다

$$p(t|\mathbf{x}, \mathbf{t}, \alpha, \beta) = \mathcal{N}(t|\mathbf{m}_N^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}), \sigma_N^2(\mathbf{x}))$$

4.5.2 예측 분포

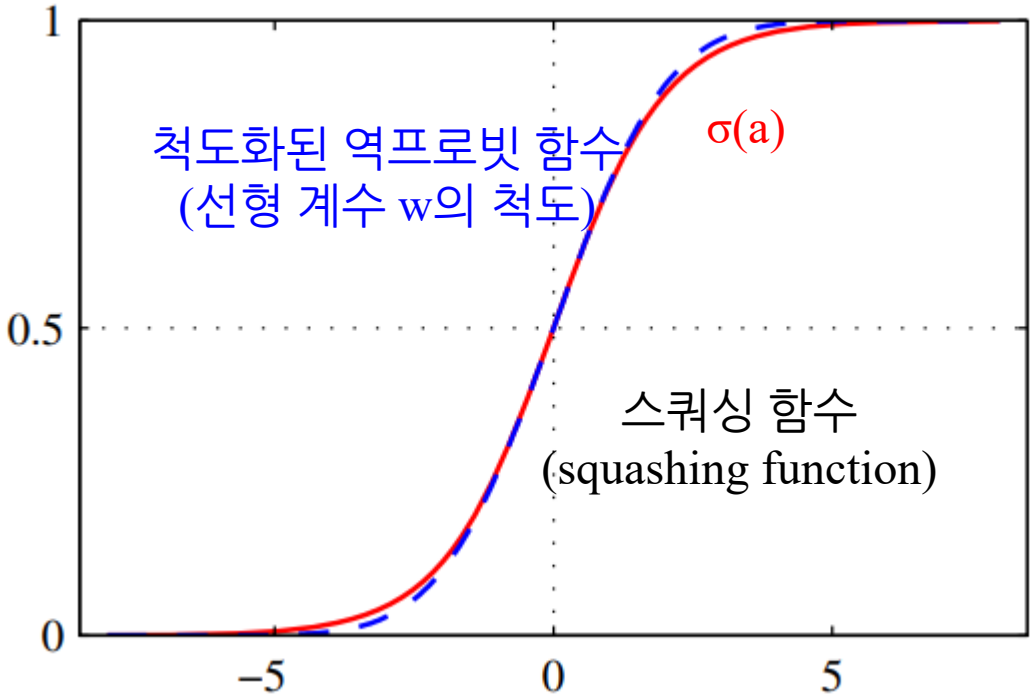
$$p(t|\mathbf{x}, \mathbf{t}, \alpha, \beta) = \mathcal{N}(t|\mathbf{m}_N^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}), \sigma_N^2(\mathbf{x}))$$

a에 대한 적분은 가우시안을 시그모이드와 콘볼루션 한 것

해석적으로 계산 불가
→ 로지스틱 시그모이드 함수와 역프로빗 함수의 유사성 이용해서 유사 분석 가능

$$\lambda^2 = \pi/8.$$

로지스틱 함수의 최고 근사치를 얻기 위해 가로축의 척도 변경



4.5.2 예측 분포

역프로빗 함수와 가우시안 콘볼루션
→ 다른 역프로빗 함수로 표현 가능

$$\int \Phi(\lambda a)\mathcal{N}(a|\mu,\sigma^2)\,\mathrm{d}a = \Phi\left(\frac{\mu}{(\lambda^{-2}+\sigma^2)^{1/2}}\right).$$

$$\sigma(a) \simeq \Phi(\lambda a)$$

양변에 대입

로지스틱 시그모이드와 가우시안의 콘볼루션에 대한 다음의 근사 얻기 가능

$$\int \sigma(a)\mathcal{N}(a|\mu,\sigma^2)\,\mathrm{d}a \simeq \sigma\left(\kappa(\sigma^2)\mu\right)$$

$$\kappa(\sigma^2) = (1+\pi\sigma^2/8)^{-1/2}.$$

여기에 적용

$$p(\mathcal{C}_1|\mathbf{t}) = \int \sigma(a)p(a)\,\mathrm{d}a = \int \sigma(a)\mathcal{N}(a|\mu_a,\sigma_a^2)\,\mathrm{d}a.$$

$$p(\mathcal{C}_1|\boldsymbol{\phi},\mathbf{t}) = \sigma\left(\kappa(\sigma_a^2)\mu_a\right)$$

예측 분포에 대한 근사치

4.5.2 예측 분포

Thank you