기계학습

2강 선형회귀(1)

장필훈 교수

학습목차

01

개요

1-1 개요

- 목표
  - 입력이 주어지면 그에 해당하는 타겟변수를 예측하는 것

$$y = f(x)$$

  - 입력의 차원이 큰 경우가 다수

$$\mathbf{x}: (x_1, x_2, x_3, \ldots, x_n)$$

02
기저함수모델

2-1 선형회귀모델

- $y(\mathbf{x}, \mathbf{w}) = w_0 + w_1 x_1 + \cdots + w_D x_D$

- $y(\mathbf{x}, \mathbf{w}) = w_o + \sum_{J=1}^{M-1} w_j \phi_j(\mathbf{x})$

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

2-1 선형회귀모델

- 기저함수 basis function $\boldsymbol{\phi}$

$$\sum_{j=0}^{M} w_j x^j = w + w_1 x + w x^2 + \cdots + w_M x^M = y(x, \boldsymbol{w})$$

- 비선형함수들을 사용해서 복잡한 형태를 표현할 수 있다.

2-1 선형회귀모델

- 다항기저함수의 한계점
  - 한 영역의 변화가 다른 영역까지 영향을 미친다
  - 대안: 스플라인 함수 spline function
    - 입력공간을 여러 영역들로 나눔
    - 각 영역별로 다른 다항식 피팅

2-1 선형회귀모델

- 다양한 기저함수

$$\phi_j(x) = \exp\left(-\frac{\left(x-\mu_j\right)^2}{2s^2}\right)$$

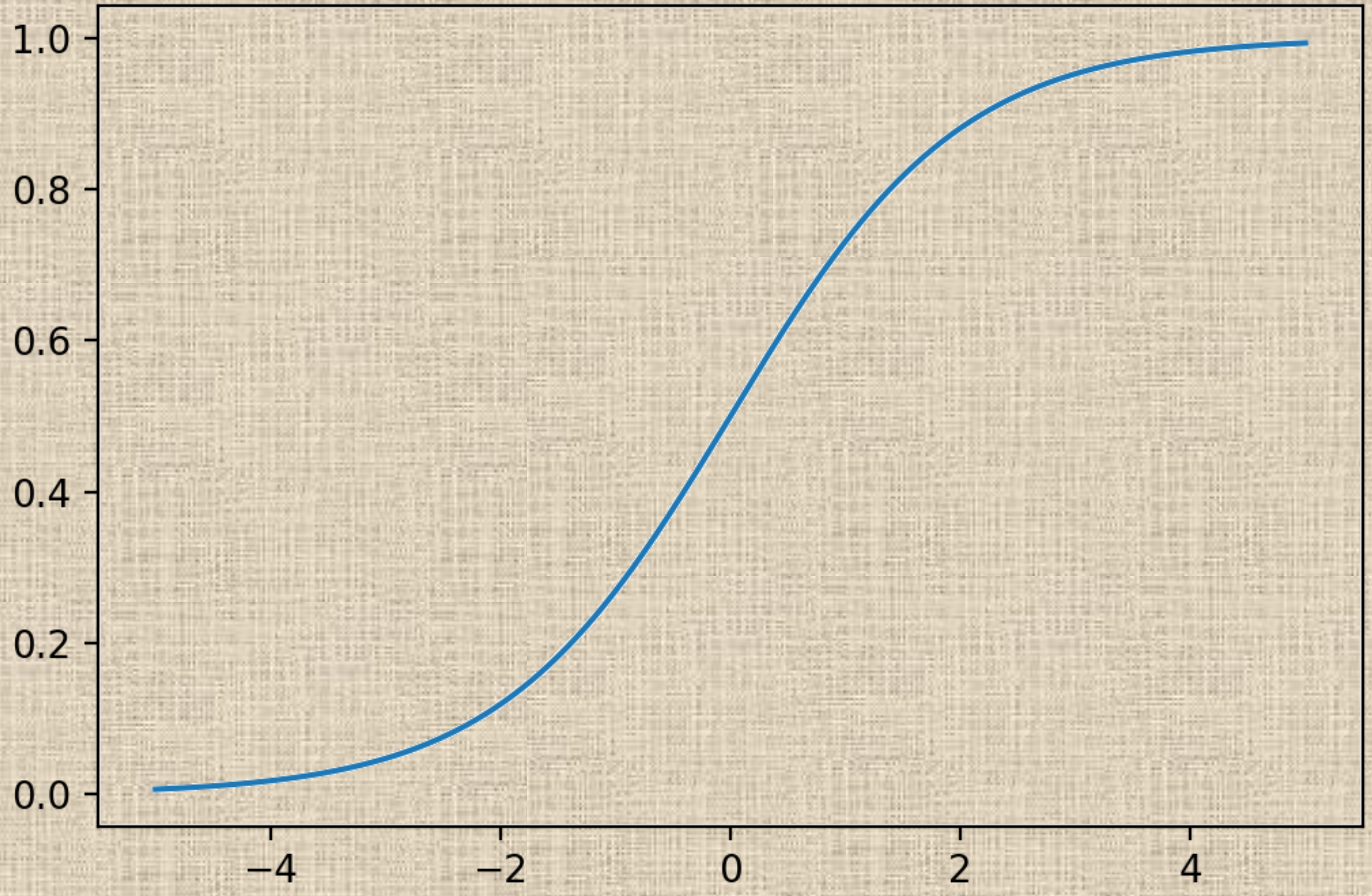
$$\phi_j(x) = \sigma\left(\frac{x-\mu_j}{s}\right)$$

$$\sigma(x) = \frac{1}{1+e^{-x}}$$

2-1 선형회귀모델

- sigmoid

$$\sigma(x) = \frac{1}{1+e^{-x}}$$



2-2 회귀에서 손실함수

- 입력 x에 대해 t의 추정값 $y(x)$를 선택한다고 할때, 손실 $L(t, y(x))$가 발생한다고 생각하면, 기대손실은

$$E[L] = \iint L(t, y(x)) p(x, t) dx dt$$

2-2 회귀에서 손실함수

- 제곱손실을 사용하면,

$$L\big(t, y(x)\big) = \{y(x) - t\}^2$$

$$E[L] = \iint \{y(x) - t\}^2 p(x, t)\, dx dt$$

$E[L]$을 최소화하는 $y(x)$를 선택하자.

2-2 회귀에서 손실함수

$$\frac{\delta E[L]}{\delta y(x)} = 2\int \{y(x) - t\}p(x,t)dt = 0$$

$$\int y(x)p(x,t)dt \; - \int tp(x,t)dt = 0$$

$$y(x)\int p(x,t)dt = \int tp(x,t)dt = E_t[t|x] \qquad y(x) = E_t[t|x]$$

x가 주어졌을 때 t의 조건부평균 : 최적의 예측값은 조건부**평균**

2-2 최대가능도와 최소제곱

$$t = y(x, w) + \epsilon$$

$$\epsilon \sim N(0, \beta^{-1})$$

- 주어진 x값에 대한 t값이

$y(x, w)$를 평균으로 하는 가우시안 분포를 따르면,

$$p(t|x, w, \beta) = \mathrm{N}(\mathrm{t}|\mathrm{y}(\mathrm{x}, \mathrm{w}), \beta^{-1}), \qquad \beta = \frac{1}{\sigma^2}$$

$$p(t|\boldsymbol{x}, w, \beta) = \prod_{n=1}^{N} N(t_n|y(x_n, w), \beta^{-1})$$

(1강 참고)

2-2 최대가능도와 최소제곱

$$p(t|\boldsymbol{x}, w, \beta) = \prod_{n=1}^{N} N(t_n|y(x_n, w), \beta^{-1}) = \prod_{n=1}^{N} N(t_n|w^T\phi(x_n), \beta^{-1})$$

기저함수

$$\ln \prod_{n=1}^{N} N(t_n|y(x_n, w), \beta^{-1})$$

$$= -\frac{\beta}{2}\sum_{n=1}^{N}\{y(x_n, w) - t_n\}^2 + \frac{N}{2}\ln\beta - \frac{N}{2}\ln(2\pi)$$

1강에서 유도

2-2 최대가능도와 최소제곱

$$= -\frac{\beta}{2}\sum_{n=1}^{N}\{t_n - w^T\phi(x_n)\}^2 + \frac{N}{2}\ln\beta - \frac{N}{2}\ln(2\pi)$$

$$= -\beta E_D(w) \qquad + \frac{N}{2}\ln\beta - \frac{N}{2}\ln(2\pi)$$

제곱합 오차함수

$$\beta\sum_{n=1}^{N}\{t_n - w^T\phi(x_n)\}\phi(x_n)^T = 0$$

w를 계산할 수 있다.

2-2 최대가능도와 최소제곱

- 편향의 의미

$$E_D(w) = \frac{1}{2}\sum_{n=1}^{N}\left\{t_n - w_0 - \sum_{j=1}^{M-1} w_j\phi_j(x_n)\}\right\}^2$$

$$\frac{d}{dw_0}E_D(w) = 0, \qquad w_0 = \bar{t} - \sum_{j=1}^{M-1} w_j\overline{\phi_j}$$

2-2 최대가능도와 최소제곱

$$\frac{d}{dw_0}\left[\frac{1}{2}\sum_{n=1}^{N}\left\{t_n - w_0 - \sum_{j=1}^{M-1} w_j\phi_j(x_n)\}\right\}^2\right] = 0$$

$$\sum_{n=1}^{N}\left(t_n - w_0 - \sum_{j=1}^{M-1} w_j\phi_j(x_n)\right)(-1) = 0$$

$$\sum_{n=1}^{N} t_n - \sum_{n=1}^{N} w_0 - \sum_{n=1}^{N}\sum_{j=1}^{M-1} w_j\phi_j(x_n) = 0$$

2-2 최대가능도와 최소제곱

$$\sum_{n=1}^{N} t_n - Nw_0 - \sum_{j=1}^{M-1}\sum_{n=1}^{N} w_j \phi_j(x_n) = 0$$

$$w_0 = \frac{1}{N}\sum_{n=1}^{N} t_n - \sum_{j=1}^{M-1}\frac{1}{N} w_j \sum_{n=1}^{N} \phi_j(x_n) = \bar{t} - \sum_{j=1}^{M-1} w_j \overline{\phi_j}$$

➢ 편향(w_0)이 훈련집합의 타겟변수들 평균($\bar{t}$)과,
기저함숫값 평균의 가중합 차이를 보상한다.

2-3 순차학습

- 전체 훈련집합을 한번에 처리해야 한다
  - 크면 불가능. 이미지, 음성등 데이터는 무조건 크다
  - 순차적으로 학습한다. = online 학습
- SGD(stochastic gradient descent)
  - 계속해서 w를 업데이트하고 실시간으로 예측

2-3 순차학습

- 각각의 데이터가 가지는 오류값(=오차함수의 결괏값)들을 합해서 여러 데이터의 오차함수값을 얻을 수 있다면, 다음과 같이 w를 업데이트 할 수 있다.

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

$$w^{(\tau+1)} = w^{\tau} - \mu\left(t_n - w^{(\tau)T}\phi_n\right)\phi_n$$

2-4 정규화된 최소제곱법

- 정규화항이 포함된 오차함수

$$E_D(\mathbf{w}) + \lambda E_W(\mathbf{w})$$

$$E_w(\mathbf{w}) = \frac{1}{2}\mathbf{w}^T\mathbf{w}$$

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

2-4 정규화된 최소제곱법

- 가중치 감쇠
  - 데이터에 의해 지지되지 않는 한 가중치가 0 에 수렴
  - parameter shrinkage의 한 예
  - 이차함수로 쓰면 최솟값을 구하기가 쉽다

2-4 정규화된 최소제곱법

- 정규화항이 포함된 오차함수 - 더 일반적인 형태

$$\frac{1}{2}\sum_{n=1}^{N}\{t_n - \mathbf{w}^T\phi(\mathbf{x}_n)\}^2 + \frac{\lambda}{2}\sum_{j=1}^{M}|w_j|^q$$

$q = 1$일 때를 lasso라고 한다.(→sparse model)

2-4 정규화된 최소제곱법

- 오차함수를 자세히 보면, 다음 제약조건 하에서 제곱합오차를 최소화 하는 것임을 알 수 있음.

$$\sum_{j=1}^{M} |w_j|^q \leq \eta$$

이유 : 라그랑주 승수법

2-4 정규화된 최소제곱법

- **라그랑주 승수법**

연속미분가능함수 $f: \mathbb{R}^D \to \mathbb{R}$과 $g: \mathbb{R}^D \to \mathbb{R}^C$

$g(x) = 0$을 제약조건으로 $f(x)$의 최대 혹은 최솟값을 찾으려면

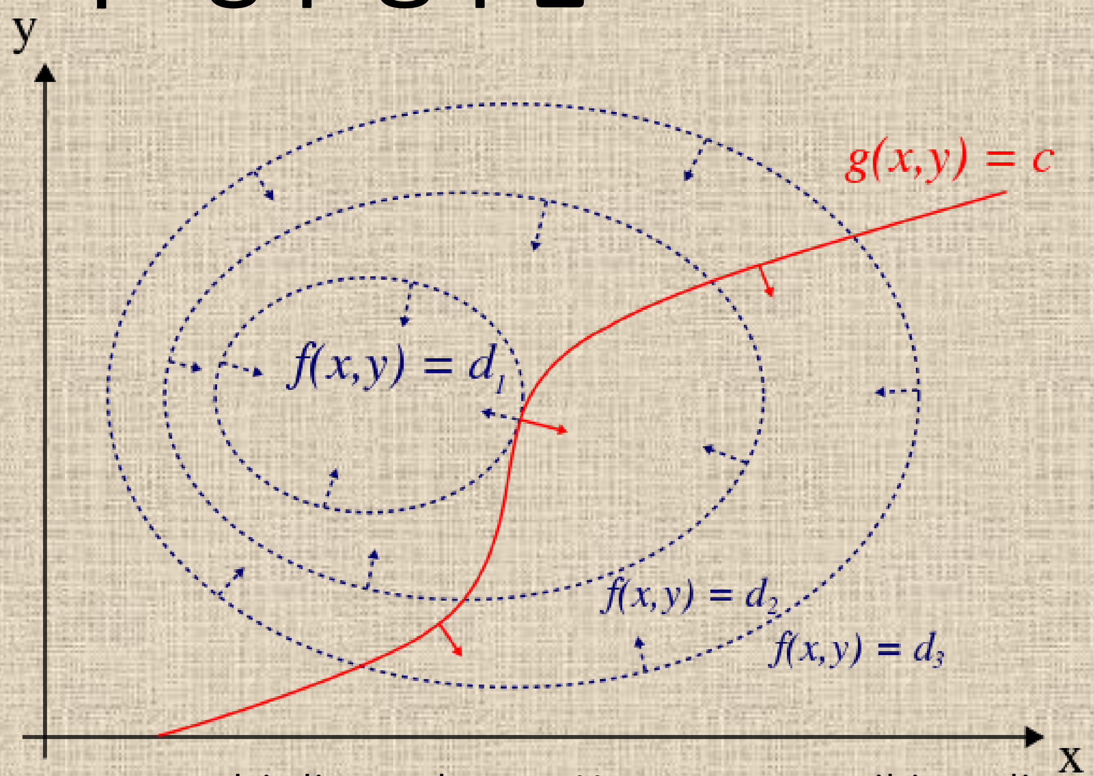
$\nabla f(x) = \lambda \nabla g(x), \quad g(x) = 0$을 푼다.

$\mathcal{L}(x, \lambda) = f(x) - \lambda g(x)$: lagrangian function

Find stationary points!

2-4 정규화된 최소제곱법

- 라그랑주 승수법



[Lagrange multiplier © https://commons.wikimedia.org/wiki/File:LagrangeMultipliers2D.svg]

2-4 정규화된 최소제곱법

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

$$\sum_{j=1}^{M}|w_j|^q \leq \eta$$

$$\mathcal{L}(w,\lambda) = \frac{1}{2}\sum_{n=1}^{N}\{t_n - \mathbf{w}^T\phi(\mathbf{x}_n)\}^2 + \frac{\lambda}{2}\left(\sum_{j=1}^{M}|w_j|^q - \eta\right)$$

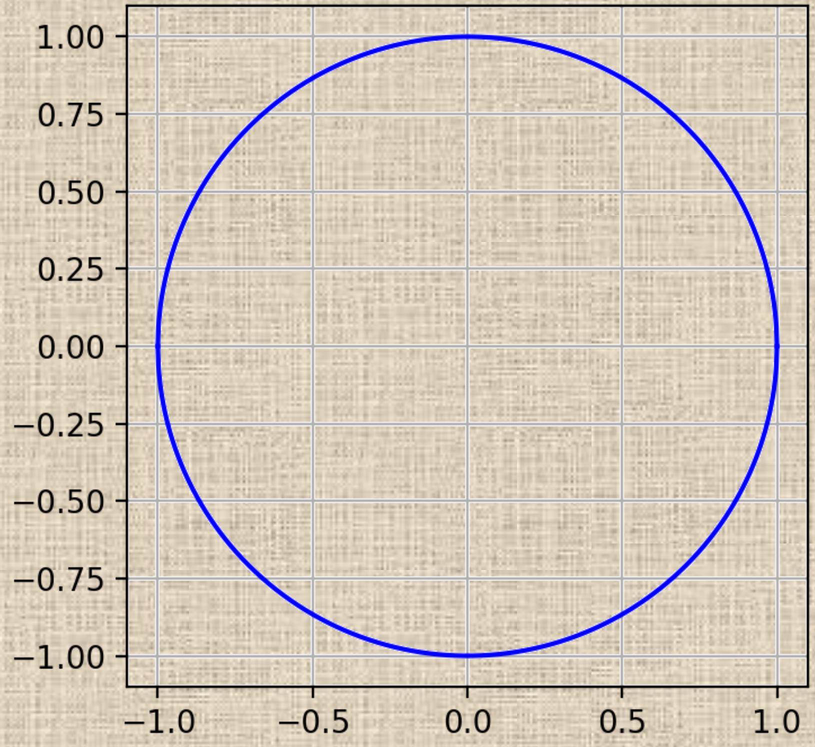
2-4 정규화된 최소제곱법

$$\mathcal{L}(w,\lambda) = \frac{1}{2}\sum_{n=1}^{N}\{t_n - \mathbf{w}^T\phi(\mathbf{x}_n)\}^2 + \frac{\lambda}{2}\left(\sum_{j=1}^{M}|w_j|^q - \eta\right)$$

$$\frac{1}{2}\sum_{n=1}^{N}\{t_n - \mathbf{w}^T\phi(\mathbf{x}_n)\}^2 + \frac{\lambda}{2}\sum_{j=1}^{M}|w_j|^q$$

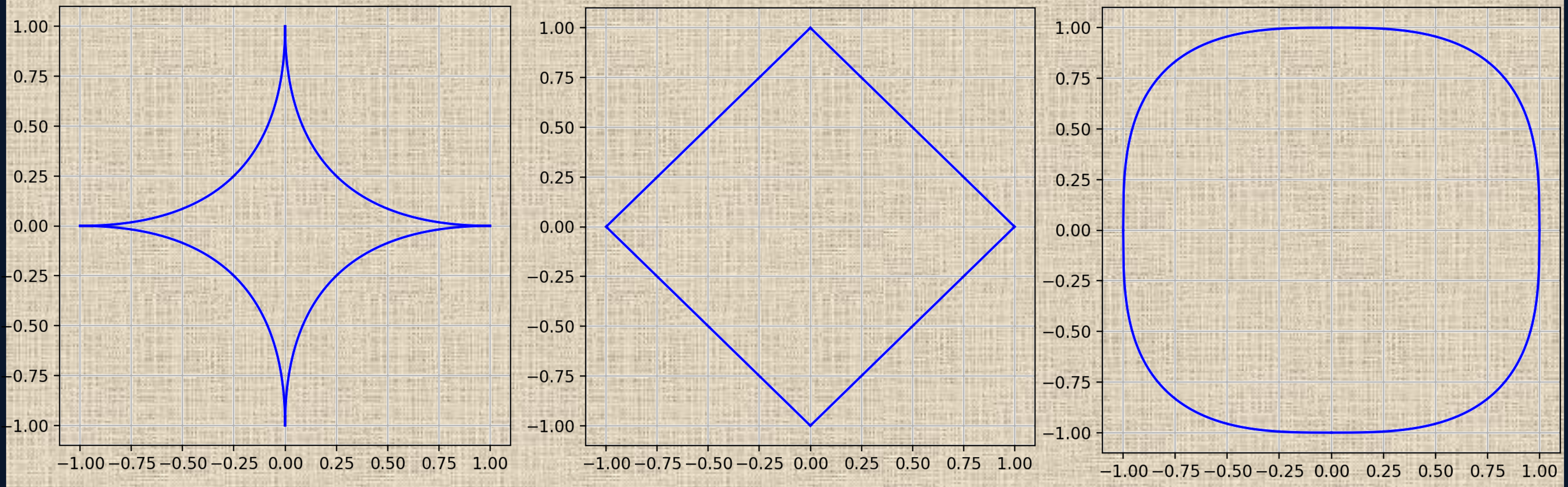
2-4 정규화된 최소제곱법

$$\sum_{j=1}^{M} |w_j|^q \leq \eta$$

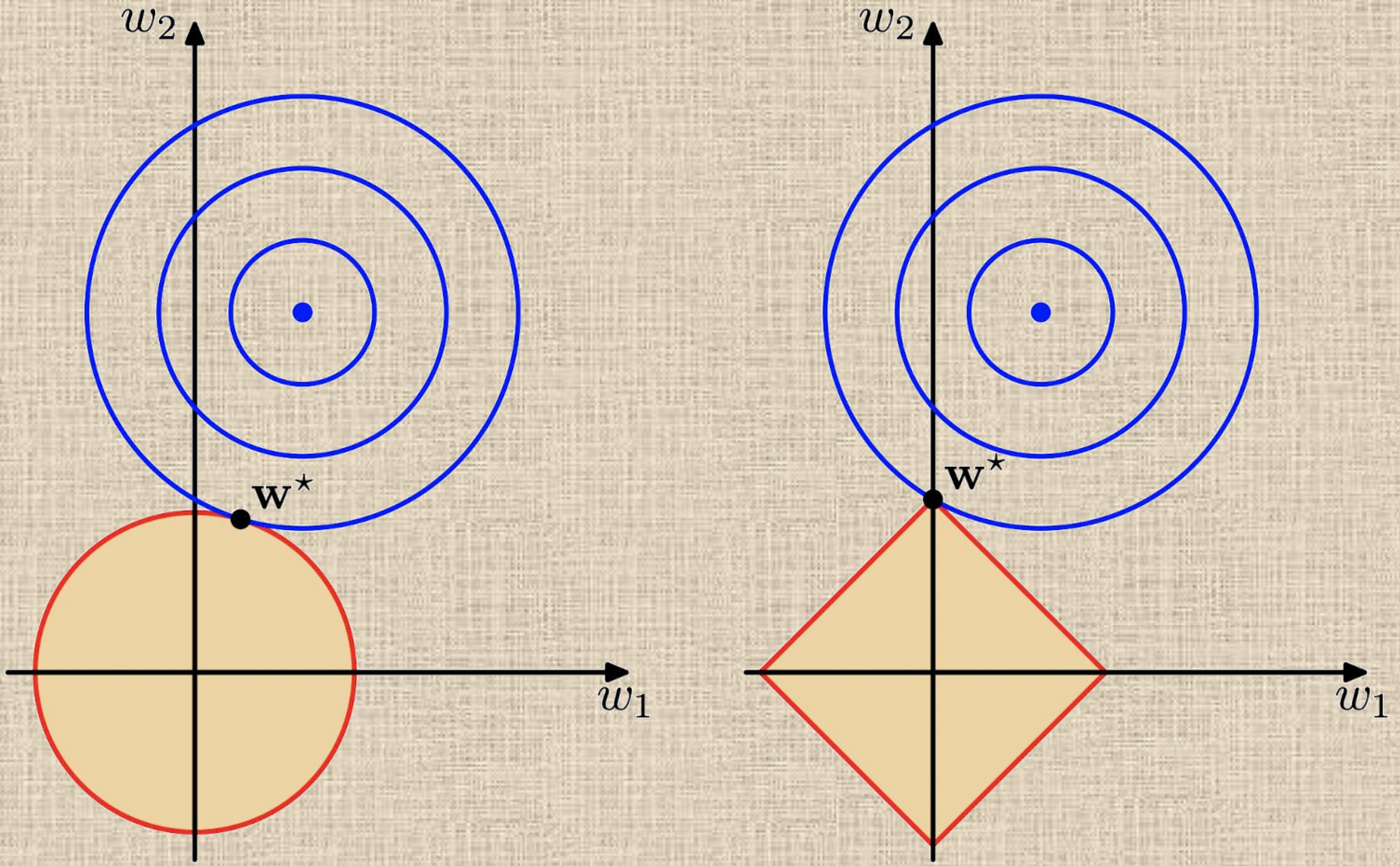


$q = 2$

2-4 정규화된 최소제곱법



2-4 정규화된 최소제곱법



Bishop, Christopher M. Pattern Recognition and Machine Learning. New York :Springer, 2006. p.146

03

편향분산분해

3 편향분산분해(1)

- 지금까지는 basis function의 수와 형태를 고정
- 너무 많이 하면 overfitting, 너무 적게 하면 underfitting
  - 많이 하고 regularization term을 사용
    - λ를 정해야 하는 새로운 문제 발생
- 모델 복잡도에 관한 문제: 편향-분산 트레이드 오프

3 편향분산분해(1)

- 기대제곱오류는 다음과 같이 쓸 수 있다.

$$\mathbb{E}[L] = \int \{y(\mathbf{x}) - h(\mathbf{x})\}^2 p(\mathbf{x})\,\mathrm{d}\mathbf{x} + \int \{h(\mathbf{x}) - t\}^2 p(\mathbf{x}, t)\,\mathrm{d}\mathbf{x}\,\mathrm{d}t$$

$$\mathbb{E}[L] = \iint \{y(\mathbf{x}) - t\}^2 p(\mathbf{x}, t)\,\mathrm{d}\mathbf{x}\,\mathrm{d}t$$

$$\begin{aligned}\{y(\mathbf{x}) - t\}^2 &= \{y(\mathbf{x}) - \mathbb{E}[t|\mathbf{x}] + \mathbb{E}[t|\mathbf{x}] - t\}^2 \\ &= \{y(\mathbf{x}) - \mathbb{E}[t|\mathbf{x}]\}^2 + 2\{y(\mathbf{x}) - \mathbb{E}[t|\mathbf{x}]\}\{\mathbb{E}[t|\mathbf{x}] - t\} + \{\mathbb{E}[t|\mathbf{x}] - t\}^2\end{aligned}$$

3 편향분산분해(1)

- 식을 손실함수에 대입하고 t에 대해 적분하면 교차항이 사라진다.

그 결과로 다음 손실함수를 얻는다.

$$\mathbb{E}[L] = \int \{y(\mathbf{x}) - \mathbb{E}[t|\mathbf{x}]\}^2 \, p(\mathbf{x}) \, \mathrm{d}\mathbf{x} + \int \{\mathbb{E}[t|\mathbf{x}] - t\}^2 p(\mathbf{x}) \, \mathrm{d}\mathbf{x}.$$

첫번째 항에서 $y(x) = E_t[t|x]$ 일때 손실기대가 최소.

앞서 얻은 결론과 같다.

3 편향분산분해(1)

- 두번째 항은 t에 대한 분포의 분산을 계산하고 x에 대해 평균.

$$\int \{\mathbb{E}[t|\mathbf{x}] - t\}^2 p(\mathbf{x}) \, \mathrm{d}\mathbf{x}$$

데이터가 가지고 있는 내재적 변동성을 표현하는 것 = 노이즈

$y(x)$에 대해 독립이기 때문에,

y를 아무리 잘 추정해도 절대로 줄일 수 없다.

3 편향분산분해(1)

- 따라서, 우리가 알아내야 할 분포를 $h(x)$라고 하면,

1. 데이터가 무한히 주어지지 않는 한 $h(x)$를 알수 없다.
2. $h(x)$로부터 추출한 데이터집합 D에 대해 $y(x)$를 추정한다.
   - 이것을 여러번 반복해서 모델의 성능을 측정한다.

3 편향분산분해(1)

- 특정 데이터셋 D 를 대상으로 했을 때, 제곱오류는

$$\{y(x;D) - h(x)\}^2$$

와 같고, 이것을 여러번 시행해서 평균을 낸다고 하면,

$$E[\{y(x;D) - h(x)\}^2]$$

를 구하는 것과 같다.

다음시간

3강

- 편향분산분해(2)
- 베이지안선형회귀
- 베이지안모델선택
- 차원의 저주