기계학습

1강 Overview of Supervised Learning

장필훈 교수

학습목차

1 개념과 범위

(기계학습, 패턴인식, 학습의 종류, 다항식 피팅)

2 확률

(베이즈 정리, 확률분포)

01

개념과 범위

1-1 기계학습

- **기계학습의 결과물**
  - ‘생각하는 기계’
  - 자율주행, 번역기
  - 기타
- **학습내용**
  - 패턴인식
  - 위의 응용 이전에 보다 근본적인 문제

1-2 패턴인식

- 패턴: 일정한 형태나 양식 또는 유형

mnist dataset

-2.631	-0.324	1.402	4.427	2.407	-0.990	8.684	0.261
-1.224	2.835	-1.355	2.523	1.664	1.969	0.780	2.607

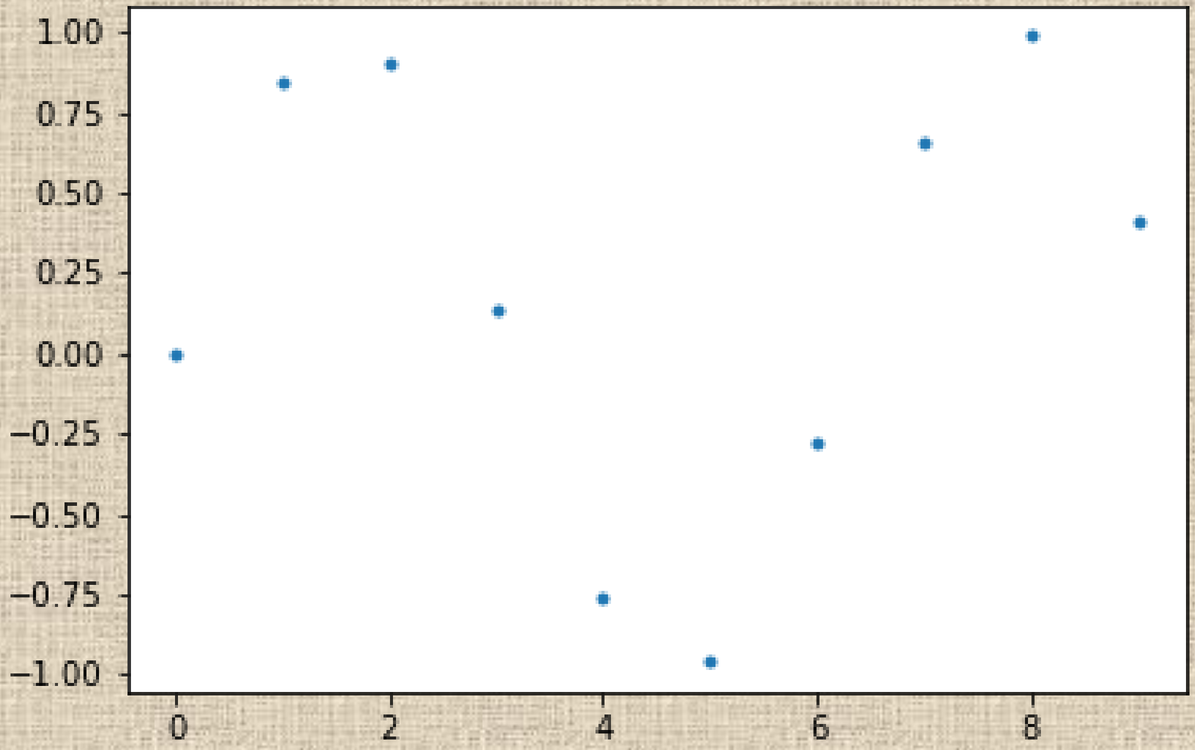
$\mu = 1, \sigma = 3$

- 패턴을 기계가 알아내도록 하는 것이 기계학습의 기본적인 목표.

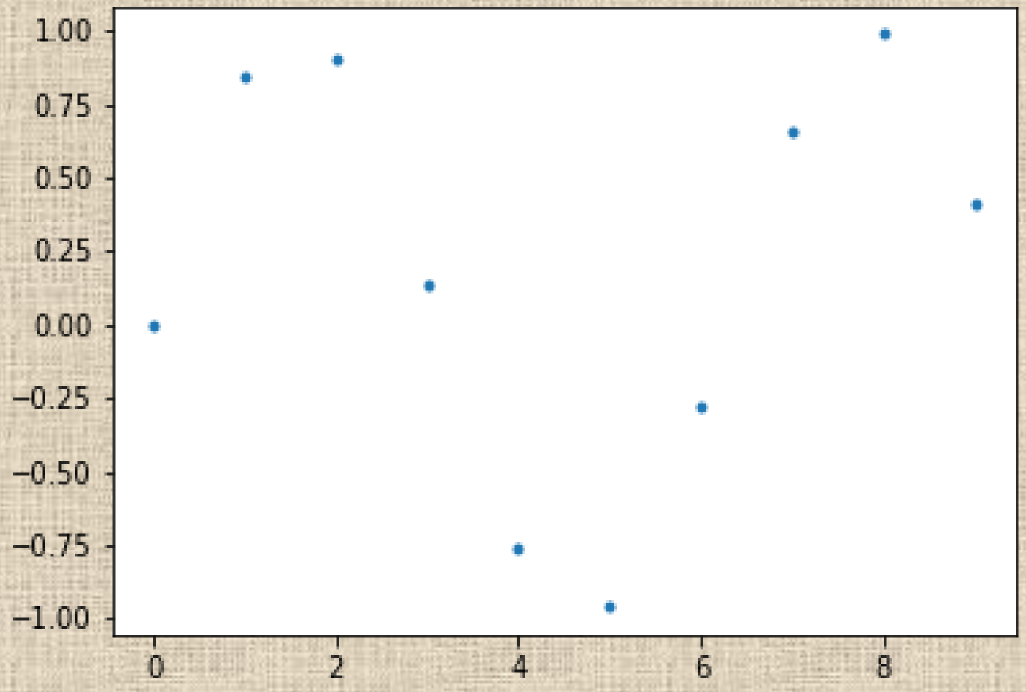
1-3 학습의 종류

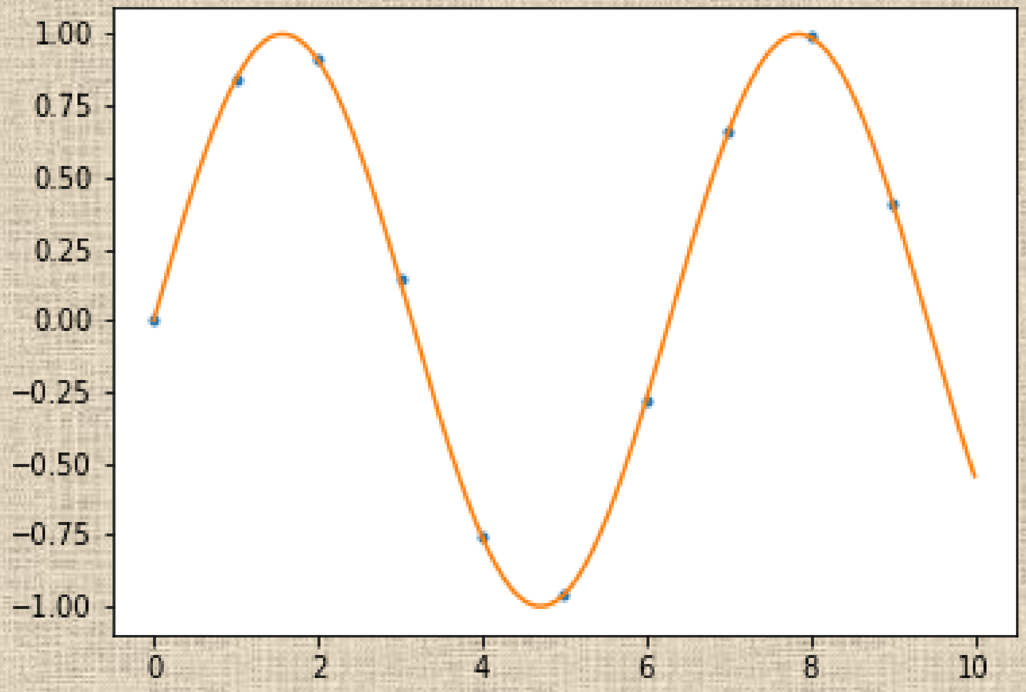
- **지도학습** supervised learning
  - 목표가 명시적으로 주어짐. 입력 하나에 답(target)하나.
  - 우리가 배울 대부분이 여기에 속함(분류, 회귀등)
- **비지도학습** unsupervised learning
  - 목표가 명시적이지 않음
  - 군집화(clustering), 밀도추정등
- **강화학습**
  - 보상을 최대화 하기 위한 행동을 찾음

1-4 다항식 곡선 피팅



1-4 다항식 곡선 피팅

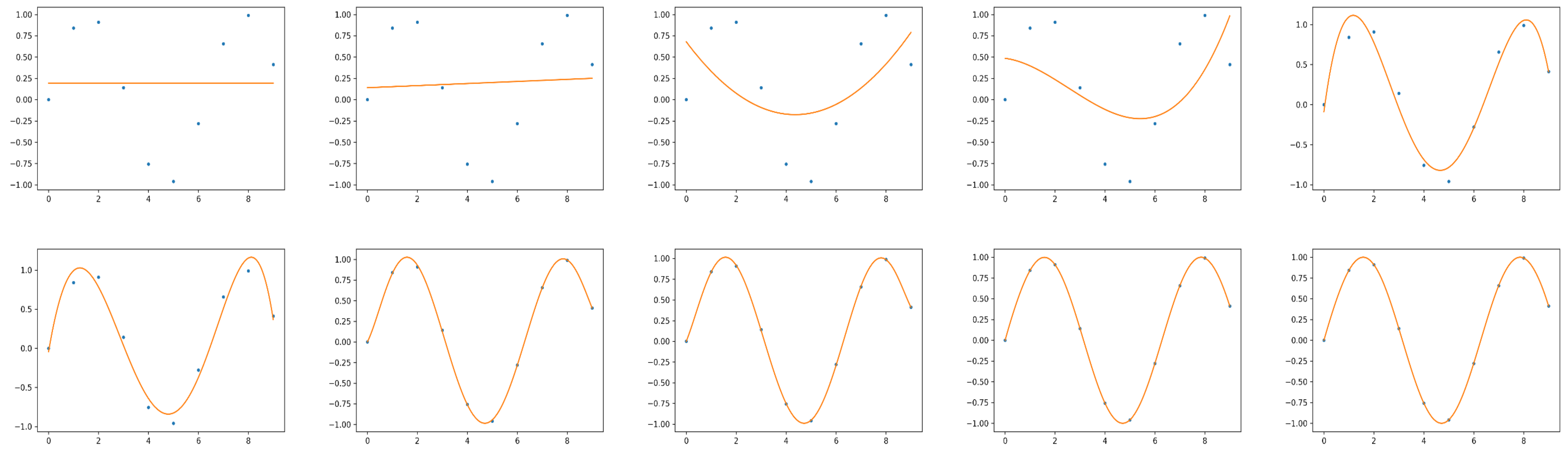




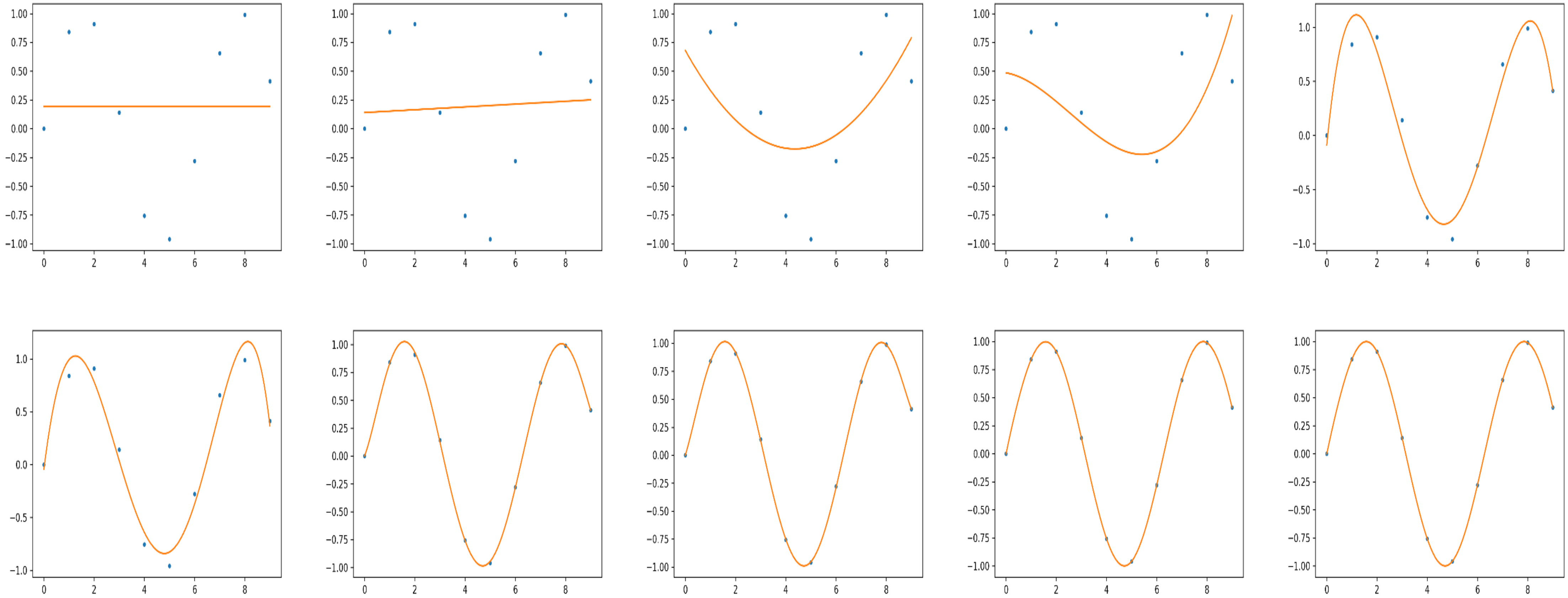
$$y = \sin x$$

1-4 다항식 곡선 피팅

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



1-4 다항식 곡선 피팅



1-4 다항식 곡선 피팅

- **모델선택**: 어떤 모델을 쓸 것인가
  - 다항식 모델이 타당한가
  - 다항식 모델을 사용한다면 M을 어떻게 선택할 것인가

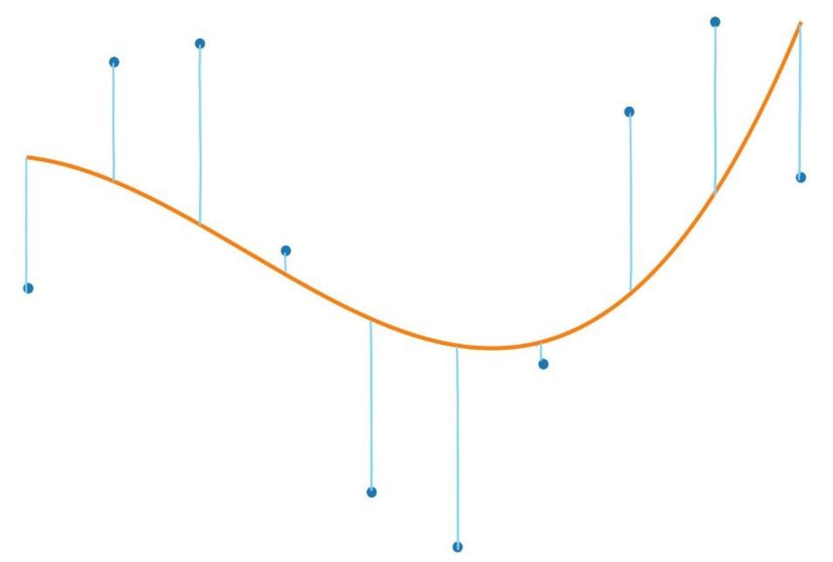
오차함수를 정의해야 비교가 가능하다.

$$\sum_{n=1}^{N} \{y(x_n, w) - t_n\}^2 \quad \text{or} \quad \sum_{n=1}^{N} |y(x_n, w) - t_n| \quad \cdots$$

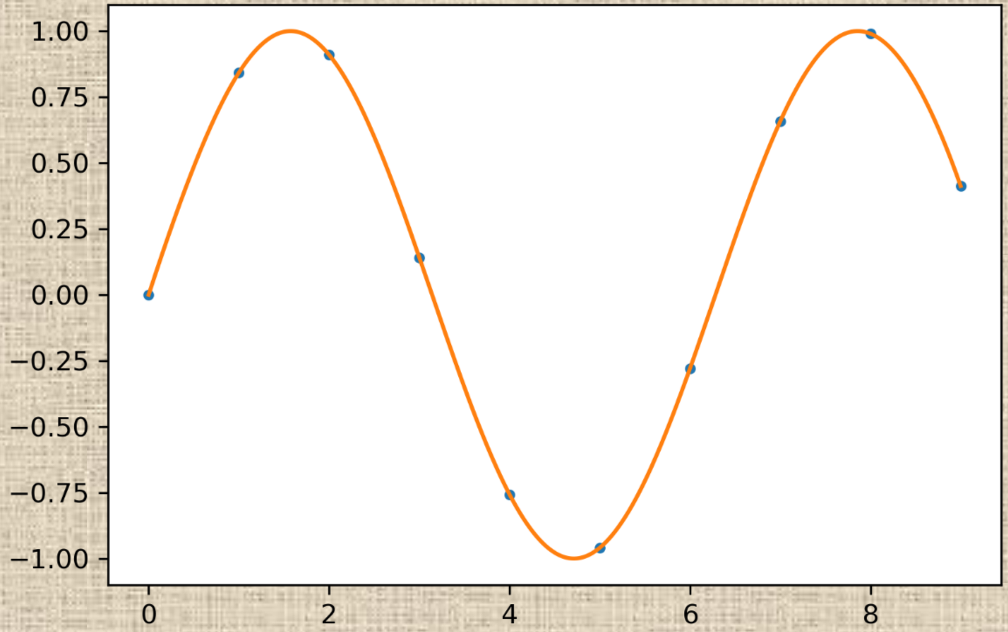
실제 데이터는 노이즈를 포함한다.

$\Rightarrow$ 모델, 오차함수의 선택은 더 어려워진다.

1-4 다항식 곡선 피팅



$M = 3$

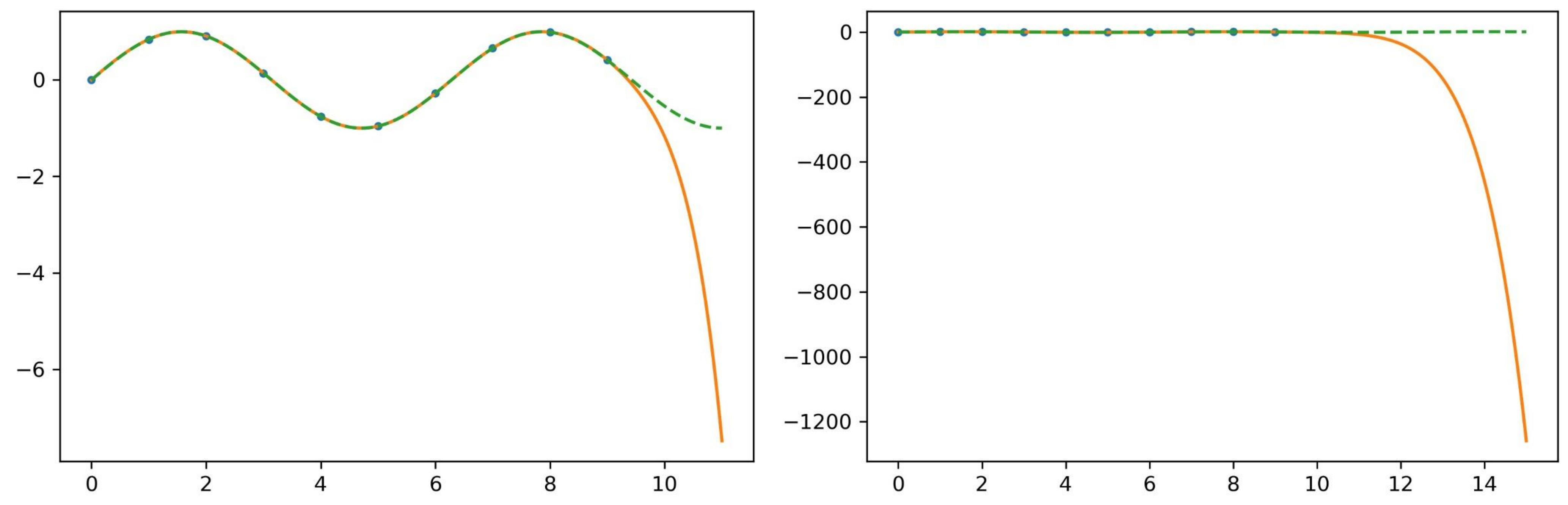


$M = 9$

어떤 오차함수를 쓰든 $M = 9$ 에서 완벽하게 fit하지만

$y = \sin x$를 표현해내는 것은 아니다. : 과적합 **over−fitting**

1-4 다항식 곡선 피팅



1-4 다항식 곡선 피팅

❖ 과적합을 막기 위한 대표적인 방법 : **정규화**

$$\tilde{E}(w) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n, w) - t_n\}^2 + \frac{\lambda}{2}\|w\|^2$$

$$\|w\|^2 = w_0^2 + w_1^2 + w_2^2 + \cdots + w_M^2$$

$$\sum \frac{\lambda}{2}|w| \quad ?$$

02

확률

2-1 베이즈정리

$$p(X,Y) = p(Y|X)p(X)$$

$$p(X) = \sum_{Y} p(X,Y)$$

$$p(Y|X) = \frac{p(X|Y)p(Y)}{p(X)}$$

$$p(X,Y) = p(Y,X) = p(Y|X)p(X) = p(X|Y)p(Y)$$

2-1 베이즈정리

likelihood(우도)

$$p(Y|X) = \frac{p(X|Y)p(Y)}{p(X)}$$

prior

evidence

posterior

$$p(Y|X) \propto p(X|Y)p(Y)$$

베이지안 정리를 이용해서 사전확률을 사후확률로 바꿀 수 있다.

2-1 베이즈정리

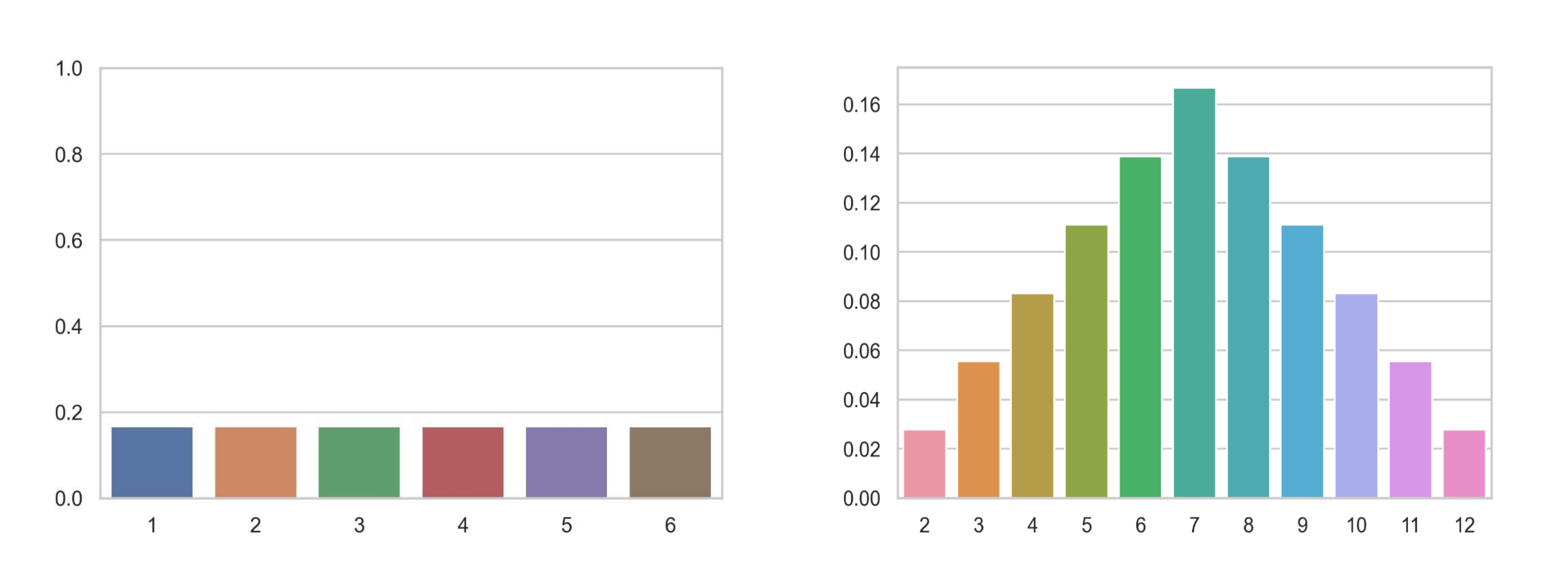
◆ 앞선 문제에서, 적합한 매개변수 w를 정할 때, 데이터를 바탕으로 w의 분포를 추정할 수 있다.

D를 관측한 후 w에 대한 불확실성을 표현

$$p(w|D) = \frac{p(D|w)p(w)}{p(D)}$$

최대가능도(maximum likelihood) : $p(D|w)$의 최대화

2-2 확률 분포

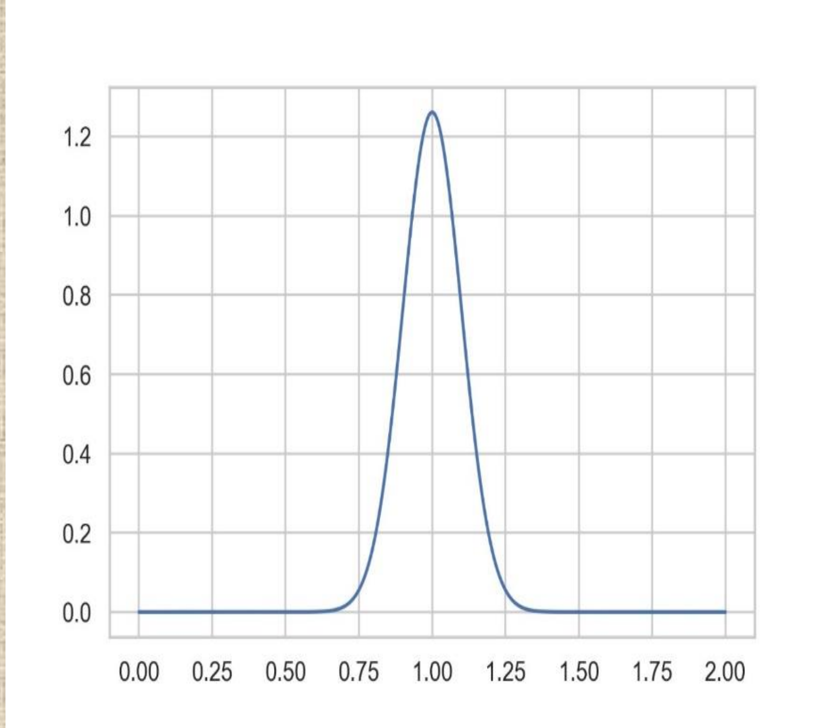


합이 언제나 1

2-2 확률 분포

$$N(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$\frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$



$\mu = 1, \sigma = 0.1$

2-2 확률 분포

노이즈가 가우시안 분포를 따른다고 할때,

어떻게 해야 가장 '좋은' w를 찾을 수 있겠는가?

- 가우시안 분포의 평균과 분산 확인
  - ➢ 가우시안 적분을 이용해서 가우시안 분포의 정규화 확인
- 가능도 함수의 최대화

2-2 확률 분포

- 가우시안 분포의 정규화 확인

$$\int_{-\infty}^{\infty} N(x|\mu, \sigma^2)\, \mathrm{d}x = 1$$

$$\left(\int_{-\infty}^{\infty} \mathrm{e}^{-x^2}\, \mathrm{d}x\right)^2 = \int_{-\infty}^{\infty} \mathrm{e}^{-x^2}\, \mathrm{d}x \int_{-\infty}^{\infty} \mathrm{e}^{-y^2}\, \mathrm{d}y = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mathrm{e}^{-(x^2+y^2)}\, \mathrm{d}x\, \mathrm{d}y$$

$$\mathrm{e}^{-(x^2+y^2)} = \mathrm{e}^{-r^2}$$

2-2 확률 분포

$$\iint_{\mathbb{R}^2} e^{-(x^2+y^2)}\,dx\,dy = \int_0^{2\pi}\int_0^{\infty} e^{-r^2}\,r\,dr\,d\theta$$

$$= 2\pi \int_0^{\infty} re^{-r^2}\,dr$$

$$\int_0^{\infty} r\,e^{-r^2} dr = \int_0^{\infty} \frac{1}{2} e^{-r^2} \cdot 2r\,\mathrm{d}r = \int_{\infty}^{0} \frac{1}{2} e^{-r^2} \cdot (-2r)\,\mathrm{d}r$$

$$= 2\pi \int_{-\infty}^{0} \tfrac{1}{2} e^s\,ds \qquad s = -r^2$$

$$= \pi \int_{-\infty}^{0} e^s\,ds$$

$$\int_{-\infty}^{\infty} e^{-a(x+b)^2}\,dx = \sqrt{\frac{\pi}{a}}.$$

$$= \pi\left(e^0 - e^{-\infty}\right)$$

$$= \pi,$$

2-2 확률 분포

$$E(x) = \int_{-\infty}^{\infty} N(x|\mu, \sigma^2)\, x \, \mathrm{d}x = \mu$$

$$\frac{1}{\sigma\sqrt{2\pi}} \int_{\infty}^{\infty} x \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \mathrm{d}x$$

$$= \frac{\sqrt{2}\sigma}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} \left(\sqrt{2}\sigma t + \mu\right) \exp\left(-t^2\right) \mathrm{d}t \qquad t = \frac{x-\mu}{\sqrt{2}\sigma}$$

$$= \frac{1}{\sqrt{\pi}} \left(\sqrt{2}\sigma \int_{-\infty}^{\infty} t \exp\left(-t^2\right) \mathrm{d}t + \mu \int_{-\infty}^{\infty} \exp\left(-t^2\right) \mathrm{d}t\right)$$

$$= \frac{1}{\sqrt{\pi}} \left(\sqrt{2}\sigma \left[-\frac{1}{2} \exp\left(-t^2\right)\right]_{-\infty}^{\infty} + \mu\sqrt{\pi}\right) = \frac{\mu\sqrt{\pi}}{\sqrt{\pi}} = \mu$$

2-2 확률 분포

$$\int x e^{x^2}\,dx = \int \frac{e^t}{2}\,dt = \frac{1}{2}\int e^t\,dt = \frac{1}{2}e^t + C = \frac{1}{2}e^{x^2} + C$$

$$t = x^2$$
$$dt = 2x\,dt$$

$$E[x^2] = \int_{-\infty}^{\infty} N(x|\mu,\sigma^2)x^2\,dx = \mu^2 + \sigma^2$$

2-2 확률 분포

$$\mathrm{Var}(x) = E[(x-\mu)^2] = E[x^2 - 2\mu x + \mu^2]$$
$$= E[x^2] - 2\mu E[x] + \mu^2 = E[x^2] - 2\mu^2 + \mu^2$$
$$= E[x^2] - E[x]^2$$
$$= \mu^2 + \sigma^2 - \mu^2 = \sigma^2$$

노이즈가 가우시안 분포를 따른다고 할때,

어떻게 해야 가장 '좋은' w를 찾을 수 있겠는가?

MLE(maximum likelihood estimation)

2-2 확률 분포

◆ 만약, 가우시안분포에서 무작위로 데이터를 추출한다면 μ, σ^2이 주어졌을 때 조건부 확률은 다음과 같다.

$$p(X|\mu,\sigma^2) = \prod_{n=1}^{N} N(x_n|\mu,\sigma^2)$$

$$X = x_1, x_2, \dots, x_N$$

위 가능도 함수(우도)를 최대로 하는 것이 목표.

2-2 확률 분포

우도와 확률은 다르다.

동전던지기에서, 동전의 앞면과 뒷면이 나올 확률이 같다($P_H = P_T = 0.5$)고 가정하면, 동전을 두 번 던졌을 때 앞면이 두 번 나올 확률은 $0.5 \times 0.5 = 0.25$입니다. 이것을 가능도(우도)함수에 관해 쓰면,

$$L(P_H = 0.5|HH) = 0.25$$

위와 같고, HH(두번 앞면)를 관찰했을 때, 동전의 앞면이 나올 확률이 0.5인 likelihood는 0.25라고 말합니다. '$P_H = 0.5$일 확률이 0.25다'는 틀린 말입니다. likelihood는 총 합이 1이 되지 않거나 넘어도 됩니다.

주어진 x값에 대한 t값이 $y(x,w)$를 평균으로 하는 가우시안 분포를 따르면,

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

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

로그함수를 사용함

1. 단조증가 함수이므로 로그를 취한 뒤 최댓값을 구하는 것과 원래 함수의 최댓값을 찾는것은 같다.
2. 곱이 합으로 바뀌므로 중간에 지나치게 작은 값이 들어갔을 때 발생하는 numeric error를 방지

2-2 확률 분포

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

$$= \ln \prod_{n=1}^{N} \frac{\beta}{\sqrt{2\pi}} \exp\left\{-\frac{\beta^2}{2}\left(t - y(x, w)\right)^2\right\}$$

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

결국 $\sum_{n=1}^{N} \{y(x_n, w) - t_n\}^2$을 최소화 하면 된다.

제곱합 오차함수

다음시간

2강

- 선형회귀
  - 선형기저함수모델
  - 편향 분산 분해
  - 베이지안 선형회귀