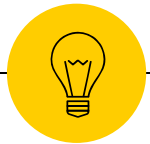
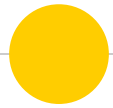


Chapter 2



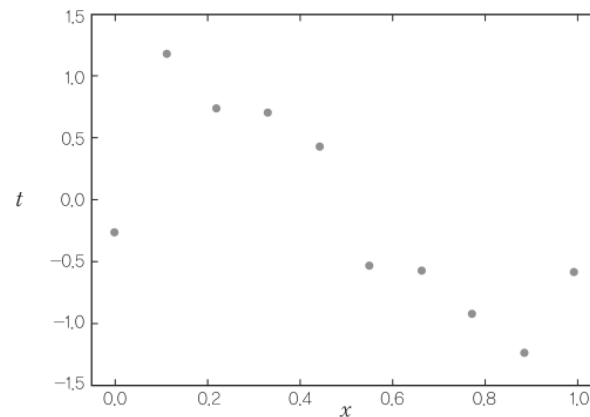
2

최소제곱법 (Least Square Method, LSM)

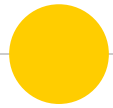


Objective

- By what function are the data points generated?

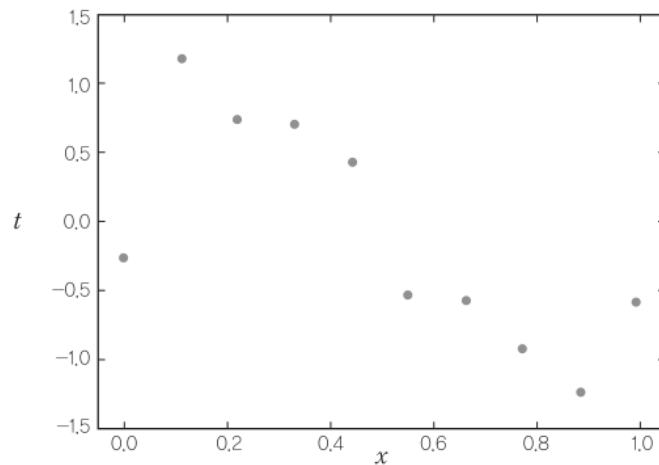


- 1) Learn the function \rightarrow 2) predict the future (build model)
- The function is found in ways to minimize error between the real observed data and the predicted data

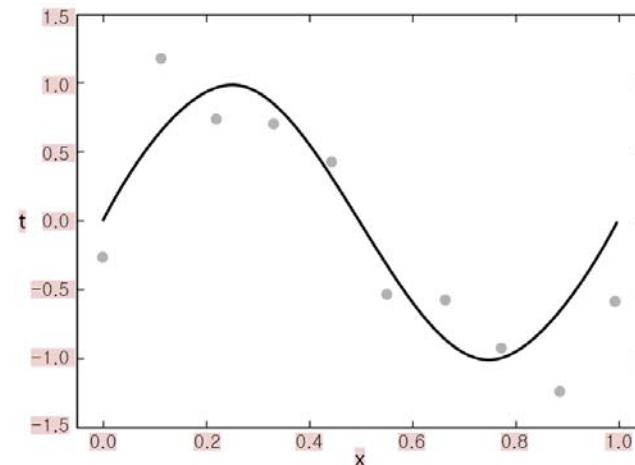


Training set

- Training set is the data set used to learn the function of the given data
- Example
 - Data: 10 data points on a t by x dimension



Learn the data to
infer the function





Training in LSM

Objective:

- How well can we predict t (objective variable, 목적변수) using x (explanatory variable, 설명변수)
- Use multinomial expression (다항식)

$$\begin{aligned} f(x) &= w_0 + w_1x + w_2x^2 + \cdots + w_Mx^M \\ &= \sum_{m=0}^M w_m x^m \end{aligned}$$

Known: x
Unknown: w, M

- Estimate error of prediction
 - Predicted t – observed value x
 - $= \{f(x_1) - t_1\}^2 + \{f(x_2) - t_2\}^2 + \cdots + \{f(x_{10}) - t_{10}\}^2$
- Final goal is to predict the optimal $\{w_m\}_{m=0}^M$



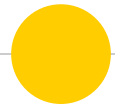
Measuring error of a model

- The error of a learned model is measured by

$$E_D = \frac{1}{2} \sum_{n=1}^N \{f(x_n) - t_n\}^2$$

$$E_D = \frac{1}{2} \sum_{n=1}^N \left\{ \sum_{m=0}^M w_m x^m - t_n \right\}^2$$

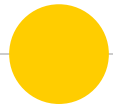
- The difference of predicted and observed value is squared to remove any negative values. Hence the name “Least square method”
- Our objective is to minimize E_D



Minimizing error E_D

- [수학을 배우는 작은방] pg 50
- Differentiate E_D by w , $\frac{\partial E_D}{\partial w} = 0$ ($m = 0, \dots, M$)
 - w is a vector, $w = (w_0, \dots, w_M)^T$
 - $\frac{\partial E_D}{\partial w} = \sum_{m'} w_{m'} \sum_{n=1}^N x_n^{m'} x_n^m - \sum_{n=1}^N t_n x_n^m = 0$
 - If x_n^m represented as $N \times (M+1)$ matrix θ
 - $w^T \theta^T \theta - t^T \theta = 0$
 - $w = (\theta^T \theta)^{-1} \theta^T t$
 - Where $t = (t_1, \dots, t_N)^N$ vector of observed real values

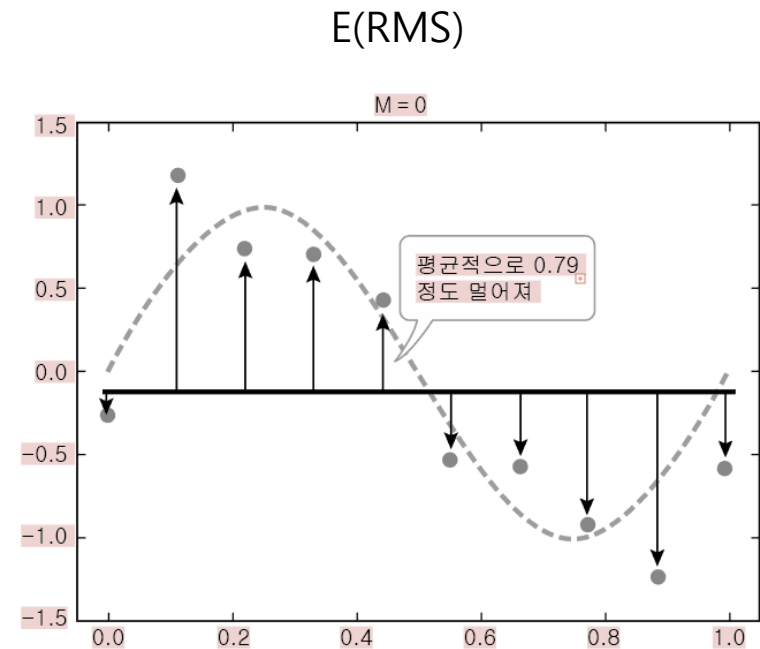
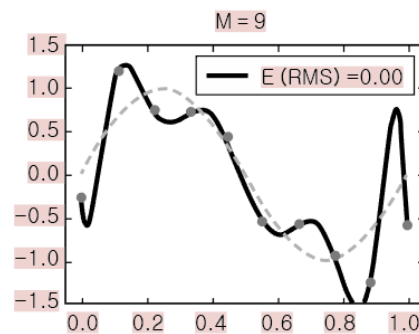
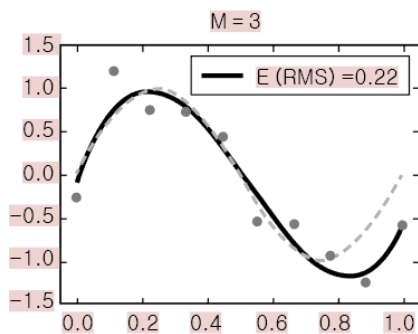
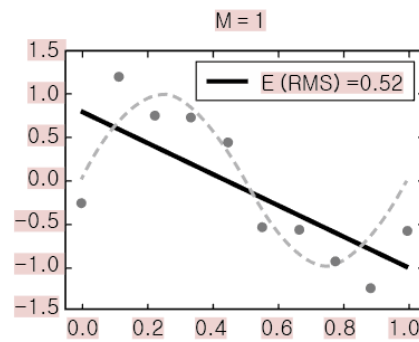
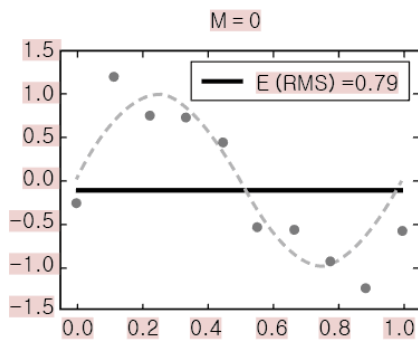
$$\theta = \begin{pmatrix} x_1^0 & x_1^1 & \cdots & x_1^M \\ x_2^0 & x_2^1 & \cdots & x_2^M \\ \vdots & \vdots & \ddots & \vdots \\ x_N^0 & x_N^1 & \cdots & x_N^M \end{pmatrix}$$

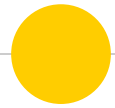


Results

When $M=\{0, 1, 3, 9\}$ the learned functions are as follows

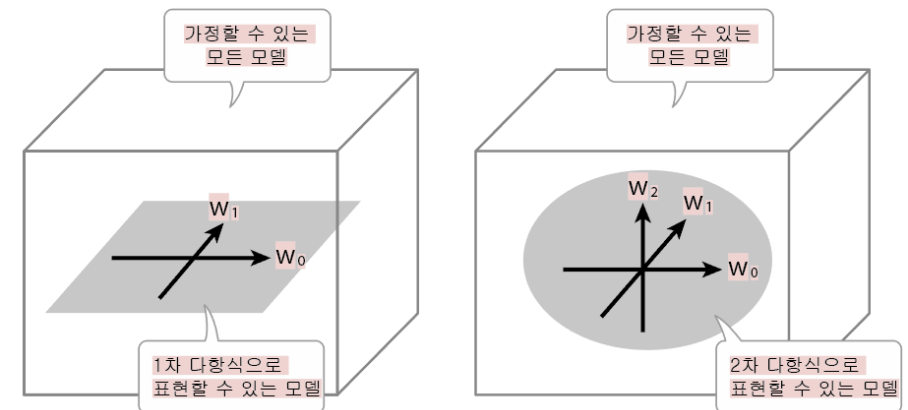
- $E(\text{RMS}-\text{Root Mean Square})=\sqrt{\frac{2E_D}{N}}$, which is the average error of the predicted values





The effect of M selection

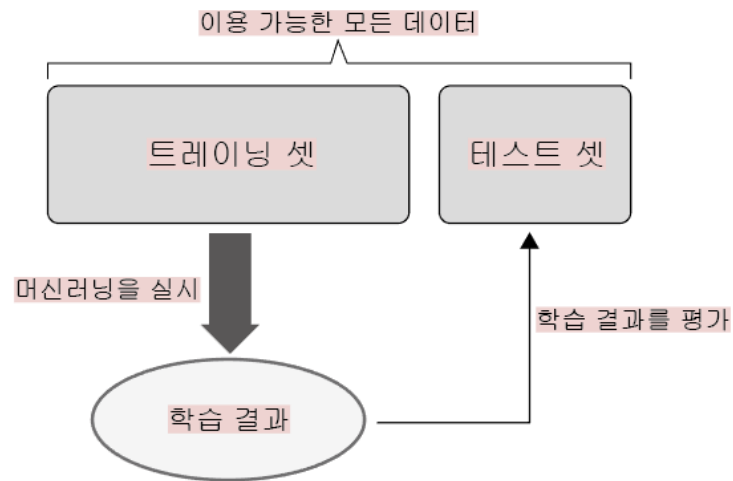
- Too small M may not represent the data well
- A large M may well represent the data
- However, a large M may cause overfitting
 - Overfitting is problem when random error or noise is represented instead of the underlying relationship
 - This is caused by lack of data or excessive use of parameters ($M=9$)
- Very important to select a suitable M
- Use test sets and cross validation

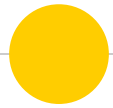




Test set

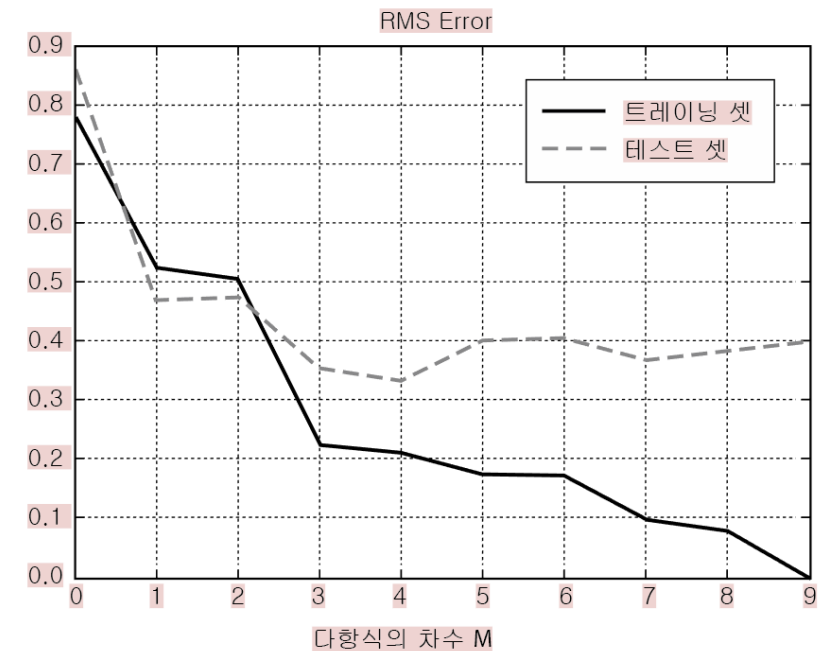
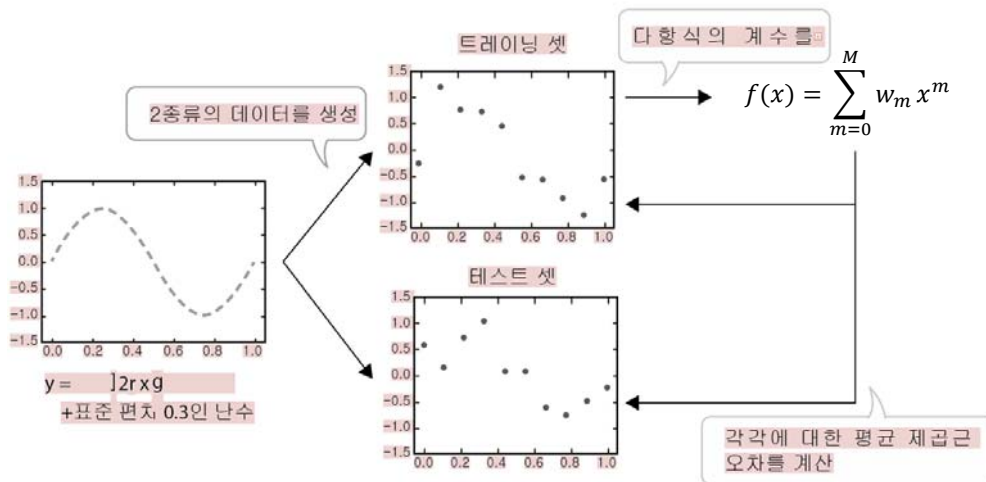
- From the collected data, split it into a training set and test set
- Use the training set to learn the function
- Use the test set to decide if overfitting has occurred

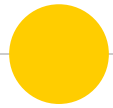




Selecting optimal M

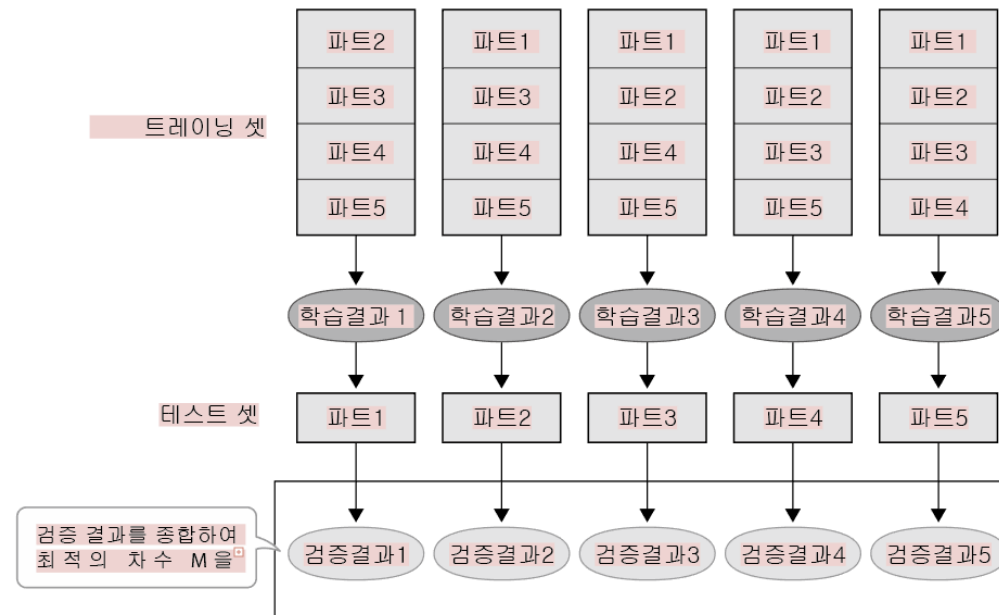
- Measure RMS of the training set and the test set independently
- With $M > 4$, the RMS increases in the test set





Cross validation (교차검증)

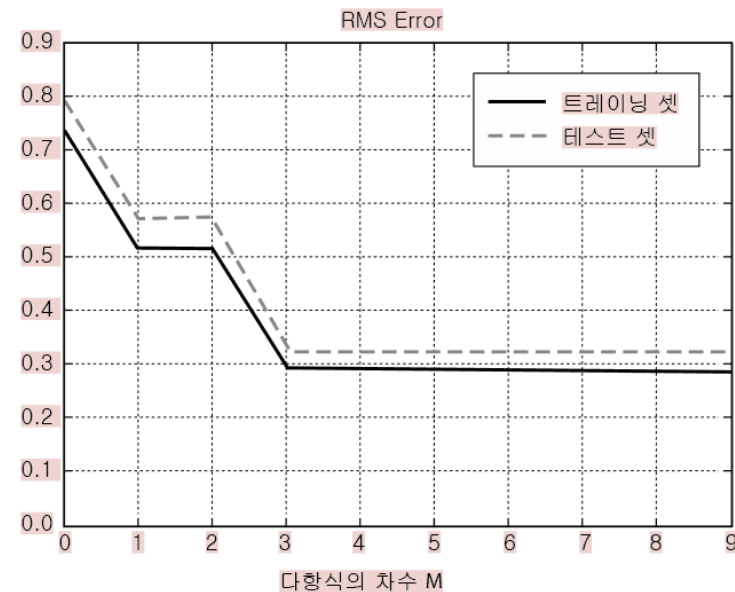
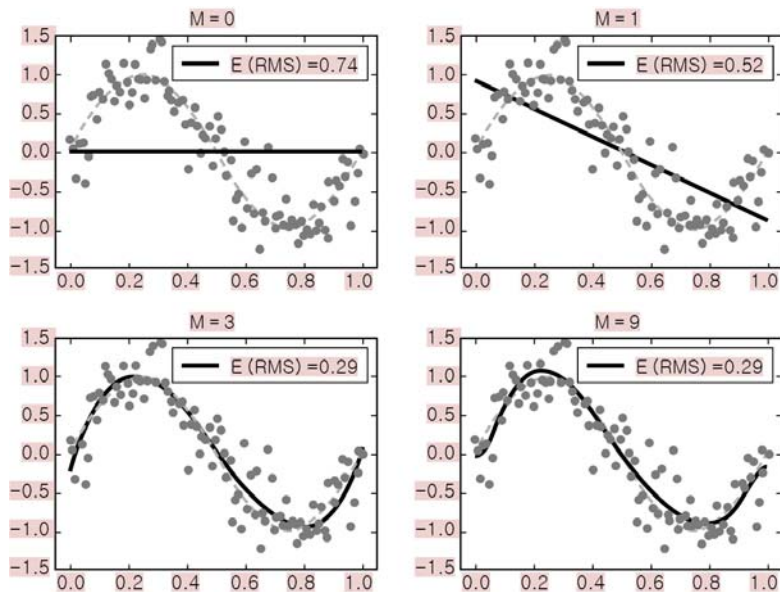
- Repeat the procedure of training/test set measurement by dividing the data set into K parts
- For each k , select k as test set and the remaining k' as training sets and measure RMS
- Select the best M , in our example it is $M=4$



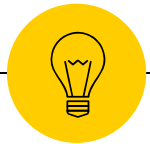


Discussion on overfitting

- If data is large enough, we may choose large M without encountering the overfitting problem
- So the more data we have, the higher chance is that LSM will model the underlying relationship instead of random error or noise
- That's why big data is important but difficult to analyze

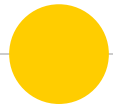


Chapter 3



3

최우추정법 (Maximum Likelihood Estimation)



Probability(확률)

- ⦿ The measure of the likelihood that an event will occur.
- ⦿ Example: normal 6-sided dice.
 - Outputs: 1, 2, 3, 4, 5, 6
 - Events: 1; 6; even
 - Probability:
 - $P(A) = \frac{\text{\# of ways event can occur}}{\text{\# of possible outcomes}}$
 - $P(1) = P(6) = 1/6$
 - $P(\text{even}) = 3/6 = 1/2$



Probability Distribution(확률 분포)

- ◉ Probability distribution: the assignment of a probability to each outcome.
 - Sum of the probabilities of all possible outcomes must be 1.
 - $\sum P(x) = 1$
- ◉ Example: normal 6-sided dice.
 - Outcomes: 1, 2, 3, 4, 5, 6
 - Probability distribution:
 - $P(x) = \frac{1}{6}, x = \{1, \dots, 6\}$
 - $\sum P(x) = 1$



Conditional Probability(조건부확률)

- $P(i|\theta)$: the probability of event i under the condition θ .
- Example: two dices D_1, D_2
 - $P(2|D_1)$: probability of picking 2 when using dice D_1 .
 - $P(2|D_2)$: probability of picking 2 when using dice D_2 .



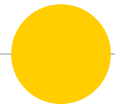
Joint Probability(결합확률)

- $P(X, Y)$: the probability of event X and Y.
- $P(X, Y) = P(X|Y)P(Y) = P(Y|X)P(X)$
- X and Y are *independent*(독립) if and only if $P(X, Y) = P(X)P(Y)$, $P(X) \neq 0, P(Y) \neq 0$.
- Example: for two dices D_1, D_2 , we first pick one dice and rolling the dice.
 - $P(D)$: probability of picking dice D_1 or D_2 .
 - Ex) $P(D_1) = 0.3, P(D_2) = 0.7$
 - $P(2|D_2)$: probability of picking 2 when using dice D_2 .
 - $P(1, D_1)$: probability of getting 1 while rolling dice D_1 .
 - $= P(1|D_1)P(D_1)$



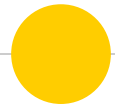
Independent Identically Distribution (독립동일분포)

- Independent Identically Distribution (i.i.d.):
 - Each random variable has the same probability distribution as the others and all are mutually independent.
- Example: When we get 1, 2 and 6 as the result of rolling normal dice three times.
 - All outputs have same probability distribution as we use the same dice for all trials.
 - All trials do not affect each other.
 - Getting 2 from second trial cannot be affected by the output 1 from previous trial or output 6 from later trial.
 - Therefore, the outputs (1, 2, 6) are independent identically distributed.



Probabilistic Model(확률 모델)

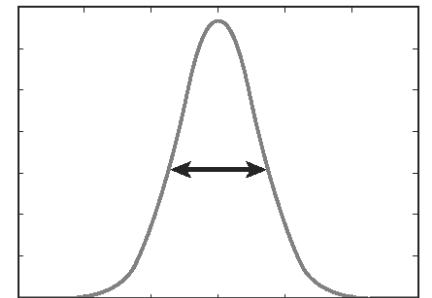
- ◉ parametric vs. non-parametric
- ◉ 3 steps of parametric model.
 1. Define model with parameters.
 2. Define the evaluation criterion.
 3. Find the best fit parameters according to evaluation criterion.



Normal Distribution(정규분포)

- Normal distribution (Gaussian distribution) is most widely used distribution because of central limit theorem.
 - Central limit theorem establishes that when independent random variables are added, their sum tends toward a normal distribution even if the original variables themselves are not normally distributed.
- The probability distribution is defined as

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





Remind: Least squares(최소제곱법)

◉ Describing Least square estimation as the 3 steps of parametric model.

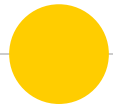
1. Define model.

- Data: $D = \{(x_n, t_n)\}_{n=1}^N$
- Parameters: $\theta = \{w_m\}_{m=0}^M$
- Model: $f(x; \theta) = \sum_{m=0}^M w_m x^m$ (M+1 차 다항식)

2. Define evaluation criterion.

- $E_D = \frac{1}{2} \sum_{n=1}^N \{f(x_n) - t_n\}^2$
- Criterion: $\underset{\theta}{\operatorname{argmin}} E_D$

3. Find parameters.



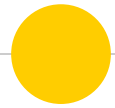
Likelihood function(우도 함수)

● Likelihood function

- Function of the parameters of a model given data which is identical to the probability of observed data given parameters.
- $L(\theta; x_1, x_2, \dots, x_n) = P(x_1, x_2, \dots, x_n | \theta)$
- If x_1, x_2, \dots, x_n are i.i.d from $P(x | \theta)$
- $L(\theta; x_1, x_2, \dots, x_n) = P(x_1, x_2, \dots, x_n | \theta) = P(x_1 | \theta) P(x_2 | \theta) \dots P(x_n | \theta) = \prod_{i=1}^n P(x_i | \theta)$

● Log likelihood function

- Since it needs a lot of multiplication of probabilities (which are usually quite small), the likelihood is likely to get very small.
- Log likelihood can alleviate this computational problem.
- $l(\theta; x_1, x_2, \dots, x_n) = \log(L(\theta; x_1, x_2, \dots, x_n)) = \sum_{i=1}^n \log(P(x_i | \theta))$



Maximum Likelihood Estimation(최우추정법)

- ⦿ If we assume the quality of observed data is sufficiently good and not biased, the observed data is the most likely to be generated from the model.
- ⦿ Therefore, maximum likelihood estimation find the parameters by maximizing the likelihood function for given observed data.



Maximum Likelihood Estimation(최우추정법)

- ◉ When we do not know parameters?
- ◉ 3 steps of parametric model.
 1. Define model.
 - Data: $D = \{(x_n, t_n)\}_{n=1}^N$
 - Parameters: θ
 - Probability distribution: $P(t_n|\theta)$ (임의의 모델)
 2. Define evaluation criterion.
 - $L(\theta; t_1, t_2, \dots, t_N) = P(t_1, t_2, \dots, t_N|\theta)$
 - Criterion: $\underset{\theta}{\operatorname{argmax}} L(\theta; t_1, t_2, \dots, t_N)$ or $\underset{\theta}{\operatorname{argmax}} \log(L(\theta; t_1, t_2, \dots, t_N))$
 3. Find parameters

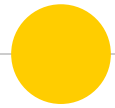


Maximum Likelihood Estimation(최우추정법)

- When we know parameters of models
- Example: two dices D_1 , D_2 whose probability distributions are
 - $P(1|D_1) = P(2|D_1) = \dots = P(6|D_1) = \frac{1}{6}$
 - $P(1|D_2) = P(2|D_2) = \dots = P(5|D_2) = \frac{1}{12}, P(6|D_2) = \frac{7}{12}$.

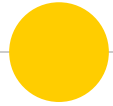
If one dice is randomly selected and the outputs of three rolling trials are (6, 6, 6), which dice is more likely to be selected?

- $L(\theta_1; 6,6,6) = P(6,6,6|\theta_1) = 3P(6|\theta_1) = 3(\frac{1}{6})^3 = 3(\frac{2}{12})^3$
 - $L(\theta_2; 6,6,6) = P(6,6,6|\theta_2) = 3P(6|\theta_2) = 3(\frac{7}{12})^3$
- Dice D_2 is more likely to be selected.



Maximum Likelihood Estimation(최우추정법)

- Example of section 3.2.1 in textbook.
- Finding parameters from given data assuming that the model is *normal distribution*.
 - Data: $D = \{(x_n, t_n)\}_{n=1}^N$
 - Parameters: $\theta = \{\mu, \sigma\}$
 - Probability distribution: $P(t_n|\theta) = N(t_n|\mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(t_n-\mu)^2}$
 - Likelihood function: $L(\theta; t_1, t_2, \dots, t_N) = P(t_1, t_2, \dots, t_N|\theta) = \prod_{n=1}^N P(t_n|\theta) = \prod_{n=1}^N N(t_n|\mu, \sigma) = \prod_{n=1}^N \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(t_n-\mu)^2}$



Maximum Likelihood Estimation(최우추정법)

- Find parameters [P.93 수학을 배우는 작은 방 참조]

- Find μ ? $\rightarrow \underset{\mu}{\operatorname{argmax}} L(\theta; t_1, t_2, \dots, t_N) = \underset{\mu}{\operatorname{argmax}} \log(L(\theta; t_1, t_2, \dots, t_N))$

- The value which satisfies $\frac{\partial L}{\partial \mu} = 0$.

- $$l = \log(L(\theta; t_1, t_2, \dots, t_N)) = \log\left(\prod_{n=1}^N \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(t_n - \mu)^2}\right) = \log\left[\left(\frac{1}{2\pi\sigma^2}\right)^{\frac{N}{2}} e^{-\frac{1}{2\sigma^2} \sum_{n=1}^N (t_n - \mu)^2}\right]$$

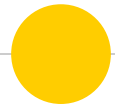
$$= -\frac{N}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{n=1}^N (t_n - \mu)^2$$

- $$\frac{\partial l}{\partial \mu} = \frac{1}{\sigma^2} \sum_{n=1}^N (t_n - \mu) = \frac{1}{\sigma^2} \sum_{n=1}^N t_n - N\mu = 0$$

- $$\therefore \hat{\mu} = \frac{1}{N} \sum_{n=1}^N t_n$$

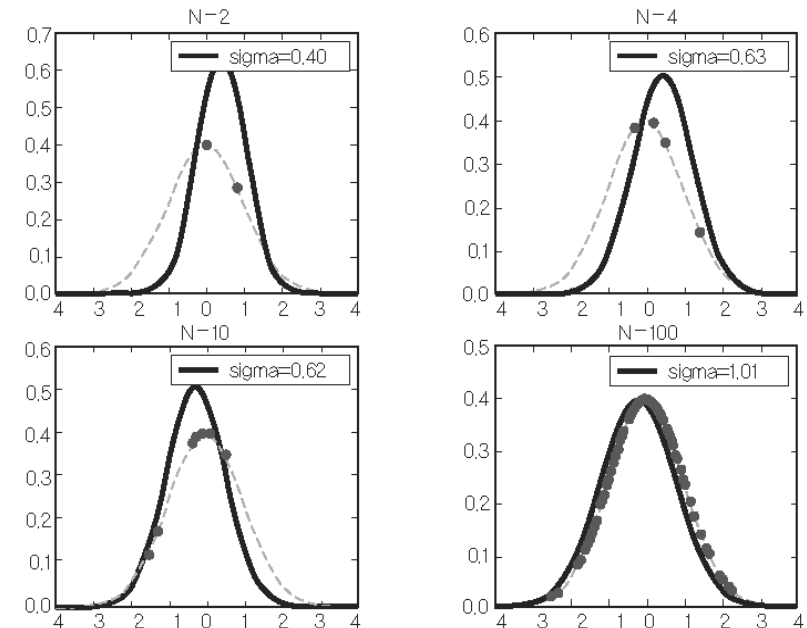
- TODO: How to find σ ?

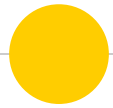
- The mean of observed data is estimated as the mean of model.



Maximum Likelihood Estimation(최우추정법)

- How well the model is estimated as the number of observed data increases.
- σ is harder to estimate for small N since values far from mean are hard to be observed for small N .





Maximum Likelihood Estimation(최우추정법)

- Example of section 3.1 in textbook.
- Finding parameters from given data whose probability distribution is defined as below.

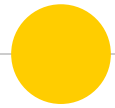
- Data: $D = \{(x_n, t_n)\}_{n=1}^N$

- Parameters: $\theta = \{\{w_m\}_{m=0}^M, \sigma\}$

- Probability distribution: $P(t_n|\theta) = N(t_n|f(x_n), \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(t-f(x_n))^2}$

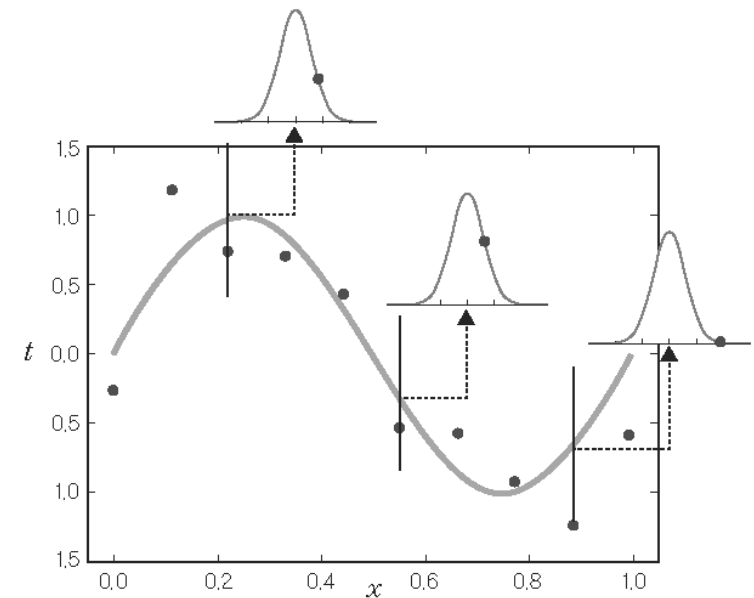
$$\text{where } f(x_n) = \sum_{m=0}^M w_m x_n^m$$

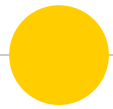
- Likelihood function: $L(\theta; t_1, t_2, \dots, t_N) = P(t_1, t_2, \dots, t_N|\theta) = \prod_{n=1}^N P(t_n|\theta) = \prod_{n=1}^N N(t_n|f(x_n), \sigma) = \prod_{n=1}^N \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(t-f(x_n))^2}$



Maximum Likelihood Estimation(최우추정법)

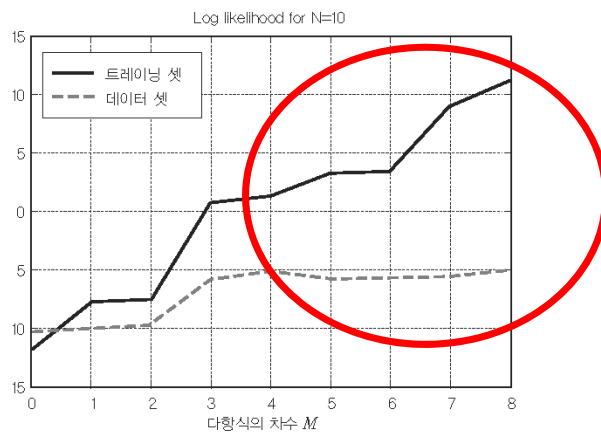
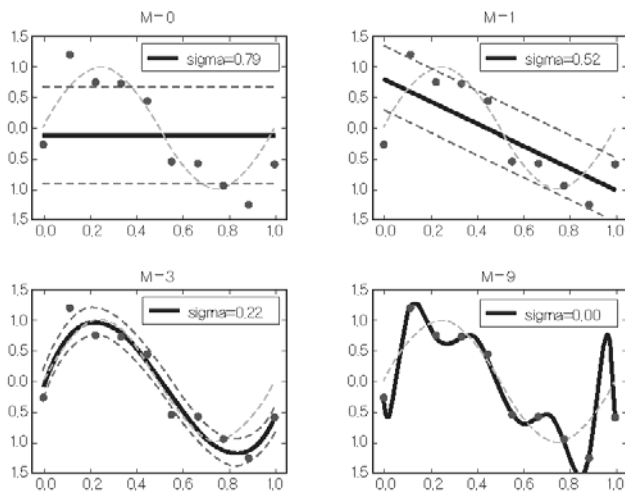
- How this distribution looks like?
- Mean points of t each x is following polynomial $f(x_n) = \sum_{m=0}^M w_m x^m$.
- For each point x , output values follow normal distribution $N(t|f(x), \sigma)$.



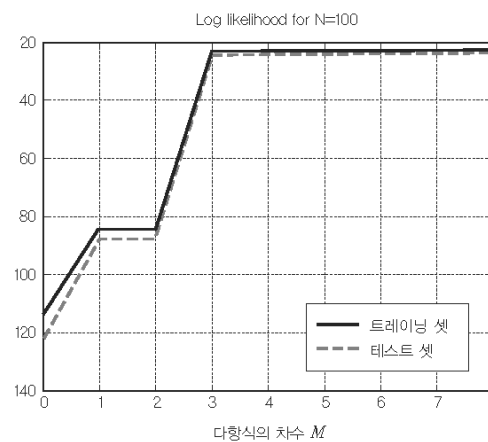
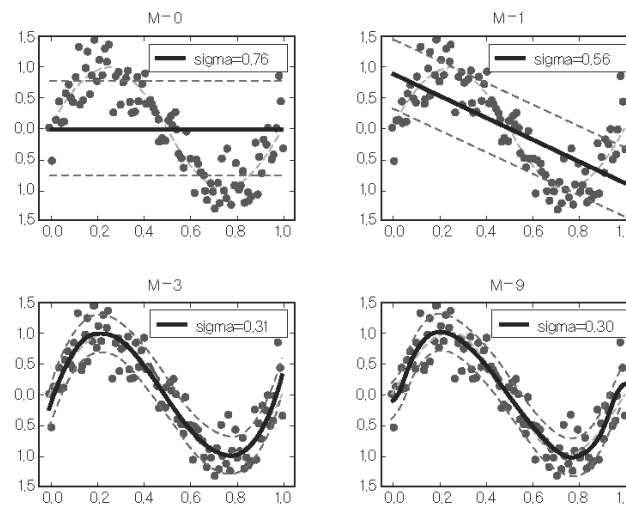


Overfitting(오버피팅)

$N = 10$



$N = 100$



The model can be overfitted if model is too complex but the number of data is not sufficient.