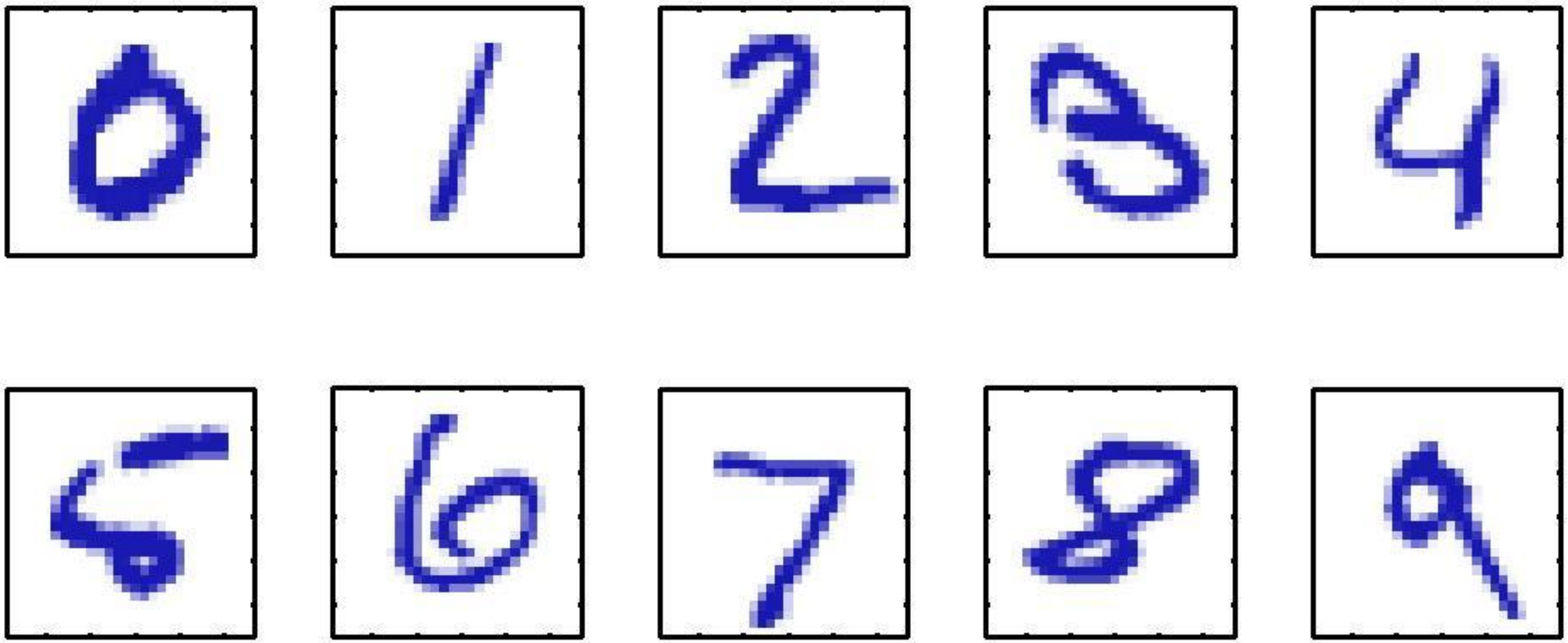
PATTERN RECOGNITION AND MACHINE LEARNING

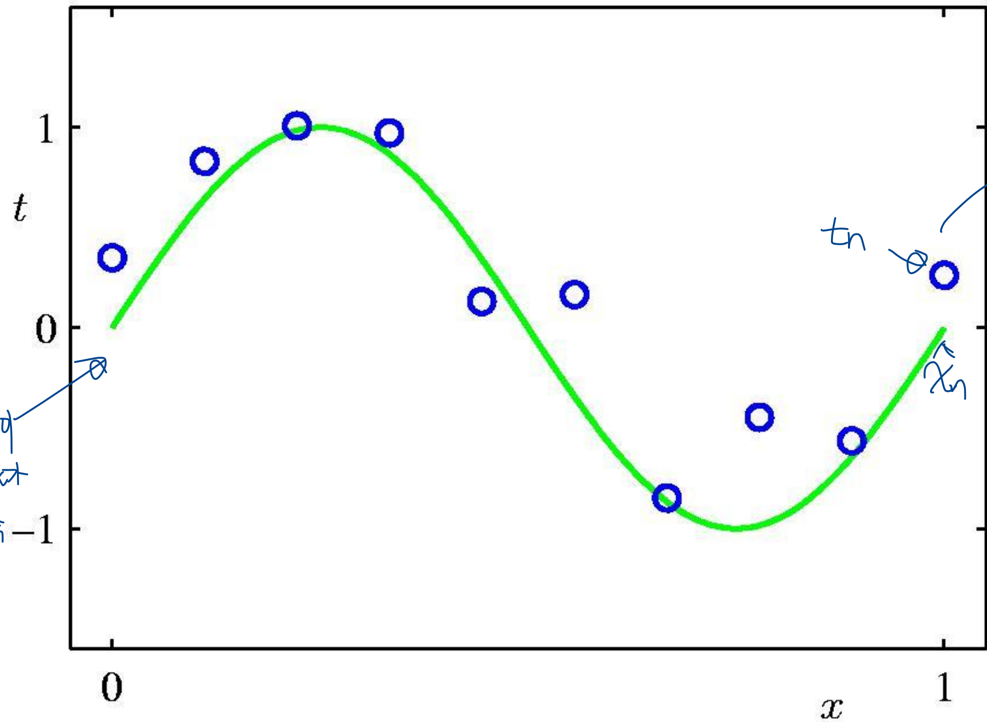
CHAPTER 1: INTRODUCTION

Example

Handwritten Digit Recognition



Polynomial Curve Fitting

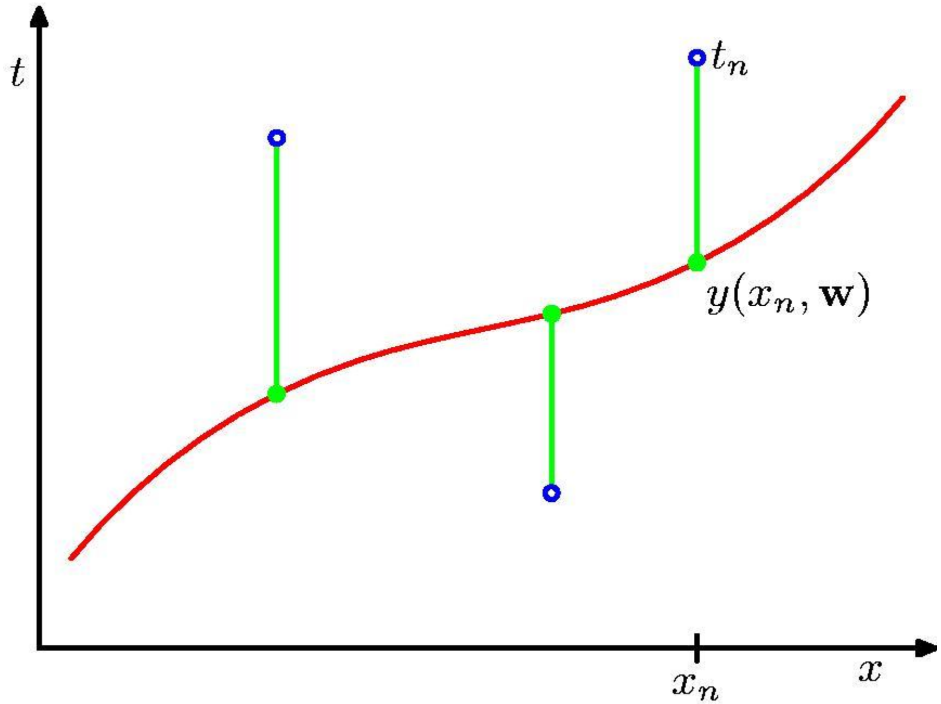


$\sin(2\pi x)$ with noise.

녹색커브 에 대한 지식이 없는 상태에서 새로운 값 x가 주어졌을때 출력값 t를 예측

$$y(x, \mathbf{w}) = w_0 + w_1 x + w_2 x^2 + \ldots + w_M x^M = \sum_{j=0}^{M} w_j x^j$$

Sum-of-Squares Error Function



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

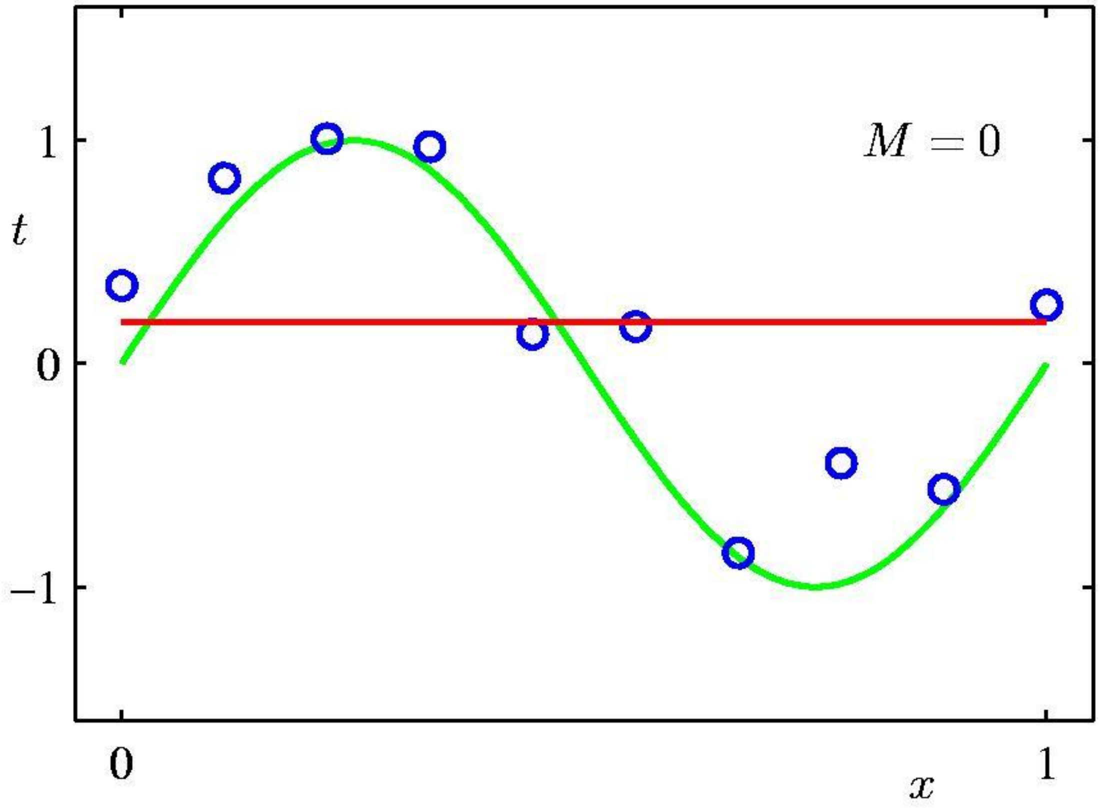
E(w)를 최소화하는 w 선택, 미분을 통해 w에 대한 선형식에서 E(w)를 최소화하는 w^*를 찾을수있다.

0th Order Polynomial

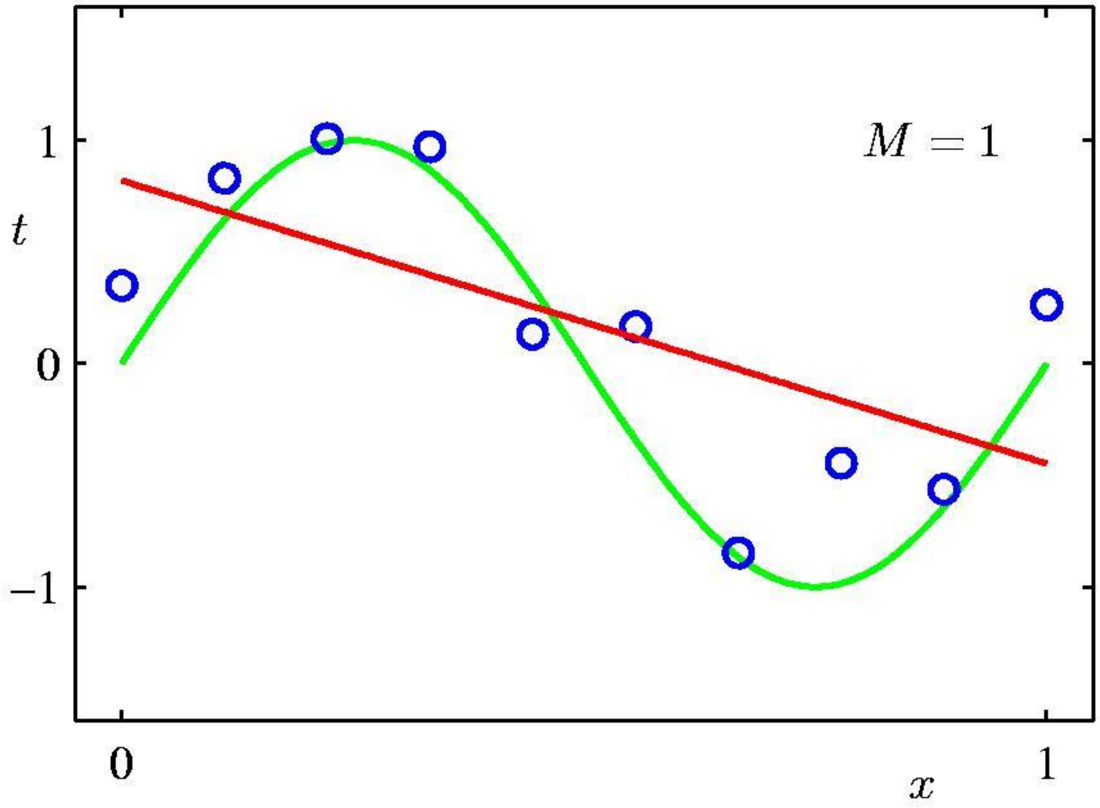
모델결정

과적합(over-fitting)

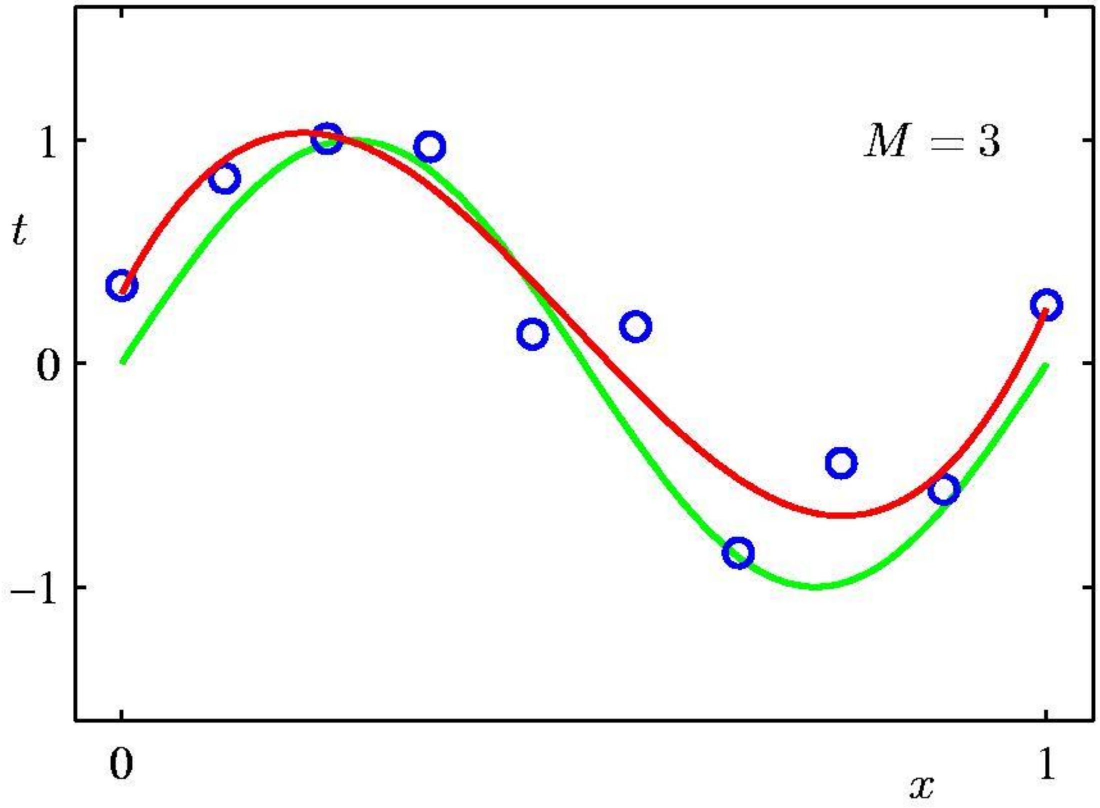
과소적합(under-fitting)



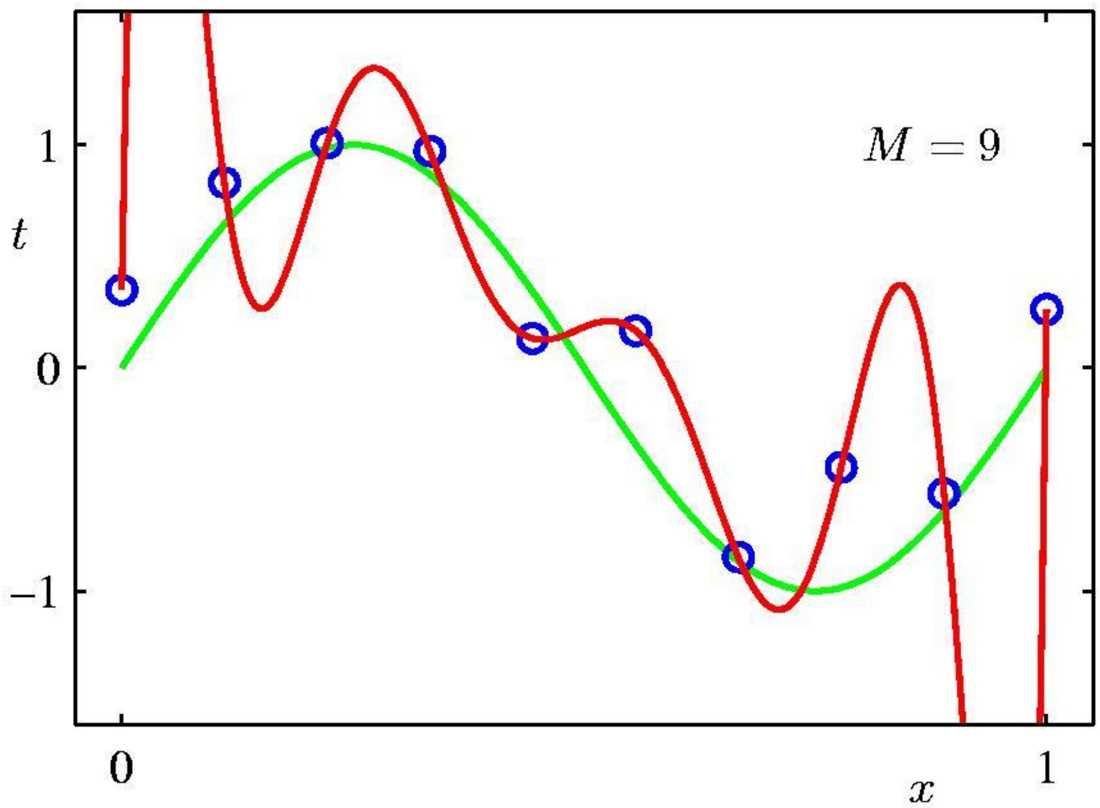
1st Order Polynomial



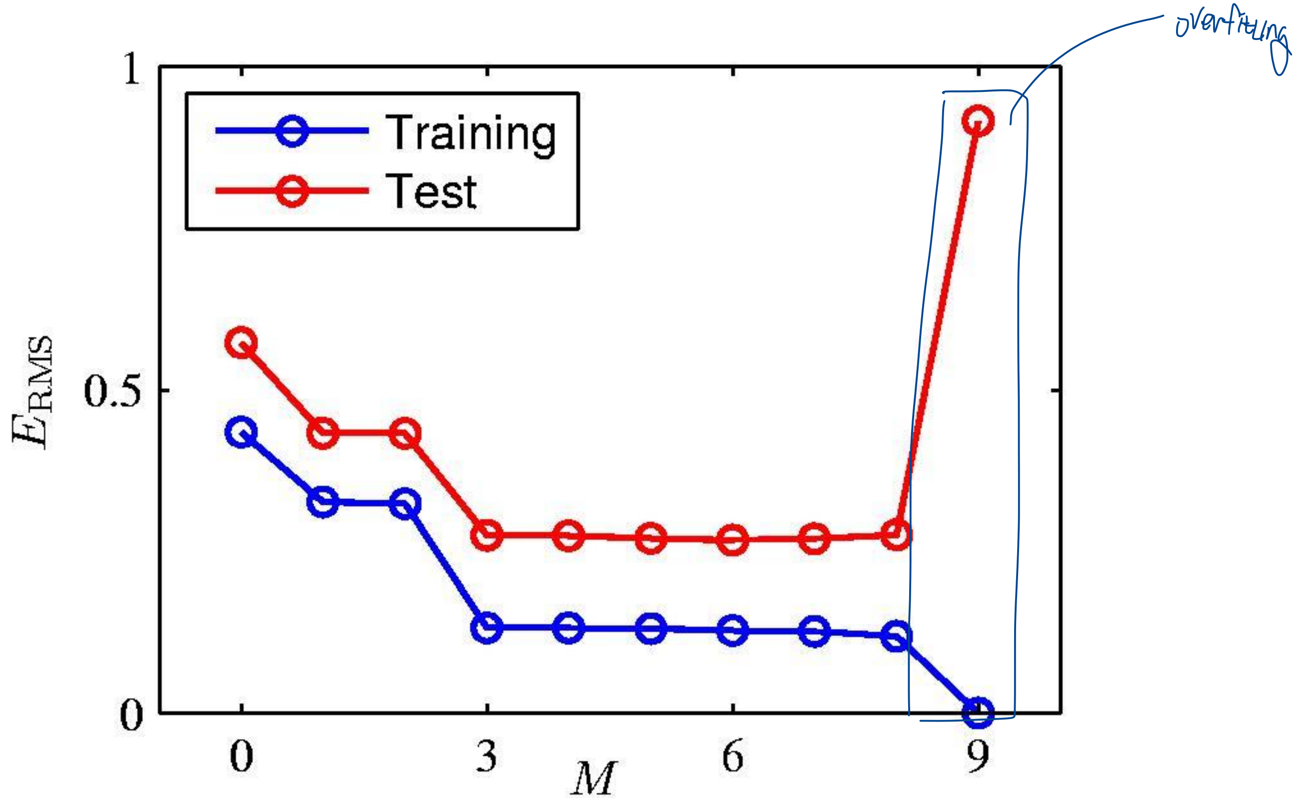
3rd Order Polynomial



9^{th} Order Polynomial



Over-fitting



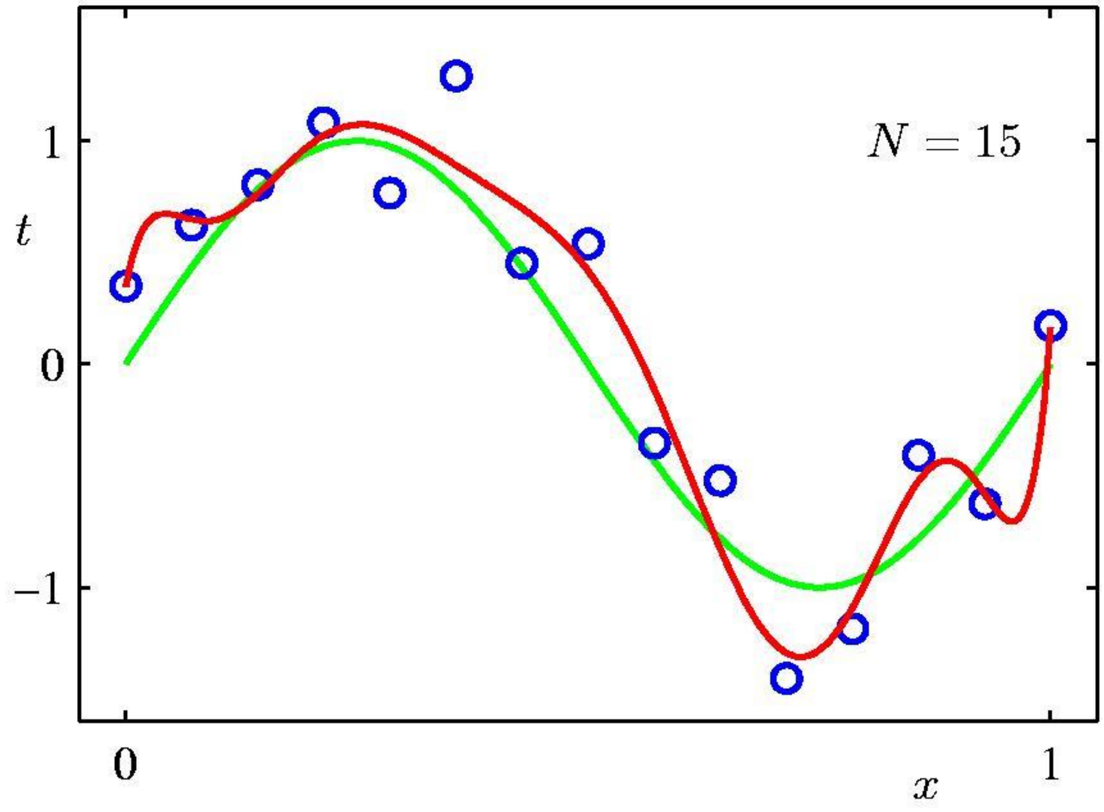
Root-Mean-Square (RMS) Error: $E_{\mathrm{RMS}} = \sqrt{2E(\mathbf{w}^\star)/N}$

Polynomial Coefficients

	$M=0$	$M=1$	$M=3$	$M=9$
$w_0^\star$	0.19	0.82	0.31	0.35
$w_1^\star$		-1.27	7.99	232.37
$w_2^\star$			-25.43	-5321.83
$w_3^\star$			17.37	48568.31
$w_4^\star$				-231639.30
$w_5^\star$				640042.26
$w_6^\star$				-1061800.52
$w_7^\star$				1042400.18
$w_8^\star$				-557682.99
$w_9^\star$				125201.43

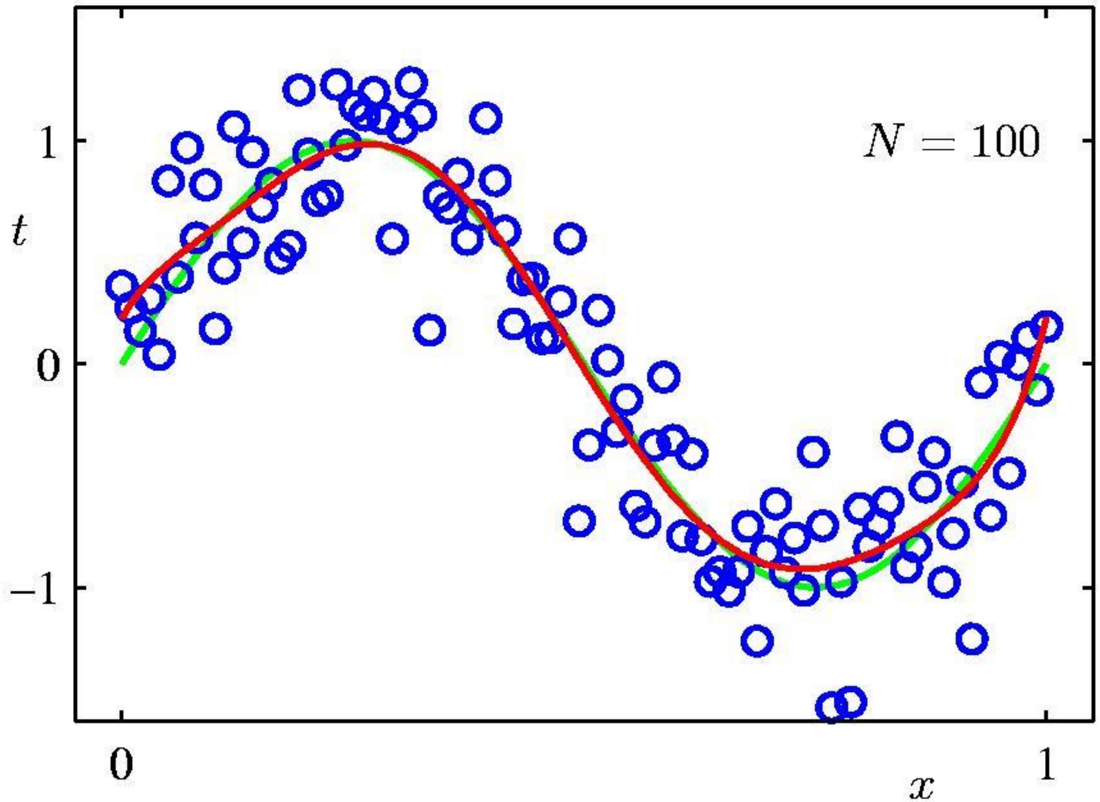
Data Set Size: $N = 15$

9th Order Polynomial



Data Set Size: $N = 100$

9th Order Polynomial 데이터를 늘리는게 과적합에 도움이된다.



Regularization 정규화

Penalize large coefficient values

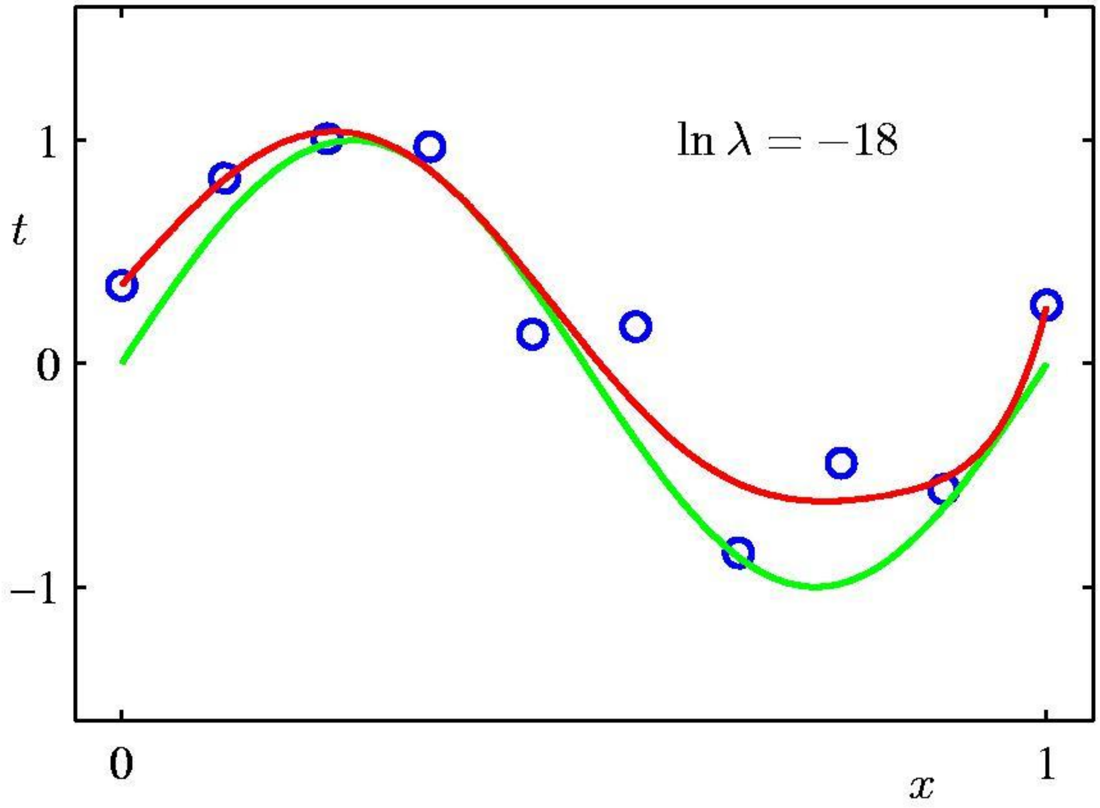
a. also called, weight decay
가중치 감쇠.

$$\widetilde{E}(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n,\mathbf{w}) - t_n\}^2 + \underbrace{\frac{\lambda}{2}\|\mathbf{w}\|^2}_{\text{정규화항}}$$

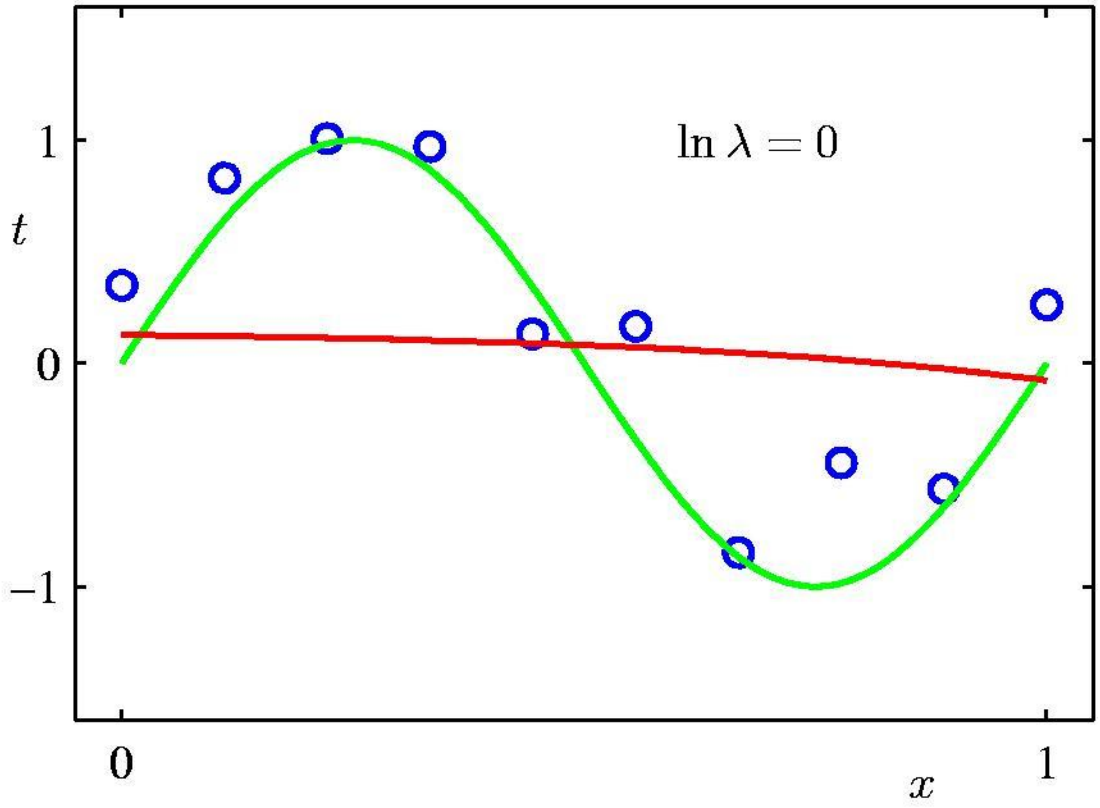
↳ 이차항 → ridge 회귀

cf. 일차항 → 라쏘회귀

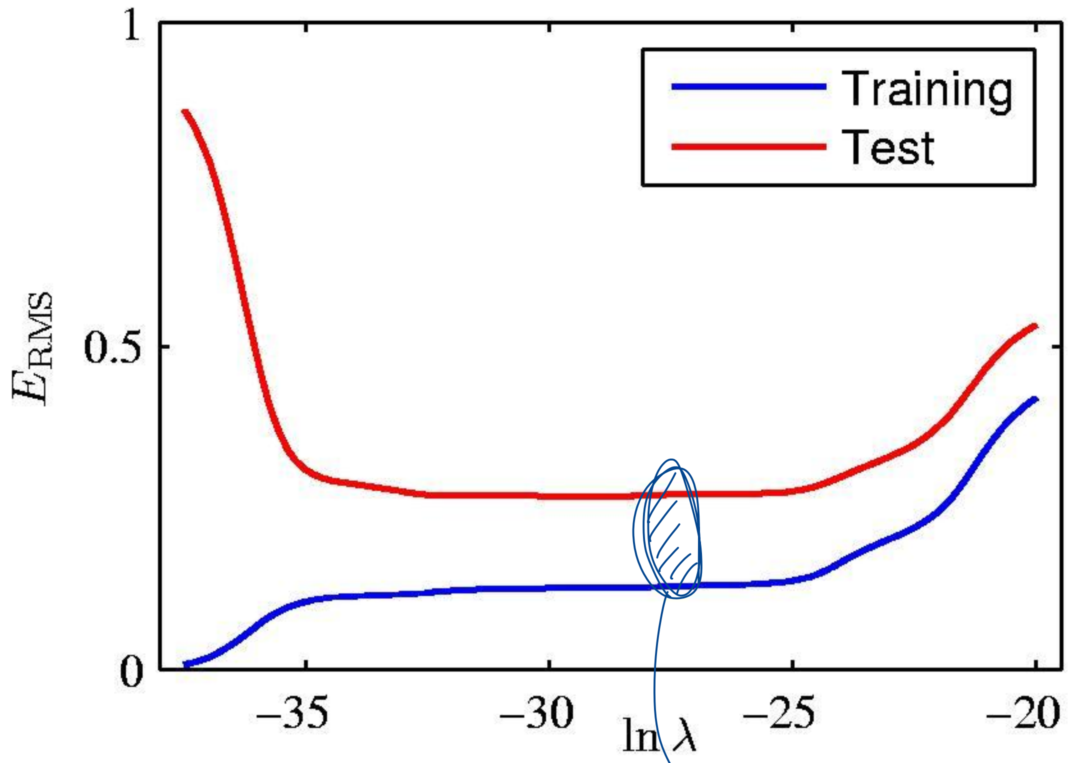
Regularization: $\ln \lambda = -18$



Regularization: $\ln \lambda = 0$



Regularization: E_{RMS} vs. $\ln \lambda$



올바른 hyperparameter를 찾아야 한다.

Polynomial Coefficients

	$\ln \lambda = -\infty$	$\ln \lambda = -18$	$\ln \lambda = 0$
$w_0^\star$	0.35	0.35	0.13
$w_1^\star$	232.37	4.74	-0.05
$w_2^\star$	-5321.83	-0.77	-0.06
$w_3^\star$	48568.31	-31.97	-0.05
$w_4^\star$	-231639.30	-3.89	-0.03
$w_5^\star$	640042.26	55.28	-0.02
$w_6^\star$	-1061800.52	41.32	-0.01
$w_7^\star$	1042400.18	-45.95	-0.00
$w_8^\star$	-557682.99	-91.53	0.00
$w_9^\star$	125201.43	72.68	0.01

Probability Theory

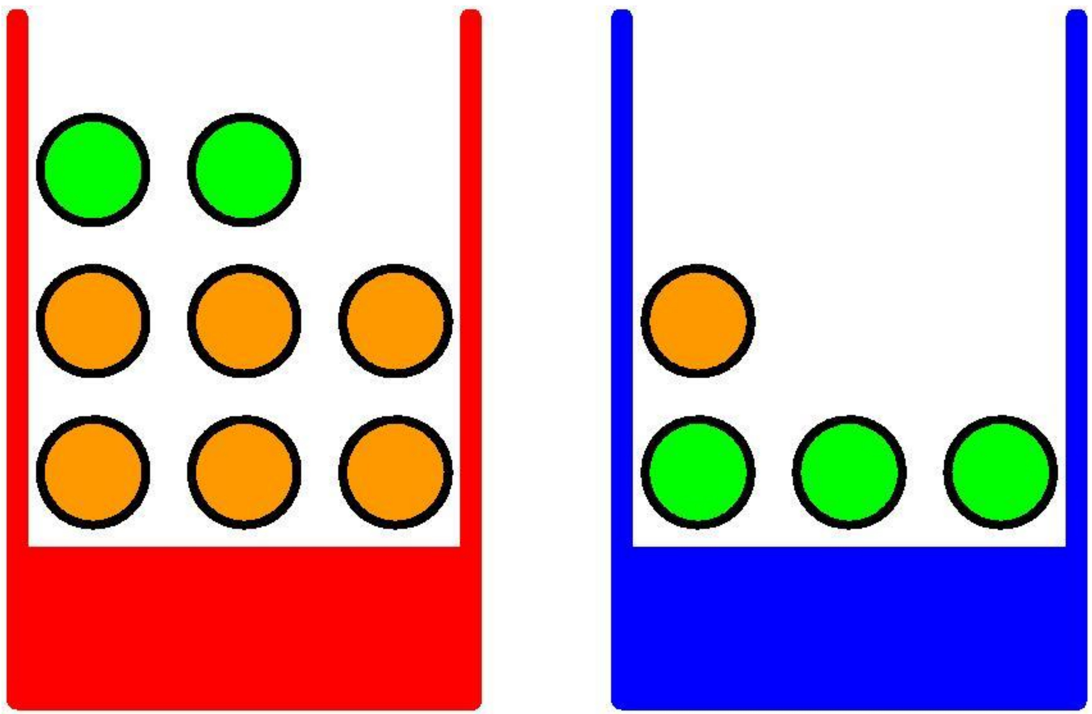
불확실성 ⇒ 확률을 통한 해소.

Apples and Oranges

B → 상자에 대한 확률변수
F → 과일을 고를 확률.

· 복원추출

· 과일의 개별확률은
모두 동일.

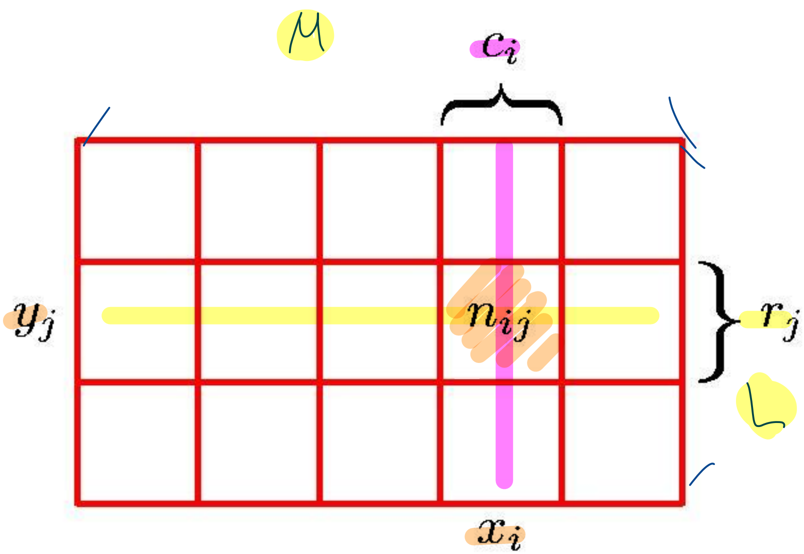


각 상자를 고를 확률.

40%

60%

Probability Theory



Marginal Probability

$$p(X = x_i) = \frac{c_i}{N}.$$

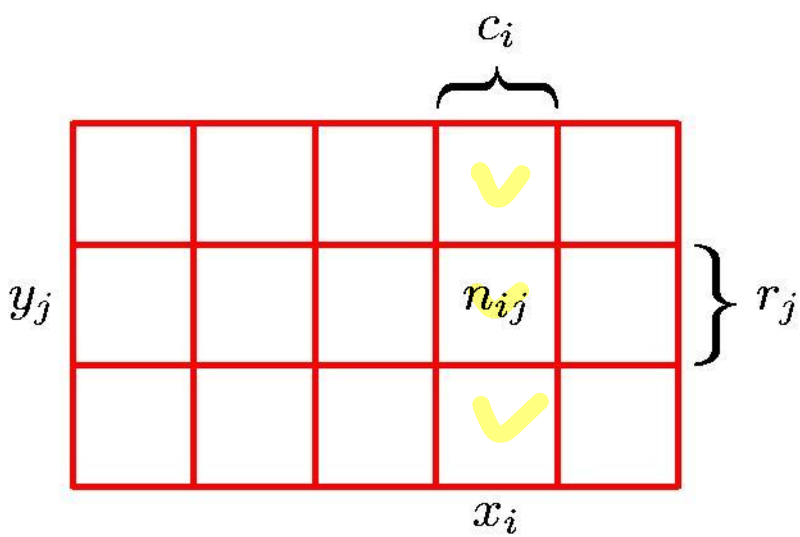
Joint Probability

$$p(X = x_i, Y = y_j) = \frac{n_{ij}}{N}$$

Conditional Probability

$$p(Y = y_j | X = x_i) = \frac{n_{ij}}{c_i}$$

Probability Theory



Sum Rule

$$p(X = x_i) = \frac{c_i}{N} = \frac{1}{N}\sum_{j=1}^{L} n_{ij}$$

$$= \sum_{j=1}^{L} p(X = x_i, Y = y_j)$$

Product Rule

$$\begin{aligned} p(X = x_i, Y = y_j) &= \frac{n_{ij}}{N} = \frac{n_{ij}}{c_i} \cdot \frac{c_i}{N} \\ &= p(Y = y_j | X = x_i)p(X = x_i) \end{aligned}$$

$$\frac{P(y_i \cap x_i)}{\cancel{P(X=x_i)}} \cdot \cancel{P(X=x_i)}$$

The Rules of Probability

주변확률(marginal probability)

Sum Rule $$p(X) = \sum_Y p(X, Y)$$

결합확률(joint probability)

Product Rule $$p(X, Y) = \underbrace{p(Y|X)}_{\text{조건부확률(conditional probability)}}\, p(X)$$

Bayes' Theorem

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

$$p(X) = \sum_Y p(X|Y)p(Y)$$

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

posterior ∝ likelihood × prior

cf. 독립

$P(X,Y) = P(X)P(Y)$

각각의 상자가 같은수의 사과와 오렌지

→ 사과와 오렌지를 고를 확률은 어떤 상자를 골랐는지와 관계없다.

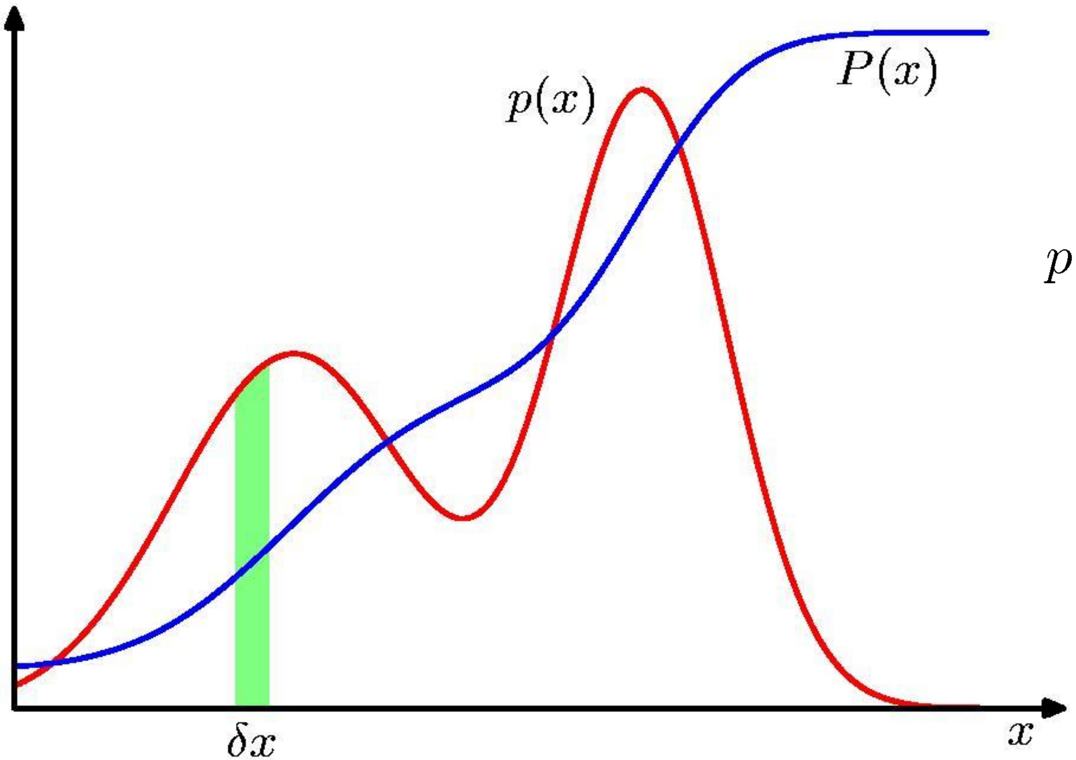
과일예제에서,

→ 어떤과일이 선택되었는지 알기전에 어떤 박스를 선택했는가? → $P(B)$

$\frac{4}{10}$ vs $\frac{6}{10}$

$P(B|F) \propto P(F|B) \times P(B)$

Probability Densities 연속적인 변수에서의 확률



$$p(x \in (a, b)) = \int_a^b p(x)\,\mathrm{d}x$$

$$P(z) = \int_{-\infty}^{z} p(x)\,\mathrm{d}x$$

cdf (누적분포함수)

$P'(x) = p(x)$

e.g. 연속에서도,

$p(x) = \int p(x, y)\, dy$

$p(x, y) = p(y|x)\, p(x)$

$$p(x) \geqslant 0 \qquad \int_{-\infty}^{\infty} p(x)\,\mathrm{d}x = 1$$

Transformed Densities

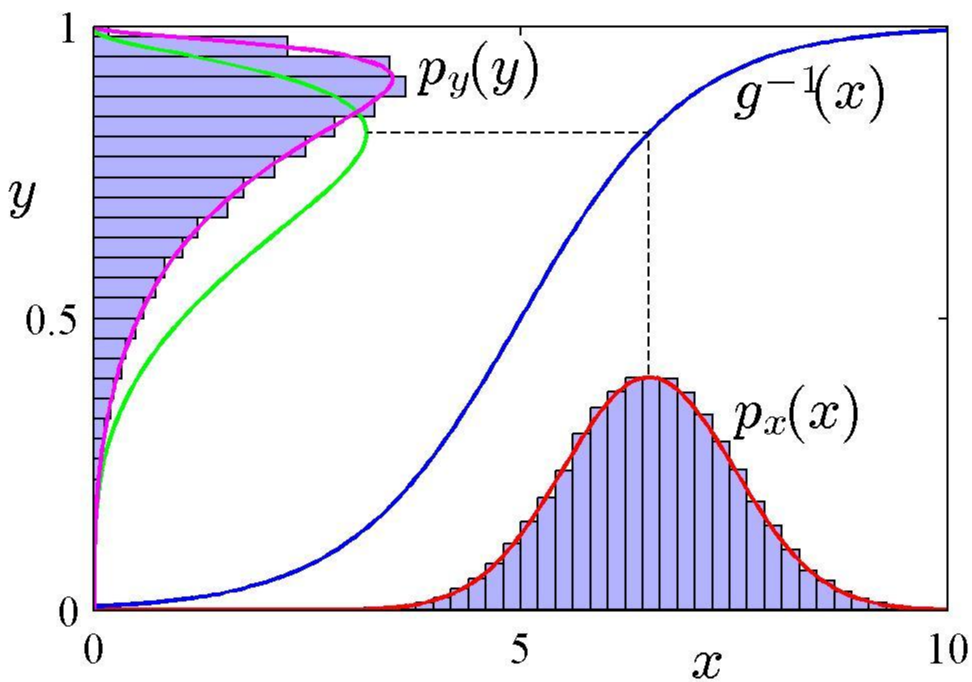
$x = g(y)$

$f(x) = f(g(y))$

$(x, x+\delta x)$

$(y, y+\delta y)$

$p_x(x)\delta x \simeq p_y(y)\delta y$



$$
\begin{aligned}
p_y(y) &= p_x(x)\left|\frac{\mathrm{d}x}{\mathrm{d}y}\right| \\
&= p_x(g(y))\,|g'(y)|
\end{aligned}
$$

확률 분포의 최대값은 어떤 변수를 선택하는지에 따라서 달라질수 있다.

Expectations

〈이산〉

확률 분포 　 어떤 함수

$$\mathbb{E}[f] = \sum_x p(x) f(x)$$

확률 분포 P(x) 하에서 어떤 함수 f(x)의 평균값은 f(x)의 기댓값이라고 한다.

각 x값에 대해 해당 확률을 가중치로 사용한 가중 평균을 구하는 것이다.

〈연속〉

$$\mathbb{E}[f] = \int p(x) f(x)\, \mathrm{d}x$$

$$\mathbb{E}_x[f|y] = \sum_x p(x|y) f(x)$$

Conditional Expectation (discrete)

$$\mathbb{E}[f] \simeq \frac{1}{N} \sum_{n=1}^{N} f(x_n)$$

Approximate Expectation (discrete and continuous)

Variances and Covariances

$$\mathrm{var}[f] = \mathbb{E}\left[\left(f(x) - \mathbb{E}[f(x)]\right)^2\right] = \mathbb{E}[f(x)^2] - \mathbb{E}[f(x)]^2$$

제곱의평균 - 평균의제곱.

cf. $Var[x] = E[x^2] - E[x]^2$

$$\begin{aligned}\mathrm{cov}[x, y] &= \mathbb{E}_{x,y}\left[\{x - \mathbb{E}[x]\}\{y - \mathbb{E}[y]\}\right] \\ &= \mathbb{E}_{x,y}[xy] - \mathbb{E}[x]\mathbb{E}[y]\end{aligned}$$

$$\begin{aligned}\mathrm{cov}[\mathbf{x}, \mathbf{y}] &= \mathbb{E}_{\mathbf{x},\mathbf{y}}\left[\{\mathbf{x} - \mathbb{E}[\mathbf{x}]\}\{\mathbf{y}^{\mathrm{T}} - \mathbb{E}[\mathbf{y}^{\mathrm{T}}]\}\right] \\ &= \mathbb{E}_{\mathbf{x},\mathbf{y}}[\mathbf{x}\mathbf{y}^{\mathrm{T}}] - \mathbb{E}[\mathbf{x}]\mathbb{E}[\mathbf{y}^{\mathrm{T}}]\end{aligned}$$

→ 두 확률변수 x와 y가 벡터일경우 공분산은 행렬이된다.

- $\mathrm{Cov}(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$[1]
  [증명]

  $(X - \mu_X)(Y - \mu_Y) = XY - \mu_X Y - \mu_Y X + \mu_X \mu_Y$를 이용하면

$$\begin{aligned}\mathrm{Cov}(X, Y) &= \mathbb{E}(XY) - \mathbb{E}(\mu_X Y) - \mathbb{E}(\mu_Y X) + \mathbb{E}(\mu_X \mu_Y) \\ &= \mathbb{E}(XY) - \mu_X \mathbb{E}(Y) - \mu_Y \mathbb{E}(X) + \mu_X \mu_Y \\ &= \mathbb{E}(XY) - \mu_X \mu_Y\end{aligned}$$

- $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(X, Y)$
  [증명]
  - 일반화: $\mathrm{Var}\left(\sum_{k=1}^{n} X_k\right) = \sum_{k=1}^{n} \mathrm{Var}(X_k) + 2\sum_{i<j} \mathrm{Cov}(X_i, X_j)$

- $|\mathrm{Cov}(X, Y)| \le \sqrt{\mathrm{Var}(X) \cdot \mathrm{Var}(Y)}$

2.1. 모공분산

[편집]

모공분산은 모집단의 공분산이다. $\mathrm{Cov}(X, Y)$ 또는 σ_{XY}로 쓴다. X와 Y는 확률 변수, N은 모집단의 표본의 개수, X_i와 Y_i는 각 확률 변수의 도수, μ는 모평균을 뜻한다.

$$\begin{aligned}\mathrm{Cov}(X, Y) &= \sigma_{XY} \\ &= \frac{1}{N}\sum_{i=1}^{n}(X_i - \mu_X)(Y_i - \mu_Y) \\ &= \mathbb{E}\{(X - \mu_X)(Y - \mu_Y)\}\end{aligned}$$

곧, 모공분산이란 X의 편차와 Y의 편차의 곱의 평균이다.

2.2. 표본공분산

[편집]

표본공분산은 표본집단의 공분산이다. S_{XY}로 쓴다. X와 Y는 확률 변수, n은 표본집단의 표본의 개수, X_i와 Y_i는 각 확률 변수의 도수, $\bar{X}$와 $\bar{Y}$는 표본평균을 뜻한다.

$$\begin{aligned}S_{XY} &= \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) \\ &= \mathbb{E}\{(X - \bar{X})(Y - \bar{Y})\}\end{aligned}$$

곧, 표본공분산이란 X의 편차와 Y의 편차의 곱의 평균이다. 주의할 점은 **(표본의 개수)$-$1**로 나눈다는 것이다. n이 아니라 $n - 1$로 나누는 것은 오차를 줄이기 위함으로, 일반적인 표본 분산의 계산법과 같다.

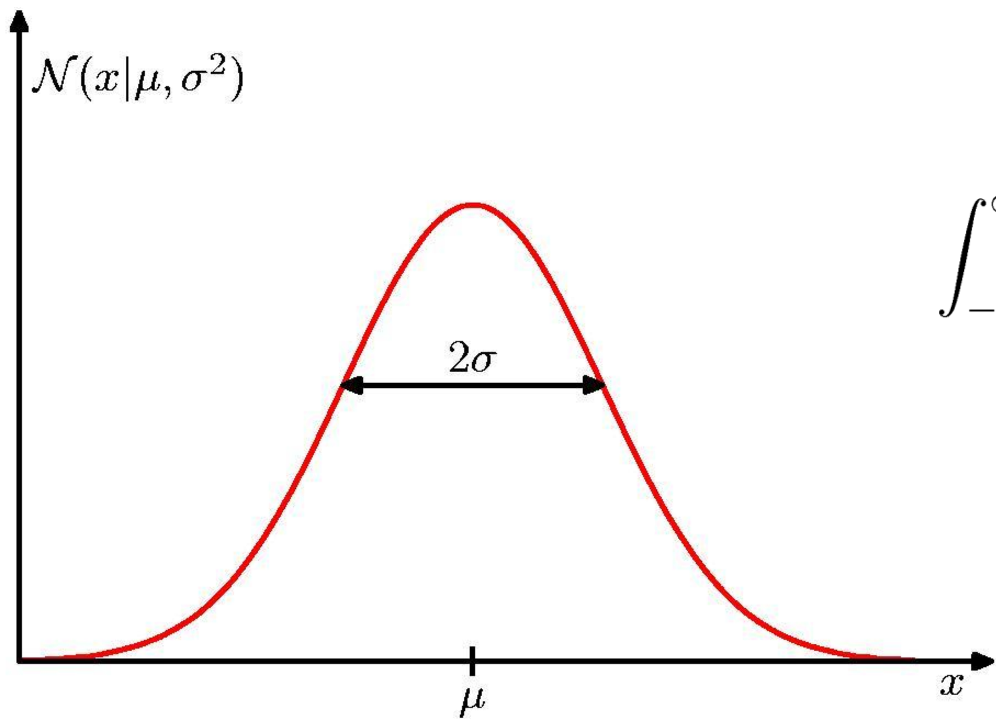
The Gaussian Distribution

e.g. $\sigma \rightarrow$ 표준편차

$\beta \rightarrow 1/\sigma^2 \rightarrow$ 정밀도

$$\mathcal{N}\left(x|\mu,\sigma^2\right) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$

평균 분산



$$\mathcal{N}(x|\mu,\sigma^2) > 0$$

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

Gaussian Mean and Variance

$$\mathcal{N}\left(x|\mu,\sigma^2\right) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$

$$\mathbb{E}[x] = \int_{-\infty}^{\infty} \mathcal{N}\left(x|\mu,\sigma^2\right) x \,\mathrm{d}x = \mu$$

$$\mathbb{E}[x^2] = \int_{-\infty}^{\infty} \mathcal{N}\left(x|\mu,\sigma^2\right) x^2 \,\mathrm{d}x = \mu^2 + \sigma^2$$

$$\mathrm{var}[x] = \mathbb{E}[x^2] - \mathbb{E}[x]^2 = \sigma^2$$

From the definition of the Gaussian distribution, X has probability density function:

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

From the definition of the expected value of a continuous random variable:

$$\mathsf{E}(X) = \int_{-\infty}^{\infty} x f_X(x) \, \mathrm{d}x$$

So:

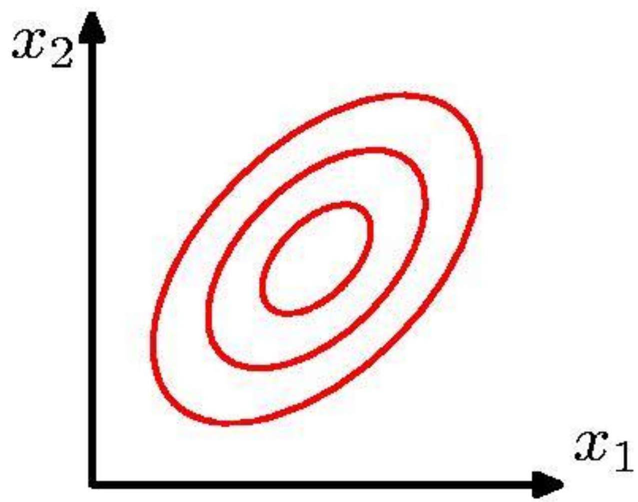
$$\begin{aligned}
\mathsf{E}(X) &= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} x \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \mathrm{d}x \\
&= \frac{\sqrt{2}\sigma}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} (\sqrt{2}\sigma t + \mu) \exp(-t^2) \, \mathrm{d}t && \text{substituting } t = \frac{x-\mu}{\sqrt{2}\sigma} \\
&= \frac{1}{\sqrt{\pi}} \left(\sqrt{2}\sigma \int_{-\infty}^{\infty} t \exp(-t^2) \, \mathrm{d}t + \mu \int_{-\infty}^{\infty} \exp(-t^2) \, \mathrm{d}t\right) \\
&= \frac{1}{\sqrt{\pi}} \left(\sqrt{2}\sigma \left[-\frac{1}{2}\exp(-t^2)\right]_{-\infty}^{\infty} + \mu\sqrt{\pi}\right) && \text{Fundamental Theorem of Calculus, Gaussian Integral} \\
&= \frac{\mu\sqrt{\pi}}{\sqrt{\pi}} && \text{Exponential Tends to Zero and Infinity} \\
&= \mu
\end{aligned}$$

$$\begin{aligned}
Var[X] &= \int_{-\infty}^{\infty} (x-\mu)^2 f_X(x) \, dx \\
&= \int_{-\infty}^{\infty} (x-\mu)^2 \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2} \, dx \\
&= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} z^2 e^{-z^2/2\sigma^2} \, dz && (\because \ let \ z = x-\mu, \ dz = dx) \\
&= \left[-\frac{1}{\sqrt{2\pi\sigma^2}} z\sigma^2 e^{-z^2/2\sigma^2}\right]_{-\infty}^{\infty} + \frac{1}{\sqrt{2\pi\sigma^2}} \sigma^2 \int_{-\infty}^{\infty} e^{-z^2/2\sigma^2} \, dz \\
&= 0 + \sigma^2 && \left(\because \ \int_{-\infty}^{\infty} e^{-z^2/2\sigma^2} \, dz = \sqrt{2\pi\sigma^2}\right)
\end{aligned}$$

The Multivariate Gaussian

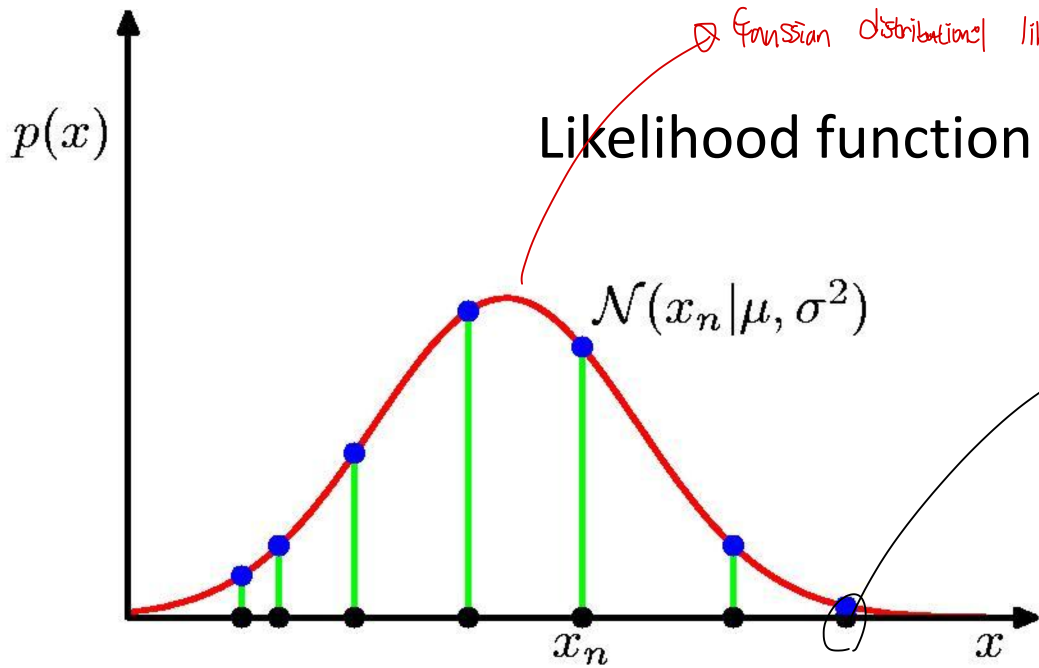
D차원 벡터 μ는 평균값, DxD 행렬 Σ는 공분산, $|\Sigma|$는 Σ의 행렬식

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



Gaussian Parameter Estimation

Gaussian distribution의 likelihood function.



검은색 : $\{x_n\}$을 값으로 가지는 데이터집합

$$p(\mathbf{x}|\mu,\sigma^2) = \prod_{n=1}^{N} \mathcal{N}\left(x_n|\mu,\sigma^2\right)$$

집합 x가 iid이다.

파란색 점들의 곱을 의미한다.

μ, σ^2 조정하여 해당 곱을 최대화함으로써 가능도를 최대화 할 수 있다.

Maximum (Log) Likelihood

log 함수는 변수에 대해 단조증가하는 함수 ∴ log를 씌워도 최댓값을 찾는데는 문제가 없다.

$$\underline{\ln p\left(\mathbf{x}|\mu, \sigma^2\right)} = -\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n - \mu)^2 - \frac{N}{2}\ln\sigma^2 - \frac{N}{2}\ln(2\pi)$$

$$\mu_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N} x_n \qquad \sigma^2_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N}(x_n - \mu_{\mathrm{ML}})^2$$

$\frac{\partial L}{\partial \mu} = 0$ $\qquad$ $\frac{\partial L}{\partial \sigma^2} = 0$

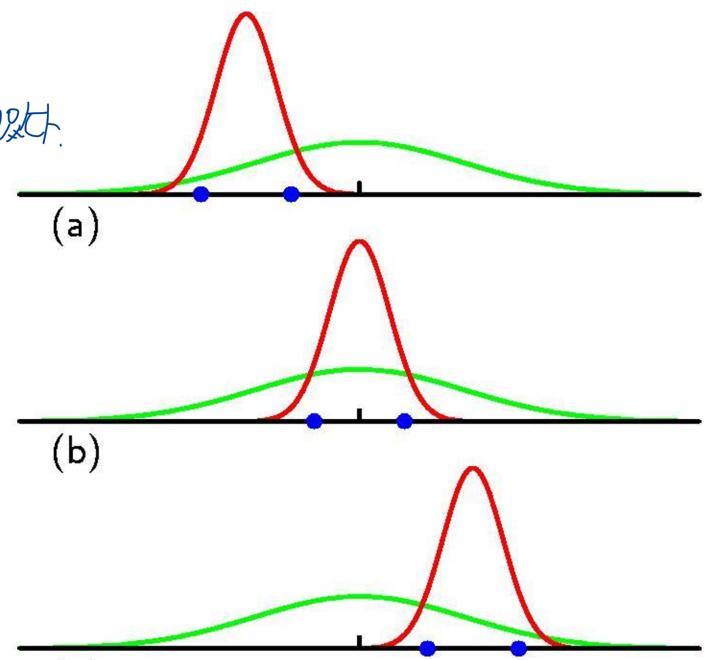
표본평균 (Sample mean) $\qquad$ 표본분산 (Sample variance)

Properties of μ_{ML} and σ^2_{ML}

$$\mathbb{E}[\mu_{\mathrm{ML}}] = \mu$$

① 분포의 분산을 $\frac{N-1}{N}$ 만큼 과소평가되어있다.

$$\mathbb{E}[\sigma^2_{\mathrm{ML}}] = \left(\frac{N-1}{N}\right)\sigma^2$$



분산이 너무 과소평가.

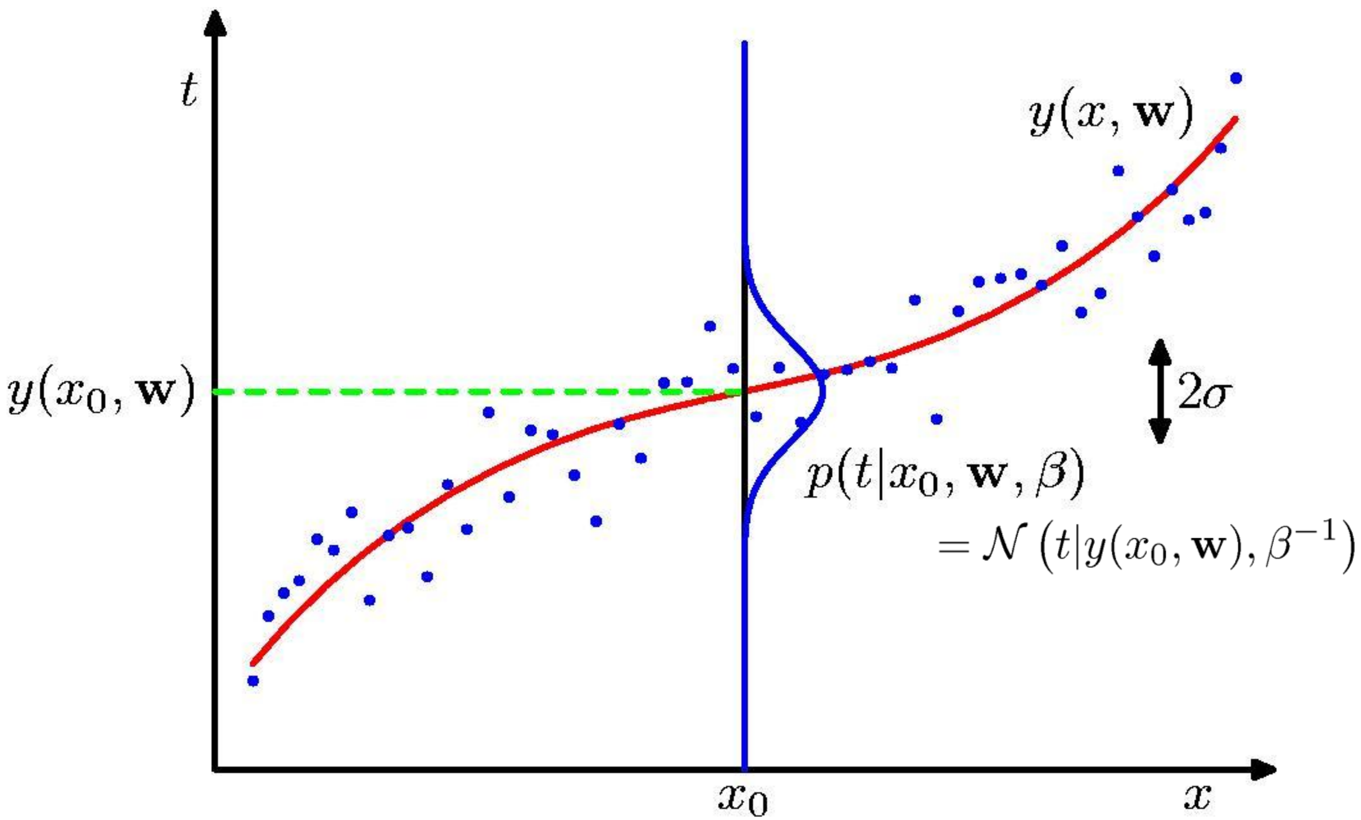
$$\begin{aligned}\widetilde{\sigma}^2 &= \frac{N}{N-1}\sigma^2_{\mathrm{ML}} \\ &= \frac{1}{N-1}\sum_{n=1}^{N}(x_n - \mu_{\mathrm{ML}})^2\end{aligned}$$

분산추정.

비편향추정.

※ 모수 추정값의 평균값이 실제 모수값과 같아지면 unbias, 아니면 bias이다.

Curve Fitting Re-visited



Maximum Likelihood

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

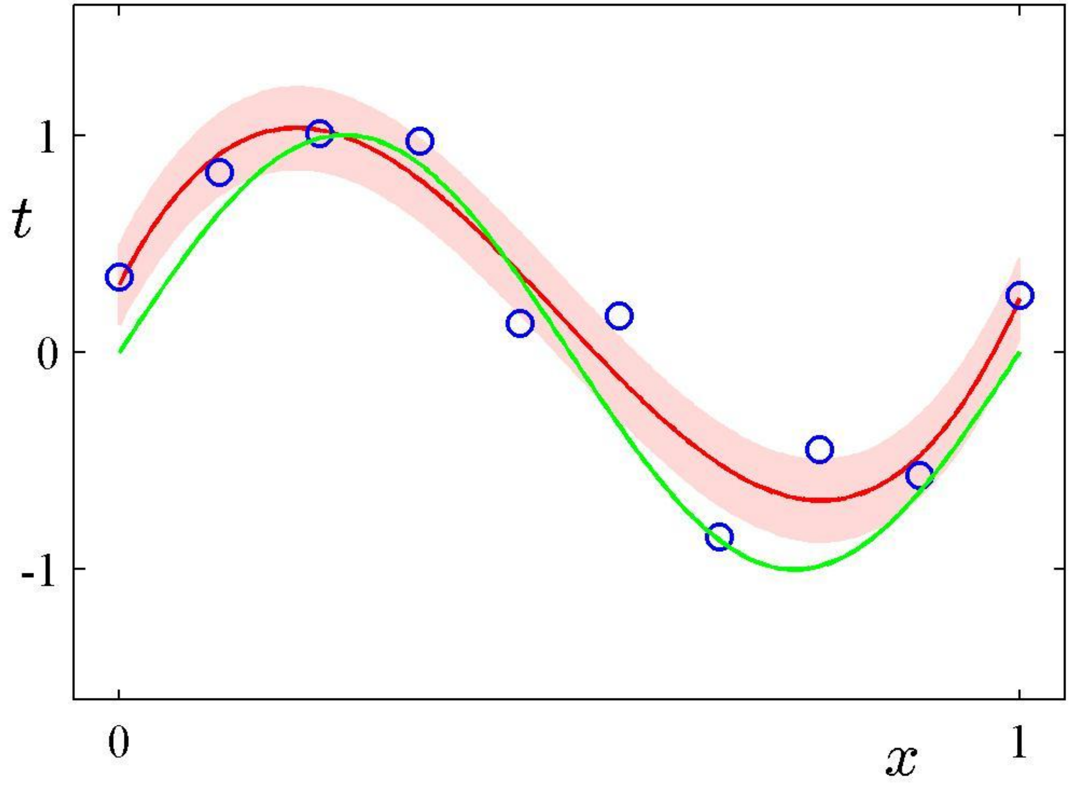
$$\ln p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) = \underbrace{-\frac{\beta}{2} \sum_{n=1}^{N} \{y(x_n, \mathbf{w}) - t_n\}^2}_{\beta E(\mathbf{w})} + \frac{N}{2} \ln \beta - \frac{N}{2} \ln(2\pi)$$

Determine $\mathbf{w}_{\mathrm{ML}}$ by minimizing sum-of-squares error, $E(\mathbf{w})$.

$$\frac{1}{\beta_{\mathrm{ML}}} = \frac{1}{N} \sum_{n=1}^{N} \{y(x_n, \mathbf{w}_{\mathrm{ML}}) - t_n\}^2$$

Predictive Distribution

$$p(t|x, \mathbf{w}_{\mathrm{ML}}, \beta_{\mathrm{ML}}) = \mathcal{N}\left(t|y(x, \mathbf{w}_{\mathrm{ML}}), \beta_{\mathrm{ML}}^{-1}\right)$$



MAP: A Step towards Bayes

$$p(\mathbf{w}|\alpha) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I}) = \left(\frac{\alpha}{2\pi}\right)^{(M+1)/2} \exp\left\{-\frac{\alpha}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}\right\}$$

$$p(\mathbf{w}|\mathbf{x}, \mathbf{t}, \alpha, \beta) \propto p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta)p(\mathbf{w}|\alpha)$$

$$\beta\widetilde{E}(\mathbf{w}) = \frac{\beta}{2}\sum_{n=1}^{N}\{y(x_n, \mathbf{w}) - t_n\}^2 + \frac{\alpha}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}$$

Determine $\mathbf{w}_{\mathrm{MAP}}$ by minimizing regularized sum-of-squares error, $\widetilde{E}(\mathbf{w})$.

Bayesian Curve Fitting

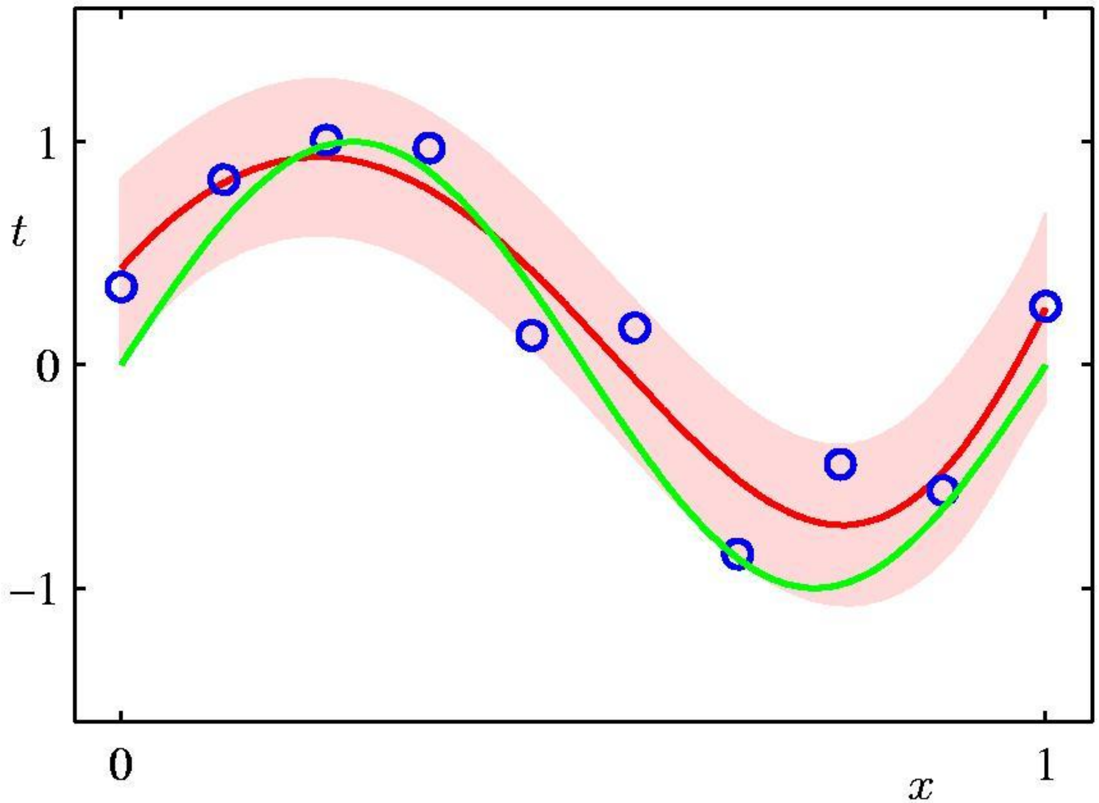
$$p(t|x, \mathbf{x}, \mathbf{t}) = \int p(t|x, \mathbf{w}) p(\mathbf{w}|\mathbf{x}, \mathbf{t}) \, \mathrm{d}\mathbf{w} = \mathcal{N}\left(t|m(x), s^2(x)\right)$$

$$m(x) = \beta \boldsymbol{\phi}(x)^{\mathrm{T}} \mathbf{S} \sum_{n=1}^{N} \boldsymbol{\phi}(x_n) t_n \qquad s^2(x) = \beta^{-1} + \boldsymbol{\phi}(x)^{\mathrm{T}} \mathbf{S} \boldsymbol{\phi}(x)$$

$$\mathbf{S}^{-1} = \alpha \mathbf{I} + \beta \sum_{n=1}^{N} \boldsymbol{\phi}(x_n) \boldsymbol{\phi}(x_n)^{\mathrm{T}} \qquad \boldsymbol{\phi}(x_n) = \left(x_n^0, \ldots, x_n^M\right)^{\mathrm{T}}$$

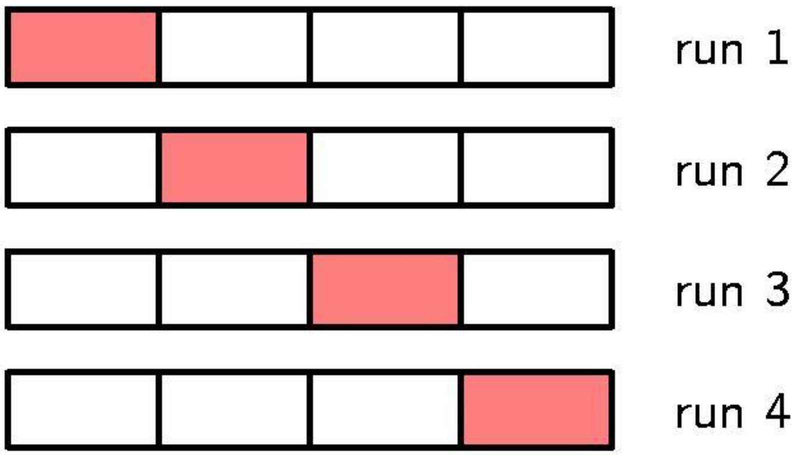
Bayesian Predictive Distribution

$$p(t|x, \mathbf{x}, \mathbf{t}) = \mathcal{N}\left(t|m(x), s^2(x)\right)$$

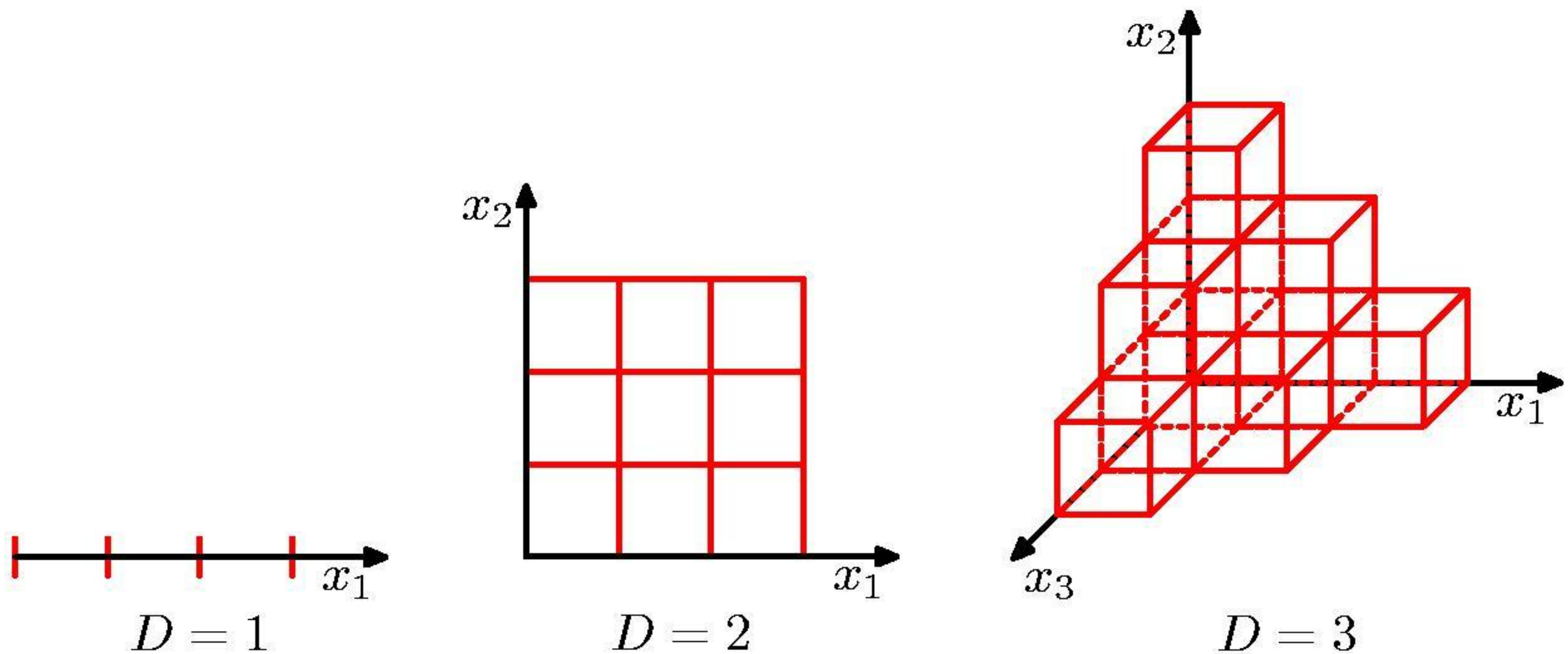


Model Selection

Cross-Validation



Curse of Dimensionality

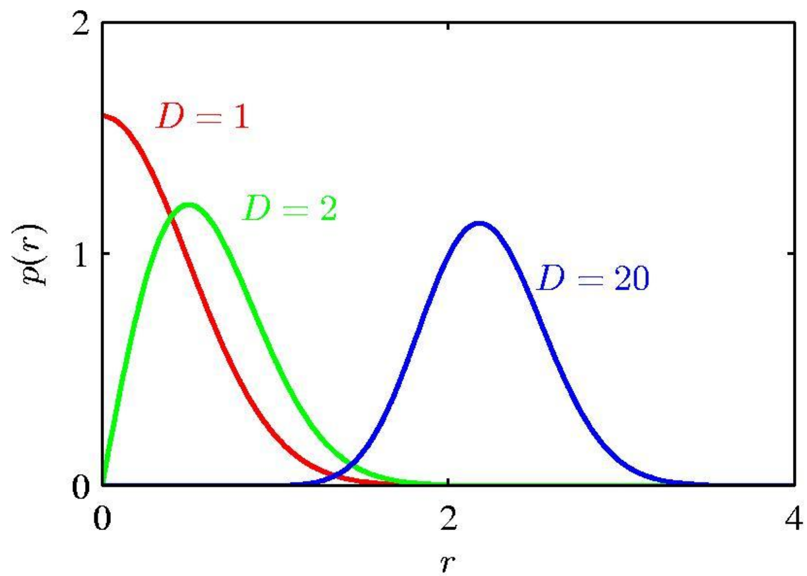


Curse of Dimensionality

Polynomial curve fitting, $M = 3$

$$y(\mathbf{x}, \mathbf{w}) = w_0 + \sum_{i=1}^{D} w_i x_i + \sum_{i=1}^{D} \sum_{j=1}^{D} w_{ij} x_i x_j + \sum_{i=1}^{D} \sum_{j=1}^{D} \sum_{k=1}^{D} w_{ijk} x_i x_j x_k$$

Gaussian Densities in higher dimensions



- 확률질량이 구의 표면 근처의 얇은 껍질에 집중되어 있음을 확인할수 있다.

- Manifold 관련.

Decision Theory

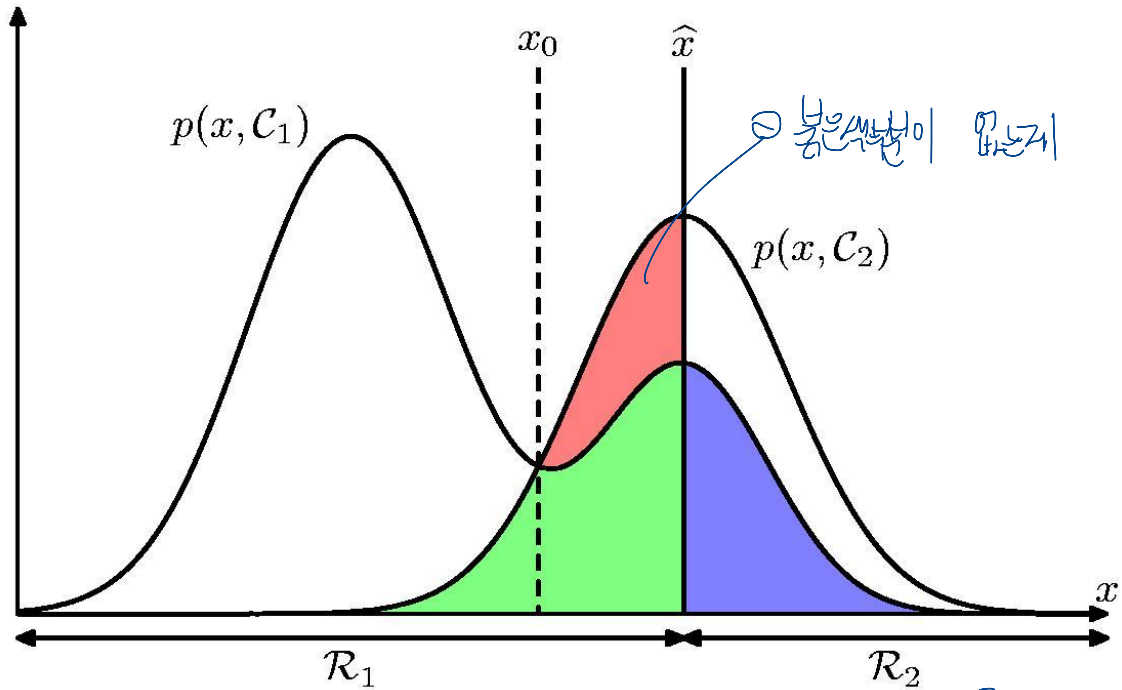
Inference step

Determine either $p(t|\mathbf{x})$ or $p(\mathbf{x}, t)$. 결합확률 분포

Decision step

For given $\mathbf{x}$, determine optimal t.

Minimum Misclassification Rate



붉은색부분이 없는게 가장 최적의 decision boundary이다.

결정구역.

$$
\begin{aligned}
p(\text{mistake}) &= p(\mathbf{x} \in \mathcal{R}_1, \mathcal{C}_2) + p(\mathbf{x} \in \mathcal{R}_2, \mathcal{C}_1) \\
&= \int_{\mathcal{R}_1} p(\mathbf{x}, \mathcal{C}_2)\,\mathrm{d}\mathbf{x} + \int_{\mathcal{R}_2} p(\mathbf{x}, \mathcal{C}_1)\,\mathrm{d}\mathbf{x}.
\end{aligned}
$$

Minimum Expected Loss

Example: classify medical images as ‘cancer’ or ‘normal’

$$
\begin{array}{ccc}
 & & \text{Decision} \\
 & & \begin{array}{cc} \text{cancer} & \text{normal} \end{array} \\
\text{Truth} & \begin{array}{c} \text{cancer} \\ \text{normal} \end{array} & \begin{pmatrix} 0 & 1000 \\ 1 & 0 \end{pmatrix}
\end{array}
$$

손실행렬.

Minimum Expected Loss

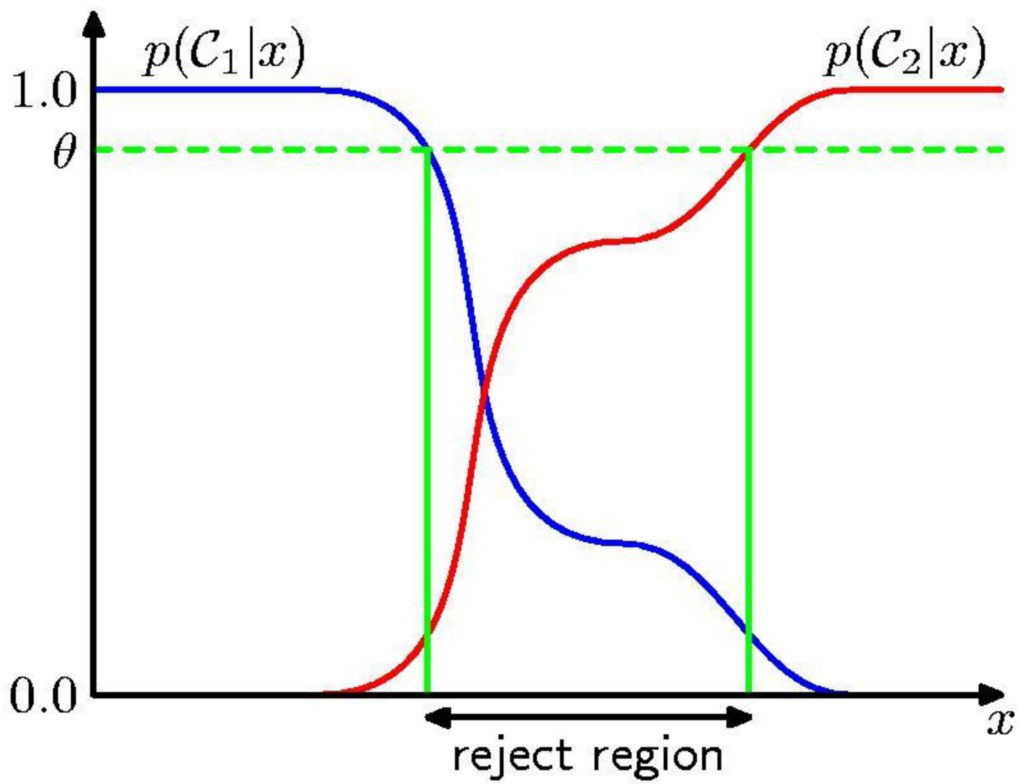
$$\mathbb{E}[L] = \sum_k \sum_j \int_{\mathcal{R}_j} L_{kj} p(\mathbf{x}, \mathcal{C}_k) \, \mathrm{d}\mathbf{x}$$

Regions $\mathcal{R}_j$ are chosen to minimize

$$\mathbb{E}[L] = \sum_k L_{kj} p(\mathcal{C}_k | \mathbf{x})$$

$p(x, C_k) = p(C_k | x) p(x)$

Reject Option



거부구역

Why Separate Inference and Decision?

- Minimizing risk (loss matrix may change over time)
- Reject option
- Unbalanced class priors
- Combining models

Decision Theory for Regression

Inference step

Determine $p(\mathbf{x}, t)$.

Decision step

For given $\mathbf{x}$, make optimal prediction, $y(\mathbf{x})$, for t.

Loss function: $$\mathbb{E}[L] = \iint \underline{L(t, y(\mathbf{x}))} p(\mathbf{x}, t) \,\mathrm{d}\mathbf{x}\,\mathrm{d}t$$

x에 대하여

t의 추정값 y(x) 선택

The Squared Loss Function

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

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

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

$$y(\mathbf{x}) = \mathbb{E}[t|\mathbf{x}]$$

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

Generative vs Discriminative

Generative approach:

Model $p(t, \mathbf{x}) = p(\mathbf{x}|t)p(t)$

Use Bayes' theorem $p(t|\mathbf{x}) = \dfrac{p(\mathbf{x}|t)p(t)}{p(\mathbf{x})}$

Discriminative approach:

Model $p(t|\mathbf{x})$ directly

Entropy : 확률변수의 상태를 결정하는 정보량의 평균

$h(x) = -\log_2 p(x)$

$h(x,y) = h(x) + h(y)$

$p(x,y) = p(x)p(y)$

$$\mathrm{H}[x] = -\sum_x p(x) \log_2 p(x)$$

$E[h(x)]$

Important quantity in

- coding theory
- statistical physics
- machine learning

Entropy

Coding theory: x discrete with 8 possible states; how many bits to transmit the state of x?

All states equally likely

$$\mathrm{H}[x] = -8 \times \frac{1}{8} \log_2 \frac{1}{8} = 3 \text{ bits.}$$

Entropy

x	a	b	c	d	e	f	g	h
$p(x)$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{64}$	$\frac{1}{64}$	$\frac{1}{64}$	$\frac{1}{64}$
code	0	10	110	1110	111100	111101	111110	111111

$$\begin{aligned} \mathrm{H}[x] &= -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{4}\log_2\frac{1}{4} - \frac{1}{8}\log_2\frac{1}{8} - \frac{1}{16}\log_2\frac{1}{16} - \frac{4}{64}\log_2\frac{1}{64} \\ &= 2 \text{ bits} \end{aligned}$$

$$\begin{aligned} \text{average code length} &= \frac{1}{2}\times 1 + \frac{1}{4}\times 2 + \frac{1}{8}\times 3 + \frac{1}{16}\times 4 + 4\times\frac{1}{64}\times 6 \\ &= 2 \text{ bits} \end{aligned}$$

Entropy

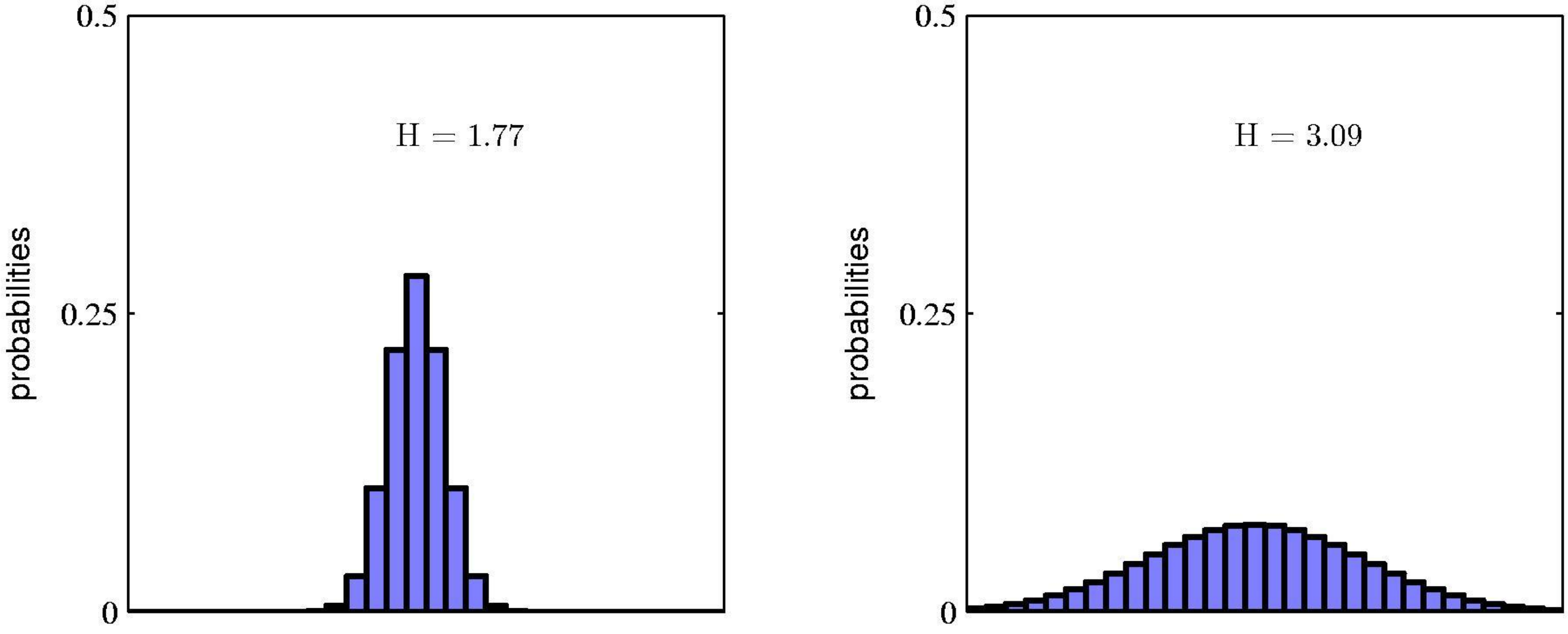
In how many ways can N identical objects be allocated M bins?

$$W = \frac{N!}{\prod_i n_i!}$$

$$\mathrm{H} = \frac{1}{N} \ln W \simeq -\lim_{N \to \infty} \sum_i \left(\frac{n_i}{N}\right) \ln \left(\frac{n_i}{N}\right) = -\sum_i p_i \ln p_i$$

Entropy maximized when $\forall i : p_i = \frac{1}{M}$

Entropy



Differential Entropy

Put bins of width Δ along the real line

$$\lim_{\Delta \to 0} \left\{ -\sum_i p(x_i)\Delta \ln p(x_i) \right\} = -\int p(x) \ln p(x) \, \mathrm{d}x$$

Differential entropy maximized (for fixed σ^2) when

$$p(x) = \mathcal{N}(x|\mu, \sigma^2)$$

in which case

$$\mathrm{H}[x] = \frac{1}{2}\left\{1 + \ln(2\pi\sigma^2)\right\}.$$

σ^2가 클수록 엔트로피가 증가한다는것을 다시 확인할수 있다.

Conditional Entropy

$$\mathrm{H}[\mathbf{y}|\mathbf{x}] = -\iint p(\mathbf{y}, \mathbf{x}) \ln p(\mathbf{y}|\mathbf{x}) \, \mathrm{d}\mathbf{y} \, \mathrm{d}\mathbf{x}$$

$$\mathrm{H}[\mathbf{x}, \mathbf{y}] = \mathrm{H}[\mathbf{y}|\mathbf{x}] + \mathrm{H}[\mathbf{x}]$$

- ※. x와 y를 특정짓기 위해 필요한 정보의 양은 x만 따로 특정짓기 위해 필요한 정보의 양과
x가 주어졌을때 y를 특정짓기 위해 필요한 정보의 양을 합친것과 같다.

The Kullback-Leibler Divergence

$$\begin{aligned}
\mathrm{KL}(p\|q) &= -\int p(\mathbf{x}) \ln q(\mathbf{x})\, \mathrm{d}\mathbf{x} - \left(-\int p(\mathbf{x}) \ln p(\mathbf{x})\, \mathrm{d}\mathbf{x}\right) \\
&= -\int p(\mathbf{x}) \ln \left\{ \frac{q(\mathbf{x})}{p(\mathbf{x})} \right\} \mathrm{d}\mathbf{x}
\end{aligned}$$

알려지지 않은 분포

modeling한 분포

$$\mathrm{KL}(p\|q) \simeq \frac{1}{N} \sum_{n=1}^{N} \{-\ln q(\mathbf{x}_n|\boldsymbol{\theta}) + \ln p(\mathbf{x}_n)\}$$

$$\mathrm{KL}(p\|q) \geqslant 0 \qquad\qquad \mathrm{KL}(p\|q) \not\equiv \mathrm{KL}(q\|p)$$

$KL(p||q)=0$ 이면 $p(x)=q(x)$

Mutual Information

$$\begin{aligned} \mathrm{I}[\mathbf{x}, \mathbf{y}] &\equiv \mathrm{KL}(p(\mathbf{x}, \mathbf{y}) \| p(\mathbf{x}) p(\mathbf{y})) \\ &= -\iint p(\mathbf{x}, \mathbf{y}) \ln \left(\frac{p(\mathbf{x}) p(\mathbf{y})}{p(\mathbf{x}, \mathbf{y})} \right) \mathrm{d}\mathbf{x}\, \mathrm{d}\mathbf{y} \end{aligned}$$

$$\mathrm{I}[\mathbf{x}, \mathbf{y}] = \mathrm{H}[\mathbf{x}] - \mathrm{H}[\mathbf{x}|\mathbf{y}] = \mathrm{H}[\mathbf{y}] - \mathrm{H}[\mathbf{y}|\mathbf{x}]$$