Mathematical notations

$\mathbf{x}$: column vector e.g. $\mathbf{x} = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} \in \mathbb{R}^3$

$\mathbf{x}^T$: transpose of $\mathbf{x}$ (row vector) $\mathbf{x}^T = (1, 2, 3)$

D : # of dim (features) i th component

N : # of samples (data), n th object

$\mathbf{X}$: data matrix with $\mathbf{x}_1, \ldots \mathbf{x}_N$

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \vdots \\ \mathbf{x}_N^T \end{pmatrix} \qquad N \times D \text{ matrix}$$

$\mathbf{X} \;//\; x_{ni}$

Chapter 1.

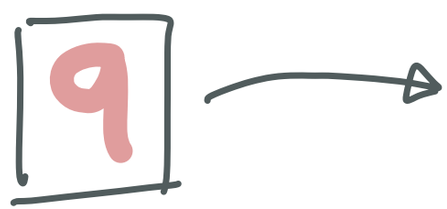
Pattern recognition

- Finding and analyzing features of data
- Decision making through prediction and classification.

Example hand-written image

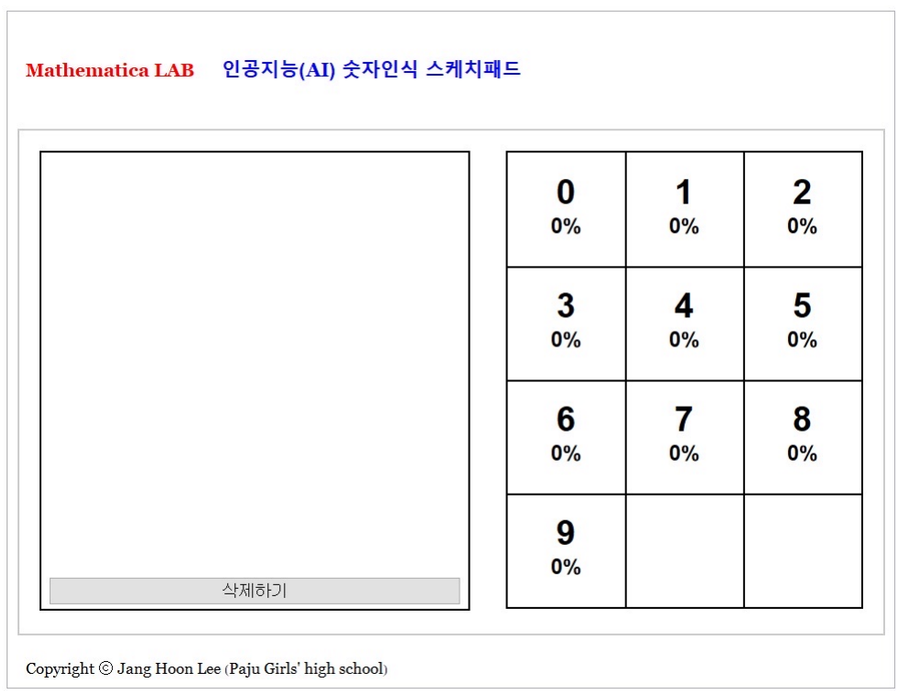
28 x 28 pixels

784-dim vector



number

9



Training

28×28 image

input vector (set) $\{\vec{x}_1, \vec{x}_2, \ldots \vec{x}_N\}$ N 개의 손글씨 image

target vector (set) $\vec{t} = \begin{pmatrix} t_1 \\ \vdots \\ t_N \end{pmatrix}$ N-dim also called label set

$y(\vec{x})$: output of ML algorithm

e.g. mean of $\vec{x}$, prediction of $\vec{t}$ (target)

training phase : process of formulation of $y(\vec{x})$

test set : data not used for training

- results of test set are important because we need to generalize our algorithm

Preprocessing

- null value
- transformation, scale conversation, normalization
- data type change
- feature extraction (특징 추출)
- dimensionality reduction (차원 축소)

Supervised learning (task) 지도학습

training data : input vectors with its target value

$x_1, \ldots x_N$ $t_1, t_2, \ldots t_N$

Find an expression (function) describing input and target

$(y(x) \sim t)$

- classification : discrete target value
- regression : continuous "

Unsupervised learning 비지도 학습

training data : only input vectors $x_1, \ldots x_N$

pattern recognition of input (feature analysis)

- clustering (군집화)
- density estimation (밀도(분포) 측정) $p(x)$
- visualization (dimensionality reduction)

1.1 Example: Polynomial curve fitting

input $x \in \mathbb{R}$

target $t \in \mathbb{R}$

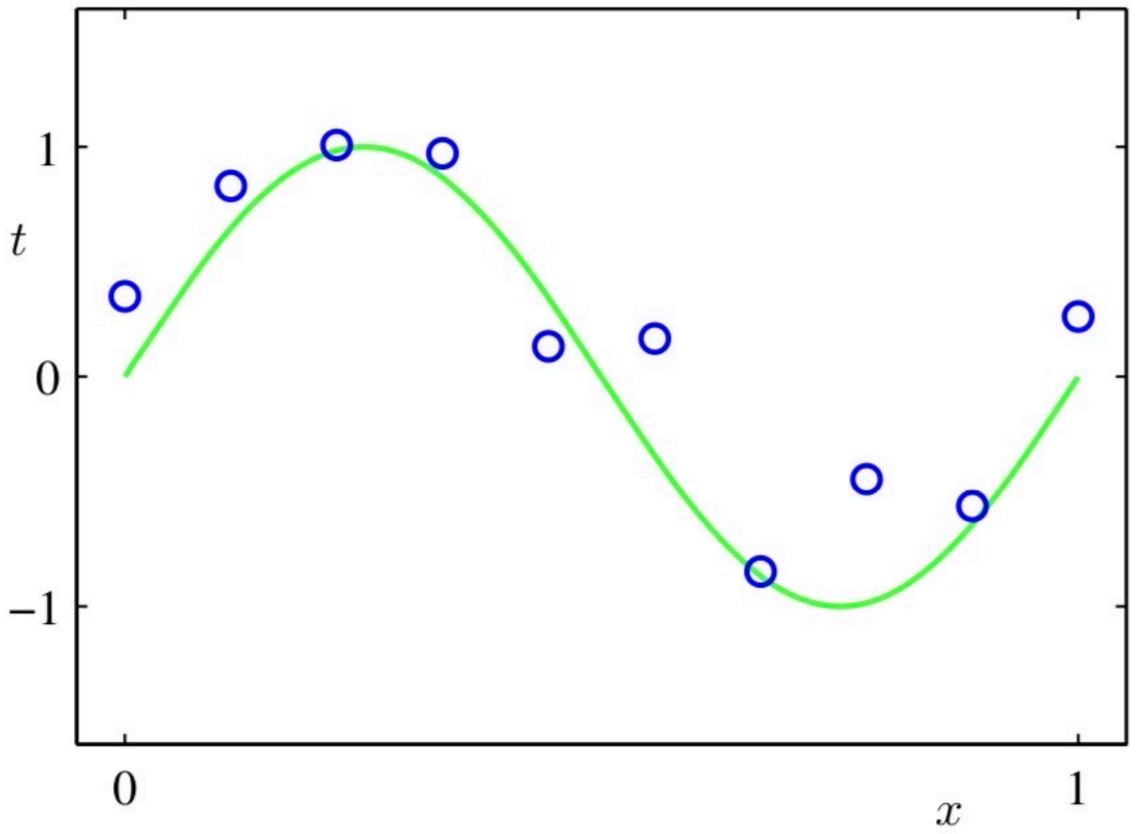
N observations $\quad \mathbb{X} := (x_1, x_2, \dots x_N)^T$ input vector

$\mathbb{t} := (t_1, t_2, \dots t_N)^T$ target vector

generated by $\sin(2\pi x)$ with small noise (perturbation)

$x_1 = 0, \quad x_N = 1, \quad x_{i+1} - x_i = \frac{1}{N-1} \quad \forall\, i = 1, \dots N-1$

Figure 1.2 Plot of a training data set of $N = 10$ points, shown as blue circles, each comprising an observation of the input variable x along with the corresponding target variable t. The green curve shows the function $\sin(2\pi x)$ used to generate the data. Our goal is to predict the value of t for some new value of x, without knowledge of the green curve.



Goal: predict $\hat{t}$ for a given $\hat{x}$ (new)

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

where M is the order of polynomial and x^j denotes x raised to the power of j

$y(x, w)$ is linear w.r.t w

not " " x

error function : measurement of the misfit between

$y(x, w)$ and target value t

SSE, sum of square error (function of w)

$$E(w) := \frac{1}{2} \sum_{n=1}^{N} \{ y(x_n, w) - t_n \}^2 \geq 0$$

$E(w) = 0$ iff $y(x_n, w) = t_n \ \forall n$ (예측값 = 실제값)

$E(w)$ is a quadratic function of w so minimization of

$E(w)$ has a unique solution ($\nabla_w E(w)$ is linear)

Let $\mathbf{w}^*$ be the sol of minimization of $E(\mathbf{w})$

$$E(\mathbf{w}^*) = \min_{\mathbf{w}} E(\mathbf{w})$$

The resulting polynomial is given by the function

$y(x, \mathbf{w}^*)$

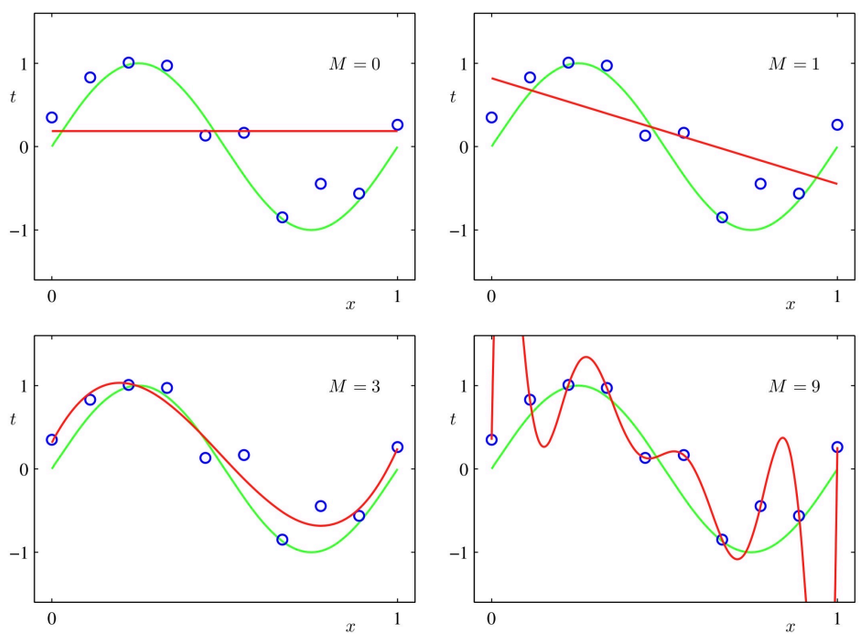
$$w_0^* + x w_1^* + x^2 w_2^* \cdots$$

How to determine M ?

model comparison (model selection)

The number of parameters (M+1) determines model complexity

Under – fitting vs over – fitting



N = 10

9 차 다항식

Figure 1.4 Plots of polynomials having various orders M, shown as red curves, fitted to the data set shown in Figure 1.2.

RMSE (root mean square error)

$$E_{RMS} = \sqrt{\frac{2E(w^*)}{N}}$$

$M=9$

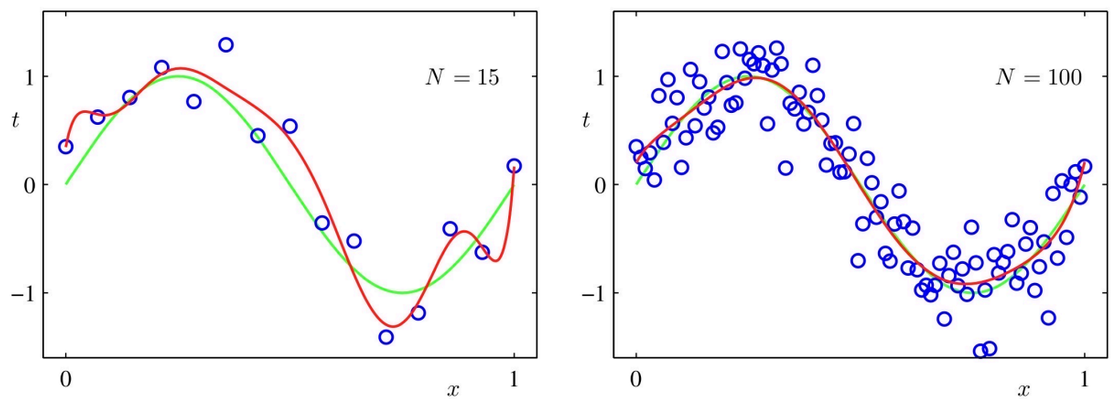


Figure 1.6 Plots of the solutions obtained by minimizing the sum-of-squares error function using the $M = 9$ polynomial for $N = 15$ data points (left plot) and $N = 100$ data points (right plot). We see that increasing the size of the data set reduces the over-fitting problem.

Remark

$N=10$

- When $M=9$, $E_{RMS} = 0$, 10 coefficients $w_0, w_1, .. w_9$ and 10 data points in training data.

Figure 1.5 Graphs of the root-mean-square error, defined by (1.3), evaluated on the training set and on an independent test set for various values of M.

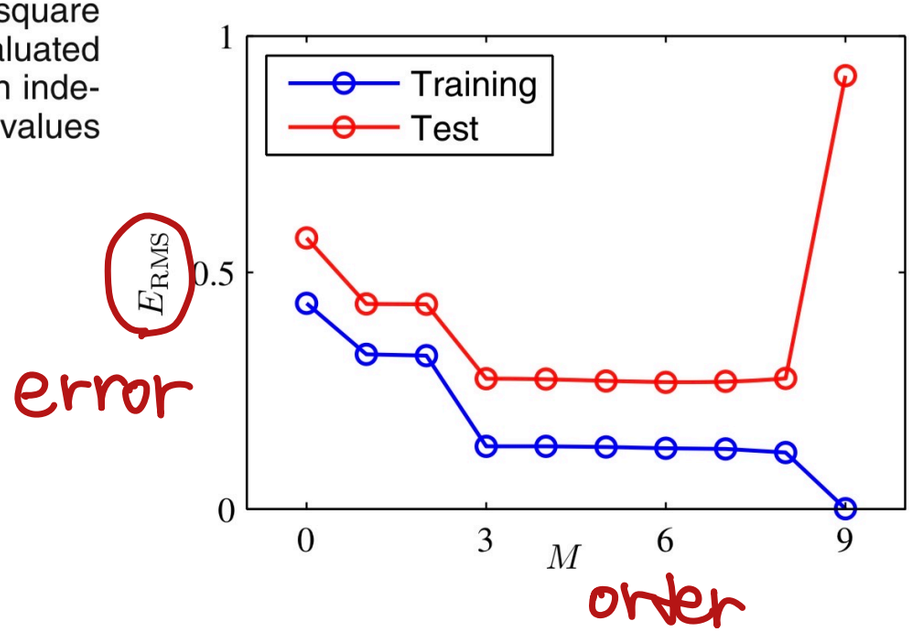


Table 1.1 Table of the coefficients $\mathbf{w}^\star$ for polynomials of various order. Observe how the typical magnitude of the coefficients increases dramatically as the order of the polynomial increases.

	$M=0$	$M=1$	$M=3$	$M=9$
$w_0^\star$	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

Remark

- When M increases, magnitude of w_i^* (solution) becomes larger
- The size of the training data set (fig. 1.6)

  over-fitting problem less serve as the size of the data set increases.

Regularization

SSE, RMSE, MSE

- prevent over-fitting
- discourage the coefficients from reaching large values

$$\tilde{E}(\mathbf{w}) = E(\mathbf{w}) + \frac{\lambda}{2}\|\mathbf{w}\|^2$$

where $\|\mathbf{w}\|^2 = \mathbf{w}^T\mathbf{w} = \sum_{i=0}^{M} w_i^2$

- Ridge regression, weight decay

N = 10

M = 9

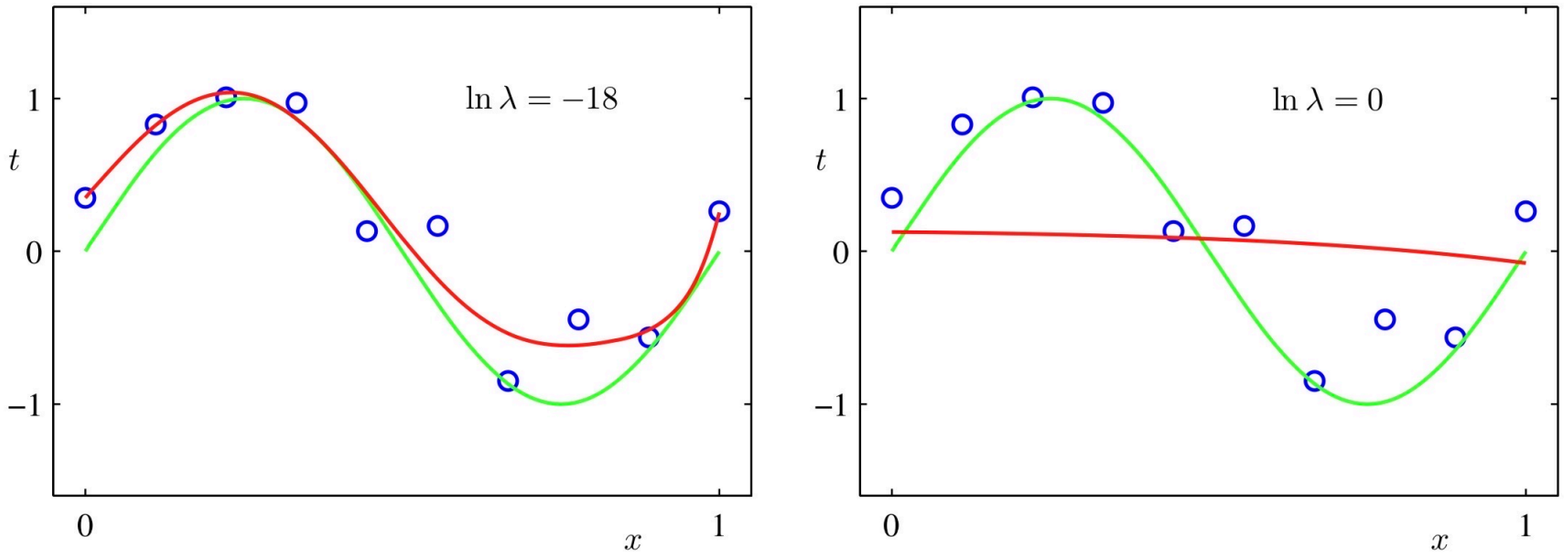


Figure 1.7 Plots of $M = 9$ polynomials fitted to the data set shown in Figure 1.2 using the regularized error function (1.4) for two values of the regularization parameter λ corresponding to $\ln \lambda = -18$ and $\ln \lambda = 0$. The case of no regularizer, i.e., $\lambda = 0$, corresponding to $\ln \lambda = -\infty$, is shown at the bottom right of Figure 1.4.

Validation set (hold-out set)

- prevent over-fitting

1.2 Probability Theory (elementary)

"uncertainty"

Sample space: set of all possible outcomes of experiment
Ω

Random variable: mathematical formalization of quantity or
X
object which depends on random events

$X : \Omega \longrightarrow E$ ($\mathbb{R}^d$) measurable space E

probability measure P on $(\Omega, \mathcal{F})$ (σ-field)

$(\Omega, \mathcal{F}, P)$

$\Omega = \{1, 2, 3, 4, 5, 6\}$ $\quad p(n) = \frac{1}{6} \quad n = 1, 2, \ldots 6$

win $+1000$ when n is even

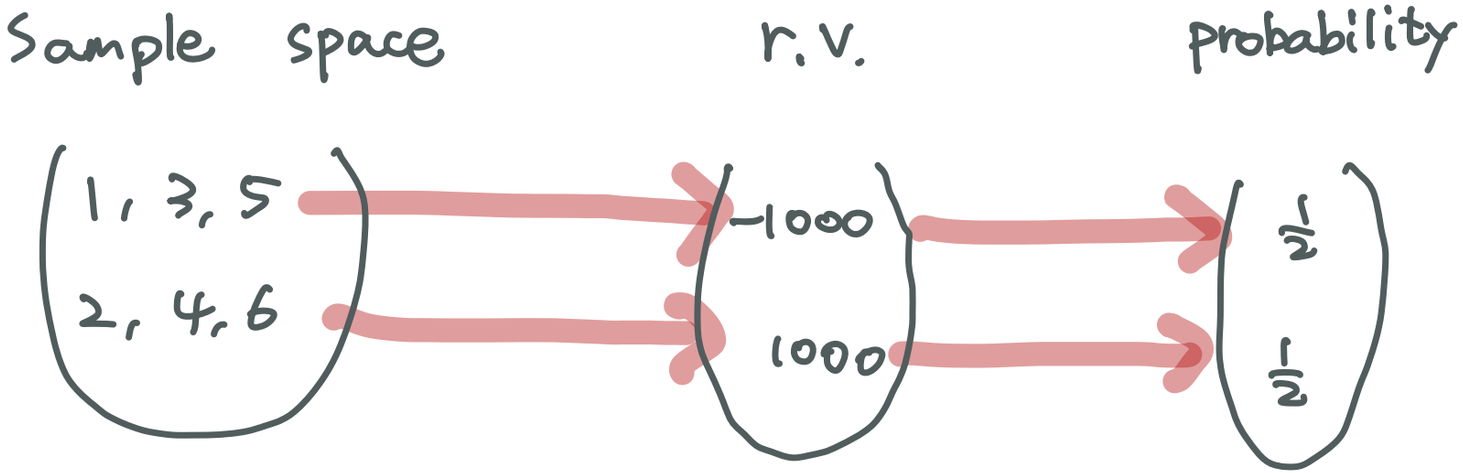
lose -1000 when n is odd

$$X(w) = \begin{cases} 1000 & w = 2, 4, 6 \\ -1000 & w = 1, 3, 5 \end{cases}$$

random variable

$$P(X = 1000) = P(\{w \in \Omega \mid X(w) = 1000\})$$

$$= P(\{2, 4, 6\}) = \frac{1}{2}$$


Sample space
r.v.
probability
1, 3, 5
2, 4, 6
-1000
1000
1/2
1/2

Sum rule and product rule

two random variables X, Y

$$X = \{x_1, \ldots x_M\}$$

$$Y = \{y_1 \ldots y_L\}$$

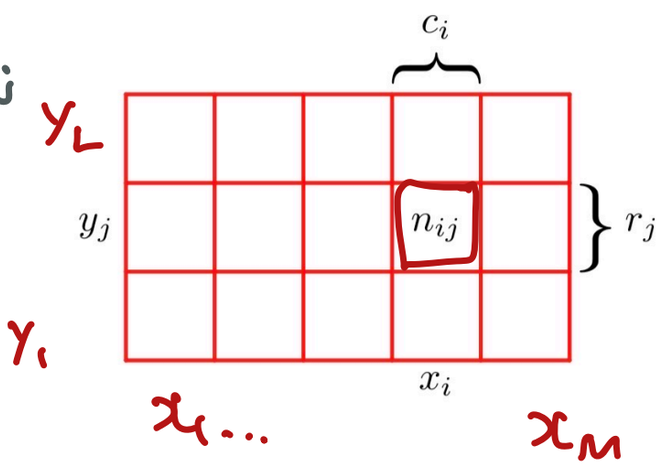
	X	Y
1	$(x_i$	$, y_j)$
2	(	)
3		
⋮		
N	(	)

Consider N trials in which we sample both X, Y

Let n_{ij} := # of trials in which $X = x_i$ and $Y = y_j$

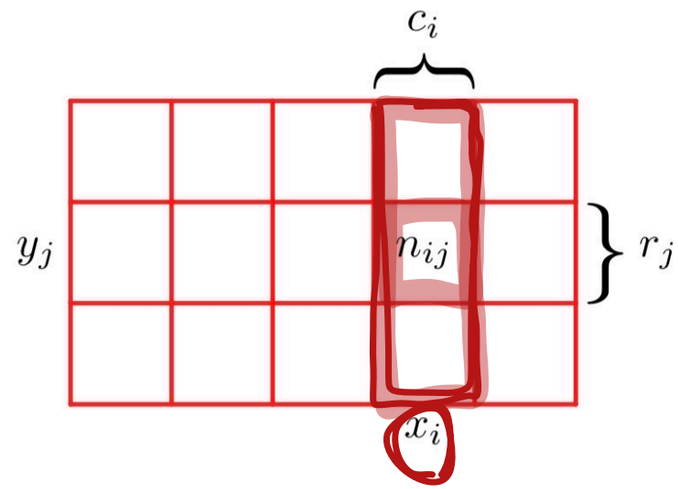
Joint probability of $X = x_i$ and $Y = y_j$

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



Let c_i := # of trials in which $X = x_i$ (regardless Y)

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



$p(X,Y) \rightarrow p(X) = \sum_Y p(X,Y)$

Sum rule

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

Conditional probability

$$p(Y = Y_j \mid X = x_i) = \frac{n_{ij}}{c_i}$$

$= p(Y = Y_j, X = x_i)$

$= p(X = x_i) > 0$ 인 경우

product rule

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

$$= p(Y = y_j \mid X = x_i)\, p(X = x_i)$$

The rules of probability

X, Y random variables

sum rule $$p(X) = \sum_Y p(X, Y)$$

product rule $$p(X, Y) = p(Y \mid X)\, p(X)$$

$$= p(X \mid Y)\, p(Y)$$

$p(X, Y)$: the probability of X and Y

$p(Y|X)$: the probability of Y given X

By symmetry property $p(X, Y) = p(Y, X)$,

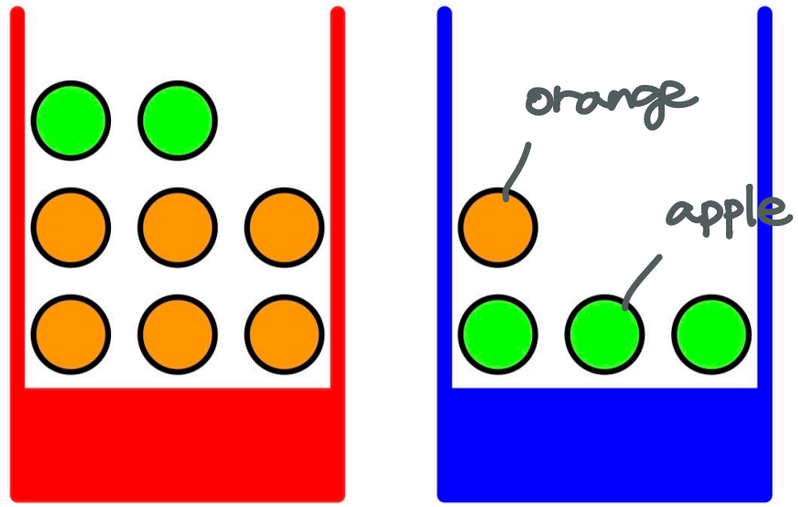
$$p(Y|X) = \frac{p(X|Y)\,p(Y)}{p(X)} \left(= \frac{p(X, Y)}{p(X)} \right)$$

Bayes's theorem

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

Example

Figure 1.9 We use a simple example of two coloured boxes each containing fruit (apples shown in green and oranges shown in orange) to introduce the basic ideas of probability.



pick red box 40% of the time

" blue " 60% "

$$p(B = \text{red}) = 4/10$$

$$p(B = \text{blue}) = 6/10$$

Suppose pick a box randomly and it turns out to be the blue box. Probability of (F = apple) given (B = blue)

$$p(F = \text{apple} \mid B = \text{blue}) = \frac{3}{4}$$

$$p(F = a \mid B = r) = \frac{1}{4}$$

$$p(F = o \mid B = r) = \frac{3}{4}$$

$$p(F = a \mid B = b) = \frac{3}{4}$$

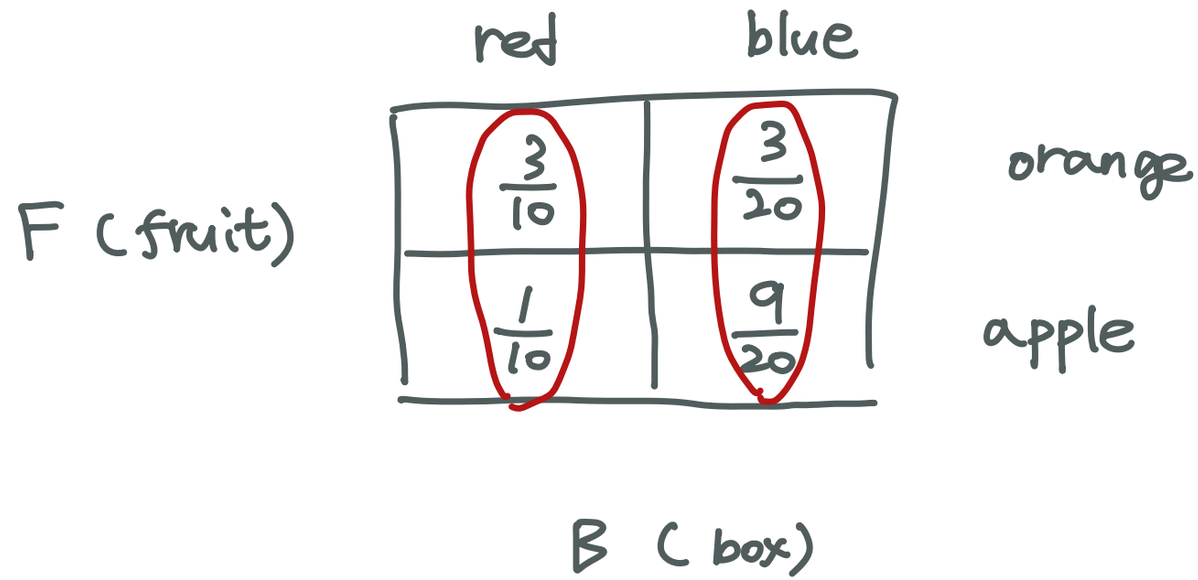
$$p(F = o \mid B = b) = \frac{1}{4}$$

$$p(F=a) = p(F=a \mid B=r)\,p(B=r) + p(F=a \mid B=b)\,p(B=b)$$

sum rule

$$= \frac{1}{4} \times \frac{4}{10} + \frac{3}{4} \times \frac{6}{10} = \frac{11}{20}$$

$$p(F=a, B=r) + p(F=a, B=b)$$



$$p(F=o) = \frac{9}{20}$$

$$p(B=r \mid F=o) = \frac{p(B=r, F=o)}{p(F=o)} = \frac{p(F=o \mid B=r)\, p(B=r)}{p(F=o)}$$

$$= \frac{2}{3}$$

Interpretation of Bayes's theorem

$p(B)$ prior probability (어떤 box를 골랐는지)

(probability available before we observe the identity of the fruit)

$p(B|F)$ posterior probability

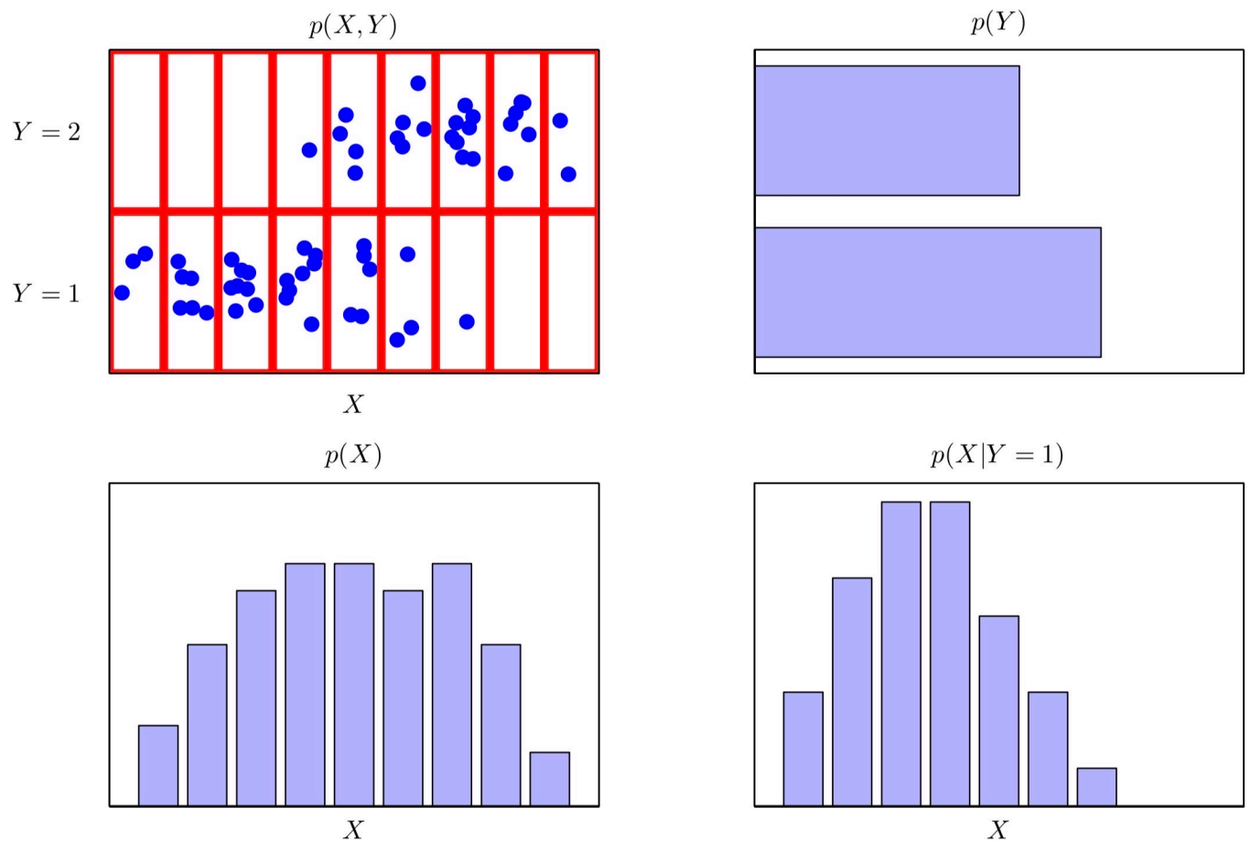
(과일 확인 후 어떤 box)

Two random variables X and Y are independent

X ~ p $\quad$ Y ~ q

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

$$\Rightarrow \quad p(Y|X) = p(Y) \quad \text{and} \quad p(X|Y) = p(X)$$



1.2.1 Probability density

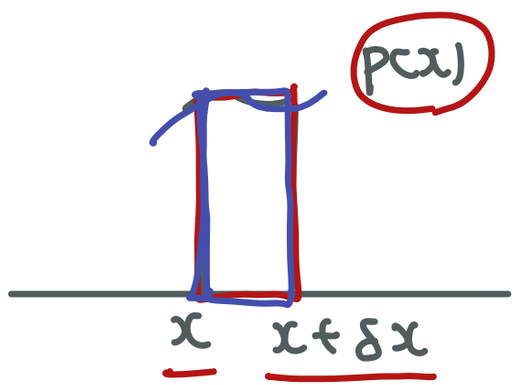
continuous variable

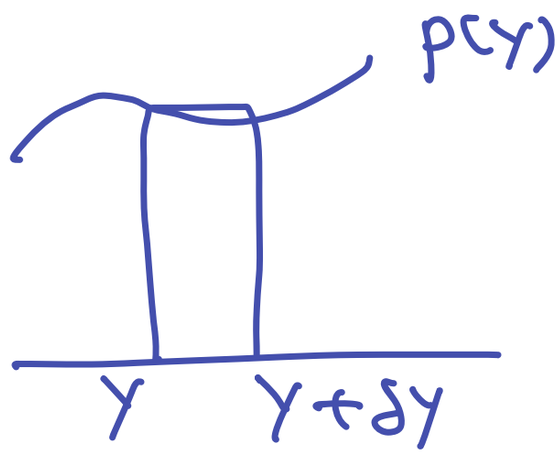
open interval

$p(\text{real valued variable falling in } (x, x+\delta x)) \simeq p(x)\delta x$

$p(x)\cdot\Delta x$

when δx is sufficiently small





$x+\Delta x$

$p(y, y+\delta y) \simeq p(y)\delta y$

$p(x)$ is called the probability density over x

$$p(x \in (a,b)) = \int_a^b p(x)\,dx$$

$$p(x) \geq 0$$

$$\int_{-\infty}^{\infty} p(x)\,dx = 1$$

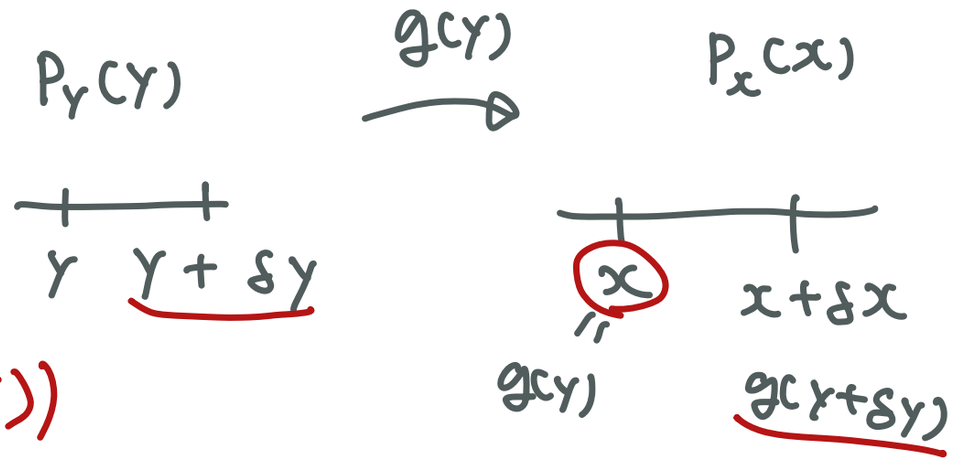
$$P_y(y)\,\delta y = P_x(x)\,\delta x$$

$$\frac{\delta x}{\delta y}$$

Consider a change of variables.

Bijective function

$$x = g(y)$$



$$P_y(y, y+\delta y) = P_x(g(y), g(y+\delta y))$$

$$P_y(y)\,\delta y \approx P_x(x, x+\delta x) = P_x(x)\,\delta x$$

$$\tilde{f}(y) = f(g(y))$$

Find the density $p_y(y)$

$(x + \delta x)$ is transformed into the range $(y, y + \delta y)$

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

$$\Rightarrow \quad p_y(y) = p_x(x)\left|\frac{dx}{dy}\right|$$

$$= p_x(g(y))\,|g'(y)|$$

Cumulative (large)

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

(small)

density

Multivariate continuous variables $x_1, .. x_D$, denoted by $\mathbf{x}$.

joint probability density $p(\mathbf{x}) = p(x_1, x_2, \ldots x_D)$ s.t

$p($ multi. variable falling in an infinitesimal volume $\delta\mathbf{x}$ containing $\mathbf{x})$ $\approx p(\mathbf{x})\,\delta\mathbf{x}$

$p(\mathbf{x})$: real valued function

Multivariate probability density

$$p(\mathbf{x} \in E) = \int_E p(\mathbf{x})\, d\mathbf{x}$$

measurable set $E \subset \mathbb{R}^D$

$$p(\mathbf{x}) \geq 0$$

$$\int_{\mathbb{R}^d} p(\mathbf{x})\, d\mathbf{x} = 1$$

We can also consider joint probability distributions over a combination of discrete and continuous

In the discrete case

$p(x)$ is called probability mass function.

1.2.2 Expectations and covariances

Variable x under a probability distribution $p(x)$

(mass
(density

Expectation of $f(x)$

In discrete case,

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

In continuous case

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

Random variable x.

Expectation of x is denoted by

discrete $\mathbb{E}[x] := \sum_x p(x)\, x$ $\quad (f(x) = x)$

continuous $= \int p(x)\, x\, dx$

$\mathbb{E}[f]$ can be approximated as a finite sum (N samples)

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

Average of the function $f(x,y)$ w.r.t distribution of x

$$\mathbb{E}_x[f(x,y)] = \int f(x,y)\, p(x)\, dx$$ (function of y)

$$\left(\text{or} \quad \sum_x f(x,y)\, p(x) \right)$$

Conditional expectation

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

(function of y)

Variance of $f(x)$ — exp. of f

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

$$\Rightarrow = \mathbb{E}[f(x)^2] - \mathbb{E}[f(x)]^2$$

As $\mathbb{E}[x]$, $\mathrm{Var}[x] := \mathbb{E}[(x - \mathbb{E}[x])^2]$ $(f(x) = x)$

Two variables x and y. covariance is defined by

$$\mathrm{cov}[x,y] = \mathbb{E}_{x,y}[\{x - \mathbb{E}[x]\}\{y - \mathbb{E}[y]\}]$$

$p(x,y)$

$$= \mathbb{E}_{x,y}[xy] - \mathbb{E}[x]\mathbb{E}[y]$$

x 와 y 가 얼마나 같이 변동하는가

If x and y are independent, $\mathrm{cov}[x,y] = 0$

$$\iint x \cdot y \, p(x,y) \, dx \, dy = \iint x \, y \, p(x) \, p(y) \, dx \, dy$$

Two vectors of random variables $\mathbf{x}$ and $\mathbf{y} \in \mathbb{R}^D$.

Let $f(\mathbf{x})$ be a multivariable function (can be vector function)

$$\mathbb{E}[f] := \sum p(\mathbf{x}) f(\mathbf{x}) = \int f(\mathbf{x}) p(\mathbf{x}) d\mathbf{x}$$

$f(\mathbf{x}) = \mathbf{x}$

If $f(\mathbf{x})$ is a vector function, $\mathbb{E}[f]$ is a vector

Covariance matrix is defined by

$$\underset{D\times D}{\mathrm{cov}[\mathbf{x},\mathbf{y}]} := \mathbb{E}_{\mathbf{x},\mathbf{y}}\left[\{\mathbf{x} - \underset{D\times 1}{\mathbb{E}[\mathbf{x}]}\}\{\underset{1\times D}{\mathbf{y}^T} - \mathbb{E}[\mathbf{y}^T]\}\right]$$

$$= \mathbb{E}_{\mathbf{x},\mathbf{y}}[\mathbf{x}\mathbf{y}^T] - \mathbb{E}[\mathbf{x}]\mathbb{E}[\mathbf{y}^T]$$

$p(\mathbf{x},\mathbf{y})$

$$\mathrm{Var}[\mathbf{x}] = \mathrm{cov}[\mathbf{x},\mathbf{x}]$$

i,j component

$= x_i, y_j$ of variance

1.2.3 Bayesian probability

$$p(Y|x) = \frac{p(x|Y)\,p(Y)}{p(x)} = p(x, Y)$$

비례

posterior 사후

$$p(Y|x) \propto p(x|Y)\,p(Y)$$

likelihood 가능도 함수

prior 사전분포

Classical frequentist vs Bayesian

- frequencies of random (repeatable events)

  데이터 수가 적을 때 over-fitting

- quantification of uncertainty 정량

  데이터 수가 적은 경우 유리

직관적인 확률은 일반적으로 classic

빨간 상자를 선택했을 확률 vs 사과가 나왔을 때 빨간 상자를 선택했을 확률

유리는 정보(data)를 활용함. 정보에 의해서 예측값 변화를 반영

Polynomial fitting curve

Aim to predict $\mathbf{w} = (w_0, w_1, \ldots w_M)$

$$y(x, \mathbf{w}) := w_0 + w_1 x + \cdots w_M x^M \quad (M\text{-th order})$$

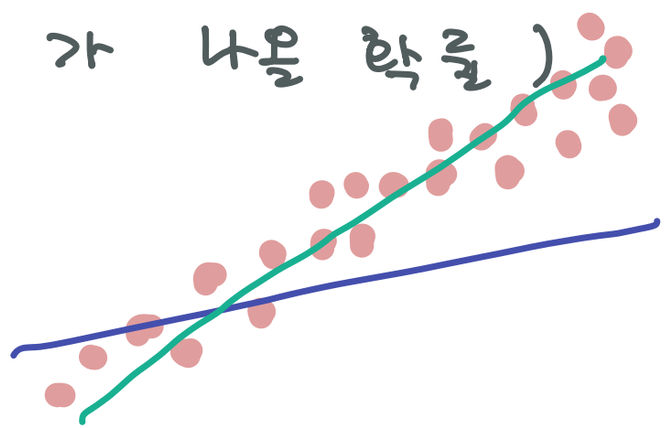
Before observing data, assume $\mathbf{w}$ is in the form $p(\mathbf{w})$ prior

The effect of $D = \{t_1 \ldots t_N\}$ is expressed through

$$p(D|\mathbf{w})$$

($\mathbf{w}$ 가 해당 데이터 D 가 나올 확률)

$$t \sim N(x | y(x, \mathbf{w}), \sigma^2)$$



Uncertainty in w after we have observed D

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

posterior: $p(w|D)$ · likelihood: $p(D|w)$ · prior: $p(w)$

Since data set D is given, $p(D|w)$ is a function of w called $p(D|w)$ likelihood function (가능도 함수)

Remark

- $p(D|w)$ is not a prob. distribution over w
- $\int p(D|w)\,dw \neq 1$

$$\int p(x|y)\,dx = 1$$

$$\int p(x|y)\,dy \neq 1$$

- $p(\mathcal{D})$ can be seen as a normalization constant
- Likelihood $p(\mathcal{D}|w)$ has an important role both approaches

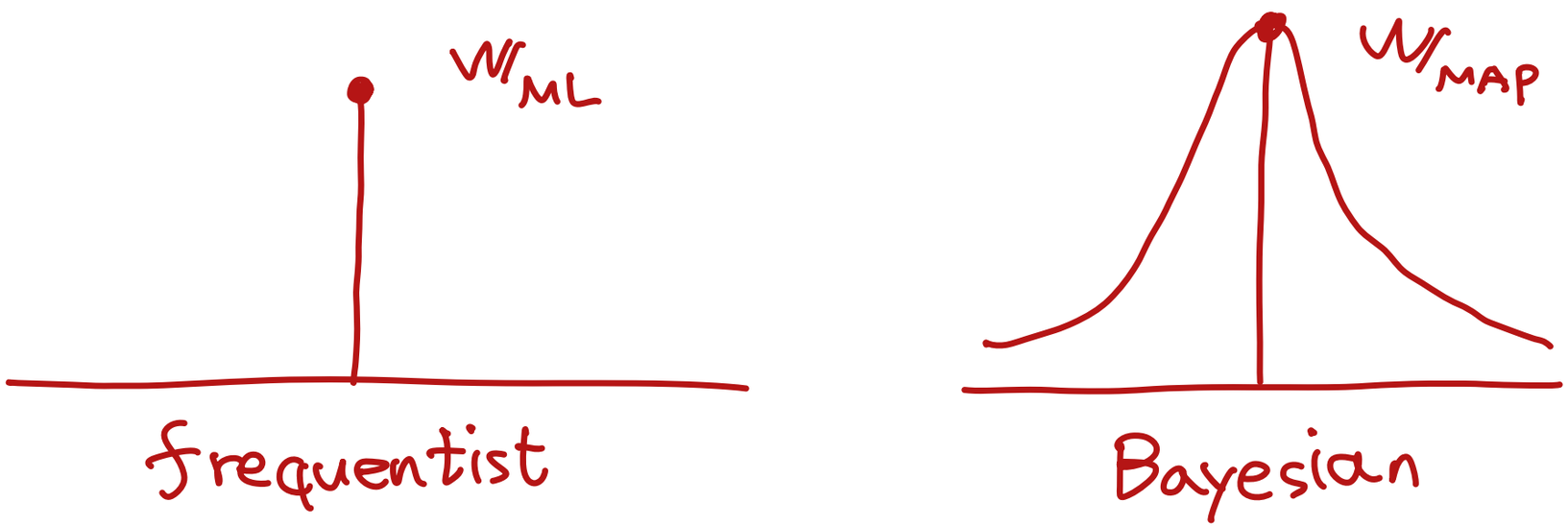
$$\text{posterior} \propto \text{likelihood} \times \text{prior}$$

Frequentist (point estimator)

- w is considered to be a fixed parameter whose value is by some form of estimator

Bayesian (distribution estimator)

- There is only a single data set D and uncertainty in the parameter is expressed through a probability distribution over w



MLE

MAP

빈도론에서 주로 사용하는 방법으로 maximum likelihood

error function = −likelihood or −log(likelihood)

Remark (베이지안 특징)

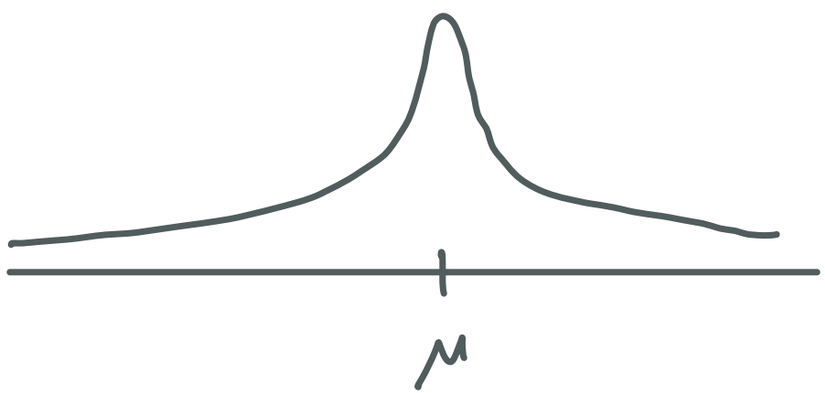
- 사전지식 (사전 분포)을 주로 과정에 반영할 수 있음
- But 일반적으로 사전 분포를 실제 사정의 믿음 보다는 수학적 편의성을 고려해서 선택 (conjugate)
- 주관적인 의견이 주로 과정에 반영
- Marginalization (주변화) over w/ 필요.

03-17

1.2.4 Gaussian distribution

Real-valued variable x, Gaussian distribution is defined by

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



unimodal (단봉)

μ: mean, σ^2: variance (2 parameters)

σ: standard deviation, $\beta = \frac{1}{\sigma^2}$: precision

Observe

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

$$\int_{-\infty}^{\infty} \mathcal{N}(x \mid \mu, \sigma^2)\, dx = 1 \qquad (*)$$

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

second moment

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

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

Let $I := \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2\sigma^2} x^2\right) dx$

$$I^2 = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2\sigma^2} x^2\right) \exp\left(-\frac{1}{2\sigma^2} y^2\right) dx\, dy$$

$$= \int_0^{2\pi} \int_0^{\infty} \exp\left(-\frac{1}{2\sigma^2} r^2\right) r\, dr\, d\theta \qquad \begin{aligned} x &= r\cos\theta \\ y &= r\sin\theta \end{aligned}$$

$$= 2\pi \int_0^{\infty} \exp\left(-\frac{1}{2\sigma^2} r^2\right) r\, dr$$

$$= 2\pi\sigma^2$$

$$\Rightarrow \quad I = \sqrt{2\pi}\,\sigma$$

D-dimensional multivariate Gaussian distribution

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

$1 \times D$ $D \times D$ $D \times 1$

determinant

$\boldsymbol{\mu}$: mean vector , $\boldsymbol{\Sigma}$: covariance matrix ($D \times D$)

(see later!)

$\mathbf{x} \in \mathbb{R}^{D}$ random vector

N scalar variables $x_1, \ldots x_N$. Let $\mathbf{x} = (x_1, \ldots x_N)^T$

Assume x_i are drawn from a normal dist.

How to find the parameters μ and σ^2

(parametric inference)

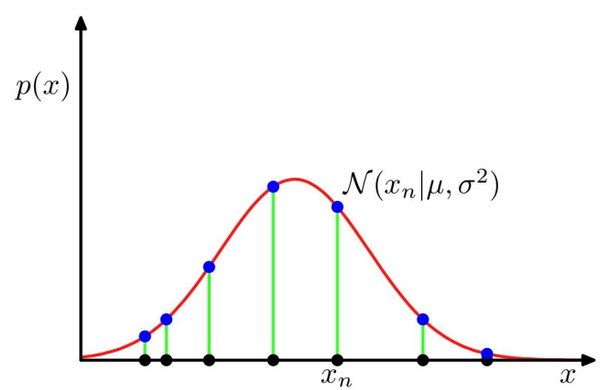
x_i are samples the same distribution independently (iid)

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

likelihood

Maximize likelihood

Figure 1.14 Illustration of the likelihood function for a Gaussian distribution, shown by the red curve. Here the black points denote a data set of values $\{x_n\}$, and the likelihood function given by (1.53) corresponds to the product of the blue values. Maximizing the likelihood involves adjusting the mean and variance of the Gaussian so as to maximize this product.



maximize likelihood $\Leftrightarrow$ maximize log (likelihood)

$$\ln p(\mathbf{x} \mid \mu, \sigma^2) = -\frac{1}{2\sigma^2} \sum_{n=1}^{N} (x_n - \mu)^2 - \frac{N}{2} \ln \sigma^2 - \underbrace{\frac{N}{2} \ln (2\pi)}_{\text{constant}}$$

Maximizing wrt μ,

$$\mu_{ML} = \frac{1}{N} \sum_{n=1}^{N} x_n \qquad \text{(sample mean)}$$

Maximizing wrt σ^2

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

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

Fix N and some dist. μ_{ML} and σ^2_{ML} are functions of the choice $x_1 \ldots x_N$

$$\mathbb{E}[\mu_{ML}] = \mu \quad \text{unbiased}$$

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

$$\mathbb{E}[\mu_{ML}] = \mathbb{E}\left[\frac{1}{N}\sum x_n\right] = \frac{1}{N}\sum \mathbb{E}[x_n] = \mu$$

1.2.5 Curve fitting (re-visited)

Probabilistic perspective

input value x, target value t

N input values $\mathbf{x} = (x_1, \ldots x_N)^T$

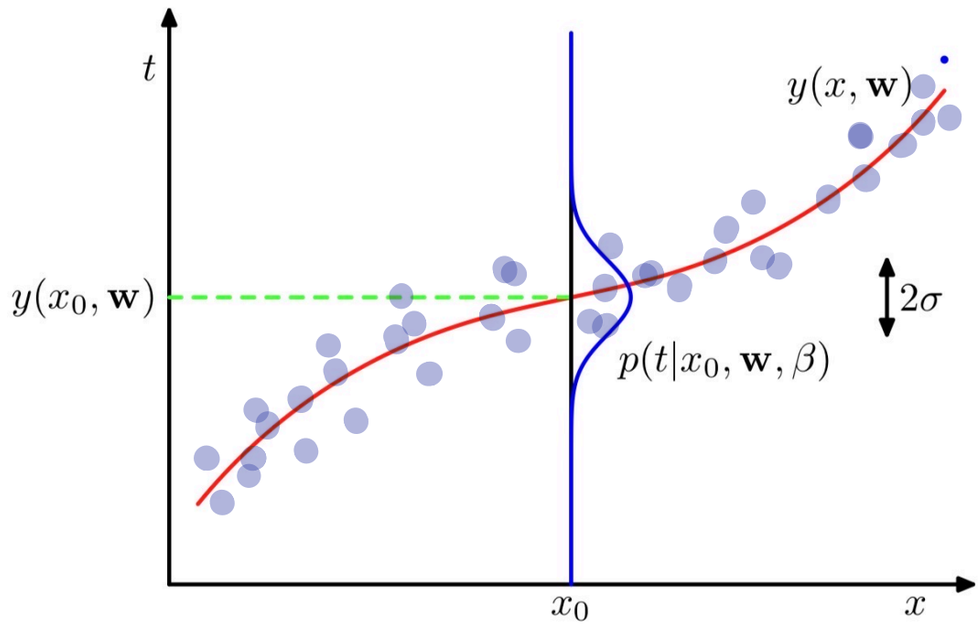
" target " $\mathbf{t} = (t_1, \ldots t_N)^T$

Predict target value t for some new input x

Express uncertainty over target t using prob. dist

Assume

$$p(t \mid x, \mathbf{w}, \beta) = \mathcal{N}\left(t \mid y(x, \mathbf{w}), \frac{1}{\beta}\right) \qquad (1.60)$$



$y(x, w)$

Figure 1.16 Schematic illustration of a Gaussian conditional distribution for t given x given by (1.60), in which the mean is given by the polynomial function $y(x, \mathbf{w})$, and the precision is given by the parameter β, which is related to the variance by $\beta^{-1} = \sigma^2$.

(t : 평균이 $y(x, w)$ 이고 분산이 $\frac{1}{\beta}$ 인 Gaussian)

Determine $\mathbf{w}$ and β by MLE.

Assume data $\mathbf{X}$ and $\mathbf{t}$ are drawn independently from (1.60)

$$p(\mathbf{t} \mid \mathbf{X}, \mathbf{w}, \beta) = \prod_{n=1}^{N} \mathcal{N}(t_n \mid y(x_n, \mathbf{w}), \frac{1}{\beta})$$

($\mathbf{w}$: assumed to be fixed)

by def

$$= \prod \left(\frac{\beta}{2\pi}\right)^{1/2} \exp\left\{-\frac{\beta}{2}\left(y(x_n, \mathbf{w}) - t_n\right)^2\right\}$$

Maximize log likelihood wrt $\mathbf{w}$ and β

$$\ln p(\mathbf{t} \mid \mathbf{X}, \mathbf{w}, \beta) = -\frac{\beta}{2} \sum_{n=1}^{N} \left\{y(x_n, \mathbf{w}) - t_n\right\}^2 + \frac{N}{2} \ln \beta - \frac{N}{2} \ln(2\pi)$$

solutions for $\mathbf{w}_{ML}$, β_{ML} !

$\mathbf{w}_{ML}$ → $y(x, \mathbf{w}_{ML})$, $\hat{x}$

Point prediction (not use β_{ML})

New input x, $t := y(x, w_{ML})$

Distribution prediction (not Bayesian)

$$p(t \mid x, w_{ML}, \beta_{ML}) = \mathcal{N}(t \mid y(x, w_{ML}), \beta_{ML}^{-1})$$

Take a step towards a more Bayesian approach.

prior distribution over $\mathbf{w}$ as the form $\begin{pmatrix} \alpha & & 0 \\ & \ddots & \\ 0 & & \alpha \end{pmatrix}$

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

where α is a precision and $M+1$ is # of coeff. $\mathbf{w}$

($y(x, \mathbf{w})$ M th order polynomial, $w_0, w_1, \ldots w_M$)

α is called hyperparameter (parameter of parameter)

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

likelihood

데이터 $\mathbf{x}, \mathbf{t}$ 에 대해 가능성이 높은 $\mathbf{w}$ 를 찾기.

Maximize posterior distribution (MAP, maximum posterior)

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

see later!

$\mathbf{w}_{ML}$ $\mathbf{w}_{MAP}$

$y(x, \mathbf{w}_{MAP})$ $\hat{t}$ 예측

$\hat{x}$

1.2.6 Bayesian curve fitting

We still make a point estimate of w.

In a fully Bayesian approach, we should apply sum and product rules of probability. So we need to integrate over all values of w. (Marginalization)

Compared with MAP (maximize posterior), we will use posterior $p(w \mid x, t)$ (x, t training data set)

all information

Predict the value of t for a new test point x

⇒ Evaluate $p(t \mid x, \mathbf{x}, \mathbf{t})$

(새로운 x 값에 대한 예측값 t 의 분포, 기존데이터 정보 활용)

Now α and β are assumed to be given and fixed.

$$p(t \mid x, \mathbf{x}, \mathbf{t}) = \int \underbrace{p(t \mid x, w)}_{(1.60)} \underbrace{p(w \mid \mathbf{x}, \mathbf{t})}_{\text{posterior}} \, dw$$

$$\mathcal{N}(t \mid y(x, w), \beta^{-1})$$

assumption

We will see and perform analytically later. The result is given by a Gaussian

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

where

$$m(x) = \beta \boldsymbol{\phi}(x)^T \mathbf{S} \sum_{n=1}^{N} \boldsymbol{\phi}(x_n) t_n$$

$$s^2(x) = \beta^{-1} + \boldsymbol{\phi}(x)^T \mathbf{S} \boldsymbol{\phi}(x)$$

Here matrix $\mathbf{S}$ is given by

$$\mathbf{S}^{-1} = \alpha \mathbf{I} + \beta \sum_{n=1}^{N} \boldsymbol{\phi}(x_n) \boldsymbol{\phi}(x_n)^T$$

$\mathbf{I}$: unit matrix ($M+1$ dim), $\boldsymbol{\phi}(x) = (\phi_0(x), \ldots \phi_M(x))^T$. $\phi_i(x) = x^i$

we will generalize later

Recall assumptions

- Prior distribution over w: $p(w|\alpha) = \mathcal{N}(w|0, \alpha^{-1}I)$ $M+1$
- likelihood: $p(t|X, w, \beta) = \prod_{n=1}^{N} \mathcal{N}(t_n | y(x_n, w), \beta^{-1})$
- posterior $p(w|x, t, \alpha, \beta) \propto p(t|X, w, \beta)\, p(w|\alpha)$

$$p(t|x, X, t) = \int p(t|x, w)\, p(w|x, t)\, dw$$

likelihood of single point

dependent of x

Curve fitting

$$y(x, \mathbf{w}) := \sum_{j=1}^{M} x^j w_j \qquad (M \text{ fixed})$$

Set any error function

Find $\mathbf{w}^*$ s.t.

minimize error function

$y(x, \mathbf{w}^*)$

model assumption

Assume $t \sim \mathcal{N}(t \mid y(x, \mathbf{w}), \frac{1}{\beta})$

likelihood $\prod_{n=1}^{N} \mathcal{N}(t_n \mid y(x_n, \mathbf{w}), \frac{1}{\beta})$

prior $\mathbf{w} \sim \mathcal{N}(\mathbf{w} \mid 0, \alpha^{-1} I)$

$\Rightarrow$ posterior $p(\mathbf{w} \mid \mathbf{x}, \mathbf{t}, \alpha, \beta)$

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

$\Rightarrow p(t \mid x, \mathbf{x}, \mathbf{t})$

$= \int p(t \mid x, \mathbf{w})\, p(\mathbf{w} \mid \mathbf{x}, \mathbf{t})\, d\mathbf{w}$

1.3 Model selection

of parameter (coefficients) ≈ model complexity

The performance on the training set is not a good indicator of predictive performance (because of over-fit)

S-fold cross-validation

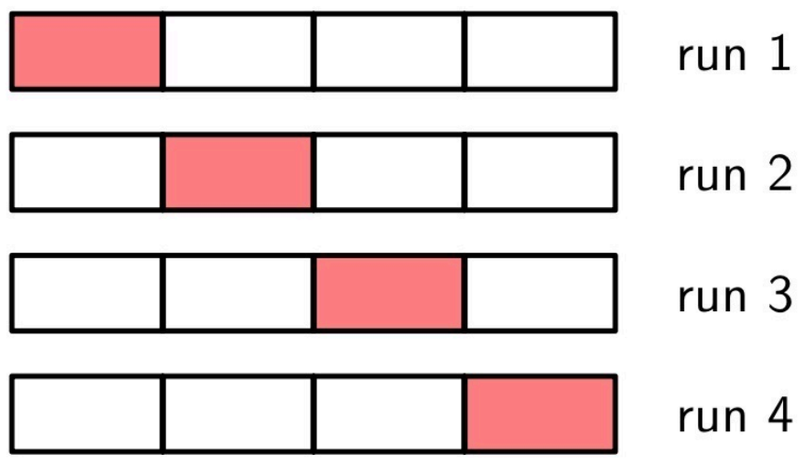


Figure 1.18 The technique of S-fold cross-validation, illustrated here for the case of $S = 4$, involves taking the available data and partitioning it into S groups (in the simplest case these are of equal size). Then $S - 1$ of the groups are used to train a set of models that are then evaluated on the remaining group. This procedure is then repeated for all S possible choices for the held-out group, indicated here by the red blocks, and the performance scores from the S runs are then averaged.

When S increases, % of training runs is increases

Another method is to add penalty term

E.g. Akaike information criterion

$$\ln p(D \mid w_{ML}) - M$$

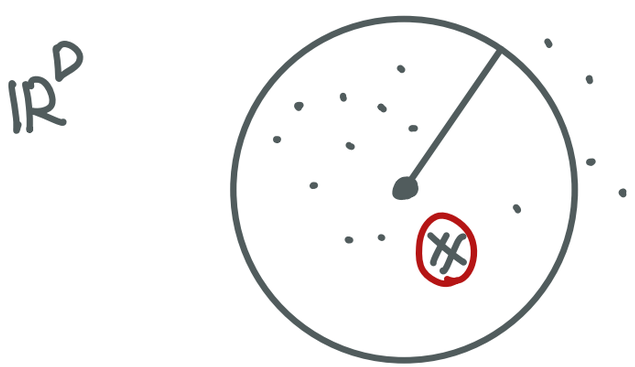
where M is % of parameters

1.4 The curse of Dimensionality

The dimension of input vector x

결론 : 차원이 높다고 (feature 의 수가 많다고) 무조건 좋은건 X

- 높은 차원 만큼 많은 데이터가 필요함
- 차원수 (또는 D^M) 에 비례한 coefficients 를 찾아야 함
- 고차원 에서의 직관은 저차원의 직관과 매우 다름



$$V_D(r) = K_D r^D$$

$$\frac{V_D(1) - V_D(1-\epsilon)}{V_D(1)} = 1 - (1-\epsilon)^D$$

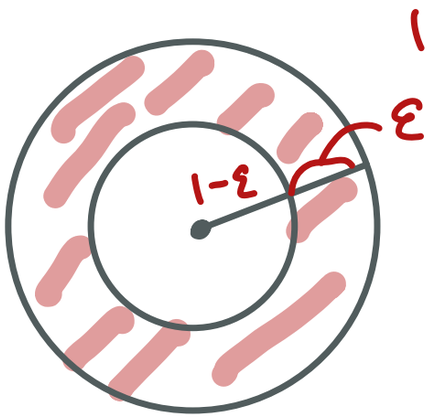
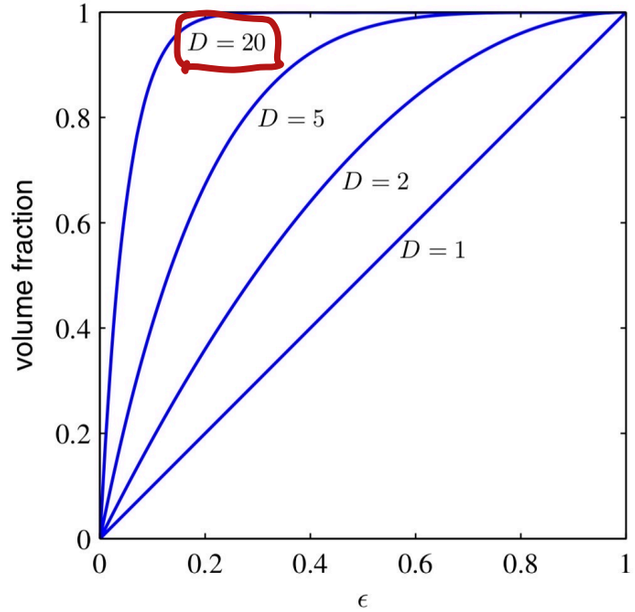


Figure 1.22 Plot of the fraction of the volume of a sphere lying in the range $r = 1-\epsilon$ to $r = 1$ for various values of the dimensionality D.



Generally, real data is confined to a region of the space having lower dimensionality (manifold)

1.5 Decision Theory

$p(x,t)$ $p(t|x)$?

Input vector x, target vector t

For regression, t comprise continuous value 회귀

For classification, t represent class label

Joint prob. $p(x,t)$ provides a 'complete' summary of the uncertainty associated with x, t.

Determination of $p(x,t)$ is an typical example. (difficult)

Input image x (x-ray)

binary classification

Class label C_1, C_2

Inference: modeling $p(x, C_k)$

Decision: for given x, decide which of two classes x is.

We are interested in $p(C_k \mid x)$ ✓

By Bayes's theorem,

class density

prior

$$P(C_k \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid C_k)\, P(C_k)}{p(\mathbf{x})}$$

posterior

Any quantities appearing in theorem can be obtained from $p(\mathbf{x}, C_k)$ by either marginalizing or conditioning

Minimize the chance to assigning $\mathbf{x}$ to wrong class

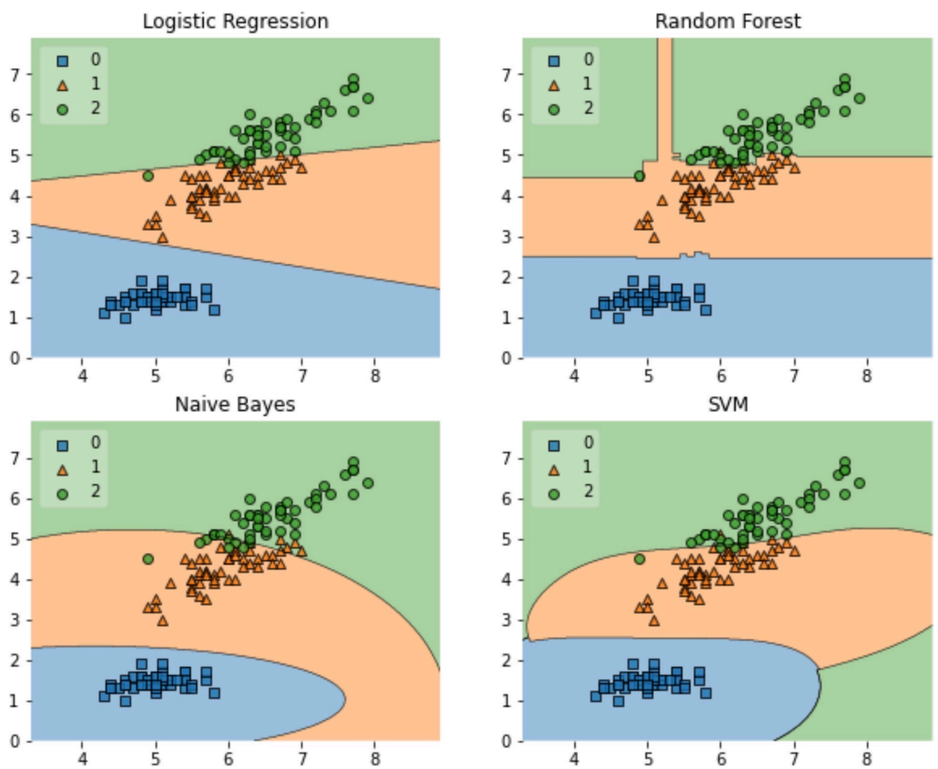
$\Rightarrow$ Choose the class having the higher posterior prob.

1.5.1 Minimizing the misclassification rate

Decision regions R_k if $x \in R_1$ then x

- Regions in domain space assigning points to class C_k

Decision boundary: boundaries between R_k



Suppose our goal is to make as few misclassification as possible.

$$p(\text{mistake}) = p(x \in R_1, C_2) + p(x \in R_2, C_1)$$

$$\text{error} = \int_{R_1} p(x, C_2)\, dx + \int_{R_2} p(x, C_1)\, dx$$

Minimize $p(\text{mistake})$!!

If $p(x, C_1) > p(x, C_2)$, assign that x to C_1 ($x \in R_1$)

if $p(x, C_2) > p(x, C_1)$, " x to C_2 ($x \in R_2$)

$$p(x, C_1) = p(C_1 \mid x)\, p(x), \qquad p(x, C_2) = p(C_2 \mid x)\, p(x)$$

In order to minimize $p(\text{mistake})$,

assign x to the class for which $p(C_k \mid x)$ is largest.

General case of K classes

$$p(\text{correct}) = \sum_{k=1}^{K} p(x \in R_k, C_k)$$

$$= \sum_{k=1}^{K} \int_{R_k} p(x, C_k)\, dx$$

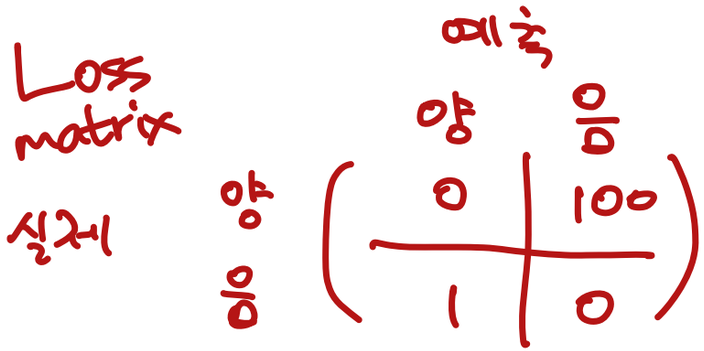
$p(x, C_k)$ ($\propto p(C_k \mid x)$) 가 최대가 R_k 로 배정

1.5.2 Minimizing the expected loss

cost function (loss)

The optimal solution is the one which minimizes the loss function

Loss matrix



Confusion matrix

E.g. loss matrix

Let L_{kj} be the loss in case that true class is C_k but it is assigned to class C_j (generally $L_{kk} = 0$)

$$\mathbb{E}[L] = \sum_k \sum_j \int_{R_j} L_{kj}\, p(\mathbf{x}, C_k)\, d\mathbf{x}$$

C_k에 있는데 C_j로 판정한 경우

Minimize $\mathbb{E}[L]$ iff minimize $\sum_k L_{kj}\, p(C_k \mid \mathbf{x})$

1.5.3 The reject option

$p(\mathbf{x}, C_k)$

$\propto p(C_k \mid \mathbf{x})$

1.5.4 Inference and decision

Inference stage : model for $p(C_k|x)$ (target: C_k, input: x) $p(x, C_k)$

Decision stage : function mapping x into decision

(decision or discriminant)

Identify three distinct approaches

(a) Find $p(C_k \mid x)$

- class conditional densities $p(x \mid C_k)$
- prior class probabilities $p(C_k)$ (simple)
- $$p(C_k \mid x) = \frac{p(x \mid C_k)\, p(C_k)}{p(x)}$$ $p(x, C_k)$
- $$p(x) = \sum_k p(x \mid C_k)\, p(C_k)$$
- Equivalently, model the joint distribution $p(x, C_k)$ directly and then obtain $p(C_k \mid x)$
- Generative model: modeling the distributions of x, C_k

(b) Find $p(C_k|x)$ directly, apply decision theory

- Call it discriminative model
- Almost deep learning models

5 : 5 와닿음

x test

$p(C_k|x)$??

(c) Find discriminant function

- mapping each x directly onto a class label
- probabilities play no role

Compensating for class prior

- Imbalanced data classification (difficult to generalize)
- $$P(C_k|x) = \frac{P(x|C_k)\,P(C_k)}{P(x)}$$

Combining models

E.g. Medical diagonosis

- blood test, X-ray image
- 두 정보를 통합하기 vs 개별 진단 (사후 확률) 모델 만들기

Conditionally independent

A and B are conditionally independent given C $(A \perp\!\!\!\perp B \mid C)$

$$P(A, B \mid C) = P(A \mid C) \cdot P(B \mid C)$$

prop $A \perp\!\!\!\perp B \mid C \iff P(A \mid B, C) = P(A \mid C)$

양 or 정상

Suppose $x_I \perp\!\!\!\perp x_B \mid C_k$ (assumptions of Naive Bayes classifier)

$$\Rightarrow \quad P(C_k \mid x_I, x_B) \propto P(x_I, x_B \mid C_k)$$

$$= P(x_I \mid C_k)\, P(x_B \mid C_k)\, P(C_k)$$

$$\propto P(C_k \mid x_I)\, P(C_k \mid x_B) / P(C_k)$$

1.5.5 Loss functions for regression

Prediction $y(\mathbf{x})$ of the value t for each $\mathbf{x}$

Let $L(t, y(\mathbf{x}))$ be the loss function. (depends on $y(\mathbf{x})$)

$$\mathbb{E}[L] = \iint L(t, y(\mathbf{x}))\, p(\mathbf{x}, t)\, d\mathbf{x}\, dt \approx \frac{1}{N} \sum L(t_n, y(\mathbf{x}_n))$$

In case $L(t, y(\mathbf{x})) := \| y(\mathbf{x}) - t \|^2$

$$\mathbb{E}[L] = \iint \| y(\mathbf{x}) - t \|^2\, p(\mathbf{x}, t)\, d\mathbf{x}\, dt \approx \frac{1}{N} \sum (t_n - y(\mathbf{x}_n))^2$$

MSE

$y(\mathbf{x})$

Goal: choose $y(x)$ as so minimize $\mathbb{E}[L]$

Assume $y(x)$ is completely flexible

$y \to L \to \mathbb{E}[L]$ (expected loss)

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

Functional F, $F[y] = \int G(y(x), y'(x), x) \, dx$ (D.5)

$$\Rightarrow \frac{\delta F}{\delta y} = \frac{\partial G}{\partial y} - \frac{d}{dx}\left(\frac{\partial G}{\partial y'}\right)$$

Since $\mathbb{E}[L] = \int \underbrace{\int \{ y(x) - t \}^2 \, p(x,t) \, dt}_{G(x, y, \dot{y})} \, dx$

$$\frac{\delta E[L]}{\delta y(x)} = \frac{\partial G}{\partial y} - \frac{d}{dx}\left(\frac{\partial G}{\partial \dot{y}}\right) = 2\int \{y(x) - t\}\, p(x,t)\, dt$$

In case $\{y(x)-t\}^2$

Continue to minimize $E[L]$ for $y(x)$

$$y(x) = \frac{\int t\, p(x,t)\, dt}{p(x)} = \int t\, p(t|x)\, dt =: \underline{E_t[t|x]}$$

function of x

We used $\int p(x,t)\, dt = p(x)$.

For multivariate target $\#$, the optimal solution can be expressed as

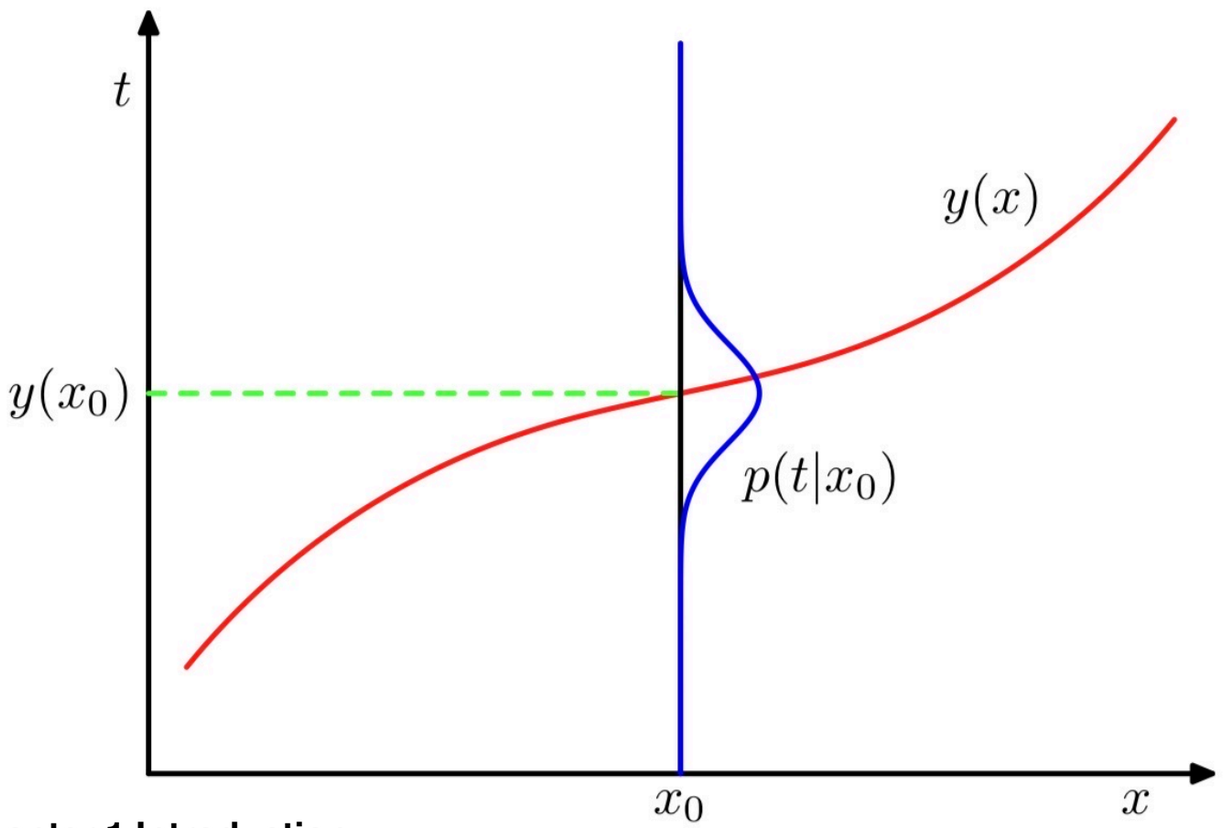
$$E_{\#}[\#|x]$$

$L = \{y(x) - t\}^2$ 인 경우 (MSE) $\mathbb{E}[L]$ (expected loss) 를

최소화 하는 $y(x) = \mathbb{E}_t[t|x] = \int t\, p(t|x)\, dt$ (conditional exp.)

x 에 대한 함수

Figure 1.28 The regression function $y(x)$, which minimizes the expected squared loss, is given by the mean of the conditional distribution $p(t|x)$.



Note that

$$\{y(\mathbf{x}) - t\}^2 = \{y(\mathbf{x}) \pm \mathbb{E}[t|\mathbf{x}] - t\}^2$$

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

$$+ \{\mathbb{E}[t|\mathbf{x}] - t\}^2$$

(1) (2) (3)

① $\iint \{y(x) - E[t|x]\}^2 p(x,t)\,dx\,dt$ $\qquad p(x,t) = p(t|x)\cdot p(x)$

$= \int \{y(x) - E[t|x]\}^2 p(x)\,dx \cdot \underbrace{\int p(t|x)\,dt}_{=1}$

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

$= \int \{y(x) - E[t|x]\} \underbrace{\left(\int \{E[t|x] - t\}\, p(t|x)\,dt \right)}\, p(x)\,dx$

$= E[t|x] \int p(t|x)\,dt - \int t\, p(t|x)\,dt$

$= 0$

③ $\iint \{E[t|x] - t\}^2 \, p(x,t) \, dx \, dt = \iint \{E[t|x] - t\}^2 \, p(t|x) \, p(x) \, dx \, dt$

mean

$= \int \left(\int \{E[t|x] - t\}^2 \, p(t|x) \, dt \right) p(x) \, dx$

$= \int \text{var}[t|x] \, p(x) \, dx$

typo

Thus,

minimum of E[L]

noise

$$E[L] = \int \{y(x) - E[t|x]\}^2 \, p(x) \, dx + \int \text{var}[t|x] \, p(x) \, dx$$

without $y(x)$

(optimal least squares predictor is given by the conditional mean)

Three distinct approaches to solving regression problems

$$y(x) = \mathbb{E}_t[t|x] = \int t\, p(t|x)\, dt \qquad (1.89)$$

(a) Determining $p(x,t)$, normalizing to find $p(t|x)$.

Calculating the conditional mean given by (1.89)

(b) Solving the inference of determining $p(t|x)$ directly.

Calculating the conditional mean given by (1.89)

(c) Find $y(x)$ directly from the training data.

classic

Bayesian

Minkowski loss

$$\mathbb{E}[L_q] = \iint |y(\mathbf{x}) - t|^q \, p(\mathbf{x}, t) \, d\mathbf{x} \, dt$$

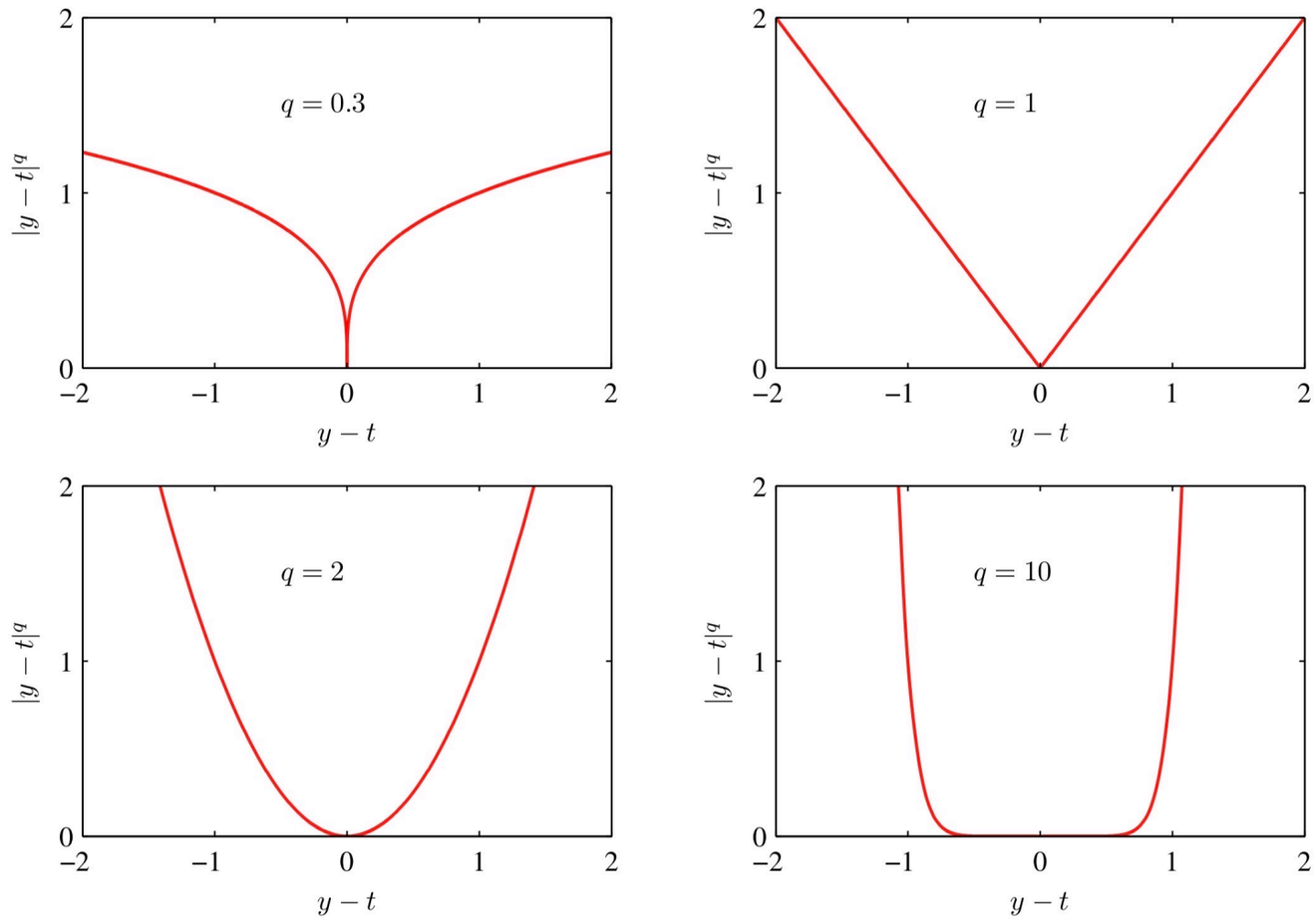


Figure 1.29 Plots of the quantity $L_q = |y - t|^q$ for various values of q.

1.6 Information Theory

Discrete random variable x

The amount of information can be view as 'degree of suprise' on learning x. (가능성 낮은 사건에 크게 놀람)

우리나라가 월드컵 우승할 확률 vs 브라질이 우승할 확률

사건이 발생 했을 때 놀라는 정도

Measure of information depends on $p(x)$

$h(x)$ express the information content

Consider the unrelated (independent) events x, y

$$h(x,y) = h(x) + h(y)$$ 각각 따로 일어났을 때 정보량의 합

함께 일어났을 때 얻는 정보량

(When $x \perp\!\!\!\perp y$, $p(x,y) = p(x) \cdot p(y)$)

Let

$$h(x) := -\log_2 p(x)$$ (amount of information)

$-\int p(x) \ln p(x) \, dx$

Entropy

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

정보 전송에 필요한 평균 정보량

$\parallel$

$\mathbb{E}[h(x)]$

Take $p(x)\log_2 p(x) := 0$ when $p(x) = 0$.

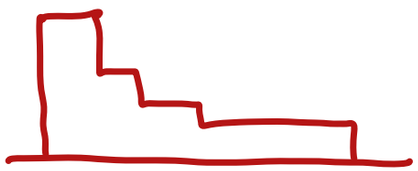
Eg. discrete random variable x having 8 possible states

(length 3 bits)

① uniform prob $(\frac{1}{8}, \frac{1}{8}, \dots \frac{1}{8})$ case

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

② prob $(\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64})$



$$H[x] = 2$$

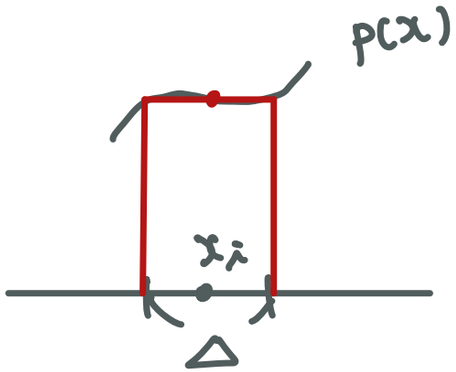
We see that non uniform distribution has a smaller entropy than uniform one.

When x is discrete rv, uniform distribution has the maximum entropy.

Introduction to entropy analysis

Let us define entropy of conti. rv x with pdf $p(x)$.

Divide x into bins of width Δ and assume $p(x)$ is continuous. By mean value theorem, $\exists\ x_i \in [i\Delta, (i+1)\Delta]$ s.t

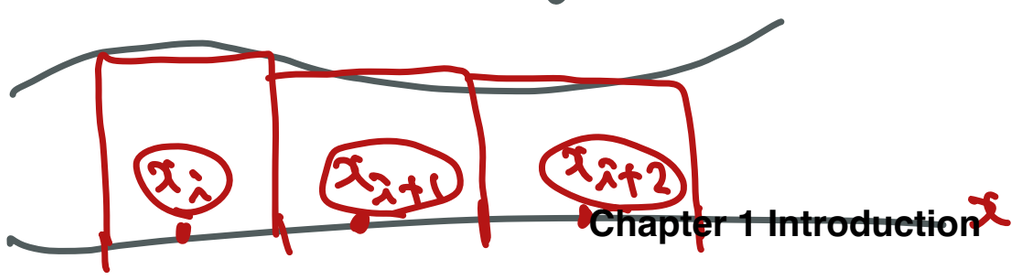


$$\int_{i\Delta}^{(i+1)\Delta} p(x)\,dx = p(x_i)\Delta$$

(convert to step function)

Assigning any value $x \in i$th bins to x_i

$\Rightarrow$ the prob. of observing the values in i^{th} bins = $p(x_i)\Delta$



Entropy of such discrete distribution $\log_2 \to \ln$

probability

$$H_\Delta = -\sum_i p(x_i)\Delta \ln(p(x_i)\Delta) = -\sum_i p(x_i)\Delta \ln p(x_i) - \ln\Delta$$

where we used $\sum_i p(x_i)\Delta = \int p(x)\,dx = 1$.

Now we omit $-\ln\Delta$ and consider $\Delta \to 0$

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

$\mathbb{E}[h(x)]$

called 'differential entropy' (continuous form of entropy)

Multivariate continuous variable

$$H[x] := -\int p(x) \ln p(x) \, dx$$

$x \longrightarrow x$

In the case of discrete distributions, uniform distribution has the maximum entropy.

Let us consider the maximum entropy configuration for a continuous variable.

$(-\infty, \infty)$ r.v x

Constraints

$p(x) = ?$

$$\int p(x)\,dx = 1$$

$$\int x\,p(x)\,dx = \mu$$

$$\int (x-\mu)^2\,p(x)\,dx = \sigma^2$$

fix

Appendix D

Using Lagrange multiplier and calculus of variations,

we set the derivative of this functional to zero giving

(*) $$p(x) = \exp\{-1 + \lambda_1 + \lambda_2 x + \lambda_3 (x-\mu)^2\}$$

Apply $p(x)$ to constraints

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

(*)

$H[x]$

We need to maximize the following functional w.r.t $p(x)$

$$-\int p(x) \ln p(x)\, dx + \lambda_1 \left(\int p(x)\, dx - 1 \right) + \lambda_2 \left(\int x\, p(x)\, dx - \mu \right)$$
$$+ \lambda_3 \left(\int (x-\mu)^2 p(x)\, dx - \sigma^2 \right)$$

Let

$$F(p(x)) := \int \{ -p(x) \ln p(x) + \lambda_1 p(x) + \lambda_2 x\, p(x) + \lambda_3 (x-\mu)^2 p(x) \}\, dx$$

(functional)

$$+ (-\lambda_1 - \lambda_2 \mu - \lambda_3 \sigma^2)$$

$$= \int G(p(x))\, dx + c$$

$$\frac{\delta F(p(x))}{\delta p(x)} = \frac{\partial G(p)}{\partial p} = -\ln p(x) - p(x)\cdot\frac{1}{p(x)} + \lambda_1 + \lambda_2 x + \lambda_3 (x-\mu)^2$$

$$= 0$$

$$\ln p(x) = \lambda_1 + \lambda_2 x + \lambda_3 (x-\mu)^2 - 1$$

(appendix D)

변분법

Differential entropy of the Gaussian

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

$$\sum -p(x)\ln p(x)$$

$$\geq 0$$

x, y two random variables with joint prob $p(x,y)$

$-\ln p(y|x)$: additional information needed to specify y, when x is given.

$p(y|x)$

Conditional entropy of y given x

$$H[y|x] := -\iint p(y,x) \ln p(y|x)\, dy\, dx$$

We have

$$H[x,y] = H[y|x] + H[x]$$

(x, y)의 정보량
= x 정보량 +
x가 주어졌을 때
y로부터 얻는 정보량

1.6.1 Relative entropy and mutual information

Unknown prob. dist. $p(x)$ (true dist.)

Approximating prob. dist. $q(x)$

Amount of additional information needed to send x using $q(x)$

$$KL(p\|q) := -\int p(x)\ln q(x)\,dx - \left(-\int p(x)\ln p(x)\,dx\right)$$

p 정보량

unknown

approximate

$$= -\int p(x)\ln\left\{\frac{q(x)}{p(x)}\right\}dx$$

Kullback – Leibler divergence (relative entropy)

Remark

- $KL(p \| q) \neq KL(q \| p)$ not symmetric
- $KL(p \| q) \geq 0$ (*)
- $KL(p \| q) = 0$ iff $p = q$
- measurement of diffrence between two prob. dists.

(*) φ: convex function, f: a function of x

prob. measure

$$\varphi\left(\int f(x)\, p(x)\, dx\right) \leq \int \varphi(f(x))\, p(x)\, dx$$

Jensen's inequality

put $\varphi(x) := -\ln(x)$. $f(x) = q(x)/p(x)$

Now suppose $\mathbf{X}$ is generated from unknown $p(\mathbf{x})$.

We try to approximate $p(\mathbf{x})$ using $q(\mathbf{x}|\theta)$ (e.g. parametric dist.)

and observed N points $\mathbf{x}_1, \ldots \mathbf{x}_N$

Monte Carlo Integration

$$\int f(x)\, p(x)\, dx \simeq \frac{1}{N} \sum f(x_n)$$

$$KL(p\|q) = -\int p(\mathbf{x}) \ln\left\{\frac{q(\mathbf{x})}{p(\mathbf{x})}\right\} d\mathbf{x} \simeq \sum_{n=1}^{N} \left\{-\ln q(\mathbf{x}_n|\theta) + \underbrace{\ln p(\mathbf{x}_n)}_{\text{indep. of } q}\right\}$$

To minimize $KL(p\|q)$, we need to maximize

$$\sum \ln q(\mathbf{x}_n|\theta) \quad (\text{log likelihood})$$

θ로 매개화된 q 분포에서 $\mathbf{x}_1 \ldots \mathbf{x}_N$이 나올 확률

$$\ln\left(\prod q(\mathbf{x}_n|\theta)\right) \quad \text{likelihood}$$

Degree of Independence

$$I[X, Y] := KL\left(p(x, y) \,\|\, p(x)\, p(y)\right)$$

mutual information between X, Y

$$0 \leq I[X, Y] = H[X] - H[X|Y]$$

$$= H[Y] - H[Y|X]$$

$X \perp\!\!\!\perp Y \Rightarrow H[X] = H[X|Y]$

The reduction in uncertainty about X as a consequence of the new Y.