$\arg\max_\theta P(D|\theta) = \arg\max_\theta \prod_t p(x^t|\theta)$

likelihood

$= \arg\max_\theta \sum_t \left[-\frac{1}{2}\log 2\pi - \frac{1}{2}\log \sigma^2 - \frac{1}{2}\frac{(x^t-m)^2}{\sigma^2} \right]$

$\frac{dL}{dm} \to \sum_t \frac{(x^t-m)}{\sigma^2} = 0 \to \mu = \frac{\sum x^t}{N} = m$ (true) (sample)

$\frac{dL}{d\sigma^2} \to \sum_t \left[-\frac{1}{2}\frac{1}{\sigma^2} + \frac{1}{2}\frac{(x^t-m)^2}{\sigma^4} \right] = 0 \to \sigma^2 = \frac{\sum (x^t-m)^2}{N} = s^2$ (sample)

input: $D = \{x^t, r^t\}_{t=1}^N$

for each $x^t, r^t \in D$.

$K \leftarrow k$ if $r_k^t = 1$.

$N_k \leftarrow N_k + 1$.

for each dimension i.

$m_{ki} \leftarrow m_{ki} + x_i^t$

$s_{ki} \leftarrow s_{ki} + (x_i^t)^2$

end

end

for each class k.

$P(C_k) \leftarrow \frac{N_k}{N}$

for each dimension i

$M_{ki} \leftarrow \frac{m_{ki}}{N_k}$

$S_{ki} \leftarrow \frac{s_{ki}}{N_k} - M_{ki}^2$.

end

end

$r_i^t = \begin{cases} 1 & \text{if } x^t \in C_i \\ 0 & \text{if } x^t \in C_j,\ j \neq i \end{cases}$

$\theta,\ \theta^*$ 정답.

$\hat\theta_1 \leftarrow D_1 = \{x_1^1, x_1^2, \cdots x_1^N\}$

$\hat\theta_2 \leftarrow D_2 = \{x_2^1, x_2^2, \cdots x_2^N\}$

$E[\hat\theta] - \theta^* \equiv$ Biased

$m = \frac{\sum x^t}{N}$ $E[m] = E\left[\frac{x^1 + \cdots + x^n}{N}\right]$

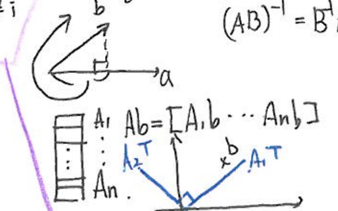
$= \frac{1}{N}(E(x^1) + \cdots + E(x^n)) = M$.

$E[m] - M = 0$ unbiased.

$E[(m - E[m])^2] = \frac{1}{N^2}E[(x^1-m)^2 + \cdots]$

$\frac{\sum x^t}{N}$ $\frac{N}{N}M$ $= \frac{1}{N}\sigma^2$.

$a^Tb = \sum_i a_i b_i$



$(AB)^T = B^TA^T$ $A \cdot A^{-1} = I$.

$(AB)^{-1} = B^{-1}A^{-1}$ $tr(A) = \sum_i A_{ii}$.

$tr(AB) = tr(BA)$

$|AB| = |A||B|$.

$|A^{-1}| = |A|^{-1}$.

$(A^T)^{-1} = (A^{-1})^T = A^{-T}$.

$\frac{d}{dx}A^Tx = \begin{bmatrix} \frac{d}{dx_1}a^Tx \\ \vdots \\ \frac{d}{dx_n}a^Tx \end{bmatrix} \to \begin{bmatrix} a_1 \\ \vdots \\ a_n \end{bmatrix} = a$. 각 변수를 미분, 벡터화.

$\begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$ $\sum_i a_i x_i$

$\frac{d}{dx}x^TAx = (A + A^T)x$

$\frac{d}{dB}tr(AB) = A^T$.

$\frac{d}{dA}\log|A| = A^{-T}$.

$y = Ax$. Matrix, vector, random variable

$E[x] = E\begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} E[x_1] \\ \vdots \\ E[x_n] \end{bmatrix} = M_x$. mean vector / $E[(x-m)(x-m)^T] = \Sigma_x$

$E[y] = AM_x$.

$E\left[\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}\right] = E\left[\begin{pmatrix} a_{11}x_1 + a_{12}x_2 \\ a_{21}x_1 + a_{22}x_2 \end{pmatrix}\right]$

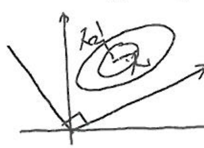
$= AE[x] = AM_x$.

$\begin{bmatrix} E[(x_1-m_1)(x_1-m_1)] & E(x_1-m_1)(x_2-m_2) \\ E(x_2-m_2)(x_1-m_1) & E(x_2-m_2)(x_2-m_2) \end{bmatrix}$

$\arg\max P(D|\theta) = \arg\max p(\theta|D) \times p(\theta)$.

$\Sigma_y = E[(y - E[y])(y - E[y])^T] = E[(Ax - AM_x)(Ax - AM_x)^T]$

$= E[A(x - M_x)(x - M_x)^TA^T] = A\Sigma_xA^T$



$\begin{bmatrix} \lambda_1 & & \\ & \lambda_2 & \\ & & \ddots \lambda_n \end{bmatrix}$ eigen value. (Matrix).

$\Phi = \{\phi_1, \phi_2, \phi_3, \cdots\}$.

$\Phi\Phi^T = I$, $\Phi^T = \Phi^{-1}$

$Z = \Phi^Tx$. $M_z = \Phi^TM_x$, $\Sigma_z = \Phi^T\Sigma_x\Phi = \Lambda$.

$p(z) = \frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma_z|^{\frac{1}{2}}} e^{-\frac{1}{2}(z-M_z)^T\Sigma_z^{-1}(z-M_z)}$

$\|$

$p(x) = \frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma_x|^{\frac{1}{2}}} e^{-\frac{1}{2}(x-M_x)\Sigma_x^{-1}(x-M_x)}$

$\frac{d}{dB}tr(AB) = A^T$.

$\frac{d}{dA}\log|A| = A^{-T}$.

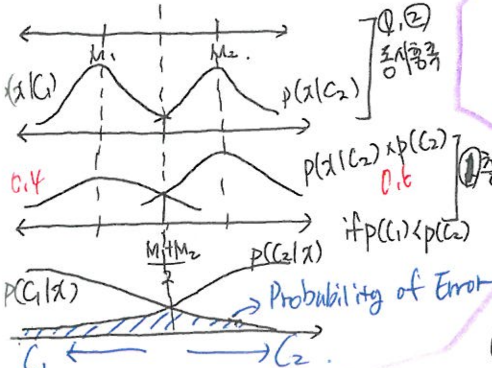
$*\ -\frac{1}{2}\log 2\pi - \frac{1}{2}\frac{(x-m_1)^2}{\sigma_1^2} + \log P(C_1) - \frac{1}{2}\log\sigma_1^2 \gtrless$

$-\frac{1}{2}\log 2\pi - \frac{1}{2}\frac{(x_2-m_2)^2}{\sigma_2^2} + \log P(C_2) - \frac{1}{2}\log\sigma_2^2$

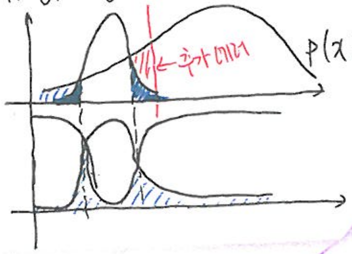
if $\sigma_1^2 = \sigma_2^2$ ①

$-\frac{1}{2}(x_1-m_1)^2 + \log P(C_1)\sigma^2 \gtrless -\frac{1}{2}(x_2-m_2)^2 + \log P(C_2)\sigma^2$.

if $P(C_1) = P(C_2)$ ② $(x_1-m_1)^2 \gtrless (x_2-m_2)^2$.



if $\sigma_1^2 \neq \sigma_2^2$.



Maximum a Posteriori $\{x_1 \cdots x_n\}$ Pos-

$\arg\max_\theta p(\theta|D) = \arg\max p(D|\theta) \cdot p(\theta)$

likelihood → random variable

$N(\theta; \theta_n, \sigma_n^2)$. $x \sim N(x; \theta, \sigma^2)$ known

$\theta \sim N(x; \theta_0, \sigma_0^2)$

$\arg\max_\theta \prod_t p(x^t|\theta) \cdot p(\theta)$

$\arg\max_\theta \alpha\, e^{-\frac{1}{2}\left[\sum_t \frac{(x^t-\theta)^2}{\sigma^2} + \frac{(\theta-\theta_0)^2}{\sigma_0^2}\right]}$

$\arg\max_\theta \sim e^{-\frac{1}{2}\frac{(\theta-\theta_n)^2}{\sigma_n^2}}$

$\begin{bmatrix} \frac{1}{\sigma_n^2} = \frac{N}{\sigma^2} + \frac{1}{\sigma_0^2} \\ \frac{\theta_n}{\sigma_n^2} = \frac{\sum x^t}{\sigma^2} + \frac{\theta_0}{\sigma_0^2} \end{bmatrix}$ $\theta_n = \frac{\sigma_0^2 \sum x^t}{N\sigma_0^2 + \sigma^2} + \frac{\sigma^2}{N\sigma_0^2 + \sigma^2}\theta_0$

$\sigma_n^2 = \frac{\sigma^2\sigma_0^2}{N\sigma_0^2 + \sigma^2}$.

θ_{ML}. $M_{ML} = \frac{\sum x^t}{N}$ 데이터 바탕

$\rightarrow \boxed{M_{MAP} = (1-\lambda)M_{ML} + \lambda M_0} \rightarrow$ 우선적

데이터 多 → 값 0.

σ_0가 크다는 것, 값을 특정하기 어려운 것.

vice versa.

$E[M_{MAP}] - M^*$

$= E[(1-\lambda)M_{ML} + \lambda M_0] - M^*$

$= \lambda(M_0 - M^*)$.

$p(x|C_k) \approx \prod p(x_i|C_k)$ 소간라. *독립 가정

$\begin{bmatrix} x_1 \\ x_n \end{bmatrix}$

$p(x) = p_1(x_1) \times p_2(x_2) = N(x_1; M_1, \sigma_1^2)$

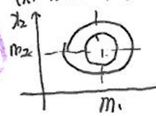
$N(x_2; M_2, \sigma_2^2)$

$= \frac{1}{(\sqrt{2\pi})^2\sigma_1\sigma_2} e^{-\frac{1}{2}\left[\frac{(x_1-m_1)^2}{\sigma_1} + \frac{(x_2-m_2)^2}{\sigma_2^2}\right]}$

$\frac{(x_1-m_1)^2}{\sigma_1^2} + \frac{(x_2-m_2)^2}{\sigma_2^2} = C$.

$*$ if $\sigma_1 = \sigma_2 = C$

$(x_1-m_1)^2 + (x_2-m_2)^2 = C\sigma^2$.



(m_1, m_2)로부터 멀어지면 확률 ↓.

$\begin{bmatrix} M_1 \\ \vdots \\ M_n \end{bmatrix}$

$*$ if $\sigma_1 \neq \sigma_2 \neq$

$\frac{(x_1-m_1)^2}{\sigma_1^2 C} + \frac{(x_2-m_2)^2}{\sigma_2^2 C} = 1$.

a^2 b^2

독립이 아닌경우.

learning.

for each class k.

$N_k \leftarrow \sum_t r_k^t$

$P(C_k) \leftarrow \frac{N_k}{N}$

for each dimension i

for each value index V_j

$p(x_i = V_j|C_k) \leftarrow \frac{\sum_t \mathbb{1}(x_i^t = V_j)r_k^t}{N_k}$

end

end output $P(C_k)$; $P(x_i = V_j|C_k)$ for each

$D = \{x^t\}_{t=1}^N$. $\begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$ $N(x; M, \Sigma)$.

$\arg\max_\theta \log P(D|\theta)$

$= \arg\max_\theta \prod_t p(x^t|\theta)$

$= \arg\max_{M, \Sigma} \sum_t \left[-\frac{n}{2}\log 2\pi + \frac{1}{2}\log|\Sigma|^{-1} - \frac{1}{2}\underbrace{(x-M)^T\Sigma^{-1}(x-M)}_{tr(x-M)(x-M)^T\Sigma^{-1}} \right]$

$\frac{dL}{dM} \Rightarrow \sum_t (\Sigma^{-1} + \Sigma^{-T})(x^t - M)\left(-\frac{1}{2}\right) = 0$ $\sum_t x^t = \sum_t M$ $\boxed{M} = \frac{\sum x^t}{N}$ ← vector

sample mean vector.

$\frac{dL}{d\Sigma^{-1}} \Rightarrow \sum_t \left[\left(\frac{1}{2}\Sigma^T - \frac{1}{2}(x^t-M)^T(x^t-M)\right]\right] \stackrel{=0}{\Rightarrow} \frac{\sum(x^t-M)(x^t-M)^T}{N} = \boxed{\Sigma}$

sample covariance

$\arg\max_k P(C_k|x) = \arg\max_k P(x|C_k)P(C_k)$ (vector)

$= \arg\max_k \underbrace{\log P(x|C_k)P(C_k)}_{g(k)}$

discriminant function. $\leftarrow g(k) \rightarrow$

$g_k(x) = -\frac{n}{2}\log 2\pi - \frac{1}{2}\log|\Sigma_k| - \frac{1}{2}(x-M_k)^T\Sigma_k^{-1}(x-M_k) + \log P(C_k)$

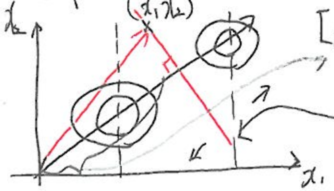
$\rightarrow -\frac{1}{2}x^T\Sigma_k^{-1}x + M_k^T\Sigma_k^{-1}x - \frac{1}{2}M_k^T\Sigma_k^{-1}M_k - \frac{1}{2}\log|\Sigma_k| + \log P(C_k)$

$\rightarrow x^T\alpha_k x + B_k^T x + \gamma_k$.

*Binary class.

$g_1(x) \gtrless g_2(x) \rightarrow x^T\alpha x + B^T x + \gamma \gtrless 0$.

Decision Boundary.



$[M_1 - M_2]\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \sigma^2\gamma \gtrless 0$.

결과

*regression.

$D = \{x^t, r^t\}_{t=g}^{N}$ $\quad r^t = h(x^t) + \epsilon$ $\quad \epsilon \sim N(m, \sigma^2)$

약간의 오차를 허용

Numeric value.

$E[h|D] = \sum_t (r^t - h(x^t))^2$. 얼마나 차이가 나는가?

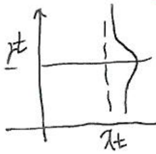
emperical error. palametric method.

$\arg\max_\theta P(D|\theta) = \arg\max_\theta \prod_t p(x^t, r^t|\theta)$

$= \arg\max_h \sum_t \log \cancel{p(x^t|h)} + \log P(r^t|x^t, h)$

x^t는 h의 값에 따라 불변, 상수.

$N(r^t; h(x^t), \sigma^2)$



$\arg\max_h \sum_t \left(-\frac{1}{2}\frac{(r^t - h(x^t))^2}{\sigma^2}\right)$

$\rightarrow \arg\min_h \sum_t (r^t - h(x^t))^2$.

$* \ h(x) = w_1 x + w_0$.

$\arg\min_{w_1 w_0} \underbrace{\sum_t (r^t - w_1 x^t - w_0)^2}_{L}$.

$\frac{L}{dw_1} = 0 \rightarrow \sum_t (r^t - w_1 x^t - w_0)x^t = 0$

$\frac{L}{dw_0} = 0 \rightarrow \sum_t (r^t - w_1 x^t - w_0) = 0$

식 2개, 문자 2개.

*polynomial.

$h(x) = w_n x^{tn} + w_{n-1}x^{t(n-1)} + \ldots$

$\arg\min_{w_n \ldots w_0} \sum_t (r^t - h(x^t))^2$.

$\left\{ \frac{dL}{dw_0} = 0, \ \vdots, \ \frac{dL}{dw_n} = 0 \right\}$ 식 $n+1$개, 미지수 $n+1$개

$A \begin{bmatrix} & \end{bmatrix}\begin{bmatrix} w_0 \\ \vdots \\ w_n \end{bmatrix} = \begin{bmatrix} b_0 \\ \vdots \\ b_n \end{bmatrix}$

$\begin{bmatrix} 1 & x_1^1 & x_2^1 & x_3^1 & \cdots & x_d^1 \\ 1 & & & & & \vdots \\ 1 & & & & & \\ 1 & \cdots & & & \cdots & x_d^N \end{bmatrix}\begin{bmatrix} w_0 \\ \vdots \\ w_n \end{bmatrix} = \begin{bmatrix} r_0 \\ \vdots \\ r_n \end{bmatrix}$

$X^T X w = X^T r$.

$w = (X^T X)^{-1} X^T r$.

$\arg\max_k P(C_k|x)$ $\quad x = (x_1 \cdots x_n)$

multiple dimension

$= \arg\max_k P(x_1 \cdots x_n|C_k) \cdot P(C_k)$

$\approx \arg\max_k \prod_i (x_i|C_k) \cdot P(C_k)$ if independent.

학습데이터.

x_1	x_2	C_1	C_2
150	40	1	0
160	50	1	1
160	60	1	1
170	60	1	3
180	70	0	1

X_1	$P(X_1 = v \mid C_1)$
150	1/4
160	1/4
170	1/4
180	1/4

x_2	$P(X_2 = v \mid C_1)$
40	1/4
50	1/4
60	3/4
70	0

X_1	$P(X_1 = v \mid C_2)$
150	0
160	2/6
170	3/6
180	1/6

X_2	$P(X_2 = v \mid C_2)$
40	0
50	1/6
60	4/6
70	1/6

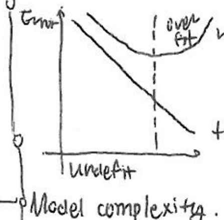
class validation

	$C(x)$ 1	0
$h(x)$ 1	TP	FP
0	FN	TN

- Error Rate
  $= (FP + FN) / (FP + FN + TP + TN) \times 2$.
- False Alarm Rate (type 1 error)
  $= FP / FP + FN$.
- Miss rate
  $= FN / TP + FN$.
- Hit Rate
  $= 1 -$ miss rate
- Precision
  $= TP / (TP + FP)$

Datatest ~ N points. (labled in 2^N ways) $\oplus, \ominus$

If for any of these problems, hypothesis $h \in H$ that can separates $\oplus, \ominus$. H shatters N points. Any learning Problem definable by N examples can be learned with no error by a hypothesis drawn from H. VC(H), measure the capacity of H. $VC(H) = n$. $\{$ i) $VC(H) \geq n$ ii) $VC(H) < n+1$.

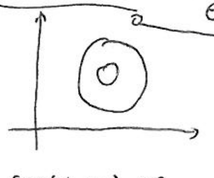


(model 복잡도 ↑ error ↑.

(실제 학습데이터 ↑) Error ↓

$P(C|x) = \frac{P(C) \cdot P(x|C)}{P(x)}$

posterior $= \frac{\text{prior} \times \text{likelihood}}{\text{evidence}}$



$Cov(x_1, x_2) = 0$
$Var(x_1) = Var(x_2)$

$Cov(x_1, x_2) = 0$
$Var(x_1) > Var(x_2)$

$Cov(x_1, x_2) > 0$

$Cov(x_1, x_2) < 0$.

Expected risk $\quad \lambda_{ik} =$ risk of α_i if $x \in C_k$.

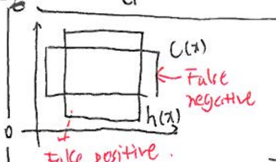
$R(\alpha_i|x) = \sum_k \lambda_{ik} P(C_k|x)$.

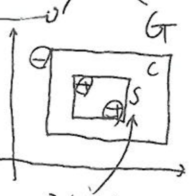
$x \in C_k$ if $k = \arg\min_i R(\alpha_i|x)$.

$\lambda_{11} P(C_1|x) + \lambda_{12} P(C_2|x) \gtrless_{C_1}^{C_2} \lambda_{21} P(C_1|x) + \lambda_{22} P(C_2|x)$.

$P_2(\alpha_1|x)$.

$P(C_1|x)\{\lambda_{11} - \lambda_{21}\} \gtrless_{C_1}^{C_2} P(C_2|x)(\lambda_{22} - \lambda_{12})$





$P(X_2|C_1)$

$P(C_1|X_1)$.

$= \frac{P(\theta|D) \cdot P(D)}{P(\theta)}$

$P(D|\theta) P(\theta)$

$P(\theta)$

$\downarrow$

$p(x_1|\theta) \times p(x_2|\theta)$

$P(C_1|X)$.

$= P(X|C_1) \times p(C_1$