L08-1 : Central limit Th

前言：Gaussin 的优点

① symmetric

② unimodal ： singal peak , mean = mode

③ localization ：有極值 at μ , Distance (x_i, μ) ↑ $P(x_i)$ ↓

Central limit theorem 中央極限定理

- $X \sim D(\mu, \sigma^2)$ 取 n 个 的 sample mean $\bar{X}$

  ⇒ $Y \sim N(\mu_Y, \sigma_Y)$? $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$ ⇒  if $n \to \infty$

  (D 為任一分佈 such as  ,  

  var

- that is ： sample 1 的 mean = $\bar{x}_1$ sample 1 = $\{x_{11} \dots x_{1n}\}$

  sample 2 的 mean = $\bar{x}_2$ sample 2 = $\{x_{21} \dots x_{2n}\}$

  ...

  $\bar{x}_1, \bar{x}_2, \dots \bar{x}_k \sim$ Normal Distribution

- 補充 MGF (moment generating function)

  把机率描述成 1st moment + 2nd moment + ...

  所以(likelihood, prior) 可寫成 1 function ⇒ 作 conjugate

Students T distribution

- 

  一個 Var 更大(more tolerant)的 distribution

- 因 Z-test 假設 $N \sim (\mu, \sigma^2)$ 需要知道 μ, σ^2

  但取小樣本時，只用 ex 10 人建分佈很不準

  所以使用 T = $\int N(x|\mu, \sigma^2) \cdot T(\sigma^2|a,b)\, d\,ab$

- 應用：t-SNE ? $\int N(x|\mu, \gamma^{-1}) \underbrace{\Gamma}_{\text{Gamma}} (\gamma|a,b)\, d\gamma$

multivariate

- univariate : $N=\frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-1}{2\sigma^2}(x-\mu)^2}$
- multivariate : $P(\mu,\overset{\text{半正定}}{\Sigma})=\frac{1}{\sqrt{2\pi}^k|\Sigma|^{\frac{1}{2}}} e^{\frac{-1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)}$ Λ precision matrix

$$\Sigma = \text{covariance matrix} = \begin{array}{c|cccc} & x_1 & x_2 & \cdots & x_k \\ \hline x_1 & \frac{(x_1-\mu)^2}{n} & & & \\ x_2 & & \frac{(x_2-\mu)^2}{n} & & \bigcirc \\ \vdots & & & \ddots & \\ x_k & & & & \frac{(x_k-\mu)^2}{n} \end{array}$$

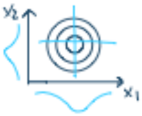
$\bigcirc$ otherwise $\frac{(x_m-\mu_m)(x_n-\mu_n)}{N}$

對角線 variance

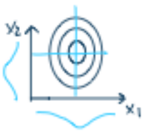
$\Rightarrow$ If Σ = diagonal matrix = $\begin{bmatrix} \diagdown & 0 \\ 0 & \diagdown \end{bmatrix}$, it means anti-corelated

shape of data

- Isotropic
  $\Sigma = I$



- Orthogonal
  Σ = Diagonal



- $P(\mu,\Sigma)=\frac{1}{\sqrt{2\pi}^k|\Sigma|^{\frac{1}{2}}} e^{\frac{-1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)}$ Euclidean distance 歐式距離
- 2 變數獨立

- General
  $\Sigma = \begin{bmatrix} \diagdown & x \\ x & \diagdown \end{bmatrix}$ non-zero

- $P(\mu,\Sigma)=\frac{1}{\sqrt{2\pi}^k|\Sigma|^{\frac{1}{2}}} e^{\frac{-1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)}$ Mahalanobis distace 馬式距離
- 2 變數不獨立

Affine property : linear trans (放大小/rotate) + 平移 = Affine trans

- $X \sim N(\mu, \overset{C}{\sigma^2}) \xrightarrow[\text{affine transformation}]{f(x)} Ax+b \sim N(A\mu+b, ACA^T)$

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

$\Rightarrow E(Ax+b) = \int (Ax+b)\,p(x)\,dx$
$= A\int x p(x)\,dx + b\int p(x)\,dx$
$= AE(x) + b\cdot 1$
$= A\mu + b$

- Ex :

$\longrightarrow$

Standard Normal
$X \sim N(0, I) \longrightarrow Ax+\mu \sim N(\mu, C)$

$Cov(x) = \Sigma = E\{(x-\mu)(x-\mu)^T\}$
$Cov(Ax+b) = E\{[(Ax+b)-(A\mu+b)][(Ax+b)-(A\mu+b)]^T\}$
$= E\{[A(x-\mu)][A(x-\mu)]^T\}$
$= E\{A(x-\mu)(x-\mu)^T A^T\}$
$= A\,E\{(x-\mu)(x-\mu)^T\}A^T$
$= A\Sigma A^T$

- Any Gaussin can be derived from Isotropic by

$\Rightarrow$

scale

rotate

relocate

L08-3 : Univariate ⇔ Multivariate Gaussian

Univariate to multivariate Gaussion

可以線性變換要滿足 2 條件

$T(X+Y) = T(X) + T(Y)$ 、 $T(aX) = aT(X)$

● ① $N(X+Y) = N(X) + N(Y)$

$X_1 \sim N(\mu_1, \sigma_1^2)$ $X_2 \sim N(\mu_2, \sigma_2^2)$ $Y = X_1 + X_2$

令 $Y = X_1 + X_2 \Rightarrow X = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$, $E(X) = \mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}$ $\Sigma = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}$

取 $A = [1 \ 1]$, $b = 0$

則 by Affine : $AX + b = [1 \ 1]\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + 0 = x_1 + x_2 \sim N(A\mu + b, A\Sigma A^T)$

$\because A\mu + b = [1 \ 1]\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} + 0 = \mu_1 + \mu_2$

$A\Sigma A^T = [1 \ 1]\begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}\begin{bmatrix} 1 \\ 1 \end{bmatrix} = \sigma_1^2 + \sigma_2^2$

$\therefore Y = X_1 + X_2 \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$

$\Rightarrow Y = \Sigma X_i \sim N(\Sigma \mu_i, \Sigma \sigma_i^2)$

● ② $N(Ax) = AN(x)$

令 $Y = b_1 X_1 + b_2 X_2 + \dots$

取 $A = [b_1 \ b_2 \ \dots]$

則 $AX + b = Y \sim N(A\mu + b, A\Sigma A^T)$

$\therefore Y \sim N(b_1\mu_1 + b_2\mu_2 \dots, b_1\sigma_1^2 + b_2\sigma_2^2 + \dots)$

$\Rightarrow Y \sim N(B\Sigma\mu_i, B\Sigma\sigma_i^2)$

Marginal Gaussin (muti to univariate)

令 $X = \begin{bmatrix} X_a \\ X_b \end{bmatrix} \begin{matrix} \rightarrow X_a = \begin{bmatrix} x_1 \\ \vdots \\ x_k \end{bmatrix} \\ \rightarrow X_b = \begin{bmatrix} x_{k+1} \\ \vdots \\ x_n \end{bmatrix} \end{matrix}$ $\mu = \begin{bmatrix} \mu_a \\ \mu_b \end{bmatrix}$ $\Sigma = \begin{bmatrix} \sigma_a^2 & \sigma_{ab}^2 \\ \sigma_{ba}^2 & \sigma_b^2 \end{bmatrix}$

取 $A = [I \ 0]$, $b = 0$

則 $AX + b = [I \ 0]\begin{bmatrix} X_a \\ X_b \end{bmatrix} = X_a \sim N(A\mu + b, A\Sigma A^T)$

$A\mu + b = \mu_a$

$A\Sigma A^T = [I \ 0]\begin{bmatrix} \sigma_a^2 & x \\ x & x \end{bmatrix}\begin{bmatrix} I \\ 0 \end{bmatrix} = \sigma_a^2$

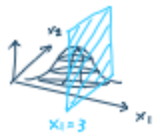
$\therefore X_a \sim N(\mu_a, \sigma_a^2)$

$X_b \sim N(\mu_b, \sigma_b^2)$

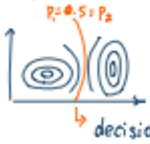
§ Conditional Gaussin

multivariate gaussin 中，一部份變數為定值

其餘變數形成之分佈仍為 Gaussin



註：If conditional independence



↳ decision boundary
(a quadratic line)

令 $X=\begin{bmatrix}x_a\\x_b\end{bmatrix}$，$P(\mu,\Sigma)=\frac{1}{\sqrt{2\pi}^k|\Sigma|^{0.5}}e^{\frac{-1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)}$

1。 $\frac{-1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)=\frac{-1}{2}\begin{bmatrix}x_a-\mu_a & x_b-\mu_b\end{bmatrix}\begin{bmatrix}\Lambda_{aa} & \Lambda_{ab}\\ \Lambda_{ba} & \Lambda_{bb}\end{bmatrix}\begin{bmatrix}x_a-\mu_a\\x_b-\mu_b\end{bmatrix}$

$$=\frac{-1}{2}\left[(x_a-\mu_a)\Lambda_{aa}+(x_b-\mu_b)\Lambda_{ba},\ (x_a-\mu_a)\Lambda_{ab}+(x_b-\mu_b)\Lambda_{bb}\right]\begin{bmatrix}x_a-\mu_a\\x_b-\mu_b\end{bmatrix}$$

$$=\frac{-1}{2}(x_a-\mu_a)\Lambda_{aa}(x_a-\mu_a)+\underline{\frac{-1}{2}(x_b-\mu_b)^T\Lambda_{ba}(x_a-\mu_a)+\frac{-1}{2}(x_a-\mu_a)^T\Lambda_{ab}(x_b-\mu_b)}+\text{const}$$

$$=\frac{-1}{2}x_a^T\Lambda_{aa}x_a+x_a^T\Lambda_{aa}\mu_a-x_a^T\Lambda_{ba}x_b+x_a^T\Lambda_{ba}\mu_b+\text{const}$$

2。 $\frac{-1}{2}(x-\mu)^T\Sigma_{x_a|x_b}^{-1}(x-\mu)$

$$=\frac{-1}{2}x^T\Sigma_{x_a|x_b}^{-1}x+x^T\Sigma_{x_a|x_b}^{-1}\mu+\text{const}$$

3。 比較 1。、2。得 $\begin{cases}\Sigma_{x_a|x_b}^{-1}=\Lambda_{aa}\Rightarrow\Sigma_{x_a|x_b}=\Lambda_{aa}^{-1}\\ x^T\Sigma_{x_a|x_b}^{-1}\mu=x_a^T\Lambda_{aa}\mu_a-x_a^T\Lambda_{ab}(x_b-\mu_b)\end{cases}$

$\therefore\ \Sigma_{x_a|x_b}^{-1}\mu=\Lambda_{aa}\mu_a-\Lambda_{ab}(x_b-\mu_b)$

$\Rightarrow\mu=\underbrace{\Sigma_{x_a|x_b}}_{}^{\Lambda_{aa}^{-1}}(\Lambda_{aa}\mu_a-\Lambda_{ab}(x_b-\mu_b))$

$=\mu_a-\Lambda_{aa}^{-1}\Lambda_{ab}(x_b-\mu_b)$

4。 $\begin{bmatrix}A & B\\C & D\end{bmatrix}^{-1}=\begin{bmatrix}M & -MBD^{-1}\\ \times & \times\end{bmatrix}$ $M=(A-BD^{-1}C)^{-1}$

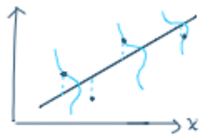
$\begin{bmatrix}\Sigma_{aa} & \Sigma_{ab}\\ \Sigma_{ba} & \Sigma_{bb}\end{bmatrix}^{-1}=\begin{bmatrix}\Lambda_{aa} & \Lambda_{ab}\\ \Lambda_{ba} & \Lambda_{bb}\end{bmatrix}\Rightarrow\begin{cases}\Lambda_{aa}=M\Rightarrow(\Lambda_{aa})^{-1}=M^{-1}\\ \Lambda_{ab}=-MBD^{-1}\end{cases}$

5。 結合 3。、4。得 $\mu=\mu_a-(-BD^{-1})(x_b-\mu_b)=\mu_a+\Sigma_{ab}\Sigma_{bb}(x_b-\mu_b)$

L09-1 Bayesian Linear Regression

§ Probability view of linear regression

- LSE 前情提要



Design matrix (feature space)

$\Phi = [\phi_1 \ \phi_2 \cdots \phi_k]$ → x_1, x_2

Ex: $\phi_1 = \sin(x)$, $\phi_2 = x^3 \ldots$

Line : $\omega^T \Phi(x)$, $D = \{(x_1, y_1) \ldots (x_d, y_d)\}$

$$y = \omega^T \Phi(x) + \varepsilon \sim N \sim (y \mid \underbrace{w^T \Phi(x)}_{0 \to w^T \Phi(x) = \text{line}}, \sigma^2)$$

or

$$y = Xw = \begin{bmatrix} 1 & x_1 & \cdots & x_1^n \\ & & \vdots & \\ 1 & x_d & \cdots & x_d^n \end{bmatrix} \begin{bmatrix} w_0 \\ \vdots \\ w_n \end{bmatrix}$$

- 重點提要

frequentist	bayesian
LSE ⟺	MLE
rLSE ⟺	MAP
	fully - gaussian ⇓ Predictive distribution

§ LSE ⟺ MLE

○ 目標 ：找出最大可能 fit 目前 data 的 ω

上課版: Likelihood $= P(D \mid w)$

$$= \prod_{\forall x} N(w^T \Phi(x_d), \sigma^2)$$

$$= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \prod_D e^{\frac{-1}{2\sigma^2}[y_d - w^T \Phi(x_d)]^2} \quad \propto \prod_D e^{\frac{-1}{2\sigma^2}[y_d - w^T \Phi(x_d)]^2}$$

$$\xrightarrow{\log} \propto \sum_D [\underbrace{y_k}_{b} - \underbrace{w^T \Phi(x_d)}_{Ax}]^2$$

講義版 : Likelihood $= P(D_y \mid D_x, W)$

$$= \prod_i \frac{1}{\sqrt{2\pi}\sigma_i} e^{\frac{-1}{2\sigma_i^2}(y_i - X_i w)^2} \propto \prod_i e^{\frac{-1}{2\sigma_i^2}(y_i - X_i w)^2}$$

$$\xrightarrow{\log} \sum_i \lg \frac{1}{\sqrt{2\pi\sigma_i^2}} + \sum_i \frac{-1}{2\sigma_i^2}(y_i - X_i w)^2$$ 对 ω 微分後剩此項

求 MLE ⟺ 求 $\max\left(\sum_i (X_i w - y_i)^2\right)$ ⟺ 求 $\max(\|A\vec{x} - \vec{b}\|^2)$

§ rLSE ⟺ MAP

但比 rLSE 更好
∵ 利用 conjugate gaussian 性質
此法可持續更新 ω
直至收斂

- 目標 : $\min E(\omega) = \sum_{n=1}^{N} [y(x_n, \omega) - t_n]^2 + \lambda \|\omega\|^2$

  找出 ω 使 posterior 最大化， $\lambda\|\omega\|^2$ 相當 prior

- 前言

posterior $P(\omega|D) = \dfrac{P(D|\omega)P(\omega)}{P(D)}$

prior $P(W) \sim N(0, b^{-1}I)$，b^{-1} 為 precision matrix

b其實是 prior 之 cov-matrix 的倒數，故

$$b^{-1} = \begin{bmatrix} \sigma^2_{w_0} & 0 & \cdots & 0 \\ 0 & \sigma^2_{w_1} & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2_{w_n} \end{bmatrix}, \quad X = [1, x \ldots x^n], \quad y = \sum_{i=0}^{n} w_i x^i = X\omega$$

- 求 max(prior · likelihood)，注意 prior 為 multivar，likelihood 為 univar

已知 $\overset{\text{likelihood}}{P(D|W)} \propto e^{\sum_i \frac{-1}{2\sigma_i^2}(y_i - x_i\omega)^2}$ 注意 posterior 為 multivar gaussian 分佈

故 $P(\omega|D) \propto \overset{\text{likelihood}}{P(D|\omega)}\,\overset{\text{prior}}{P(\omega)}$ prior：MAP 和 MLE 即差在有 prior

cov-matrix 倒數

$$\propto e^{\sum_i \frac{-a}{2}(y_i - x_i\omega)^2} \cdot e^{\frac{-b}{2}(\omega-\vec{0})^T I (\omega - \vec{0})}$$

$$= e^{\sum_i \frac{-a}{2}(y_i - x_i\omega)^2 + \frac{-1}{2}\omega^T bI\omega}$$

取 ln 得 $P(\omega|D) \propto \sum_i \frac{-a}{2}(y_i - x_i\omega)^2 + \frac{-1}{2}\omega^T bI\omega$

matrix form 為 $\frac{-a}{2}\left(\|X\omega - y\|^2 + \frac{b}{a}\omega^T\omega\right)$

故可發現 $\frac{b}{a}$ 相當 rLSE 之 λ

remind that x 解 $= (A^TA + \lambda I)^{-1}A^Tb$

λ大 ⇒ b大 (var小)　λ小 ⇒ b小 (var大) ∵ b^{-1} = var

- 求 $a\|X\omega - y\|^2 + b\omega^T\omega$ quadratic form

是為了証明 posterior 為 multivar gaussian distribution

$a\|X\omega - y\|^2 + b\omega^T\omega$

$= a(X\omega - y)^T(X\omega - y) + b\omega^T\omega$

$= a(\omega^TX^TX\omega - 2\omega^TX^Ty + y^Ty) + b\omega^T\omega$

$= \omega^T(aX^TX + bI)\omega - 2a\omega^TX^Ty + ay^Ty$

對照 quadratic form $(x-\mu)^T\Lambda(x-\mu)$

$= x^T\Lambda x - 2x^T\Lambda\mu + \mu^T\mu$

得出 $\Lambda = aX^TX + bI$

$\mu = a\Lambda^{-1}X^Ty$ $\because a\omega^TX^Ty = \omega^T\Lambda\mu$

故 poterior $\sim N(\mu, \Lambda^{-1})$

註① $e^{\omega^T(aX^TX + bI)\omega - 2a\omega^TX^Ty + ay^Ty}$

$= e^{(\omega^T\Lambda\omega - 2\omega^T\Lambda\mu + \mu^T\mu)\underbrace{-\mu^T\mu + ay^Ty}_{\text{当係數}}}$

$= A\, e^{(\omega-\mu)^T\Lambda(\omega-\mu)}$

註② $\mu = a\Lambda^{-1}X^Ty$

$= a(aX^TX - bI)^{-1}X^Ty$

$= (\frac{a}{a}X^TX - \frac{a}{b}I)X^Ty$

$= (X^TX + \lambda I)X^Ty$

故 rLSE 即 frequentist 版的 MAP

L09-2 : online learning

Iter 1： 令 prior $\sim N(0, b^{-1}I)$，b^{-1}可隨意假設

則 posterior $\sim N(\mu, \Lambda^{-1})$

其中 $\begin{cases} \Lambda = aX^TX + bI \\ \mu = a\Lambda^{-1}X^Ty \end{cases}$

Iter 2： Prior 更新為 (m, S^{-1})

則 posterior $\sim N(\mu', \Lambda'^{-1})$

其中 $\begin{cases} \Lambda' = aX^TX + bI \\ \mu' = \Lambda'^{-1}(aX^Ty + S\cdot m) \end{cases}$

推導过程：

$P(w|D) \propto P(D|w)\, P(w)$

$$\propto e^{\frac{-a}{2}\sum_i (x_i w - y_i) + \frac{-1}{2}(w-m)^T S (w-m)}$$

取 ln 得：$\frac{-a}{2}\sum_i (x_i w - y_i) + \frac{-1}{2}(w-m)^T S (w-m)$

矩陣 form：$\frac{-a}{2}\left[\|Xw - y\|^2 + \frac{1}{a}(w-m)^T S (w-m) \right]$

忽略係數：$a(w^TX^TXw - 2w^TX^Ty + y^Ty) + (w^TSw - 2w^TSm + m^TSm)$

$$= w^T(aX^TX + S)w - 2w^T(aX^Ty + Sm) + ay^Ty + m^TSm$$

对照 quadratic form of $(w-\mu')^T\Lambda'(w-\mu')$

$$= w^T\Lambda' w - 2w^T\Lambda'\mu' + \mu'^T\Lambda'\mu'$$

故得 $\begin{cases} \Lambda' = aX^TX + S \\ \mu' = \Lambda'^{-1}(aX^Ty + Sm) \end{cases}$

註：$(aX^Ty - Sm) = \Lambda'\mu'$

L10 : Fully Bayesian

Fully Baysian (Prediction distribution)

- review $\begin{cases} \text{MLE} \\ \text{MAP} \end{cases}$

- 

MLE: point estimation: $x \to y$

MAP: $x \to$ select $p(x)$ 最大者的 y

$$P(y|D) = \int \underbrace{P(y|w,P)}_{\text{likelihood}}\ \underbrace{P(w|P_x)}_{\text{prior}}\, dw$$

① By Affine, $P(y|D)$ is gaussin $\Leftarrow$ $P(y|w,P)$ is gaussin

② w 變數代表不同 possible line

③ marginalize w 得 $P(y|D)$

$$= \int N(y|Xw, \cancel{\sigma^2}\ a^{-1})\, N(w|\mu, \Lambda^{-1})\, dw \quad (\text{MAP})$$

$$\propto \int e^{\frac{-a}{2}(Xw-y)^2}\, e^{\frac{-1}{2}(w-\mu)^T\Lambda(w-\mu)}\, dw$$

$$= \int e^{\frac{-a}{2}(w^TX^TXw - 2w^TX^Ty + y^Ty) + \frac{-1}{2}(w^T\Lambda w - 2w^T\Lambda\mu + \mu^T\Lambda\mu)}\, dw$$

$$= \int e^{\frac{-1}{2}[w^T(aX^TX+\Lambda)w - 2w^T(aX^Ty+\Lambda\mu) + ay^Ty + \mu^T\Lambda\mu]}\, dw$$

對照 quadratic form of $(x-\mu)^T\Lambda(x-\mu)$

$= x^T\Lambda x - 2x^T\Lambda\mu + \mu^T\Lambda\mu$

可令 $\begin{cases} C = aX^TX + \Lambda \\ \mu' = C^{-1}(aX^Ty + \Lambda\mu) \end{cases}$

$$= \int e^{\frac{-1}{2}[(w-\mu')^TC(w-\mu') - \mu'^TC\mu' + ay^Ty + \mu^T\Lambda\mu]}\, dw$$

$$= \underbrace{\int e^{\frac{-1}{2}[(w-\mu')^TC(w-\mu')]}}_{\text{multivar gaussian 積分後}=1} \cdot \underbrace{e^{\frac{1}{2}(\mu'^TC\mu' - ay^Ty - \mu^T\Lambda\mu)}}_{\text{當常數項提出}}\, dw$$

$$= e^{\frac{-1}{2}(-\mu'^TC\mu' + ay^Ty + \mu^T\Lambda\mu)}$$

$$= e^{\frac{-1}{2}(y-\mu'')^TC'(y-\mu'')}$$

其中 $\mu'' = X\mu$, $(C')^{-1} = \frac{1}{a} + X\Lambda^{-1}X^T$

故分佈為 $N \sim (X\mu, \frac{1}{a} + X\Lambda^{-1}X^T)$

指數項展開：

$$-[C^{-1}(ax^Ty+\Lambda\mu)]^T \cancel{C}[\cancel{C}^{-1}(ax^Ty+\Lambda\mu)] + ay^Ty + \mu^T\Lambda\mu$$

$$= -(ax^Ty+\Lambda\mu)^T \underbrace{(C^{-1})^T}_{=C^{-1}} (ax^Ty+\Lambda\mu) + ay^Ty + \mu^T\Lambda\mu$$

$$= -a^2y^Tx C^{-1}x^Ty - 2ay^Tx C^{-1}\Lambda\mu - \mu^T\Lambda^T C^{-1}\Lambda\mu + ay^Ty + \mu^T\Lambda\mu$$

$$= y^T(a-a^2xC^{-1}x^T)y - 2ay^TxC^{-1}\Lambda\mu - \mu^T\Lambda^TC^{-1}\Lambda\mu + \mu^T\Lambda\mu$$

對照 quadratic form of $(y-\mu'')C'(y-\mu'')$

可得
$$\begin{cases} C' = a - a^2xC^{-1}x^T \\ \mu'' = C'^{-1}(axC^{-1}\Lambda\mu) = aC'^{-1}x(ax^Tx+\Lambda)^{-1}\Lambda\mu = X\mu \end{cases}$$

$\because C'\mu'' = axC^{-1}\Lambda\mu$

利用 Sherman-Morrison formula

若 $C = ax^Tx + \Lambda$

則 $C^{-1} = \Lambda^{-1} - \dfrac{\Lambda^{-1}ax^Tx\Lambda^{-1}}{1+ax\Lambda^{-1}x^T}$　check：$CC^{-1}=1$

故 $C' = a - a^2xC^{-1}x^T$

$$= a - a^2x\left(\Lambda^{-1} - \frac{\Lambda^{-1}ax^Tx\Lambda^{-1}}{1+a\underbrace{x\Lambda^{-1}x^T}_{令為\alpha}}\right)x^T$$

$$= a - a^2x\left(\Lambda^{-1} - \frac{\Lambda^{-1}ax^T\alpha(x^T)^{-1}}{1+a\alpha}\right)x^T$$

$$= a - a^2\left(\overset{\alpha}{\cancel{x\Lambda^{-1}x^T}} - \frac{x\Lambda^{-1}ax^T\alpha}{1+a\alpha}\right)$$

$$= a - a^2\left(\alpha - \frac{a\alpha^2}{1+a\alpha}\right)$$

$$= a - a^2\left(\frac{\alpha}{1+a\alpha}\right)$$

$$= \frac{a}{1+a\alpha}$$

故 $(C')^{-1} = \dfrac{1+a\alpha}{a}$ $= \dfrac{1}{a} + x\Lambda^{-1}x^T$

故 $\mu'' = a(C')^{-1}xC^{-1}\Lambda\mu$

$$= a[(C')^{-1}xC^{-1}\Lambda\mu]^T$$ 取T

$$= a[\mu^T\Lambda C^{-1}x^T(C')^{-1}]$$

$$= a\mu^T\Lambda C^{-1}x^T(C')^{-1}$$

$$= a\mu^T\Lambda\left(\Lambda^{-1} - \frac{\Lambda^{-1}ax^Tx\Lambda^{-1}}{1+a\underbrace{x\Lambda^{-1}x^T}_{\alpha}}\right)x^T(C')^{-1}$$

$$= a\mu^T\left(x^T - \frac{ax^T\alpha}{1+a\alpha}\right)\left(\frac{1+a\alpha}{a}\right)$$

$$= \cancel{a}\mu^T\left(\frac{x^T}{\cancel{1+a\alpha}}\right)\left(\frac{\cancel{1+a\alpha}}{\cancel{a}}\right) = \mu^Tx^T$$

取T

$$= x\mu$$

L11-1 : Decision Theory

Decision Theory (classification)

- review : In either MLE or MAP,
  we estimate the params of $N(\mu, \sigma^{-1})$ according to observation

Estimator

- This is a statistic (統計量) to approximate the property of a distribution
  也就是從抽樣樣本近似母体樣本

1. 抽樣樣本 random var

   ① $\mu_{MLE} = \hat{\mu} = \frac{1}{n}\sum_i x_i$ if $D = x_1, \dots x_n$ is i.i.d.

   ② $\sigma_{MLE} = \hat{\sigma}$

   $$\begin{cases} \sigma_{MLE}^2 = \hat{\sigma}^2 = \frac{1}{n}\sum(x_i - \mu)^2 \\ \sigma_{unbiased}^2 = \underbrace{\frac{1}{n-1}}_{\text{自由度}}\sum(x_i - \mu)^2 \end{cases}$$, μ 為母体平均

   Bias 來自 $E(\hat{\theta}) - \theta \neq 0$，用 $\sigma_{unbiased}$，$E(\hat{\theta}) - \theta$ 才 $= 0$

2. 近似母体

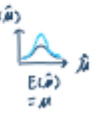
   ① $E(\hat{\mu})$ = 樣本 mean 再取 mean ⇒ $\hat{\mu}$ 呈 gaussin ∵ 中央極限 th

   $= E(\frac{1}{n}\sum X_i) = \frac{1}{n}E(\sum X_i) = \frac{1}{n} n\mu = \mu$

   ② $Var(\hat{\mu})$ = 樣本 mean 再取 var

   $= Var(\frac{1}{n}\sum X_i) = Var(\frac{X_1}{n} + \dots \frac{X_n}{n}) = \frac{1}{n^2} Var(X_1 + \dots X_n) = \frac{1}{n^2} n Var(X)$

   $= \frac{Var(X)}{n} = \frac{\sigma^2}{n}$

   ∴ 取樣次数越多，$\hat{\mu}$ 分佈越陡 

   越能精估母体 μ

L11-2 : Bias

Bias

- Estimator 和母体差 $E(\hat{\theta}) - \theta$

  $E(\hat{\mu})$ 前面推过, $= E(\frac{1}{n}\sum X_i) = \frac{1}{n}\sum E(X_i) = \frac{1}{n} n\mu = \mu$

  $\therefore$ bias $= 0$

  $Var(\hat{\mu})$ 前面推过 $= Var(\frac{X_1 + \dots X_n}{n}) = \frac{1}{n^2} Var(X_1 + \dots X_n)$

  $= \frac{1}{n^2}(n\sigma^2) = \frac{\sigma^2}{n}$

  $E(\sigma^2_{MLE}) = E[\frac{1}{n}\sum(x_i - \hat{\mu})^2] = \frac{1}{n}E[\sum x_i^2 - 2\hat{\mu}\underbrace{\sum x_i}_{n\hat{\mu}} + \hat{\mu}^2]$

  $= \frac{1}{n}E[\sum x_i^2 - n\hat{\mu}^2]$

  $= \frac{1}{n}E(\sum x_i^2) - E(\hat{\mu}^2)$

  $= \underbrace{E(x_i^2)}_{Var(x_i) + E^2(x_i)} - \underbrace{E(\hat{\mu}^2)}_{Var(\hat{\mu}) + E^2(\hat{\mu})}$

  $= (\sigma^2 + \mu^2) - (\frac{\sigma^2}{n} + \mu^2)$

  $= \sigma^2 - \frac{\sigma^2}{n}$ ⇒ 總之 $E(\sigma^2_{MLE}) = \sigma^2 - \frac{\sigma^2}{n} \neq \sigma^2$

  故改為 $E(\sigma^2_{unbiased}) = E[\frac{1}{n-1}\sum(x_i - \hat{\mu})^2] = \frac{1}{n-1}E(\sum x_i^2 - n\hat{\mu}^2)$

  $= \frac{1}{n-1} n(\sigma^2 + \hat{\mu}^2) - \frac{1}{n-1} n(\frac{\sigma^2}{n} - \hat{\mu}^2)$

  $= \frac{1}{n-1}(n-1)\sigma^2 = \sigma^2$ ⇒ 總之 $E(\sigma^2_{unb}) = \sigma^2$

  $\therefore$ bias $= 0$

$MSE(\hat{\theta}) = bias^2(\hat{\theta}) + Var(\hat{\theta})$

- $E[(\hat{\theta} - \theta)^2] = E[(\hat{\theta} - \mu - (\theta - \mu))^2]$

  $= E[(\hat{\theta} - \mu)^2] - 2\underbrace{E[(\hat{\theta} - \mu)(\theta - \mu)]}_{} + E[(\theta - \mu)^2]$

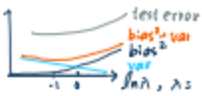
  $= E[(\hat{\theta} - \mu)](\theta - \mu)$

  $= [\cancel{E(\hat{\theta}) - \mu}](\theta - \mu)$ 0

  $= E[(\hat{\theta} - \mu)^2] + (\theta - \underbrace{\mu}_{E(\hat{\theta})})^2$

  $= Var(\hat{\theta}) + bias^2(\hat{\theta})$

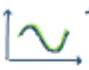
  $= Var(\hat{\theta})$ if $bias(\hat{\theta}) = 0$

  $\ln\lambda$, $\lambda = \frac{b}{a}$ $(\frac{\gamma}{\omega}?)$ ⇒ 總之 λ⊕ $\begin{cases} \text{bias} ⊕ \\ \text{var} ⊖ \end{cases}$

Bias-Variance tradeoff

$\ln\lambda = 2.6$
bias 大
var 小 : var ∝ sensitivity to training data



$\ln\lambda = -0.31$
bias 小
var 大 : complex model with many features $\Phi = \phi_1(x), \dots \phi_n(x)$

⇒ need ① dimension reduction 、② regularization

結論 : ① low variance : underfitting (simple)

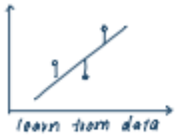
- Regression
- Naive Bayes
- Linear Model

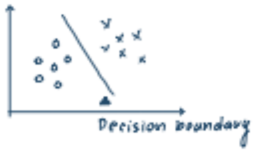
② low bias : overfitting (complex)

- non-linear
- non parametric (no assumption)
- KNN

L12-1：檢驗名詞

前言





- Confusion matrix

θ_{label} , $\hat{\theta}_{model}$

	$\hat{\theta}=Y$	$\hat{\theta}=N$	
$\theta=Y$	TP	FN	Yes
$\theta=N$	FP	TN	No
	P	N	

$$\text{accuracy} = \frac{TP+TN}{\text{total}}$$

$$1-\text{accuracy} = \frac{FP+FN}{\text{total}} = \text{error rate}$$

$$MCC : \frac{TP\cdot TN - FP\cdot FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

$$\text{specifility} = \frac{TN}{TN+FP} \quad (\text{No})$$

$$\underset{(\text{power}/\text{recall})}{\text{sensitivity}} = \frac{TP}{TP+FN} \quad (\text{Yes})$$

$$\Rightarrow F_1\text{-score} = 2\,\frac{\text{Precision}\cdot\text{Recall}}{\text{Precision}+\text{Recall}}$$

$$\underset{FPR}{\text{False positive rate}} \ \frac{TN}{TN+FP} \qquad \underset{(\text{Precision})}{\text{Positive Prediction Rate}} = \frac{TP}{TP+FP}$$

- ROC, AUC

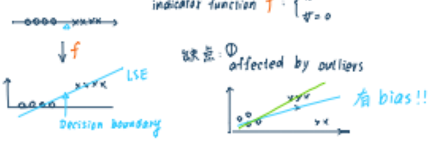
L12-2 : Regressian to classification

Regression to classification

- one-coding

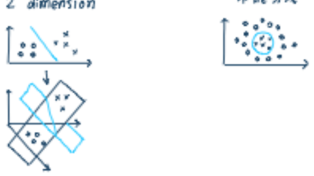
indicator function f : $\begin{cases} 男=1 \\ 女=0 \end{cases}$

缺點：① affected by outliers

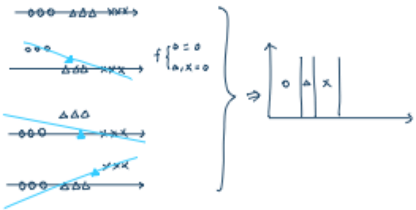


- 2 dimension

② 不能分成



- one-k-coding : multi class



課本：

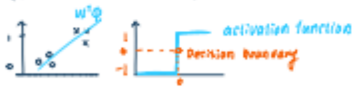
總之, we found that linear model is not that good

Loss Function Alternative

- Fisher Linear Discriminant (FLD)

Perceptron

perceptron ⟶ Logistic regression → nested regression
(used in neural network)



- Perceptron criterium

$$J = \sum \max[\underbrace{w^T\phi \cdot (\overset{t=1\vee-1}{-t})}_{\text{wrong}}, \underbrace{0}_{\text{right}}]$$

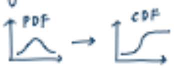
∴ J 代表 wrong prediction

註：NN 

Logistic Regression

- 利用 sigmoid 作 activation function

  想法：gaussion CDF



  但太複雜

  總之找到 logistic function $f(x) = \frac{1}{1+e^{-kx}}$

- Probability point of view

  ○ Given $D = \{(x_i, y_i) \mid y = \{0,1\}\}$ s.t. $y_i \sim \text{Bernoulli}(f(\overset{Xw}{w^T\phi}))$

  ○ 藉找 MLE，找 w：

  MLE function $= \prod \text{Bernoulli}(y \mid \overset{\theta}{w}; \phi)$

1. $\underset{w}{\arg\max}\, P(D \mid \overset{\theta}{w})$

$$= \underset{w}{\arg\max} \prod_i \left[\left(\frac{1}{1+e^{-x_iw}}\right)^{y_i} \left(\frac{e^{-x_iw}}{1+e^{-x_iw}}\right)^{1-y_i} \right]$$

2. $\Rightarrow J = \sum_{i=1}^{n} \left[y_i \log\left(\frac{1}{1+e^{-x_iw}}\right) + (1-y_i) \log\left(\frac{e^{-x_iw}}{1+e^{-x_iw}}\right) \right]$

$$\Rightarrow \frac{\delta J}{\delta w_j} = \sum \left[y_i \frac{\delta}{\delta w_j} \log\left(\frac{1}{1+e^{-x_iw}}\right) + (1-y_i) \frac{\delta}{\delta w_j} \log \frac{e^{-x_iw}}{1+e^{-x_iw}} \right]$$

① $\frac{\delta}{\delta w_j} \log\left(\frac{1}{1+e^{-x_iw}}\right) = \frac{-\delta}{\delta w_j} \log(1+e^{-x_iw}) = \frac{(x_{ij})e^{-x_iw}}{1+e^{-x_iw}}$

② $\frac{\delta}{\delta w_j} (1-y_i) \log\left(\frac{e^{-x_iw}}{1+e^{-x_iw}}\right) = (1-y_i) \frac{\delta}{\delta w_j} \left[\log e^{-x_iw} - \log(1+e^{-x_iw}) \right]$

$$= (1-y_i) \left(\frac{-x_{ij} \cancel{e^{-x_iw}}}{\cancel{e^{-x_iw}}} + \frac{x_{ij} e^{-x_iw}}{1+e^{-x_iw}} \right)$$

$$= (1-y_i) \frac{-x_{ij}}{1+e^{-x_iw}}$$

3. $\frac{\delta J}{\delta w_j} = \sum_{i=1}^{n} \left(\frac{y_i x_{ij} e^{-x_iw} - x_{ij} + y_i x_{ij}}{1+e^{-x_iw}} \right)$

$$= \sum_{i=1}^{n} \left[x_{ij} \left(y_i - \frac{1}{1+e^{-x_iw}} \right) \right]$$

4. 令 $\frac{\delta J}{\delta w_j} = 0$，以 Newton's method $x_{n+1} = x_n - H^{-1}f(x_n)\,\nabla f(x_n)$

① $\nabla f = \begin{bmatrix} \frac{\delta J}{\delta w_1} \\ \vdots \\ \frac{\delta J}{\delta w_k} \end{bmatrix} = 0$, $\quad \Phi = \begin{array}{c} \begin{matrix} \phi_1 & \cdots & \phi_d \end{matrix} \\ \begin{bmatrix} x_{11} & & x_{1d} \\ x_{21} & & x_{2d} \\ \vdots & & \vdots \\ x_{n1} & & x_{nd} \end{bmatrix} \end{array}$

$$= \Phi^T \left(\frac{1}{1 + e^{-\underset{\begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}}{w^T\phi}}} - \underset{\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}}{y} \right)$$

②

$$H = \begin{bmatrix} \frac{\delta^2 J}{\delta w_1 \delta w_1} & \frac{\delta^2 J}{\delta w_1 w_2} & \\ \frac{\delta^2 J}{\delta w_1 \delta w_2} & \frac{\delta^2 J}{\delta w_2 \delta w_2} & \\ & & \ddots \\ & & \end{bmatrix}$$

For entry $\frac{\delta}{\delta w_k}\frac{\delta}{\delta w_j} J$:

$$= \frac{\delta}{\delta w_k}\left[\sum_{i=1}^{n} x_{ij} \left(y_i - \frac{1}{1+e^{-x_i w}} \right) \right]$$

$$= \frac{-\delta}{\delta w_k} \sum_{i=1}^{n} \frac{x_{ij}}{1+e^{-x_i w}}$$

$$= \frac{-\delta}{\delta w_k} \left(\frac{x_{ij}}{1+e^{-x_{i1}w_1}} + \frac{x_{ij}}{1+e^{x_{i2}w_2}} + \cdots \frac{x_{ij}}{1+e^{-x_{ik}w_k}} + \cdots \right)$$

$$= \frac{x_{ij}\,x_{ik}\,e^{-x_{ik}w_k}}{(1+e^{-x_{ik}w_k})^2}$$

可寫成 $H = \Phi^T D \Phi$

③ $\sum_{i=1}^{n} x_{ij}\,x_{ik}\,(1+e^{-x_{ik}w_k})^{-2}\,e^{-x_{ik}w_k}$ 寫成 $H = \Phi^T D \Phi$ 的 matrix form

$$\Phi = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1d} \\ x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & & & \vdots \\ x_{d1} & \cdots & \cdots & x_{dd} \end{bmatrix} \qquad D = \begin{bmatrix} \frac{e^{-x_{i1}w_1}}{(1+e^{-x_{i1}w_1})^2} & & 0 \\ & \ddots & \\ 0 & & \frac{e^{-x_{id}w_d}}{(1+e^{-x_{id}w_d})^2} \end{bmatrix}$$

x_{ij} : jth column $\left[\; {}^{i=1,2,\ldots,d} \; \right]$ $\qquad x_{ik}$: kth column $\begin{bmatrix} x_{1k} \\ x_{2k} \\ \vdots \\ x_{dk} \end{bmatrix}$

最後取 inverse 代入 $x_{n+1} = x_n - H^{-1}f(x_n)\,\nabla f(x_n)$