L01-1: LR

LR $\begin{cases}\text{Curve fitting}\\ \text{Classification}\end{cases}$, LR also applied on non-linear thing

Data

- Diff from Algo ML is data-driven process
- In ML, express everything as **points**.
- 人 → (height, weight, age)
  - 2×2 pixels → 4 dim
- Simple model Model ← solve params

- fit the line $ax+b$ ⇒ $\begin{cases}7 = 4a+b\\ 2 = a+b\end{cases}$
- **Input Dada** $A=\begin{bmatrix}4 & 1\\ 1 & 1\end{bmatrix}$ **target** $\vec{b}=\begin{bmatrix}7\\ 2\end{bmatrix}$, **Params** $\vec{x}=\begin{bmatrix}a\\ b\end{bmatrix}$
- $A\vec{x}=\vec{b} \Rightarrow \vec{x}=A^{-1}\vec{b}$
- 解法① Gauss-Jordan $\left[\begin{array}{cc|cc}4 & 1 & 1 & 0\\ 1 & 1 & 0 & 1\end{array}\right] \rightarrow \left[\begin{array}{cc|cc}1 & 0 & \frac{1}{3} & \frac{-1}{3}\\ 1 & 1 & 0 & 1\end{array}\right] \rightarrow \left[\begin{array}{cc|cc}1 & 0 & \frac{1}{3} & \frac{-1}{3}\\ 0 & 1 & \frac{-1}{3} & \frac{4}{3}\end{array}\right]$

  Note : For $A_{N\times M}$, Time $= O(N^2 \times M)$

- 解法② LU decomposition

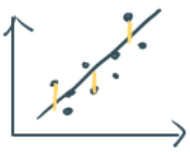
  $A\vec{x}=\vec{b} \Rightarrow LU\vec{x}=\vec{b} \Rightarrow LY=\vec{b}$ 解 $Y \Rightarrow U\vec{x}=Y$ 解 $\vec{x}$

  $\begin{bmatrix}1 & 0\\ \frac{1}{4} & 1\end{bmatrix}\begin{bmatrix}4 & 1\\ 0 & \frac{3}{4}\end{bmatrix}\vec{x}=\vec{b} \Rightarrow \begin{bmatrix}1 & 0\\ \frac{1}{4} & 1\end{bmatrix}\begin{bmatrix}y_1\\ y_2\end{bmatrix}=\begin{bmatrix}7\\ 2\end{bmatrix}$ 得 $\begin{bmatrix}y_1=7\\ y_2=\frac{1}{4}\end{bmatrix} \Rightarrow \begin{bmatrix}4 & 1\\ 0 & \frac{3}{4}\end{bmatrix}\begin{bmatrix}x_1\\ x_2\end{bmatrix}=\begin{bmatrix}7\\ \frac{1}{4}\end{bmatrix}$ 得 $\begin{bmatrix}x=\frac{5}{3}\\ x_2=\frac{1}{3}\end{bmatrix}$

  Note : leading submatrices 的 det 都 $\neq 0$ 才可 LU 分解

③ LL decomposition (cholesky)

LSE : least square error



- Concept : Minimum Loss (cost, distance, likelihood)
- 公式 : $\min \Sigma [f(x_i)-y_i]^2 = \min \|A\vec{x}-b\|^2$
  - 課本為 $\min\limits_{w} \{y(x_n,w)-t_n\}^2$ Note : 對 poly basis func, $y(x,w)=w_0+w_1x^1+\dots w_Mx^M$
  - 平方 比 | | 容易微分 故以此算 error Note : error 有最小值 ∵ error ≥ 0 ∴ is impossible

§ LSE by matrix caculus

- concept ： $\|A\vec{x}-\vec{b}\|^2$

$$= \left\| \left(\begin{bmatrix} x_0 & 1 \\ x_1 & 1 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} - \begin{bmatrix} y_0 \\ y_1 \end{bmatrix} \right) \right\|^2$$

$$= (ax_0+b-y_0)^2 + (ax_1+b-y_1)^2$$

- Error $\|A\vec{x}-\vec{b}\|^2$

$$L = (Ax-b)^T(Ax-b)$$

$$= (x^TA^T-b^T)(Ax-b)$$

$$= x^TA^TAx - \underbrace{\overset{\text{scalar}}{x^TA^Tb}}_{1\times n \;\; n\times m \;\; m\times 1} - \overset{\text{scalar}}{b^TAx} + b^Tb$$

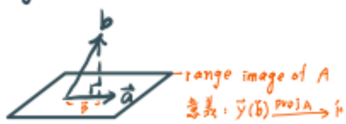
$$= x^TA^TAx - 2x^TA^Tb + b^Tb$$

$$\frac{dL}{dx} = 2A^TAx - 2A^Tb = 0$$

gram matrix，semi-positive definite ($A^TA \geq 0$, ∴ not always invertible)
∴不一定可解

$$\Rightarrow x = (A^TA)^{-1}A^Tb$$

- Geometry　Orthonormal projection 正交投影



$$a^Te = a^T(b-ax) \qquad A^T(b-Ax)=0$$

$$\Rightarrow a^Tb = a^Tax \qquad \Rightarrow A^Tb = A^TAx$$

$$\Rightarrow x = \frac{a^Tb}{a^Ta} \qquad \Rightarrow x = (A^TA)^{-1}A^Tb$$

註： $\frac{d(x^TA^TAx)}{dx}$ 推導

1. 令 A^TA 為 symmetric matrix $G = \begin{bmatrix} g_{11} & g_{12} & \cdots & g_{1n} \\ g_{21} & g_{22} & \cdots & g_{2n} \\ \vdots & & & \vdots \\ g_{n1} & g_{n2} & \cdots & g_{nn} \end{bmatrix}$

2. $\frac{d(x^TGx)}{dx} = \begin{bmatrix} \frac{\partial}{\partial x_1} \\ \frac{\partial}{\partial x_2} \\ \vdots \\ \frac{\partial}{\partial x_n} \end{bmatrix} [x_1 x_2 \dots x_n] \begin{bmatrix} g_{11} & \cdots & g_{1n} \\ g_{21} & \cdots & g_{2n} \\ \vdots & & \\ g_{n1} & \cdots & g_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$

$$= \begin{bmatrix} \frac{\partial}{\partial x_1} \\ \frac{\partial}{\partial x_2} \\ \vdots \\ \frac{\partial}{\partial x_n} \end{bmatrix} \left[x_1(g_{11}x_1 + \dots g_{1n}x_n) + x_2(g_{21}x_1 + \dots g_{2n}x_n) + \dots x_n(g_{n1}x_1 + \dots g_{nn}x_n) \right]$$

$$= \begin{bmatrix} 2g_{11}x_1 + (\sum_{i=1}^{n} g_{1i}x_i - g_{11}x_1) + (\sum_{i=1}^{n} g_{i1}x_i - g_{11}x_1) \\ \vdots \\ 2g_{nn}x_n (\sum_{i=1}^{n} g_{ni}x_i - g_{nn}x_n) + (\sum_{i=1}^{n} g_{in}x_i - g_{nn}x_n) \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} g_{1i}x_i \\ \vdots \\ \sum_{i=1}^{n} g_{ni}x_i \end{bmatrix} + \begin{bmatrix} \sum_{i=1}^{n} g_{i1}x_i \\ \vdots \\ \sum_{i=1}^{n} g_{in}x_i \end{bmatrix}$$

$$= G\vec{x} + G^T\vec{x} = 2G\vec{x} \quad (\because G = G^T)$$

註： $\frac{d\vec{x}^TA^T\vec{b}}{d\vec{x}}$ 推導

$$= \begin{bmatrix} \frac{\partial}{\partial x_1} \\ \vdots \\ \frac{\partial}{\partial x_n} \end{bmatrix} [x_1 \dots x_n] \underbrace{\overset{\text{令為 } A^T}{\begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & & \vdots \\ a_{n1} & \cdots & a_{nn} \end{bmatrix}}} \begin{bmatrix} b_1 \\ \vdots \\ b_n \end{bmatrix}$$

$$= \begin{bmatrix} \frac{\partial}{\partial x_1} \\ \vdots \\ \frac{\partial}{\partial x_n} \end{bmatrix} \left[x_1(\sum_{i=1}^{n} a_{1i}b_i) + x_2(\sum_{i=1}^{n} a_{2i}b_i) + x_n(\sum_{i=1}^{n} a_{ni}b_i) \right]$$

$$= \left[\sum_{i=1}^{n} a_{1i}b_i + \dots \sum_{i=1}^{n} a_{ni}b_i \right] = A^T\vec{b}$$

L01-3 : Non-linear LSE and overfitting

LSE in non-linear case

basis $\phi_0 = x^0$, $\phi_1 = x^1$, $\phi_2 = x^2$

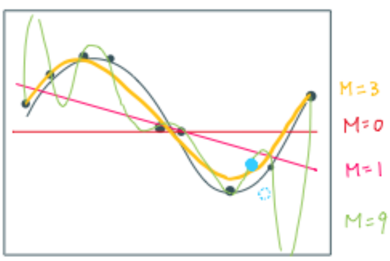
design matrix $\Phi = \begin{bmatrix} \phi_2(x_0) & \phi_1(x_0) & \phi_0(x_0) \\ \phi_2(x_1) & \phi_1(x_1) & \phi_0(x_1) \\ \vdots & \vdots & \vdots \end{bmatrix}$

Φ 和 A 概念相似，$A = \begin{bmatrix} x_0^2 & x_0 & 1 \\ x_1^2 & x_1 & 1 \\ & \vdots & \end{bmatrix} \begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} y_0 \\ y_1 \end{bmatrix}$

- Concept : $\|\Phi x - b\|^2$
- Error : $\|\Phi x - b\|^2$, similar to A, we can get $x = (\Phi^T \Phi)^{-1} \Phi^T b$
- Problem1 : Singularity ⇒ not invertible if singular
- Problem2 : Too many bases ⇒ overfitting

Overfitting

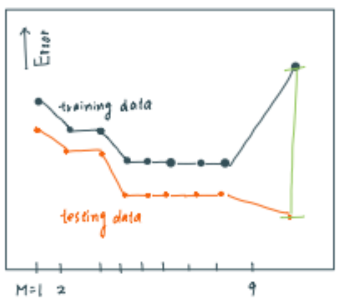
- 決定 basic function 步驟
  1. W (係數)
  2. 次方 (in poly-func case)
  3. 決定完 basic function ⇒ 對 Error $E(w)$, usually quadratic 微分 ⇒ $E'(w) = 0$ 唯一解 = LSE
- Overfitting 示範



M = 0, 1 ： underfitting，不足以描述 data

M = 9 ： overfitting ，太詳細，以致再取 sin 上一點時誤差變大

M (basis function) 太多，課本 suggest data 量的 $\frac{1}{5}$、$\frac{1}{10}$



Error 沒有越小↘ ： overfit may occur

training data fitted perfectly, but won't fit well on other data

原因 ： ① (M 太大)

② data 太少 ，(若增加上圖 ◌，倒數 no.2 轉折可能不那么大)

③ 雜訊 (上圖 •)

	M=0	M=1	M=6	M=9
w_0	0.19	0.82	0.31	0.35
w_1		1.27	7.99	232.37
w_2			-25.43	-5321.83
w_3			17.37	48568.31
w_4				-231639.30
⋮				⋮

overfit 特徵 ： function ‖係數‖ hen 大

rLSE : regularization 防 overfitting

- Concept : 加入 penalty 懲罰太大的 W
- E(w) : $\sum_{n=1}^{N}\left[y(x_n, w)-t_n\right]^2+\lambda\|w\|^2$
- 求 minimum E(W) by 微分 :

$$\frac{d}{d\vec{x}}\left(\|A\vec{x}-\vec{b}\|^2+\lambda\|w\|^2\right)$$

$$=\frac{d}{d\vec{x}}\left[(A\vec{x}-\vec{b})^T(A\vec{x}-\vec{b})+\lambda\vec{x}^T\vec{x}\right]$$

Note : $(\vec{x}^TA^T-\vec{b}^T)(A\vec{x}-\vec{b})$
$=\vec{x}^TA^TA\vec{x}-\vec{x}^TA^T\vec{b}-\vec{b}^TA\vec{x}+\vec{b}^T\vec{b}$
$=\vec{x}^TA^TA\vec{x}-2\vec{x}^TA^T\vec{b}+\vec{b}^T\vec{b}$ $\because$ $(A\vec{x})^T\vec{b}=\vec{b}^T(A\vec{x})$ 為純量

$$=\frac{d}{d\vec{x}}\left(x^TA^TA\vec{x}-2\vec{x}^TA^T\vec{b}+\vec{b}^T\vec{b}+\lambda\vec{x}^T\vec{x}\right)$$

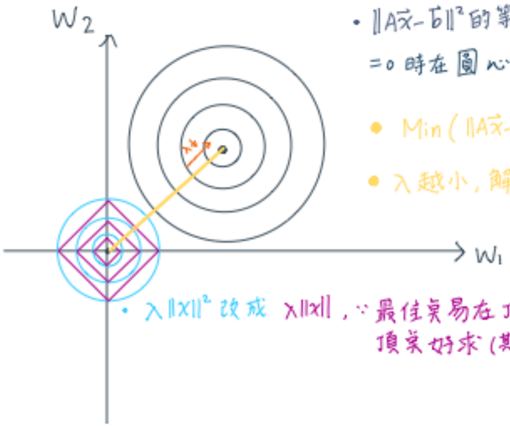
$\Rightarrow 2A^TA\vec{x}-2A^T\vec{b}+2\lambda\vec{x}=0$

$\Rightarrow A^TA\vec{x}-A^T\vec{b}+\lambda\vec{x}=0$

$\Rightarrow (A^TA+\lambda I)\vec{x}=A^T\vec{b}$

$\Rightarrow \vec{x}=(A^TA+\lambda I)^{-1}A^T\vec{b}$ $\Rightarrow$ for $A^TA\geq 0$, $A^TA+\lambda I$ 必 >0 為正定 $\Rightarrow$ 必可逆

- 求 minimum E(w) by 圖 :

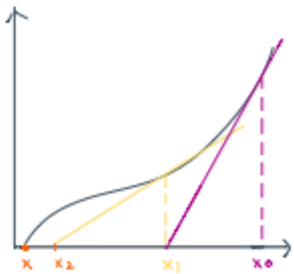


- $\|A\vec{x}-\vec{b}\|^2$ 的等高線
  $=0$ 時在圓心, $\vec{x}=(A^TA)^{-1}A^Tb$
- Min($\|A\vec{x}-\vec{b}\|^2+\lambda\|\vec{x}\|^2$) 的解 (o line)
- λ 越小, 解越接近 $\|A\vec{x}-\vec{b}\|$ 圓心, 反之同理
- $\lambda\|x\|^2$ 改成 $\lambda\|x\|$, $\because$ 最佳点易在頂点 (尤其高維)
  頂点好求 (其他維 $=0$)

L02 : Newton's Method

Newton's method : root finding ⇒ optimization

- 观点①



Step1 : choose x_0 . find $x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}$

Step2 : continue from x_1 . $x_2 = x_1 - \frac{f(x_1)}{f'(x_1)}$

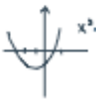
⋮

Stepn : stop at $x_n = x_{n-1} - \frac{f(x_{n-1})}{f'(x_{n-1})}$ till $(x_n - x_{n-1})$ 夠小

或

$f(x_n) = 0$

Ex :



$x^2+2x-3 = (x+3)(x-1)$

取 $x_0=2$, $x_1 = 2 - \frac{f(2)}{f'(2)} = 1.167$

$\Rightarrow x_2 = 1.167 - \frac{f(1.167)}{f'(1.167)} = 1.01$

$\Rightarrow \dots$

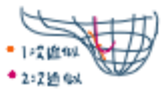
$\Rightarrow x_N$ 收斂至 1.00

- 观点② : Taylor expansion

  - $f(x) \cong f(x_0) + \frac{1}{1!}f'(x_0)(x-x_0) + \frac{1}{2!}f''(x_0)(x-x_0)^2 + \dots$

$$= \sum_{n=0}^{\infty} \frac{f^n(x_0)}{n!}(x-x_0)^n$$

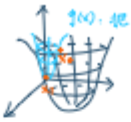
$$= g(x) \text{ such that } \begin{cases} g(x_0) = f(x_0) \\ g'(x_0) = f'(x_0) \\ \vdots \\ g^n(x_0) = f^n(x_0) \end{cases}$$



- $(x_0=0)$ $f(x) = f(x_0) + \frac{f'(x_0)}{1!}x + \frac{f''(x_0)}{2!}x^2 + \dots$ Maclarin series
- $(x_0=0)$ $e^x = 1 + x + \frac{x^2}{2!} + \dots = \sum_{n=0}^{\infty}\frac{x^n}{n!}$
- $(x_0=0)$ $\sin x = \cancel{\sin 0} + \frac{\cos 0}{1!}x + \cancel{\frac{-\sin 0}{2!}x^2} + \frac{-\cos 0}{3!}x^3 + \cancel{\frac{\sin 0}{4!}}x^4 + \dots = x - \frac{x^3}{3} + \frac{x^5}{5} - \frac{x^7}{7} + \dots$
- $(x_0=0)$ $\cos x = \cos 0 + \cancel{\frac{-\sin 0}{1!}}x + \frac{-\cos 0}{2!}x^2 + \cancel{\frac{\sin 0}{3!}}x^3 + \frac{\cos 0}{4!}x^4 + \dots = 1 - \frac{x^2}{2} + \frac{x^4}{4} - \frac{x^6}{6} + \dots$
- $(x_0=0)$ $e^{ix} = 1 + i\frac{e^{ix}}{1!}x + (-1)\frac{e^{ix}}{2!}x^2 + (-1)\frac{e^{ix}}{3!}x^3 + (1)\frac{e^{ix}}{4!}x^4 + \dots = i\sin x + \cos x$

$\Rightarrow$ for $x = \pi$, $e^{i\pi} + 1 = 0$

- Newton's Method in optimization



$g(x)$：把點 x_0 的 $f(x)$ 精簡成2次近似

$$f(x) \simeq f(x_0) + f'(x_0)\underbrace{\Delta x}_{x-x_0} + \frac{f''(x_0)}{2!}\Delta x^2 = g(x)$$

註: Hession matrix
(Hession func of $f(x)$)

$$\begin{bmatrix} \frac{\partial^2 f}{\partial x_0^2} & \frac{\partial^2 f}{\partial x_0 x_1} & \cdots \\ \frac{\partial^2 f}{\partial x_1 x_0} & \frac{\partial^2 f}{\partial x_1^2} & \cdots \\ \cdots & \cdots & \cdots \end{bmatrix}$$

- 目標：$f'(x)=0$，從 x_0 出發，找 $f'(x_0+\Delta x)=0$

$$\Rightarrow g'(x) = f'(x_0) + f''(x_0)\Delta x = 0$$

$$\Rightarrow \Delta x = \frac{-f'(x_0)}{f''(x_0)}$$

$$\therefore X_{n+1} = x_n + -\frac{f'(x_n)}{f''(x_n)}$$

$$= X_n - \underbrace{H^{-1}(f(x))}_{f''(x_n)} \underbrace{\nabla f(x_n)}_{f'(x_n) \Rightarrow \text{gradient}}$$

$\Rightarrow$ 一直取 x_{i+1} 至 $(x_i - x_{i-1})$ 夠小

- Newton's method 驗証 LSE

$$\|A\vec{x}-\vec{b}\|^2 = (A\vec{x}-\vec{b})^T(A\vec{x}-b)$$

$$= \vec{x}^TA^TA\vec{x} - 2\vec{x}^TA^T\vec{b} + \vec{b}^T\vec{b}$$

微分 $\Rightarrow 2A^TA\vec{x} - 2A^T\vec{b} = \nabla f(\vec{x})$

再微 $\Rightarrow 2A^TA = Hf(\vec{x})$

$$\therefore \vec{x}_1 = \vec{x}_0 - H^{-1}f(\vec{x}_0)\nabla f(\vec{x}_0)$$

$$= \vec{x}_0 - (2A^TA)^{-1}(2A^TA\vec{x}_0 - 2A^T\vec{b})$$

$$= \vec{x}_0 - \cancel{\tfrac{1}{2}}(A^TA)^{-1}(\cancel{2}A^TA\vec{x}_0 - \cancel{2}A^T\vec{b})$$

$$= (A^TA)^{-1}A^Tb$$

- 缺点

① Newton's method may be trapped in local

② Hessian takes $O(n^3)$, slow but takes less iteration

similar to $\begin{cases} \text{gradient descent} \\ \text{conjugate gradient descent} \end{cases}$

③ Hession 可能 invertible，解法 $\begin{cases} \text{Pseudo inverse} \\ \text{quasi-Newton approach} \end{cases}$

L03-1 : probability

Probability

Used for approximation since we know too few about the parameters that will affect the results

- 名詞 (例：擲硬幣)
  - trial：toss 2 coins
  - outcome：HH、HT、TH、TT
  - event : set of outcomes {HH, HT, TH, TT}
  - sample space：all outcomes (U)
  - random variable：mapping function Ex：HH → 0, HT → 1, TH → 2, TT → 3 ⇒ $P(x=H) = P(x=1) \xrightarrow[\text{of trials}]{\infty \text{ times}} P(x=1)=\frac{1}{2}$

- PMF、PDF、CDF

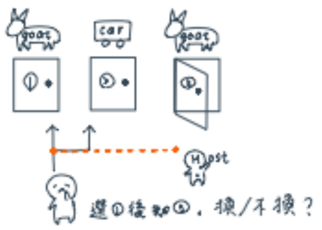
Conditional / Joint Probability

- Conditional probability $P(A|\overset{\text{sample space shrink}}{B}) = \frac{P(A,B)}{P(B)} = \frac{P(A)\,P(B|A)}{P(B)}$
- Joint probability $P(A,B) = P(A)\,P(B|A) = P(A,B|U)$，U 為宇集

- Conditional problem 三門問題

令換/不換 Y/N
得車/羊 C/G

Key : sample space changed
(shrink as ③ is opended)
主持人必選羊 (非公正)



解法1

			換	不換
Pick		host	T	F
host	Pick	host	F	T
host		Pick	T	F

∴ 換而中獎机率 $= \frac{2}{3}$

不換中獎机率 $= \frac{1}{3}$

解法2

	選	開	換	結果	P
	$\frac{1}{3}$	十羊	O $\frac{1}{2}$	必羊	$\frac{1}{6}$
			X $\frac{1}{2}$	必車	$\frac{1}{6}$
	$\frac{2}{3}$	十羊	O $\frac{1}{2}$	必車	$\frac{2}{6}$
			X $\frac{1}{2}$	必羊	$\frac{2}{6}$

∴ $P(車|換) = \frac{2/6}{3/6} = \frac{2}{3}$

$P(車|不換) = \frac{1/6}{3/6} = \frac{1}{3}$

☆ 考試可能出 4門2車等

L03-2 : Bayes Theorem

Baye's Theorem

$$\underbrace{P(A|B)}_{\text{Posterior}} = \underbrace{P(B|A)}_{\text{likelihood}} \frac{\overset{\text{prior, A is parameter}}{P(A)}}{\underset{\text{marginal}}{P(B)}} \Rightarrow P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)}, \ D = \text{event}$$

Ex: A = Vote for 馬, B = vote for KMT

$P(A) = 10\%$, $P(B) = 50\%$, $P(B|A) = 90\%$ ⇒ $P(A|B)$

Ex: A = lung cancer, B = smoke

$P(B|A)$ 好求 ∵ A人少

⇒ $P(A|B)$ 吸煙得肺癌机率 $= \dfrac{0.1 \times 0.9}{0.5} = 18\%$

註

Sum rule : $P(A) = P(A,B) + P(A \text{ not } B) = \overset{\text{marginalize}}{\sum_{v} P(A,v)}$

product rule : $P(A_4, A_3, A_2, A_1) = P(A_4, A_3, A_2 | A_1) P(A_1)$

$= P(A_4, A_3 | A_2, A_1) P(A_2 | A_1) P(A_1)$

$= P(A_4 | A_3, A_2, A_1) P(A_3 | A_2, A_1) P(A_2 | A_1) P(A_1)$

Bayesian V.S. Frequentist (見 L05 note)

Bayesian	Frequentist
prior knowledge	observation
$P(B) = \frac{1}{2}$ though 很多 BB...B	∴ 很多 BB...B ⇒ guess next is B

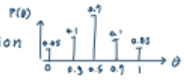
Distribution

Describe
- location
  - mean $E(x) = \overset{\text{1st moment}}{\sum p(x_i) x_i}$
  - median
  - mode (most frequent)
- Dispersion - variance $Var(x) = \overset{\text{2nd moment}}{E((x-\mu)^2)} = \frac{1}{n}\sum(x-\mu)^2 = \frac{1}{n}\sum(x^2 - 2\mu x + \mu^2) = \frac{1}{n}\sum x^2 - \frac{1}{n}2\mu\sum x + \frac{1}{n}\sum\mu^2 = \sum x^2 - \mu^2 = E(x^2) - E(x)^2$
- Skewness -  $\overset{\text{3rd moment}}{E((x-\mu)^3)}$
- Kurtosis (peakness) $\overset{\text{4th moment}}{E((x-\mu)^4)}$

L03-3 : Naïve Bayes classifier

Naïve Bayes classifier

- MLE：找 $\theta = P$ 得到 maximum likelihood　$P(D|\theta) \Rightarrow L(\theta|D)$
  Posterior

o 例：对 Bayesian 令 θ 的 distribution 

trial 1:

$$P(\theta=0|H) = \frac{P(H|\theta=0)P(0)}{P(H)} = 0$$

$$P(\theta=0.3|H) = \frac{P(H|\theta=0.3)P(0.3)}{P(H)} = \frac{0.3 \cdot 0.1}{0.5} = 0.06 \xrightarrow{HH} 0.032 \ \ldots \xrightarrow{H^{10}} 1.03E\text{-}06$$

$0.03 + 0.35 + 0.07 + 0.05 = 0.5$

$$P(\theta=0.5|H) = \frac{P(H|\theta=0.5)P(0.5)}{P(H)} = \frac{0.5 \cdot 0.7}{0.5} = 0.7 \xrightarrow{HH} 0.618 \ \ldots \xrightarrow{H^{10}} 0.03$$

$$P(\theta=0.7|H) = \frac{0.7 \cdot 0.1}{0.5} = 0.14 \xrightarrow{HH} 0.173 \ \ldots \xrightarrow{H^{10}} 0.027$$

$$P(\theta=1|H) = \frac{1 \cdot 0.05}{0.5} = 0.1 \xrightarrow{HH} 0.177 \ \ldots \xrightarrow{H^{10}} 0.97$$

☆ ∵ H 次數↑，MLE 發生點 從 $\theta=0.5 \to \theta=1$（趨 frequentist）

o 例：对 Frequentist

$P(HH|P=1) = L(P=1|HH) = 1$ MLE

$P(HH|P=0.7) = 0.49$

$P(HH|P=0.5) = 0.25$

$P(HH|P=0.1) = 0.01$

- Bayes 例題

技① $P(\theta|D) = \dfrac{P(D|\theta)P(\theta)}{P(D)}$

data 不夠假設

技② conditional independence $= \dfrac{P(d_1|\theta)P(d_2|\theta)P(d_3|\theta)P(d_4|\theta)}{P(D)}$

$$P\left(play = yes \,\middle|\, \begin{array}{l} outlook = sunny \\ Temp = cool \\ Humidity = high \\ wind = strong \end{array}\right)$$

技③ 不用贏闘 $P(D)$ ∵最後會 normalize

$$= \frac{P(O=☼, T=❄, H=high, W=strong \mid yes)\, P(yes)}{P(☼, ❄, H=high, W=strong)}$$

$$= \frac{P(☼|yes)\,P(❄|yes)\,P(H=high|yes)\,P(W=strong|yes)\,\frac{9}{14}}{P(☼, ❄, H=high, W=strong)}$$

$$= \frac{\frac{2}{9}\cdot\frac{3}{9}\cdot\frac{3}{9}\cdot\frac{3}{9}\left(\frac{9}{14}\right)}{P(☼, ❄, H=high, W=strong)}$$

∵
$$\begin{cases} P(play=Yes|D) = \dfrac{\frac{2}{9}\frac{3}{9}\cancel{\frac{3}{9}}\cancel{\frac{3}{9}}\left(\cancel{\frac{9}{14}}\right)}{P(D)} = \dfrac{1}{14}\dfrac{1}{P(D)}\dfrac{2}{3^3} \\ P(play=No|D) = \dfrac{\frac{3}{5}\frac{1}{5}\frac{4}{5}\cancel{\frac{3}{5}}\left(\cancel{\frac{5}{14}}\right)}{P(D)} = \dfrac{1}{14}\dfrac{1}{P(D)}\dfrac{36}{5^3} \end{cases}$$

$$\Rightarrow \frac{P(play=yes|D)}{P(play=No|D)} = \frac{2\cdot 5^3}{3^3\cdot 36} = \frac{125}{486}$$

$$\Rightarrow \begin{cases} P(play=yes|D) = \dfrac{125}{611} \cong 20\% \\ P(play=No|D) = \dfrac{486}{611} \cong 80\% \end{cases}$$

註：conditional independence

$P(A,B) = P(A)P(B)$

$P(A,B|U) = P(A|U)P(B|U)$

$P(A_1, A_2, A_3, A_4)$

$= P(A_1|A_2, A_3, A_4)\,P(A_2|A_3, A_4)\,P(A_3|A_4)\,P(A_4)$

If A_2, A_3, A_4 are independent

$= P(A_1|A_4)\,P(A_2|A_4)\,P(A_3|A_4)$

① ∵ $P(A_1, A_2|A_3, A_4) = P(A_1|A_3, A_4)P(A_2|A_3, A_4)$

$$\Rightarrow \frac{P(A_1,A_2,A_3,A_4)}{\cancel{P(A_3,A_4)}} = \frac{P(A_1,A_3,A_4)}{P(A_3,A_4)}\frac{P(A_2,A_3,A_4)}{\cancel{P(A_3,A_4)}}$$

$$\Rightarrow \frac{P(A_1,A_2,A_3,A_4)}{P(A_2,A_3,A_4)} = \frac{P(A_1,A_3,A_4)}{P(A_3,A_4)}$$

∴ $P(A_1|A_2,A_3,A_4) = P(A_1|A_3,A_4) \overset{②}{=} P(A_1|A_4)$

② ∵ $P(A_2, A_3|A_4) = P(A_2|A_4)P(A_3|A_4)$

$$\Rightarrow \frac{P(A_2,A_3,A_4)}{\cancel{P(A_4)}} = \frac{P(A_2,A_4)}{P(A_4)}\frac{P(A_3,A_4)}{\cancel{P(A_4)}}$$

$$\Rightarrow \frac{P(A_2,A_3,A_4)}{P(A_3,A_4)} = \frac{P(A_2,A_4)}{P(A_4)}$$

∴ $P(A_2|A_3,A_4) = P(A_2|A_4)$

L04-1 : Information Theory

Information theory

- Entropy 熵 : randomness 亂度, to describe information
  - unit of randomness / uncertainty : $-\log_2 P$, which shows how many bits to describe

    Ex: $\begin{cases} -\log\frac{1}{4} = 2 & \text{(take 2 bits)} \\ -\log\frac{1}{1024} = 10 & \text{: rare event gain more info} \end{cases}$

  - Def : Exp (info) in events

    △ review $\begin{cases} E(x) = \sum p(x)x \\ E(f(x)) = \sum p(x) f(x) \end{cases}$

    △ $H(x) = -\sum P(x) \log P(x) = -\underset{x\sim p}{E}[\log_2 P]$

    △ Example : flip 2 fair coins

    mapping : random variable X

    HH → 2
    HT → 1
    TH → 1
    TT → 0

    $H(X) = -\underbrace{\sum \frac{1}{4}}_{\text{4 outcomes}} \cdot \log_2\frac{1}{4} = 2$ (得 2 bit 資訊量)

    $H(X) = -\left[\frac{9}{16}\log_2\frac{9}{16} + 2\left(\frac{3}{16}\log_2\frac{3}{16}\right) + \frac{1}{16}\log_2\frac{1}{16}\right]$ if $p(H) = \frac{3}{4}$

    △ Note : $\begin{cases} \text{max(entropy) = base} \\ \text{max(entropy) 發生在 uniform (難 predict、越乱 → 含更多 info)} \end{cases}$

    $-\sum 1 \log 1 = 0$, 若 always happen 則無 uncertainty

- Conditional entropy

  $H(Y|X) = -\sum P(x,y) \log\frac{P(x,y)}{P(x)} = (H(X,Y) - H(X))$

  pf: $H(Y|X) = -\sum_i P(x_i) H(Y|X=x_i)$

  $= -\sum_i P(x_i) \sum_j P(y_i|x_i) \log (y_j|x_i)$

  $= -\sum_i\sum_j P(x_i, y_j) \log P(y_j|x_i)$

  $= -\sum_i\sum_j P(x_i, y_j) \log\frac{P(x_i,y_j)}{P(x_i)}$

  $= \sum_i\sum_j P(x_i, y_j) \log\frac{P(x_i)}{P(x_i,y_j)}$

- Joint entropy

  $H(X,Y) = H(X) + H(Y|X)$

  Info needed by H(x) to get H(x,y)

  pf: $H(X,Y) = -\sum_{x\in X}\sum_{y\in Y} P(x,y) \log P(x,y)$

  $= -\sum_{x\in X}\sum_{y\in Y} P(x,y) \log P(x) P(y|x)$

  $= -\sum_{x\in X}\sum_{y\in Y} P(x,y) \log P(x) - \sum_{x\in X}\sum_{y\in Y} P(x,y)\, P(y|x)$

  $= -\sum_{x\in X} P(x) \log P(x) - \sum_{x\in X}\sum_{y\in Y} P(x,y) P(Y|x)$

  $= H(X) + H(Y|X)$

- Relative entropy (KL divergence)
  - usage : Distance between 2 distribution (random var)
  - $KL(M \| N) = H_N(M) - H(M)$

    $= \left[-\sum_i M(x_i) \log N(x_i)\right] - \left[-\sum_i M(x_i) \log M(x_i)\right]$

    $= -\sum_i M(x_i) \frac{N(x_i)}{M(x_i)}$

    $= \sum_i M(x_i) \frac{M(x_i)}{N(x_i)}$

  $\therefore KL(p \| q) = \sum_{x\in X} P(x_i) \ln\frac{P(x_i)}{q(x_i)} = \underset{x\sim p}{E}[P(x) - q(x)]$

  但不好用 ∵ $KL(p\|q) \neq KL(q\|p)$

  ⇒ 改用 cross entropy $H(P,q) = \underset{x\sim p}{E}[-\ln q(x)]$

  $= -P\ln q - (1-P)\ln(1-q)$

  $= KL(P\|q) + H(P)$

- Mutual information  mutual entropy
  - usage : to show the independence

    ∴ 若 X Y independent , $I(X,Y) = 0$

  - 公式 : $I(X;Y) = H(X) - H(X|Y) \vee H(Y) - H(Y|X)$

    pf : $I(X;Y) = \sum_{x,y} P(x,y) \log\frac{P(x,y)}{P(x)P(y)}$

    $= \sum_{x,y} P(x,y) \frac{P(x|y)}{P(x)} \vee \sum_{x,y} P(x,y) \frac{P(y|x)}{P(y)}$

    $= H(X) - H(X|Y) \vee H(Y) - H(Y|X)$

  - Problem : The scale is unpredictable

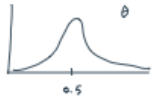
L04-2 : Maximum entropy

- Maximum Entropy Principle
  - frequentist v.s. bayeshan

$L(H \ldots H \mid \theta = 0.5) = \frac{1}{2^{10}}$

$L(H \ldots H \mid \theta = 0.9) = \frac{9^{10}}{10^{10}}$

$L(H \ldots H \mid \theta = 1) = 1$



↳ frequentist 看到会猜 coin always H | bayeshan ∵有 prior ∴猜 coin 还是 fair

  - uniform distribution → max(entropy)

Pf : $H(x) = -\int_a^b p(x) \log p(x)\,dx$, $p(x) \geq 0$ & $\int_a^b p(x)\,dx = 1$

$\Rightarrow \frac{\delta}{\delta p(x)} \left[-\int_a^b p(x) \lg p(x)\,dx + \lambda \left(\int_a^b p(x)\,dx - 1 \right) \right]$ 見 lagrange multiplier

$\Rightarrow \frac{\delta}{\delta p(x)} \left[-p(x) \lg p(x) + \lambda p(x) \right]$

$\Rightarrow \left[-\lg p(x) - \cancel{p(x)} \frac{\overset{1}{\cancel{1}}}{p(x)} + \lambda \right] = 0$

$\Rightarrow p(x) = e^{\lambda - 1}$

$\because \int_a^b p(x)\,dx = 1 \Rightarrow p(x)\, x \Big|_a^b = 1 \Rightarrow p(x) \frac{1}{b-a} = 1$

$\therefore p(x) = e^{\lambda - 1} = \frac{1}{b-a}$

註 : lagrange multiplier : f 在 constraint g 取極值

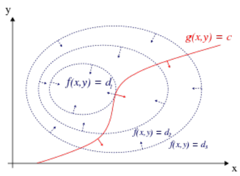
Ex : $f = 4xy$, constraint $x>0 . y>0 . \frac{x^2}{9} + \frac{y^2}{16} = 1$

$L = f - \lambda g$, $g = \frac{x^2}{9} + \frac{y^2}{16}$ Note : 最小化 : $\nabla L = \nabla f - \lambda \nabla g = 0$

最大化 : $\nabla L = \nabla f + \lambda \nabla g = 0$

$\nabla L = (4y\vec{i} + 4x\vec{j}) - \lambda \left(\frac{2x}{9}\vec{i} + \frac{y}{8}\vec{j} \right)$

$\Rightarrow \begin{cases} 4y = \lambda \frac{2}{9} x \\ 4x = \lambda \frac{1}{8} y \end{cases} \Rightarrow \frac{y}{x} = \frac{4}{3}$ 代入 $g \Rightarrow y = \pm 2\sqrt{2}$ (取正)



- Maximum Entropy (Given μ)

Given expectation : $\int_0^\infty x\, p(x)\,dx = \mu$ 求 $\max = -\int_0^\infty p(x) \ln p(x)\,dx$

令 $L = \underline{-\int_0^\infty p(x) \ln p(x)\,dx} + \underbrace{\lambda_0 \left(\int_0^\infty p(x) - 1 \right)}_{C_1} + \underbrace{\lambda_1 \left(\int_0^\infty x\, p(x) - \mu \right)}_{C_2}$

$\frac{\delta L}{\delta p(x)} = -(\ln p(x) + 1) + \lambda_0 + \lambda_1 x = 0$

$\Rightarrow \ln p(x) = \lambda_0 + \lambda_1 x - 1$

$\Rightarrow p(x) = e^{\lambda_0 + \lambda_1 x - 1}$ ⓐ1

代入 $C_1 \Rightarrow \int_0^\infty e^{\lambda_0 + \lambda_1 x - 1}\,dx = 1$

$\Rightarrow \frac{1}{\lambda_1} e^{\lambda_0 + \lambda_1 x - 1} \Big|_0^\infty = 1$

$\Rightarrow \frac{e^{\lambda_0 - 1}}{\lambda_1} (e^{\lambda_1 \infty} - 1) = 1$ $\therefore \lambda_1$ 必 ≤ 0 使 $e^{\lambda_1 \infty} = 0$

$\Rightarrow \frac{e^{\lambda_0 - 1}}{\lambda_1} = -1 \Rightarrow \lambda_1 = -e^{\lambda_0 - 1}$ ⓐ2

代入 C_2 : $\int_0^\infty x e^{\lambda_0 + \lambda_1 x - 1}\,dx = e^{\lambda_0 - 1} \int_0^\infty \underset{f}{\underline{x}}\, \underset{g'}{\underline{e^{\lambda_1 x}}}\,dx = \mu$ Note: $\int f g' = fg - \int f' g$

$\Rightarrow e^{\lambda_0 - 1} \left[\underbrace{\cancel{x \frac{e^{\lambda_1 x}}{\lambda_1} \Big|_0^\infty}}_{0-0} - \int_0^\infty \frac{e^{\lambda_1 x}}{\lambda_1}\,dx \right] = \mu$

$\Rightarrow \frac{1}{-\lambda_1} \int_0^\infty \cancel{e^{\lambda_0 + \lambda_1 x - 1}}^{\,1}\,dx = \frac{1}{-\lambda_1} = \mu$

$\Rightarrow \lambda_1 = \frac{1}{-\mu}$ ⓐ3

From ⓐ1 ⓐ2 ⓐ3 : $p(x) = e^{\lambda_0 - 1} \cdot e^{\lambda_1 x}$

$= -\lambda_1 \cdot e^{-\frac{1}{\mu} x}$

$= \frac{1}{\mu} e^{-\frac{1}{\mu} x}$ exponential distribution 

- Maximum Entropy (Given μ 、σ^2)

令 $L = -\int p(x) \ln p(x)\,dx + \underbrace{\lambda_0 \left(\int p(x)\,dx - 1 \right)}_{C_1} + \underbrace{\lambda_1 \left(\int x\, p(x)\,dx - \mu \right)}_{C_2} + \underbrace{\lambda_2 \left(\int (x-\mu)^2 p(x)\,dx - \sigma^2 \right)}_{C_3}$

$\frac{\delta L}{\delta p(x)} = -(\ln p(x) + 1) + \lambda_0 + \lambda_1 x + \lambda_2 (x-\mu)^2 = 0$

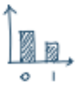
...

$\Rightarrow p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ Guassin distribution

L05-1 : Bernoulli distribution

- 整理
  - uniform distr. 
  - exponential distr. 
  - Gaussin distr. 

Bernoulli Distribution



- $P(x=0,1|\theta)=\theta^{x}(1-\theta)^{1-x}$

  $E(x)=1\cdot\theta+0\cdot(1-\theta)=\underline{\theta}$ Location

  $Var(x)=E(x^2)-E(x)^2$

  $=\theta-\theta^2$

  $=\underline{\theta(1-\theta)}$ Dispersion

- Bernoulli distribution
  - Given $[0,0,1,\dots]=[x_1,x_2,\dots x_N]=D$

  $P(D|\theta)=\prod_{i=1}^{N}\theta^{x_i}(1-\theta)^{1-x_i}$

☆ 若已知結果D，求參數θ讓D發生机率最大=MLE

即找 $\underset{\theta}{\arg\max}\prod\theta^{x_i}(1-\theta)^{1-x_i}\Rightarrow\underset{\theta}{\arg\max}\left[\sum x_i\log\theta+\sum(1-x_i)\log(1-\theta)\right]$

令 $L=\sum x_i\log\theta+\sum(1-x_i)\log(1-\theta)$

$\frac{dL}{d\theta}=\sum_{i=1}^{N}x_i\frac{1}{\theta}-\sum_{i=1}^{N}(1-x_i)\frac{1}{1-\theta}=0$ Note : $\frac{d}{dx}\log x=\frac{1}{x}$

$\Rightarrow\sum_{i=1}^{N}x_i\frac{1}{\theta}=\sum_{i=1}^{N}\frac{1}{1-\theta}-\sum_{i=1}^{N}x_i\frac{1}{1-\theta}$

$\Rightarrow\sum_{i=1}^{N}x_i\left(\frac{1}{\theta}+\frac{1}{1-\theta}\right)=\sum_{i=1}^{N}\frac{1}{1-\theta}$

$\Rightarrow\left(\frac{1}{\theta\cancel{(1-\theta)}}\right)\sum_{i=1}^{N}x_i=\frac{N}{\cancel{1-\theta}}$

$\Rightarrow\theta=\frac{\sum_{i=1}^{N}x_i}{N}$ ⇒ MLE發生在 sample mean

∴ 若10次有5次中，应猜θ為 $\frac{5}{10}$

L05-2 : Binomial distribution

Binomial distribution

- Multiple Bernoulli trial (N) ⇒ Binomial

  $P(X=m \mid N,\theta) = \binom{N}{m} \theta^{m}(1-\theta)^{N-m}$, X is # of succes in N trials

  $E(x) = N\theta$

  $Var(x) = N\theta(1-\theta)$

  $MLE = \theta = \frac{\Sigma m}{\Sigma N}$

conjugate prior

Frequentist based on MLE (Maximum Likelihood Estimation)、dataset , Bayesian based on prior ex: $P(\theta=0.5)$ = 大机率, $P(\theta=1)$ = 小机率

Frequentist 忽略其他可能，Bayesian 要看 prior $P(\theta)$ 好壞

∴ 折衷找 $P(\theta|x)$, posterior 為：

$$\overset{\text{Beta}}{P(\theta|x)} = \frac{\overset{\text{Binomial}}{P(x|\theta)}\overset{\text{Beta}}{P(\theta)}}{P(x)}$$

- conjugate prior - posterior
  - 來由：算 posterior 太累，∴找 distribution 使 prior、posterior in the same form
    此 distribution is called conjugate

Beta distribution

- Beta function $\mathcal{B}(\theta|a,b)$

$$= P^{a-1}(1-P)^{b-1} \frac{1}{\beta(a,b)}$$

$$= p^{a-1}(1-p)^{b-1} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}$$

- $E(x) = \frac{a}{a+b}$ $\quad Var(x) = \frac{ab}{(a+b)^2(a+b+1)}$

Pf :

$$E(x) = \int_0^\infty x \cdot x^{a-1}(1-x)^{b-1}\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}dx$$
$$= \frac{1}{\beta(a,b)}\int_0^\infty x^{(a+1)-1}(1-x)^{b-1}dx$$
$$= \frac{1}{\beta(a,b)} \cdot \beta(a+1,b)$$
$$= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \frac{a\Gamma(a)\Gamma(b)}{(a+b)\Gamma(a+b)} = \frac{a}{a+b}$$

$$E(x^2) = \int_0^\infty x^2 \cdot x^{a-1}(1-x)^{b-1}\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}dx$$
$$= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\int_0^\infty x^{(a+2)-1}(1-x)^{b-1}dx$$
$$= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \cdot \frac{\Gamma(a+2)\Gamma(b)}{\Gamma(a+b+2)}$$
$$= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \frac{(a+1)a\Gamma(a)\Gamma(b)}{(a+b+1)(a+b)\Gamma(a+b)} = \frac{(a+1)a}{(a+b+1)(a+b)}$$

$$Var(x) = E(x^2) - E(x)^2$$
$$= \frac{a}{a+b}\left(\frac{a+1}{a+b+1} - \frac{a}{a+b}\right) = \frac{a}{a+b}\left(\frac{b}{(a+b+1)(a+b)}\right)$$
$$= \frac{ab}{(a+b)^2(a+b+1)}$$

註: gamma function $\Gamma(x) = \int_0^\infty p^{x-1}e^{-p}dp$

性質 ① $\Gamma(x) = (x-1)\Gamma(x-1)$

Pf : $\Gamma(x) = \int_0^\infty \underbrace{p^{x-1}}_{u}\underbrace{e^{-p}}_{dv}dp$
$$= -p^{x-1}e^{-p}\Big|_0^\infty + \int_0^\infty (x-1)p^{x-2}e^{-p}dp$$
$$= (x-1)\int_0^\infty p^{(x-2)}e^{-p}dp$$
$$= (x-1)\Gamma(x-1)$$

因此推得

$$\Gamma(1) = \int_0^\infty e^{-p}dp = -e^{-p}\Big|_0^\infty = 0-(-1) = 1$$
$$\Gamma(2) = \int_0^\infty pe^{-p}dp = -pe^{-p}\Big|_0^\infty + \int 1 \cdot p^0 e^{-p}dp = 0 + 1\Gamma(1)$$
$$\Gamma(3) = \int_0^\infty p^2e^{-p}dp = -p^2e^{-p}\Big|_0^\infty + \int 2 \cdot p^1 e^{-p}dp = 0 + 2\Gamma(2)$$
$$\Gamma(x) = \begin{cases} 1 & , x=1,2 \\ (x-1)\Gamma(x-1) = (x-1)! & , \text{otherwise} \end{cases}$$

性質 ② $\int_0^\infty p^{a-1}(1-p)^{b-1}dx = \beta(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$

Pf: $\Gamma(a)\Gamma(b) = \int_0^\infty x^{a-1}e^{-x}dx \int_0^\infty y^{b-1}e^{-y}dy$
$$= \int_0^\infty\int_0^\infty e^{-(x+y)}x^{a-1}y^{b-1}dx\,dy$$

令 $x = uv$, $y = u(1-v)$ ⇒ $u = x+y$, $v = \frac{x}{x+y}$ 且 $x,y>0$ ∴ $0<u$, $0<v<1$

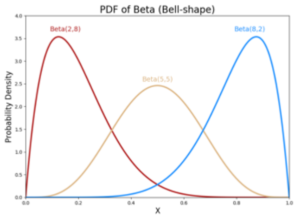
$$\Rightarrow \int_0^1\int_0^\infty e^{-u}u^{a+b-2}v^{a-1}(1-v)^{b-1}\underbrace{\left\| \begin{matrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{matrix} \right\|}_{=u} du\,dv$$
$$= \int_0^1\int_0^\infty e^{-u}u^{a+b-1}v^{a-1}(1-v)^{b-1}du\,dv$$
$$= \int_0^1 v^{a-1}(1-v)^{b-1}dv \cdot \int_0^\infty e^{-u}u^{a+b-1}du$$
$$= \beta(a,b) \cdot \Gamma(a+b)$$

$$\therefore \beta(a+b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$$

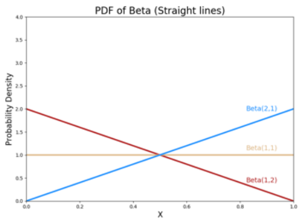
L05： 補充

The PDF of Beta distribution can be U-shaped with asymptotic ends, bell-shaped, strictly increasing/decreasing or even straight lines. As you change α or β, the shape of the distribution changes.

a. Bell-shape

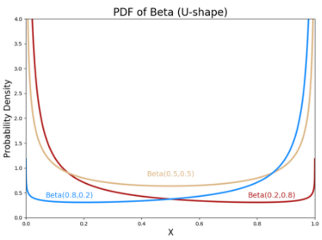


b. Straight Lines



c. U-shape

When α <1, β<1, the PDF of the Beta is U-shaped.



L05-3 : Beta-Binomial conjugate

Beta Binomial conjugate

- likelihood: $\underbrace{P(X=m \mid N,P)}_{\text{Binomial}} = \binom{N}{m} p^m (1-p)^{N-m}$

  Prior: $\underbrace{B(\theta \mid a,b)}_{\text{Beta}} = \theta^{a-1}(1-\theta)^{b-1} \underbrace{\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}}_{=\frac{1}{\beta(a,b)}}$ ，成功a次 trial a+b次

- Posterior $= \frac{\text{likelihood} \cdot \text{prior}}{\text{marginal}} = B(P \mid a+m, b+N-m) = p^{m+a-1}(1-p)^{N-m+b-1} \frac{\Gamma(a+b+N)}{\Gamma(m+a)\Gamma(N-m+b)}$

  Pf: $\dfrac{\cancel{\binom{N}{m}} p^m (1-p)^{N-m} \overbrace{\left(p^{a-1}(1-p)^{b-1} \cancel{\frac{1}{\beta(a,b)}}\right)}^{\text{prior}}}{\int_0^1 \cancel{\binom{N}{m}} \theta^m (1-\theta)^{N-m} \theta^{a-1}(1-\theta)^{b-1} \cancel{\frac{1}{\beta(a,b)}}}$

  $= \dfrac{p^{m+a-1}(1-p)^{N-m+b-1}}{\int_0^1 \theta^{m+a-1}(1-\theta)^{N-m+b-1} d\theta}$ $= \frac{\Gamma(m+a)\Gamma(N-m+b)}{\Gamma(N+a+b)}$ 見gamma 性質②

  $= p^{m+a-1}(1-p)^{N-m+b-1} \dfrac{1}{\beta(a+m, b+N-m)}$

  $= \beta(P, a+m, b+N-m)$

註: review posterior

- 应用: online (sequential) learning

  △ prior $\xrightarrow{\text{caculate}}$ Posterior

  △ 例

  5 success 6 fail → +1 success 2 fail → + 1000 success

△ 优点

得到 distribution 而非 point estimation

一 可得 mean, var

二 可從 mode 得 most likely θ

- Frequentist v.s. Bayesian

Frequentist

Predict based on Data

算 Likelihood function

⇒ MLE to the $\hat{\theta}_{MLE}$

Bayesian

Predict Based on prior ⇒ 資料少也能利用 prior (knowledge) 預測

利用 posterior distribution

⇒ MLE tothe $\underbrace{\hat{\theta}_{MAP}}_{\text{posterior}}$

$\hat{\theta}_{MAP} = \dfrac{\text{likelihood} \cdot \text{prior}}{\text{marginal}} = \dfrac{\binom{N}{m} p^m (1-p)^{N-m} \cdot \frac{1}{\beta(a,b)} p^{a-1}(1-p)^{b-1}}{\text{marginal}}$

$= \beta(p, \overbrace{a+m}^{a'}, \overbrace{b+N-m}^{b'})$ ⇒  distribution

⇒ 算 MLE of $\hat{\theta}_{MAP}$

$\dfrac{dB(p,a',b')}{dp} = 0$

$\Rightarrow \dfrac{1}{\beta(a',b')} \dfrac{d\left[p^{a'-1}(1-p)^{b'-1}\right]}{dp} = \dfrac{1}{\beta(a',b')}\left[p^{a'-2}(1-p)^{b'-1} - p^{a'-1}(1-p)^{b'-2}\right] = 0$

$\Rightarrow p^{a'-2}(1-p)^{b'-2}(1-2p) = 0$

$\Rightarrow p = \frac{1}{2}$ (?)

L05-4 : Multinomial

- Multi nomial
  - Dice 100 times for example :

$$X = [m_1, m_2, \ldots m_6] = [30, 40, \ldots]$$

$$M(x) = \binom{N}{m_1 m_2 \ldots m_i} \prod_i P_i^{m_i}$$

$$= \frac{N!}{m_1! m_2! \ldots m_i!}$$

  - Dirchlet distribution

likelihood	Model params	Conjugate Prior distribution	Posterior hyperparams	Interpretation of hyperparams	Posterior prediction
Bernoulli	P (probability)	Beta	α, β	$\alpha + \sum x_i$, $\beta + n - \sum x_i$	
Binomial	P	Beta	α, β	$\alpha + \sum x_i$, $\beta + \sum N_i - \sum x_i$	
Multi nomial	P	Dirchlet			

L06 : Gaussian distribution

Gaussin Distribution

- Univariate / Multivariate Gaussin

- Gaussin Integral : $\int_{-\infty}^{\infty} e^{-x^2} dx = \sqrt{\pi}$

Pf : $\int_{-\infty}^{\infty} e^{-x^2} dx$

$= \sqrt{\left(\int_{-\infty}^{\infty} e^{-x^2} dx\right)^2}$

$= \sqrt{\left(\int_{-\infty}^{\infty} e^{-x^2} dx\right)\left(\int_{-\infty}^{\infty} e^{-y^2} dy\right)}$

$= \sqrt{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-(x^2+y^2)} dx\, dy}$

$= \sqrt{\int_0^{2\pi}\int_0^{\infty} e^{-r^2} (r\, dr\, d\theta)}$

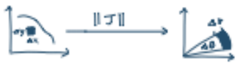
$= \sqrt{2\pi \int_0^{\infty} r e^{-r^2} dr} = \sqrt{2\pi \left(\frac{1}{2} e^{-r^2}\Big|_{\infty}^{0}\right)}$

$= \sqrt{\pi}$

註①: Polor coordinate 極座標

面積 $= \int_0^{2\pi}\int_0^{2} r\, dr\, d\theta = 2\pi \frac{r^2}{2}\Big|_0^2 = 4\pi$

註②: Jacobian determinant



$J = \det\left(\begin{bmatrix} \frac{\partial x}{\partial r} & \frac{\partial x}{\partial \theta} \\ \frac{\partial y}{\partial r} & \frac{\partial y}{\partial \theta} \end{bmatrix}\right)$

$= \det\left(\begin{bmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{bmatrix}\right) = r$

原理:

① 令 $B = [\vec{e}_1, \vec{e}_2]$ 的區塊為 $S\{x\vec{e}_1 + y\vec{e}_2 \mid 0 \le x, y \le 1\}$

經線性轉換 T 使 $A\vec{e}_i = \vec{a}_i$ 故 $T(x\vec{e}_1 + y\vec{e}_2) = A(x\vec{e}_1 + y\vec{e}_2) = x\vec{a}_1 + y\vec{a}_2$

故面積 $= \det\left(\begin{bmatrix} a_1 \\ a_2 \end{bmatrix}\right) = \det\left(\begin{bmatrix} A\vec{e}_1 \\ A\vec{e}_2 \end{bmatrix}\right) = \det((AB)^T) = \det A \det B$

② $\begin{bmatrix} du \\ 0 \end{bmatrix}$、$\begin{bmatrix} 0 \\ dv \end{bmatrix} = \begin{bmatrix} \frac{\partial x}{\partial u} du \\ \frac{\partial y}{\partial u} du \end{bmatrix}$、$\begin{bmatrix} \frac{\partial x}{\partial v} dv \\ \frac{\partial y}{\partial v} dv \end{bmatrix} \xrightarrow[\text{面積}]{\text{張開}} \det\left(\begin{bmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{bmatrix}\right)(du\, dv) = \det(J(u,v))\, du\, dv$

- Gaussian distribution 推導

△ Step 1 求 P(x)

$P(x,y) = P(x)P(y) = g(r)$

$\Rightarrow \frac{dg(r)}{d\theta} = \frac{d(P(x)P(y))}{d\theta}$

$\Rightarrow 0 = \frac{dP(x)}{d\theta} P(y) + \frac{dP(y)}{d\theta} P(x)$, $\begin{cases} x = r\cos\theta \Rightarrow \frac{dx}{d\theta} = -r\sin\theta = -y \\ y = r\sin\theta \Rightarrow \frac{dy}{d\theta} = r\cos\theta = x \end{cases}$

$\Rightarrow 0 = \frac{dP(x)}{dx} \overset{-y}{\cancel{\frac{dx}{d\theta}}} P(y) + \frac{dP(y)}{dy} \overset{x}{\cancel{\frac{dy}{d\theta}}} P(x)$

$\Rightarrow xP(x)\frac{dx}{dP(x)} = yP(y)\frac{dy}{dP(y)} = C$ ∵ 若 $x \neq y$ 等式成立 則應為常數.

$\Rightarrow \begin{cases} \frac{1}{c}x = \frac{dP(x)}{P(x)}\frac{1}{dx} \\ \frac{1}{c}y = \frac{dP(y)}{P(y)}\frac{1}{dy} \end{cases} \Rightarrow \begin{cases} \int \frac{1}{c} x\, dx = \int \frac{1}{P(x)} dP(x) \cancel{\frac{1}{dx} dx} = \ln P(x) \\ \int \frac{1}{c} y\, dy = \int \frac{1}{P(y)} dP(y) \cancel{\frac{1}{dy} dy} = \ln P(y) \end{cases}$

$\Rightarrow \begin{cases} \frac{2x^2}{c} = \ln P(x) \\ \frac{2y^2}{c} = \ln P(y) \end{cases} \Rightarrow \begin{cases} P(x) = e^{\overset{<0}{\frac{2}{c}} x^2} = e^{-kx^2} \\ P(y) = e^{\frac{2}{c} y^2} = e^{-ky^2} \end{cases}$

△ Step 2 : Normalize $P(x) \rightarrow P(x)' = AP(x)$

已知 $A\int_{-\infty}^{\infty} e^{-kx^2} = 1$, $A\int_{-\infty}^{\infty} e^{-ky^2} dy = 1$

$\Rightarrow \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-k(x^2+y^2)} dx\, dy = \frac{1}{A^2} = \int_0^{2\pi}\int_0^{\infty} e^{-kr^2} r\, dr\, d\theta$

$\Rightarrow 2\pi \left(\frac{e^{-kr^2}}{-2k}\Big|_0^{\infty}\right) = 2\pi\left(0 + \frac{1}{2k}\right) = \frac{\pi}{k} = \frac{1}{A^2}$

$\Rightarrow A = \sqrt{\frac{k}{\pi}} \Rightarrow P(x)' = \sqrt{\frac{k}{\pi}} e^{-kx^2}$

△ Step 3 代入 $\mu = 0$、σ^2

$\int_{-\infty}^{\infty} x^2 \sqrt{\frac{k}{\pi}} e^{-kx^2} dx = \sigma^2$

$\Rightarrow \sqrt{\frac{k}{\pi}} \left[x \cdot \frac{e^{-2kx^2}}{-2k}\Big|_{-\infty}^{\infty} - \int_{-\infty}^{\infty} \frac{e^{-2kx^2}}{-2k} dx \right] = \sigma^2$

$\Rightarrow \sqrt{\frac{k}{\pi}} \left(0 + \frac{1}{2k} \int_{-\infty}^{\infty} e^{\cancel{-2kx^2}} dx \right) = \sigma^2$ 有誤

$\Rightarrow \frac{1}{2k} = \sigma^2 \Rightarrow k = \frac{1}{2\sigma^2}$

$P(x) = \sqrt{\frac{k}{\pi}} e^{-kx^2}$, $k = \frac{1}{2\sigma^2} \xrightarrow[\text{given}]{\mu,\ \sigma^2} \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-x^2}{2\sigma^2}} \xrightarrow[\text{平移}]{\text{在 } \mu} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} = X \sim N(\mu, \sigma^2)$

L07 : MLE on Gaussian

MLE on Gaussian 必考

D: 一組 data $x_1, x_2, \ldots x_n$

$$L(\theta=\mu,\sigma^2 \mid D) = P(D\mid\theta) = \prod_{i=1}^{n} P(x_i\mid\theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x_i-\mu)^2}{2\sigma^2}}$$

求 MLE: $\underset{\theta}{\operatorname{argmax}} \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x_i-\mu)^2}{2\sigma^2}} \equiv \underset{\theta}{\operatorname{argmax}} \underline{\sum_{i=1}^{n} \ln\left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x_i-\mu)^2}{2\sigma^2}}\right)}$

令 $L = \sum_{i=1}^{n} \ln\left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x_i-\mu)^2}{2\sigma^2}}\right) = \frac{-1}{2}\sum_{i=1}^{n} \ln(2\pi\sigma^2) + \sum_{i=1}^{n} \frac{-(x_i-\mu)^2}{2\sigma^2}$

$= \frac{-n}{2}\ln(2\pi\sigma^2) + \sum_{i=1}^{n} \frac{-(x_i-\mu)^2}{2\sigma^2}$

① $\frac{dL}{d\mu} = \frac{d}{d\mu}\sum_{i=1}^{n} -\frac{(x_i-\mu)^2}{2\sigma^2}$

$\Rightarrow \frac{d}{d\mu}\sum_{i=1}^{n}(\cancel{x_i^2} - 2\mu x_i + \mu^2) = 0$

$\Rightarrow 2\sum_{i=1}^{n} x_i = 2\sum_{i=1}^{n}\mu = 2n\mu$

$\Rightarrow \mu = \frac{\sum_{i=1}^{n} x_i}{n}$

② $\frac{dL}{ds} = \frac{d}{ds}\left(\frac{-n}{2}\ln(2\pi s) + \sum_{i=1}^{n}\frac{(x_i-\mu)^2}{2s}\right)$

令 $s=\sigma^2$

$= \left(\frac{-n}{2}\frac{2\pi}{2\pi s}\right) + \sum_{i=1}^{n}\frac{(x_i-\mu)^2}{2s^2} = 0$

$= \frac{n}{2s} = \frac{\sum(x_i-\mu)^2}{2s^2}$

$\Rightarrow s = \frac{\sum(x_i-\mu)^2}{n}$

Conjugate Prior for Gaussin

Gaussin 也有 conjugate 性質: Prior: Gaussin $\longrightarrow$ Posterior: Gaussin

令 prior: Gaussin ~ $N(\mu\mid\mu_0, \sigma_0^2)$

likelihood: Gaussin ~ $N(D, \mu, \sigma^2)$, σ^2 is fix

則 posterior: Gaussin ~ $N(\mu\mid \bigcirc\bigcirc)$?

解:

① $P(\theta\mid D) = \frac{P(D\mid\theta)P(\theta)}{P(D)} = \frac{P(D\mid\mu)P(\mu)}{P(D)}$

② $P(D\mid\mu)P(\mu) = \prod_{i=1}^{n} P(x_i\mid\mu)\, P(\mu\mid\mu_0, \sigma_0^2)$

$$= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n e^{\frac{-1}{2\sigma^2}\sum(x_i-\mu)^2} \cdot \frac{1}{\sqrt{2\pi\sigma_0^2}} e^{-\frac{1}{2\sigma_0^2}(\mu-\mu_0)^2}$$

$$= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^{n+1} \underline{e^{\frac{-\sum(x_i-\mu)^2}{2\sigma^2} + \frac{-(\mu-\mu_0)^2}{2\sigma_0^2}}}$$

→ 目標: 符合 Gaussin form

$A e^{\frac{-(x-\mu)^2}{B}} = e^{\frac{-(x-\mu)^2}{B} + c}$

③ $$\frac{-\sum(x_i-\mu)^2}{2\sigma^2}+\frac{-(\mu-\mu_0)^2}{2\sigma_0^2}=\frac{-1}{2}\left[\frac{\sum x_i^2}{\sigma^2}-\frac{2\mu\sum x_i}{\sigma^2}+\frac{\overset{n\mu^2}{\cancel{\sum \mu^2}}}{\sigma^2}+\frac{\mu^2}{\sigma_0^2}-\frac{2\mu\mu_0}{\sigma_0^2}+\frac{\mu_0^2}{\sigma_0^2}\right]$$

假設：

$k' = \left(\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}\right)$ $=\sigma_n^2$

$\mu_n = \frac{\left(\frac{\sum x_i}{\sigma^2}+\frac{\mu_0}{\sigma_0^2}\right)}{k'}$

$k = \frac{k'}{2}$

$$=\frac{-1}{2}\left[\mu^2\left(\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}\right)-2\mu\left(\frac{\sum x_i}{\sigma^2}+\frac{\mu_0}{\sigma_0^2}\right)+\left(\frac{\sum x_i^2}{\sigma^2}+\frac{\mu_0^2}{\sigma_0}\right)\right]$$

$$=\frac{-k'}{2}\left[\mu^2-2\mu\mu_n+\mu_n^2-\mu_n^2+\left(\frac{\frac{\sum x_i^2}{\sigma^2}+\frac{\mu_0^2}{\sigma_0^2}}{k'}\right)\right]$$

$$=-k\left[(\mu-\mu_n)^2\right]+C\ ,\quad C=-k\left(\mu_n^2-\frac{\frac{\sum x_i^2}{\sigma^2}+\frac{\mu_0^2}{\sigma_0^2}}{k'}\right)$$

$$\therefore\ e^{\frac{-\sum(x_i-\mu)^2}{2\sigma^2}+\frac{-(\mu-\mu_0)^2}{2\sigma_0^2}}=e^{-k(\mu-\mu_n)^2+c}=Ae^{-k(\mu-\mu_n)^2},\ A=e^c$$

④ Marginal $= P(D)=\int_{-\infty}^{\infty}P(D|\mu')P(\mu')d\mu'$

$$=\int_{-\infty}^{\infty}Ae^{-k(\mu'-\mu_n)^2}d\mu' = \int_{-\infty}^{\infty}Ae^{-k\mu'^2}d\mu'$$

$$=A\sqrt{\frac{\pi}{k}}$$

③④ $\Rightarrow$ $$\frac{P(D|\mu)P(\mu)}{P(D)}=\frac{Ae^{-k(\mu-\mu_n)^2}}{A\sqrt{\frac{\pi}{k}}}=\frac{e^{\frac{-k'}{2}(\mu-\mu_n)^2}}{\sqrt{2\pi}\sqrt{\frac{1}{k'}}}$$

$$=\frac{e^{\frac{-1}{2\sigma_n^2}(\mu-\mu_n)^2}}{\sqrt{2\pi\sigma_n^2}}=N\sim(\mu_n,\sigma_n^2)$$

假設 $\sigma_n^2=\frac{1}{k'}=\frac{1}{\left(\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}\right)}$

已知 $\mu_n=\sigma_n^2\left(\frac{\sum x_i}{\sigma^2}+\frac{\mu_0}{\sigma_0^2}\right)$

$$=\frac{\sigma_n^2 n}{\sigma^2}\mu_{MLE}+\frac{\sigma_n^2}{\sigma_0^2}(\mu_0)$$

$\Rightarrow$ 介於 μ_{MSE} 和 μ_0 間

$\because\ \frac{\sigma_n^2 n}{\sigma^2}+\frac{\sigma_n^2}{\sigma_0^2}=1$

μ_{MSE} current data　μ_n posterior　μ_0 prior

- 分析

$$P(\theta|D)=N\sim(\mu_n,\sigma_n^2)=\frac{e^{\frac{-1}{2\sigma^2}(\mu-\mu_n)^2}}{\sqrt{2\pi\sigma_n^2}}$$

其中 $\sigma_n^2=\frac{1}{\left(\frac{n}{\sigma_n^2}+\frac{1}{\sigma_0^2}\right)}$, $\mu_n=\sigma_n^2\left(\frac{n\mu_{MLE}}{\sigma^2}+\frac{\mu_0}{\sigma_0^2}\right)$

$\therefore$ 當 $n\to 0$ ： $\sigma_n^2=\frac{1}{0+\frac{1}{\sigma_0^2}}=\sigma_0^2$, $\mu_n=\sigma_0^2\left(0+\frac{\mu_0}{\sigma_0^2}\right)=\mu_0$　資料量=0時 $\rightarrow N\sim(\mu_0,\sigma_0^2)$

當 $n\to\infty$ ： $\sigma_n^2=0$, $\mu_n=\left(\frac{1}{\frac{n}{\sigma^2}+\cancel{\frac{1}{\sigma_0^2}}}\right)\left(\frac{n\mu_{MLE}}{\sigma_2}+\cancel{\frac{\mu_0}{\sigma_0^2}}\right)=\mu_{MLE}=\frac{\sum x_i}{n}$　資料量和 sample space 一樣大 $\Rightarrow$ 不需要机率了