unsupervised learning

complete data

先看 complete 怎麼做，我們再來看 incomplete data 要如何處理

one coin tossing

$$X:\{H,H,T,T,H\} \sim Bernoulli\left(\begin{matrix}H\\T\end{matrix}\middle| P\right)$$

$$P_{MLE}=\frac{\sum_{i=1}^{N} x_i}{N}$$

two coin tossing

有兩種不同的銅板，假設為 C_0 及 C_1，我們並不知道兩個銅板擲出正面的機率(P_0 及 P_1)

我們擲了銅板十次，我們也知道每次投擲的是哪個銅板(supervise learning)
假設結果如下

$$X:\{H,T,\ T,H,\ H,T,\ H,H,\ T,\ T\}$$
$$Z:\{C_0,C_1,C_0,C_0,C_1,C_1,C_0,C_0,C_1,C_0\}$$

我們想要知道 P_0 及 P_1 之外，我們還想知道 C_0 出現的機率 λ

$$P(X\mid P_0,P_1,Z,\lambda)$$
$$=\prod_{i=1}^{n}(\lambda {P_0}^{x_i}(1-P_0)^{1-x_i})^{1-z_i}\cdot((1-\lambda){P_1}^{x_i}(1-P_1)^{1-x_i})^{z_i}$$

第 i 次事件 x_i 是 C_0 的機率

第 i 次事件 x_i 是 C_1 的機率

若是套用在這個例子上

$$P(X\mid P_0,P_1,Z,\lambda)$$
$$=\lambda P_0\cdot(1-\lambda)P_1\cdot\lambda(1-p_0)...$$

我們要找 MLE 時各個參數為何，需要做微分，故一樣做法，我們還是先取 log

$$J = \log P(X \mid P_0, P_1, Z, \lambda)$$

$$= \sum_{i=1}^{n} \log(\lambda P_0^{x_i}(1-P_0)^{1-x_i})^{1-z_i} + \sum_{i=1}^{n} \log((1-\lambda)P_1^{x_i}(1-P_1)^{1-x_i})^{z_i}$$

$$= \sum_{i=1}^{n}(1-z_i)(\log\lambda + x_i \log P_0 + (1-x_i)\log(1-P_0)) + \sum_{i=1}^{n} z_i(\log(1-\lambda) + x_i \log P_1 + (1-x_i)\log(1-P_1))$$

先來看 MLE 時的 λ

$$\frac{\partial J}{\partial \lambda} = \sum_{i=1}^{n}(1-z_i)\frac{1}{\lambda} + \sum_{i=1}^{n} z_i \frac{-1}{1-\lambda} = 0$$

$$\Rightarrow \sum_{i=1}^{n}(1-z_i)\frac{1}{\lambda} = \sum_{i=1}^{n} z_i \frac{1}{1-\lambda}$$

$$\Rightarrow (1-\lambda)\sum_{i=1}^{n}(1-z_i) = \lambda \sum_{i=1}^{n} z_i$$

$$\Rightarrow \sum_{i=1}^{n}(1-z_i) = \lambda \sum_{i=1}^{n} z_i + \lambda \sum_{i=1}^{n}(1-z_i) = \lambda \sum_{i=1}^{n} 1 = n\lambda$$

$$\Rightarrow \lambda = \frac{\sum_{i=1}^{n}(1-z_i)}{n}$$

1-z_i 為出現 coin 0 的事件，則 $\lambda_{MLE} = \frac{coin\,0\text{出現次數}}{\text{總試驗次數}}$，蠻合乎我們的常理的

再來看 MLE 時的 P_0

$$\frac{\partial J}{\partial P_0} = \sum_{i=1}^{n}(1-z_i)(x_i \frac{1}{P_0} + (1-x_i)\frac{-1}{1-P_0}) = 0$$

$$\Rightarrow \sum_{i=1}^{n}(1-z_i)\frac{x_i}{P_0} = \sum_{i=1}^{n}(1-z_i)\frac{1-x_i}{1-P_0}$$

$$\Rightarrow (1-P_0)\sum_{i=1}^{n}(1-z_i)x_i = P_0 \sum_{i=1}^{n}(1-z_i)(1-x_i)$$

$$\Rightarrow \sum_{i=1}^{n}(1-z_i)x_i = P_0 \sum_{i=1}^{n}(1-z_i)x_i + P_0 \sum_{i=1}^{n}(1-z_i)(1-x_i) = P_0 \sum_{i=1}^{n}(1-z_i)$$

$$\Rightarrow P_0 = \frac{\sum_{i=1}^{n}(1-z_i)x_i}{\sum_{i=1}^{n}(1-z_i)}$$

則 $P_{0,MLE} = \frac{coin\,0\text{出現且出現正面的次數}}{coin\,0\text{出現次數}}$，蠻合乎我們的常理的

再來看 MLE 時的 P_1

$$\frac{\partial J}{\partial P_1}=\sum_{i=1}^{n} z_i\left(x_i\frac{1}{P_1}+(1-x_i)\frac{-1}{1-P_1}\right)=0$$

$$\Rightarrow \sum_{i=1}^{n} z_i\frac{x_i}{P_1}=\sum_{i=1}^{n} z_i\frac{1-x_i}{1-P_1}$$

$$\Rightarrow (1-P_1)\sum_{i=1}^{n} z_i x_i = P_1\sum_{i=1}^{n} z_i(1-x_i)$$

$$\Rightarrow \sum_{i=1}^{n} z_i x_i = P_1\sum_{i=1}^{n} z_i x_i + P_1\sum_{i=1}^{n} z_i(1-x_i) = P_1\sum_{i=1}^{n} z_i$$

$$\Rightarrow P_{1,MLE}=\frac{\sum_{i=1}^{n} z_i x_i}{\sum_{i=1}^{n} z_i}$$

則 $P_1=\frac{coin\,1\text{出現且出現正面的次數}}{coin\,1\text{出現次數}}$，蠻合乎我們的常理的

如此，我們便得到了 MLE 時的 P_0,P_1 和 λ

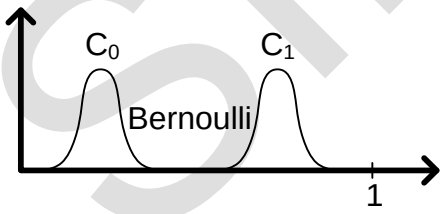
incomplete data

相較於 supervised learning，unsupervised learning 沒有提供每組 data 的 label，也就是 incomplete data，如果是一樣用兩個銅板的例子來看

$X:\{H,T,\ H,H,\ T,T,\ H,H,\ T,\ T\}$

$Z:\{C_?,C_?,C_?,C_?,C_?,C_?,C_?,C_?,C_?,C_?\}$

supervised learning 例子中公式的 z_i 我們已經無法使用，雖然我們並不知道每次擲的銅板是哪一個，不過我們至少可以知道兩個銅板正反面機率的分布應該是不太一樣的



退而求其次，我們可以求得"該次是某個銅板的機率"，假設我們將某次擲出 coin 0 的機率設為 w_0，有時會稱為 responsibility，則擲出 coin 1 的機率為 1-w_0，注意，這裡的 w_0 和 λ 定義不一樣，雖然中文的意思好像差不多。如果還是搞不清楚，可以想想原本 z_i 的定義是甚麼，我們需要先知道該次銅板是甚麼，我們才能去算出"出現該次銅板"的機率為何，也就是 λ。這裡 w_0 的意義也是一樣，我們要先知道某一次的銅板為何，也就是 w_0 表現的意義，知道這項資訊後，我們才能知道要代入哪個銅板的公式(λ or 1-λ)為何。

首先，我們就要先定義出 w_0，假設該次 outcome 為 x_i(也就是正面會反面)，要注意的是，其實 w_0 仍會受該次 outcome 為何而影響，我們想像一種情況，如果有一個銅板 X 正面機率較高，另一個 Y 較低，則如果出現正面，那擲出的是 X 的機率應該較高

$$w_{0i} = \frac{\text{出現}coin\ 0\text{且}outcome\text{為}x_i\text{的機率}}{\text{出現}coin\ 0\text{且}outcome\text{為}x_i\text{的機率} + \text{出現}coin\ 1\text{且}outcome\text{為}x_i\text{的機率}} = \frac{P(z_i = C_0, x_i \mid \theta)}{P(z_i = C_0, x_i \mid \theta) + P(z_i = C_1, x_i \mid \theta)}$$

其中

$$P(z_i = C_0, x_i \mid \theta) = \lambda P_0^{x_i} (1 - P_0)^{1 - x_i}$$
$$P(z_i = C_1, x_i \mid \theta) = (1 - \lambda) P_1^{x_i} (1 - P_1)^{1 - x_i}$$

我們往後將 w_{0i} 改為 w_i ，以便和 supervised learning 公式的 z_i 對齊

則 weighted likelihood 為

$$P(X \mid P_0, P_1, w_i, \lambda)$$
$$= \prod_{i=1}^{n} (\lambda P_0^{x_i} (1 - P_0)^{1 - x_i})^{w_i} \cdot ((1 - \lambda) P_1^{x_i} (1 - P_1)^{1 - x_i})^{1 - w_i}$$

$$J = \log P(X \mid P_0, P_1, w_i, \lambda)$$
$$= \sum_{i=1}^{n} w_i (\log \lambda + x_i \log P_0 + (1 - x_i) \log(1 - P_0)) + \sum_{i=1}^{n} (1 - w_i)(\log(1 - \lambda) + x_i \log P_1 + (1 - x_i) \log(1 - P_1))$$

和 supervised learning 的推導一模一樣，只是要將(1-z_i)改為 w_i ，z_i 改為 $(1 - w_i)$ 而已

則

$$\lambda = \frac{\sum_{i=1}^{n} w_i}{n}$$

$$P_0 = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

$$P_1 = \frac{\sum_{i=1}^{n} (1 - w_i) x_i}{\sum_{i=1}^{n} (1 - w_i)}$$

EM algorithm

先敘述該如何執行 EM algorithm，其正確性等一下再證明

初始

隨意猜測θ，也就是 P_0,P_1 和 λ

E(expectation) step

利用已知的θ，計算 responsibility w_i

M(maximization) step

利用 E step 算出的 w_i，推得該 w_i 下的 MLE P_0,P_1 和 λ

往後重複 E step 和 M step，直至參數都收斂至一定值為止

e.g.

假設事件為 $X:\{H,T,H,T,H,H,T,H,T,H\}$

初始

假設 $P_0=0.7,\ P_1=0.6,\ \lambda=0.3$

第一次疊代

E step:

出現 H 時

為 coin 0 的機率為

$$w_i=\frac{P(z_i=C_0,\ x_i=H\mid\theta)}{P(z_i=C_0,\ x_i=H\mid\theta)+P(z_i=C_1,\ x_i=H\mid\theta)}=\frac{\lambda P_0}{\lambda P_0+(1-\lambda)P_1}=\frac{0.3\times0.7}{0.3\times0.7+0.7\times0.6}=0.33$$

為 coin 1 的機率為

$1-w_i=0.67$

出現 T 時

為 coin 0 的機率為

$$w_i=\frac{P(Z_i=C_0,\ x_i=T\mid\theta)}{P(z_i=C_0,\ x_i=T\mid\theta)+P(z_i=C_1,\ x_i=T\mid\theta)}=\frac{\lambda(1-P_0)}{\lambda(1-P_0)+(1-\lambda)(1-P_1)}=\frac{0.3\times0.3}{0.3\times0.3+0.7\times0.4}=0.24$$

為 coin 1 的機率為

$1-w_i=0.76$

M step:

$$\lambda=\frac{\sum_{i=1}^{n}w_i}{n}=\frac{0.33+0.24+0.33+0.24+0.33+0.33+0.24+0.33+0.24+0.33}{10}=0.29$$

$$P_0 = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i} = \frac{0.33 \times 6}{2.94(\lambda\text{已算過})} = 0.67$$

$$P_1 = \frac{\sum_{i=1}^{n} (1-w_i) x_i}{\sum_{i=1}^{n} (1-w_i)} = \frac{0.67 \times 6}{0.67 \times 6 + 0.76 \times 4} = 0.57$$

第二次疊代

E step:

出現 H 時

為 coin 0 的機率為

$$w_i = \frac{P(z_i = C_0, x_i = H \mid \theta)}{P(z_i = C_0, x_i = H \mid \theta) + P(z_i = C_1, x_i = H \mid \theta)} = \frac{\lambda P_0}{\lambda P_0 + (1-\lambda) P_1} = \frac{0.29 \times 0.67}{0.29 \times 0.67 + 0.71 \times 0.57} = 0.32$$

為 coin 1 的機率為

$1 - w_i = 0.68$

出現 T 時

為 coin 0 的機率為

$$w_i = \frac{P(z_i = C_0, x_i = T \mid \theta)}{P(z_i = C_0, x_i = T \mid \theta) + P(z_i = C_1, x_i = T \mid \theta)} = \frac{\lambda(1-P_0)}{\lambda(1-P_0) + (1-\lambda)(1-P_1)} = \frac{0.29 \times 0.33}{0.29 \times 0.33 + 0.71 \times 0.43} = 0.24$$

為 coin 1 的機率為

$1 - w_i = 0.76$

M step:

$$\lambda = \frac{\sum_{i=1}^{n} w_i}{n} = \frac{0.32 + 0.24 + 0.32 + 0.24 + 0.32 + 0.32 + 0.24 + 0.32 + 0.24 + 0.32}{10} = 0.29$$

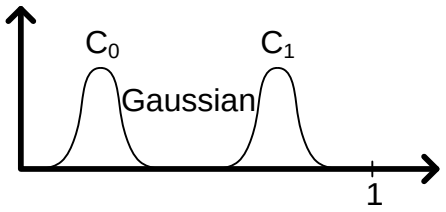
$$P_0 = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i} = \frac{0.32 \times 6}{2.88(\lambda\text{已算過})} = 0.67$$

$$P_1 = \frac{\sum_{i=1}^{n} (1-w_i) x_i}{\sum_{i=1}^{n} (1-w_i)} = \frac{0.68 \times 6}{0.68 \times 6 + 0.76 \times 4} = 0.57$$

P_0,P_1 和 λ 已不太變化，收斂，此 P_0,P_1 和 λ 就是 MLE 的 P_0,P_1 和 λ

Gaussian mixture model

前面的推導的分布都是 Bernoulli，我們可以改為 Gaussian distribution



若有兩類，我們就需要五個參數 $\lambda, \mu_0, \mu_1, \sigma_0 (or\ \Sigma_0), \sigma_1 (or\ \Sigma_1)$

Special case (univariate, variance=1)

complete data

$$P(X \mid \lambda, \mu_0, \mu_1, z)$$

$$= \prod_{i=1}^{n} (\lambda \frac{1}{\sqrt{2\pi}} e^{-(x_i-\mu_0)^2})^{1-z_i} \cdot ((1-\lambda) \frac{1}{\sqrt{2\pi}} e^{-(x_i-\mu_1)^2})^{z_i}$$

$$J = \log(P(X \mid \lambda, \mu_0, \mu_1))$$

$$= \sum_{i=1}^{n} (1-z_i)(\log \frac{1}{\sqrt{2\pi}} + \log \lambda - (x_i-\mu_0)^2) + \sum_{i=1}^{n} z_i (\log \frac{1}{\sqrt{2\pi}} + \log(1-\lambda) - (x_i-\mu_1)^2)$$

推導 λ_{MLE} 的公式和前面 complete data 的推導一模一樣

$$\frac{\partial J}{\partial \lambda} = \sum_{i=1}^{n} (1-z_i) \frac{1}{\lambda} + \sum_{i=1}^{n} z_i \frac{-1}{1-\lambda} = 0$$

$$\Rightarrow \sum_{i=1}^{n} (1-z_i) \frac{1}{\lambda} = \sum_{i=1}^{n} z_i \frac{1}{1-\lambda}$$

$$\Rightarrow (1-\lambda) \sum_{i=1}^{n} (1-z_i) = \lambda \sum_{i=1}^{n} z_i$$

$$\Rightarrow \sum_{i=1}^{n} (1-z_i) = \lambda \sum_{i=1}^{n} z_i + \lambda \sum_{i=1}^{n} (1-z_i) = \lambda \sum_{i=1}^{n} 1 = n\lambda$$

$$\Rightarrow \lambda_{MLE} = \frac{\sum_{i=1}^{n} (1-z_i)}{n}$$

$$\frac{\partial J}{\partial \mu_0}$$

$$= \sum_{i=1}^{n} (1-z_i)(x_i - \mu_0) = 0$$

$$\Rightarrow \sum_{i=1}^{n} (1-z_i) x_i - \sum_{i=1}^{n} (1-z_i) \mu_0 = 0$$

$$\Rightarrow \sum_{i=1}^{n}(1-z_i)\mu_0 = \sum_{i=1}^{n}(1-z_i)x_i$$

$$\Rightarrow \mu_{0,MLE} = \frac{\sum_{i=1}^{n}(1-z_i)x_i}{\sum_{i=1}^{n}(1-z_i)}$$

$$\frac{\partial J}{\partial \mu_1}$$

$$= \sum_{i=1}^{n} z_i(x_i - \mu_1)$$

$$\Rightarrow \sum_{i=1}^{n} z_i x_i - \sum_{i=1}^{n} z_i \mu_1 = 0$$

$$\Rightarrow \mu_1 \sum_{i=1}^{n} z_i = \sum_{i=1}^{n} z_i x_i$$

$$\Rightarrow \mu_{1,MLE} = \frac{\sum_{i=1}^{n} z_i x_i}{\sum_{i=1}^{n} z_i}$$

General GMM

complete data

$$P(X \mid \lambda, \mu_0, \mu_1, \sigma_0, \sigma_1, z)$$

$$= \prod_{i=1}^{n} (\lambda \frac{1}{\sqrt{2\pi}\sigma_0^{\,2}} e^{\frac{-(x_i-\mu_0)^2}{2\sigma_0^{\,2}}})^{1-z_i} \cdot ((1-\lambda)\frac{1}{\sqrt{2\pi}\sigma_1^{\,2}} e^{\frac{-(x_i-\mu_1)^2}{2\sigma_1^{\,2}}})^{z_i}$$

$$J = \log(P(X \mid \lambda, \mu_0, \mu_1, \sigma_0, \sigma_1))$$

$$= \sum_{i=1}^{n}(1-z_i)(\log\lambda - \log(\sqrt{2\pi}\sigma_0^{\,2}) + \frac{-(x_i-\mu_0)^2}{2\sigma_0^{\,2}}) + \sum_{i=1}^{n} z_i(\log(1-\lambda) - \log(\sqrt{2\pi}\sigma_1^{\,2}) + \frac{-(x_i-\mu_1)^2}{2\sigma_1^{\,2}})$$

其實都在做類似的事情，微分=0 找 MLE

$$\frac{\partial J}{\partial \lambda} = \sum_{i=1}^{n}(1-z_i)\frac{1}{\lambda} + \sum_{i=1}^{n} z_i \frac{-1}{1-\lambda} = 0$$

$$\Rightarrow \sum_{i=1}^{n}(1-z_i)\frac{1}{\lambda} = \sum_{i=1}^{n} z_i \frac{1}{1-\lambda}$$

$$\Rightarrow (1-\lambda)\sum_{i=1}^{n}(1-z_i) = \lambda \sum_{i=1}^{n} z_i$$

$$\Rightarrow \sum_{i=1}^{n}(1-z_i) = \lambda \sum_{i=1}^{n} z_i + \lambda \sum_{i=1}^{n}(1-z_i) = \lambda \sum_{i=1}^{n} 1 = n\lambda$$

$$\Rightarrow \lambda_{MLE} = \frac{\sum_{i=1}^{n}(1-z_i)}{n}$$

$$\frac{\partial J}{\partial \mu_0}$$

$$= \frac{1}{2\sigma_0^{\,2}} \sum_{i=1}^{n}(1-z_i)(x_i - \mu_0) = 0$$

$$\Rightarrow \sum_{i=1}^{n}(1-z_i)x_i - \sum_{i=1}^{n}(1-z_i)\mu_0 = 0$$

$$\Rightarrow \sum_{i=1}^{n}(1-z_i)\mu_0 = \sum_{i=1}^{n}(1-z_i)x_i$$

$$\Rightarrow \mu_{0,MLE} = \frac{\sum_{i=1}^{n}(1-z_i)x_i}{\sum_{i=1}^{n}(1-z_i)}$$

$$\frac{\partial J}{\partial \mu_1}$$

$$= \frac{1}{2\sigma_1^{\,2}} \sum_{i=1}^{n} z_i(x_i - \mu_1)$$

$$\Rightarrow \sum_{i=1}^{n} z_i x_i - \sum_{i=1}^{n} z_i \mu_1 = 0$$

$$\Rightarrow \mu_1 \sum_{i=1}^{n} z_i = \sum_{i=1}^{n} z_i x_i$$

$$\Rightarrow \mu_{1,MLE} = \frac{\sum_{i=1}^{n} z_i x_i}{\sum_{i=1}^{n} z_i}$$

$$\frac{\partial J}{\partial \sigma_0}$$

$$= \sum_{i=1}^{n}(1-z_i)\left(\frac{-2\sqrt{2\pi}\sigma_0}{\sqrt{2\pi}\sigma_0^{\,2}} + \frac{(x_i-\mu_0)^2}{\sigma_0^{\,3}}\right) = 0$$

$$\Rightarrow \sum_{i=1}^{n}(1-z_i)\frac{2}{\sigma_0} = \sum_{i=1}^{n}(1-z_i)\frac{(x_i-\mu_0)^2}{\sigma_0^{\,3}}$$

$$\Rightarrow 2\sigma_0^{\,2} \sum_{i=1}^{n}(1-z_i) = \sum_{i=1}^{n}(1-z_i)(x_i-\mu_0)^2$$

$$\Rightarrow {\sigma_{0,MLE}}^2 = \frac{\sum_{i=1}^{n}(1-z_i)(x_i-\mu_0)^2}{2\sum_{i=1}^{n}(1-z_i)}$$

$$\frac{\partial J}{\partial \sigma_1}$$

$$= \sum_{i=1}^{n} z_i \left(\frac{-2\sqrt{2\pi}\sigma_1}{\sqrt{2\pi}{\sigma_1}^2} + \frac{(x_i-\mu_1)^2}{{\sigma_1}^3}\right) = 0$$

$$\Rightarrow \sum_{i=1}^{n} z_i \frac{2}{\sigma_1} = \sum_{i=1}^{n} z_i \frac{(x_i-\mu_1)^2}{{\sigma_1}^3}$$

$$\Rightarrow 2{\sigma_1}^2 \sum_{i=1}^{n} z_i = \sum_{i=1}^{n} z_i (x_i-\mu_1)^2$$

$$\Rightarrow {\sigma_{1,MLE}}^2 = \frac{\sum_{i=1}^{n} z_i (x_i-\mu_1)^2}{2\sum_{i=1}^{n} z_i}$$

若為多維的 multivariate Gaussian distribution，只需更改$\Sigma_{0,MLE}$及$\Sigma_{1,MLE}$

$${\Sigma_{0,MLE}}^2 = \frac{\sum_{i=1}^{n}(1-z_i)(x_i-\mu_0)(x_i-\mu_0)^T}{\sum_{i=1}^{n}(1-z_i)}$$

$${\Sigma_{1,MLE}}^2 = \frac{\sum_{i=1}^{n} z_i (x_i-\mu_1)(x_i-\mu_1)^T}{\sum_{i=1}^{n} z_i}$$

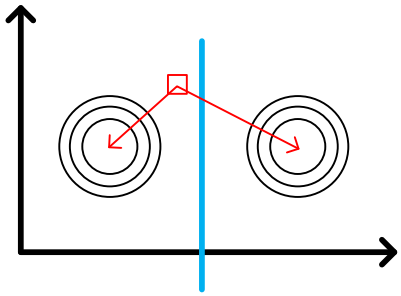
incomplete data

$$w_i = \frac{\lambda \frac{1}{\sqrt{2\pi}{\sigma_0}^2} e^{\frac{-(x_i-\mu_0)^2}{2{\sigma_0}^2}}}{\lambda \frac{1}{\sqrt{2\pi}{\sigma_0}^2} e^{\frac{-(x_i-\mu_0)^2}{2{\sigma_0}^2}} + (1-\lambda)\frac{1}{\sqrt{2\pi}{\sigma_1}^2} e^{\frac{-(x_i-\mu_1)^2}{2{\sigma_1}^2}}}$$

其他的做法和 EM algorithm 一樣，照舊就好

k-means clustering

在 incomplete data 時，沒有像 EM algorithm 的 w_i 的模糊空間，會定一個閾值(threshold)，當 w_i 的機率大於閾值時，outcome 設為 0，若小於閾值時，outcome 設為 1。



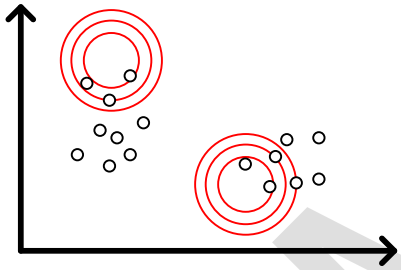
像是上圖，方框 data 就會屬於 C_0，不會有像是 60%機率是 C_0，40%機率是 C_1 這種事

此後，我們就可以用 complete data 的方式來求解，每次得到 w 時就轉為 z，一直重複疊代。

稍為圖示一下該怎麼做 k-mean clustering

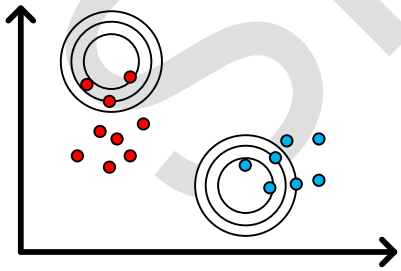
Step1

如同 EM algorithm 的初始化，假設一兩分類之分布，但必須是 isotropic Gaussian distribution(通常只給 mean)



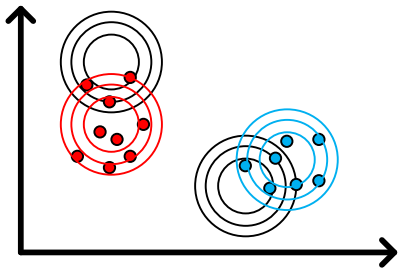
Step2

計算每個 data 和兩個 cluster mean 的距離，哪個較近就歸哪一類。



Step3

如此就得到了 Z，就可以做 complete data 的 EM algorithm 的 M step，可得到一新分布(只需要 mean)



如此一直重複 step2 和 step3，直至收斂為止

K-means clustering 其實是簡化版的 GMM，因為增加了許多限制和前提

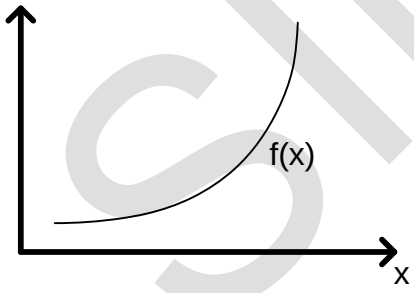
1. 不計算 w，以 activation function 將 w 轉為 z(但其實只需算距離，根本不需要算 w)
2. 分布的假設永遠都是 isotropic Gaussian distribution，所以只需要算距離就可以，Gaussian distribution 中的 $e^{\frac{-(x_i-\mu)^2}{2\sigma^2}}$ 只需要知道指數項中誰較大，就代表誰的距離較遠，直線距離越遠，機率一定越低

k-means 雖然計算較簡單，但是其限制較多，也無法模擬 skew 的 Gaussian distribution，能用 k-means 時，都一定可以用 GMM，故如果可以，通常都使用 GMM。

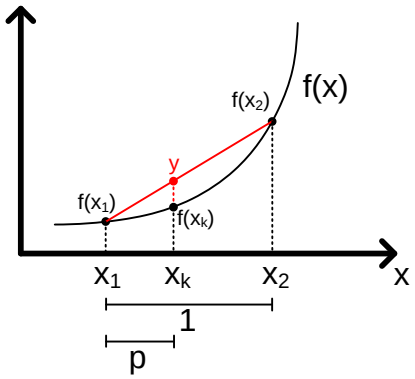
Why EM works?

EM 和相關的演算法都已經講完了，接下來就要證明為何隨機假設及每次疊代，都能得到更好的結果，直到收斂一定值。

前情提要(Jensen's inequality)
若一函數是凸函數(convex function)，像是下圖



取函數上兩點 x_1, x_2，並取兩點間一點 x_k，令 $p=\frac{x_k - x_1}{x_2 - x_1}$



則 $x_k = px_1 + (1-p)x_2$

Jensen's inequality:

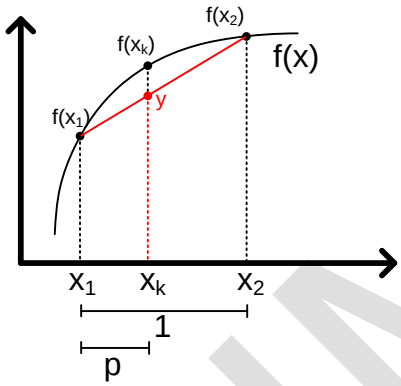
$y = pf(x_1) + (1-p)f(x_2) \geq f(px_1 + (1-p)x_2)$

若用多個點表達 xk

$y = \sum_i p_i f(x_i) = E(f(x)) \geq f(\sum_i p_i x_i) = f(E(x))$

$\Rightarrow E(f(x)) \geq f(E(x))$

若是凹函數(concave function)，符號就會相反



Jensen's inequality:

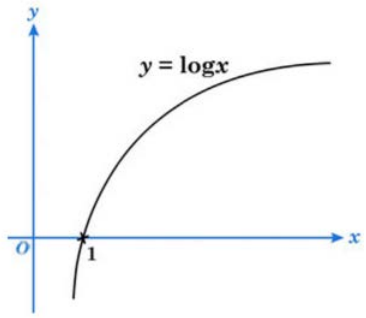
$y = pf(x_1) + (1-p)f(x_2) \leq f(px_1 + (1-p)x_2)$

$E(f(x)) \leq f(E(x))$

有了工具後，我們就能來驗證 EM algorithm for incomplete data 的正確性了

首先，先確定目標，也就是找出 $\arg\max_{\theta} P(X \mid \theta)$

而為找極大值，我們會使用 log，再微分=0，而 log 是一個 concave function，先記著就好



圖片來源：http://developer.hanluninfo.com:8088/2005/hkcee/math/other/summary_08.htm

$$\log(P(X \mid \theta))$$
$$= \log \sum_i P(X, z_i \mid \theta)$$

因為我們沒有任何 Z(label)的資訊，推到這一步就推不下去了，但我們可以用一技巧繼續下去，我們假設一個未知的分布 q(z)

$$\log(P(X \mid \theta))$$
$$= \log \sum_i P(X, z_i \mid \theta)$$
$$= \log (\sum_i q(z_i) \frac{P(X, z_i \mid \theta)}{q(z_i)})$$
$$= \log(E_q(\frac{P(X, z_i \mid \theta)}{q(z_i)}))$$

這裡的 $E_q(\frac{P(X, z_i \mid \theta)}{q(z_i)})$ 指的是將 $\frac{P(X, z_i \mid \theta)}{q(z_i)}$ 代入 q(z)分布的期望值

我們可以使用 Jensen's inequality，注意，log 函數是 concave function，故 $f(E(x)) \geq E(f(x))$

$$\log(P(X \mid \theta))$$
$$= \log(E_q(\frac{P(X, z_i \mid \theta)}{q(z_i)}))$$
$$\geq E_q(\log \frac{P(X, z_i \mid \theta)}{q(z_i)}) = E_q(\log \frac{P(z_i \mid X, \theta) P(X \mid \theta)}{q(z_i)})$$
$$= \sum_i q(z_i) \log P(X \mid \theta) + \sum_i q(z_i) \log \frac{P(z_i \mid X, \theta)}{q(z_i)}$$
$$= \sum_i q(z_i) \log P(X \mid \theta) - \sum_i q(z_i) \log \frac{q(z_i)}{P(z_i \mid X, \theta)}$$

當我們希望得到 $\max \log(P(X \mid \theta))$ ，有個拖油瓶，就是 $\sum_i q(z_i) \log \frac{q(z_i)}{P(z_i \mid X, \theta)}$ ，而這個拖油瓶剛好就是 relative entropy(KL divergence)， $\sum_i q(z_i) \log \frac{q(z_i)}{P(z_i \mid X, \theta)} = KL(q \| p)$ ，KL divergence 是描述兩個

分布的距離，必為一非負值， $KL(q \| p) \geq 0$ ，會降低我們的 lower bound，我們若將其變為 0，就可以將 lower bound 提升，讓 $\log(P(X \mid \theta))$ 可能有更大值。

$$\sum_i q(z_i) \log \frac{q(z_i)}{P(z_i \mid X, \theta)} = KL(q \| p) = 0$$

由於 q(z)必為大於等於 0 的值，故唯一能使 $KL(q \| p) = 0$ 的可能只有

$$\log \frac{q(z_i)}{P(z_i \mid X, \theta)} = 0$$

$$\Rightarrow \frac{q(z_i)}{P(z_i \mid X, \theta)} = 1$$ ，其實就是 q 與 p 兩分布相同時 KL divergence 為 0

$$\Rightarrow q(z_i) = P(z_i \mid X, \theta)$$

而我們在做" $KL(q \| p) = 0$ "這件事時，就是在做 E step!
我們來看看 w_i 的公式

$$w_i = \frac{\text{出現}coin\ 0\text{且}outcome\text{為}x_i\text{的機率}}{\text{出現}coin\ 0\text{且}outcome\text{為}x_i\text{的機率} + \text{出現}coin\ 1\text{且}outcome\text{為}x_i\text{的機率}} = \frac{P(z_i = C_0, x_i \mid \theta)}{P(z_i = C_0, x_i \mid \theta) + P(z_i = C_1, x_i \mid \theta)}$$

$$= \frac{P(z_i, X \mid \theta)}{\sum_i P(z_i, X \mid \theta)} = \frac{P(z_i, X \mid \theta)}{P(X \mid \theta)} = P(z_i \mid X, \theta)$$

$q(z_i)$就是 w_i ！

給定 $q(z_i)$後，再來看看我們最初的目標 $\max \log(P(X \mid \theta))$

$$J = \log(P(X \mid \theta))$$

$$\geq \sum_i q(z_i) \log \frac{P(X \mid \theta)}{q(z_i)} - \sum_i q(z_i) \log \frac{q(z_i)}{P(z_i \mid X, \theta)}$$

$$= \sum_i q(z_i) \log \frac{P(X \mid \theta)}{q(z_i)}$$

$$= \sum_i q(z_i) \log P(X \mid \theta) - \sum_i q(z_i) \log q(z_i)$$

當我們要找適當的 θ 使得 J 最大時，會做 $\frac{\partial J}{\partial \theta}$ ，此時上式第二項 $\sum_i q(z_i) \log q(z_i)$ 微分後=0，故可以直接忽略不會影響結果

故 $\arg\max_\theta \log(P(X \mid \theta)) = \arg\max_\theta \sum_i q(z) \log P(X \mid \theta)$

看看我們的 M step 在做甚麼事

$$\arg\max_{\theta} J$$

$$J = \log P(X \mid P_0, P_1, w_i, \lambda)$$

$$= \sum_{i=1}^{n} w_i(\log\lambda + x_i \log P_0 + (1-x_i)\log(1-P_0)) + \sum_{i=1}^{n}(1-w_i)(\log(1-\lambda) + x_i \log P_1 + (1-x_i)\log(1-P_1))$$

$$= \sum_i w_i P(X \mid P_0, P_1, \lambda) = \sum_i q(z) P(X \mid \theta)$$

我們 EM algorithm 實際上就是每次將 KL divergence 設為 0(E step)，再最大值的θ求出來，此時 $P(X \mid \theta)$ 增大。而因為更新了θ，KL divergence 又不為 0，再次將 KL divergence 設為 0，可以得到更大的 $P(X \mid \theta)$，然後再次更新θ，我們每次都將 lower bound 提高，不然就是至少持平，故只可能更好，不可能更糟，一直做到更新θ後，會逐漸趨近於一個最佳的，θKL divergence 趨近於 0，就會到達收斂。