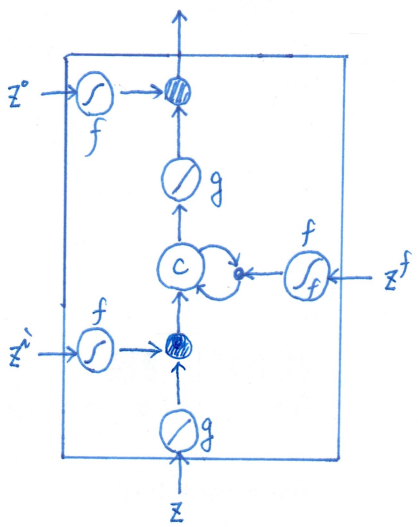
MLHW4

Q1. Given LSTM model and
- z_i, z_f, z_o → gate control input
- c → cell memory.
- f, g, h → activation function (f: sigmoid / g, h are linear)
- w, w_i, w_f, w_o → weight / b, b_i, b_f, b_o → bias
- updated cell memory → $c' = f(z_i)g(z) + cf(z_f)$

Input Sequence
$x^1 = [0, 1, 0, 3]$
$x^2 = [1, 0, 1, -2]$
$x^3 = [1, 1, 1, 4]$
$x^4 = [0, 1, 1, 0]$

Weight
$w = [0, 0, 0, 1]$
$w_i = [100, 100, 0, 0]$
$w_f = [-100, -100, 0, 0]$
$w_o = [0, 0, 100, 0]$

Bias
$b = 0$
$b_i = -10$
$b_f = 110$
$b_o = -10$



- $x^1 = [0, 1, 0, 3]$, $c = 0$

$z = wx^1 + b = [0, 0, 0, 1]\begin{bmatrix}0\\1\\0\\3\end{bmatrix} + 0 = \boxed{3} \Rightarrow g(z) = 3$

$z^i = w^i x^1 + b_i = [100, 100, 0, 0]\begin{bmatrix}0\\1\\0\\3\end{bmatrix} + (-10) = \boxed{90} \Rightarrow f(z^i) = \boxed{1}$

$z^f = w^f x^1 + b^f = [-100, -100, 0, 0]\begin{bmatrix}0\\1\\0\\3\end{bmatrix} + 110 = \boxed{10} \Rightarrow f(z^f) = \boxed{1}$

$z^o = w^o x^1 + b^o = [0, 0, 100, 0]\begin{bmatrix}0\\1\\0\\3\end{bmatrix} - 10 = \boxed{-10} \Rightarrow f(z^o) = \boxed{0}$

$c' = f(z^i)g(z) + cf(z^f) = 1 \cdot 3 + 0 \cdot 1 = \boxed{3} \Rightarrow h(c') = 3$

$y = f(z^o)h(c') = \boxed{0}$

- $x^2 = [1, 0, 1, -2]$, $c = 3$

$z = \boxed{-2} \Rightarrow g(z) = -2$

$z^i = \boxed{90} \Rightarrow f(z^i) = \boxed{1}$

$z^f = \boxed{10} \Rightarrow f(z^f) = \boxed{1}$

$z^o = \boxed{90} \Rightarrow f(z^o) = \boxed{1}$

$c' = 1 \times -2 + 3 \cdot 1 = \boxed{1}$

$h(c') = 1$

$y = \boxed{1}$

- $x^3 = [1, 1, 1, 4]$, $c = 1$

$z = \boxed{4} \Rightarrow g(z) = 4$

$z^i = \boxed{190} \Rightarrow f(z^i) = \boxed{1}$

$z^f = \boxed{-90} \Rightarrow f(z^f) = \boxed{0}$

$z^o = \boxed{90} \Rightarrow f(z^o) = \boxed{1}$

$c' = \boxed{4}$, $h(c') = 4$, $y = \boxed{4}$

- $x^4 = [0, 1, 100, 0]$, $c = 4$

$z = \boxed{0} \Rightarrow g(z) = 0$

$z^i = \boxed{90} \Rightarrow f(z^i) = \boxed{1}$

$z^f = \boxed{10} \Rightarrow f(z^f) = \boxed{1}$

$z^o = \boxed{90} \Rightarrow f(z^o) = \boxed{1}$

$c' = \boxed{4}$, $h(c') = 4$, $y = \boxed{4}$

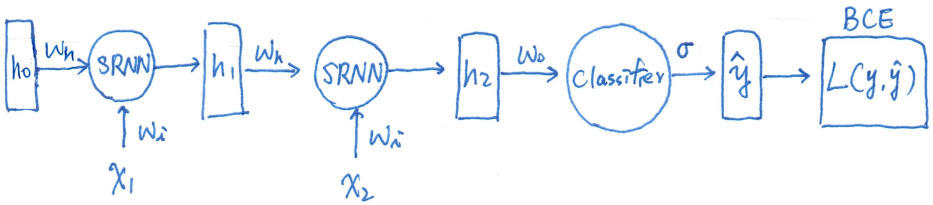
t	z^i	$f(z^i)$	z^f	$f(z^f)$	z^o	$f(z^o)$	z	c'	y
1	90	1	10	1	-10	0	3	3	0
2	90	1	10	1	90	1	-2	1	1
3	190	1	-90	0	90	1	4	4	4
4	90	1	10	1	90	1	0	4	4

※

Q2. Given Objective function is Binary Cross Entropy $L(y, \hat{y})$ where y is label

$$\hat{y} = \sigma(w_o h_2) = \frac{1}{1+\exp(-w_o h_2)}$$

$$h_t = \tanh(w_i x_t + w_h h_{t-1})$$



Goal: derive $\frac{\partial L(y,\hat{y})}{\partial w_o}$, $\frac{\partial L(y,\hat{y})}{\partial w_h}$, $\frac{\partial L(y,\hat{y})}{\partial w_i}$

$L(y, \hat{y})$, $\hat{y}(w_o, h_2)$, $h_2(w_i, x_2, w_h, h_1)$, $h_1(w_i, x_1, w_h, h_0)$, $h_0 = 0$

(1) $\frac{\partial L}{\partial w_o} = \frac{\partial L}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial w_o}$

(2) $\frac{\partial L}{\partial w_h} = \frac{\partial L}{\partial h_2} \frac{\partial h_2}{\partial w_h} + \frac{\partial L}{\partial h_1} \frac{\partial h_1}{\partial w_h} = \frac{\partial L}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial h_2} \frac{\partial h_2}{\partial w_h} + \frac{\partial L}{\partial h_2} \frac{\partial h_2}{\partial h_1} \frac{\partial h_1}{\partial w_h}$

$= \frac{\partial L}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial h_2} \frac{\partial h_2}{\partial w_h} + \frac{\partial L}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial h_2} \frac{\partial h_2}{\partial h_1} \frac{\partial h_1}{\partial w_h} = \frac{\partial L}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial h_2} \left[\frac{\partial h_2}{\partial w_h} + \frac{\partial h_2}{\partial h_1} \frac{\partial h_1}{\partial w_h} \right]$

(3) $\frac{\partial L}{\partial w_i} = \frac{\partial L}{\partial h_2} \frac{\partial h_2}{\partial w_i} + \frac{\partial L}{\partial h_1} \frac{\partial h_1}{\partial w_i} = \frac{\partial L}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial h_2} \frac{\partial h_2}{\partial w_i} + \frac{\partial L}{\partial h_2} \frac{\partial h_2}{\partial h_1} \frac{\partial h_1}{\partial w_i}$

$= \frac{\partial L}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial h_2} \frac{\partial h_2}{\partial w_i} + \frac{\partial L}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial h_2} \frac{\partial h_2}{\partial h_1} \frac{\partial h_1}{\partial w_i} = \frac{\partial L}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial h_2} \left[\frac{\partial h_2}{\partial w_i} + \frac{\partial h_2}{\partial h_1} \frac{\partial h_1}{\partial w_i} \right]$

---

(1) $\frac{\partial \hat{y}}{\partial w_o} = \frac{h_2 \exp(-w_o h_2)}{[1+\exp(-w_o h_2)]^2}$

※ $\begin{cases} h_2 = \tanh(w_i x_2 + w_h h_1) \\ h_1 = \tanh(w_i x_1 + w_h h_0) \end{cases}$

(2) $\frac{\partial \hat{y}}{\partial h_2} = \frac{w_o \exp(-w_o h_2)}{[1+\exp(-w_o h_2)]^2}$, $\frac{\partial h_2}{\partial w_h} = h_1 \operatorname{sech}^2(w_i x_2 + w_h h_1)$

$\frac{\partial h_2}{\partial h_1} = w_h \operatorname{sech}^2(w_i x_2 + w_h h_1)$, $\frac{\partial h_1}{\partial w_h} = h_0 \operatorname{sech}^2(w_i x_1 + w_h h_0) = 0$

(3) $\frac{\partial h_2}{\partial w_i} = x_2 \operatorname{sech}^2(w_i x_2 + w_h h_1)$, $\frac{\partial h_1}{\partial w_i} = x_1 \operatorname{sech}^2(w_i x_1 + w_h h_0)$

$$\frac{\partial L}{\partial \hat{y}} = \frac{\partial -[y\log\hat{y} + (1-y)\log(1-\hat{y})]}{\partial \hat{y}} = -\left[\frac{y}{\hat{y}} - \frac{1-y}{1-\hat{y}}\right] = \frac{1-y}{1-\hat{y}} - \frac{y}{\hat{y}}$$

(1) $$\frac{\partial L}{\partial w_o} = \left(\frac{1-y}{1-\hat{y}} - \frac{y}{\hat{y}}\right) \frac{h_2 \exp(-w_o h_2)}{[1+\exp(-w_o h_2)]^2}$$

(2) $$\frac{\partial L}{\partial w_h} = \left(\frac{1-y}{1-\hat{y}} - \frac{y}{\hat{y}}\right) \frac{w_o \exp(-w_o h_2)}{[1+\exp(-w_o h_2)]^2} \left[h_1 \operatorname{sech}^2(w_i x_2 + w_h h_1) + 0\right]$$

(3) $$\frac{\partial L}{\partial w_i} = \left(\frac{1-y}{1-\hat{y}} - \frac{y}{\hat{y}}\right) \frac{w_o \exp(-w_o h_2)}{[1+\exp(-w_o h_2)]^2} \left[x_2 \operatorname{sech}^2(w_i x_2 + w_h h_1) + w_h \operatorname{sech}^2(w_i x_2 + w_h h_1) \cdot x_1 \operatorname{sech}^2(w_i x_1 + w_h h_0)\right]$$

Q.E.D.

Q3. Given input space $\mathcal{X}$, # of classes K, training data set $\{(x_i, \hat{y}_i)\}_{i=1}^m$ ($\hat{y}_i \in [1,K]$, $x_i \in \mathcal{X}$), collection of multiclass classifiers $\mathcal{F}$

We want to find function $g_{T+1}^k(x) = \sum_{t=1}^{T} \alpha_t f_t^k(x)$, $k \in [1,K]$ where $f \in \mathcal{F}$, $\alpha_t \in \mathbb{R}$, $t \in [1,T]$

Objective: to minimize $L(g_{T+1}^1, \cdots, g_{T+1}^K) = \sum_{i=1}^m \exp\left(\frac{1}{K-1} \sum_{k \neq \hat{y}_i} g_{T+1}^k(x_i) - g_{T+1}^{\hat{y}_i}(x_i)\right)$

and show how use f_t and α_t to achieve this goal.

<sol> In Binary classifier, that is $\mathcal{Y} = \{\pm 1\}$, we aim to minimize $\hat{R}_S(g)$ just like this problem.

※ determine direction $f_t = \underset{f \in H}{\operatorname{argmin}} \frac{\partial}{\partial \alpha} \hat{R}_S(g_t + \alpha f)\Big|_{\alpha=0}$

※ determine step-size $\alpha_t = \underset{\alpha \in \mathbb{R}}{\operatorname{argmin}} \hat{R}_S(g_t + f_t)$

$f_t = \underset{f \in H}{\operatorname{argmin}} \frac{\partial}{\partial \alpha} \sum_{i=1}^m \exp\left(\frac{1}{K-1} \sum_{k \neq \hat{y}_i} g_{T+1}^k(x_i) - g_{T+1}^{\hat{y}_i}(x_i)\right)\Big|_{\alpha=0}$ take $g_t + \alpha f$ in Loss function

$\Rightarrow f_t = \underset{f \in H}{\operatorname{argmin}} \frac{\partial}{\partial \alpha} \sum_{i=1}^m \exp\left(\frac{1}{K-1} \sum_{k \neq \hat{y}_i} \left(g_t^k(x_i) + \alpha f^k(x_i)\right) - \left(g_t^{\hat{y}_i}(x_i) + \alpha f^{\hat{y}_i}(x_i)\right)\right)\Big|_{\alpha=0}$

$= \underset{f \in H}{\operatorname{argmin}} \frac{\partial}{\partial \alpha} \sum_{i=1}^m \exp\left[\frac{1}{K-1} \sum_{k \neq \hat{y}_i} \left(g_t^k(x_i) + \alpha f^k(x_i)\right)\right] \exp\left(-g_t^{\hat{y}_i}(x_i) - \alpha f^{\hat{y}_i}(x_i)\right)\Big|_{\alpha=0}$

$= \underset{f \in H}{\operatorname{argmin}} \sum_{i=1}^m \exp\left[\frac{1}{K-1} \sum_{k \neq \hat{y}_i} \left[g_t^k(x_i) + \alpha f^k(x_i)\right]\right] \cdot \left[\frac{1}{K-1} \sum_{k \neq \hat{y}_i} f^k(x_i)\right] \exp\left(-g_t^{\hat{y}_i}(x_i) - \alpha f^{\hat{y}_i}(x_i)\right)$

$+ \exp\left[\frac{1}{K-1} \sum_{k \neq \hat{y}_i} \left(g_t^k(x_i) + \alpha f^k(x_i)\right)\right] \cdot \exp\left(-g_t^{\hat{y}_i}(x_i) - \alpha f^{\hat{y}_i}(x_i)\right) \cdot \left(-f^{\hat{y}_i}(x_i)\right)\Big|_{\alpha=0}$

$= \underset{f \in H}{\operatorname{argmin}} \sum_{i=1}^m \exp\left[\frac{1}{K-1} \sum_{k \neq \hat{y}_i} \left[g_t^k(x_i) + \alpha f^k(x_i)\right]\right] \exp\left(-g_t^{\hat{y}_i}(x_i) - \alpha f^{\hat{y}_i}(x_i)\right) \cdot \left[\frac{1}{K-1} \sum_{k \neq \hat{y}_i} f^k(x_i) - f^{\hat{y}_i}(x_i)\right]\Big|_{\alpha=0}$

$= \underset{f \in H}{\operatorname{argmin}} \sum_{i=1}^m \exp\left[\frac{1}{K-1} \sum_{k \neq \hat{y}_i} \left[g_t^k(x_i) + \alpha f^k(x_i)\right] - \left(g_t^{\hat{y}_i}(x_i) + \alpha f^{\hat{y}_i}(x_i)\right)\right] \cdot \left[\frac{1}{K-1} \sum_{k \neq \hat{y}_i} f^k(x_i) - f^{\hat{y}_i}(x_i)\right]\Big|_{\alpha=0}$

$= \underset{f \in H}{\operatorname{argmin}} \sum_{i=1}^m \exp\left[\frac{1}{K-1} \sum_{k \neq \hat{y}_i} g_t^k(x_i) - g_t^{\hat{y}_i}(x_i)\right] \left[\frac{1}{K-1} \sum_{k \neq \hat{y}_i} f^k(x_i) - f^{\hat{y}_i}(x_i)\right]$

Let $Z_t = \sum_{i=1}^{m} \exp\left(\frac{1}{K-1}\sum_{k\neq \hat{y}_i} g_t^k(x_i) - g_t^{\hat{y}_i}(x_i)\right)$, $D_t(i) = \frac{\exp\left[\frac{1}{K-1}\sum_{k\neq \hat{y}_i} g_t^k(x_i) - g_t^{\hat{y}_i}(x_i)\right]}{Z_t}$

$\Rightarrow \underset{f\in H}{\operatorname{argmin}} \sum_{i=1}^{m} Z_t D_t(i)\left[\frac{1}{K-1}\sum_{k\neq \hat{y}_i} f^k(x_i) - f^{\hat{y}_i}(x_i)\right]$

$= \underset{f\in H}{\operatorname{argmin}}\ \underbrace{Z_t}_{\text{fixed}}\ \mathbb{E}\left[\frac{1}{K-1}\sum_{k\neq \hat{y}_i} f^k(x_i) - f^{\hat{y}_i}(x_i)\right]$

this expectation just has 2 values including {correct, wrong}

$\Rightarrow \mathbb{P}_{i\sim D_t}\left[\frac{1}{K-1}\sum_{k\neq \hat{y}_i} f^k(x_i) - f^{\hat{y}_i}(x_i) = \text{判對}\right] \times$ value of correct

$+ \mathbb{P}_{i\sim D_t}\left[\frac{1}{K-1}\sum_{k\neq \hat{y}_i} f^k(x_i) - f^{\hat{y}_i}(x_i) = \text{判錯}\right] \times$ value of wrong

$\Rightarrow$ 1 - 判對的機率 = 判錯的機率

$\Rightarrow \left[1 - \mathbb{P}_{i\sim D_t}\left[\frac{1}{K-1}\sum_{k\neq \hat{y}_i} f^k(x_i) - f^{\hat{y}_i}(x_i) = \text{判錯}\right]\right] \times$ value of correct

$+ \mathbb{P}_{i\sim D_t}\left[\frac{1}{K-1}\sum_{k\neq \hat{y}_i} f^k(x_i) - f^{\hat{y}_i}(x_i) = \text{判錯}\right] \times$ value of wrong

$= \underbrace{\left[(\text{value of correct}) + (\text{value of wrong})\right]}_{\text{fixed value}} \times \mathbb{P}_{i\sim D_t}\left[\frac{1}{K-1}\sum_{k\neq \hat{y}_i} f^k(x_i) - f^{\hat{y}_i}(x_i) = \text{判錯}\right] + \underbrace{\boxed{\text{value of correct}}}_{\text{fixed value}}$

相當於 $\Rightarrow \underset{f\in H}{\operatorname{argmin}}\ \mathbb{P}_{k\sim D_t}\left[\frac{1}{K-1}\sum_{k\neq \hat{y}_i} f^k(x_i) - f^{\hat{y}_i}(x_i) = \text{判錯}\right]$ Q.E.D.

In Binary classifier, $\alpha_t = \underset{\alpha\in\mathbb{R}}{\operatorname{argmin}}\ \hat{R}_S(g_t + \alpha f_t)$

$\Rightarrow \alpha_t = \underset{\alpha\in\mathbb{R}}{\operatorname{argmin}} \sum_{i=1}^{m} \exp\left(\frac{1}{K-1}\sum_{k\neq \hat{y}_i}\left[g_t^k(x_i) - \alpha f_t^k(x_i)\right] - \left(g_t^{\hat{y}_i}(x_i) + \alpha f_t^{\hat{y}_i}(x_i)\right)\right)$

$= \sum_{i: f_t^{\hat{y}_i}(x_i)\text{判錯}} \exp\left(\frac{1}{K-1}\sum_{k\neq \hat{y}_i}\left[g_t^k(x_i) - \alpha f_t^k(x_i)\right] - \left(g_t^{\hat{y}_i}(x_i) + \alpha f_t^{\hat{y}_i}(x_i)\right)\right) +$

$\sum_{i: f_t^{\hat{y}_i}(x_i)\text{判對}} \exp\left(\frac{1}{K-1}\sum_{k\neq \hat{y}_i}\left[g_t^k(x_i) - \alpha f_t^k(x_i)\right] - \left(g_t^{\hat{y}_i}(x_i) + \alpha f_t^{\hat{y}_i}(x_i)\right)\right)$, if value of correct is C and value of wrong is W

$= \sum_{i: f_t^{\hat{y}_i}(x_i) = W} Z_t D_t(x_i) \exp\left(-g_t^{\hat{y}_i}(x_i) + \alpha W\right) + \sum_{i: f_t^{\hat{y}_i}(x_i) = C} Z_t D_t(x_i) \exp\left(-g_t^{\hat{y}_i}(x_i) + \alpha C\right)$

Let $\varepsilon_t = \mathbb{P}_{i\sim D_t}\left[f_t^{\hat{y}_i}(x_i) = \text{判錯}\right] = Z_t \varepsilon_t \exp\left(-g_t^{\hat{y}_i}(x_i) + \alpha W\right) + Z_t(1-\varepsilon_t)\exp\left(-g_t^{\hat{y}_i}(x_i) + \alpha C\right)$

Let $\frac{\partial}{\partial \alpha}\left[\varepsilon_t \exp\left(-g_t^{\hat{y}_i}(x_i) + \alpha W\right) + (1-\varepsilon_t)\exp\left(-g_t^{\hat{y}_i}(x_i) + \alpha C\right)\right]$

$= W\underbrace{\left[\varepsilon_t \exp\left(-g_t^{\hat{y}_i}(x_i)\right)\right]}_{=G}\exp(\alpha W) + C\exp\left(-g_t^{\hat{y}_i}(x_i)\right)\exp(\alpha C) - C\varepsilon_t \exp\left(-g_t^{\hat{y}_i}(x_i)\right)\exp(\alpha C)$

$= G\left[W\exp(\alpha W) + \frac{C}{\varepsilon_t}\exp(\alpha C) - C\exp(\alpha C)\right]$ Q.E.D.

MLHW Q4.

- Let X be observed variable ⇒ 外在特徵

  Z be latent variable (X 分別屬於哪一群) ⇒ 內在性格 (不易觀察)

- Want to do maximum likelihood estimation $\theta^* = \underset{\theta \in \Theta}{\operatorname{argmax}}\, p(X;\theta)$ where $\{P(\cdot;\theta) : \theta \in \Theta\}$ is a collection of probability models

- EM algorithms

  Initial $\theta^{(1)}$, then iterate 1, 2, 3 ... as follow:

  ▷ E-Step: Derive

$$Q(\theta|\theta^{(t)}) = \sum_Z \underbrace{P(Z|X;\theta^{(t)})}_{\text{根據當前的 probability model 和外在特徵 ⇒ 猜測內在性格}} \log P(X,Z;\theta)$$

  ▷ M-Step: Find

$$\theta^{(t+1)} = \underset{\theta \in \Theta}{\operatorname{argmax}}\, Q(\theta|\theta^{(t)})$$

Example:

If $X = \{x_1, x_2 \dots x_N\}$ (N training data), $Z = \{z_1, z_2 \dots z_N\}$ (corresponding latent variable), $P_\theta(X,Z;\theta) = \prod_{i=1}^{N} P(x_i, z_i;\theta)$ (data are i.i.d generated) $\forall\, \theta \in \Theta$

Then, $Q(\theta|\theta^{(t)}) = \mathbb{E}_{Z\sim P(\cdot|X;\theta^{(t)})}[\log P(X,Z;\theta)]$

$$= \sum_{i=1}^{N} \mathbb{E}_{z\sim P(\cdot|x;\theta^{(t)})}[\log P(x_i, z_i;\theta)] \quad \text{by independence}$$

Denote the parameter estimate at t as: $\theta^{(t)} = \{(\pi_k^{(t)}, \mu_k^{(t)}, \Sigma_k^{(t)})\}_{k=1}^{K}$

$$= \sum_{i=1}^{N} \mathbb{E}_{z_i\sim P(\cdot|x_i;\theta^{(t)})}[\log P(x_i, z_i;\theta)]$$ (因為是獨立的，所以推測 z_i 只要知道 x_i 就好)

$P(\cdot|x_i;\theta^{(t)})$ ↙ 要算的東西

Apply to GMM

- In GMM with K gaussian distributions, $P(X, z=k;\theta) = \pi_k \mathcal{N}(X;\mu_k,\Sigma_k)$

  ↳ data是從 k-th gaussian distribution 產生且外在特徵是 X

  where $\theta = \{(\pi_k, \mu_k, \Sigma_k)\}_{k=1}^{K}$

  ↳ 從 k-th gaussian dist. 發生的機率 (sampling from k-th G.D 的 P.)

  ▷ $P(z_i=k|x_i;\theta^{(t)}) = \dfrac{P(z_i=k, x_i;\theta^{(t)})}{\sum_{j=1}^{k} P(z_i=j, x_i;\theta^{(t)})}$ ← 貝氏定理

$$= \frac{\pi_k^{(t)} \mathcal{N}(x_i;\mu_k^{(t)},\Sigma_k^{(t)})}{\sum_{j=1}^{k} \pi_k^{(t)} \mathcal{N}(x_i;\mu_j^{(t)},\Sigma_j^{(t)})} = \delta_{ik}^{(t)}$$ (可以算出固定的值)

where $\mathcal{N}(x;\mu,\Sigma) = \frac{1}{\sqrt{(2\pi)^m|\Sigma|}} \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)$ is the p.d.f of

Gaussian distribution, with mean μ, and covariance matrix Σ

▷ $\log P(x_i, z_i=k ; \theta^{(t)}) = \log\left(\pi_k \frac{1}{\sqrt{(2\pi)^m |\Sigma_k|}} \exp\left(-\frac{1}{2}(x_i-\mu_k)^T \Sigma_k^{-1}(x_i-\mu_k)\right)\right)$

$= \log \frac{\pi_k}{\sqrt{(2\pi)^m |\Sigma_k|}} - \frac{1}{2}(x_i-\mu_k)^T \Sigma_k^{-1}(x_i-\mu_k)$ (A very simple function w.r.t. θ as desired)

• E-Step:

$$Q(\theta|\theta^{(t)}) = \sum_{i=1}^{N} \mathbb{E}_{z_i \sim P(\cdot|x_i;\theta^{(t)})}\left[\log P(x_i, z_i ; \theta)\right]$$

$$= \sum_{i=1}^{N}\sum_{k=1}^{K} P(z_i=k | x_i ; \theta^{(t)}) \log P(x_i, z_i=k ; \theta)$$

$$= \sum_{i=1}^{N}\sum_{k=1}^{K} \delta_{ik}^{(t)} \left[\log \frac{\pi_k^{(t)}}{\sqrt{(2\pi)^m |\Sigma_k|}} - \frac{1}{2}(x_i-\mu_k)^T \Sigma_k^{-1}(x_i-\mu_k)\right]$$

• M-Step: To maximize $Q(\theta|\theta^{(t)})$, we compute $\nabla_{\mu_k} Q(\theta|\theta^{(t)})$ and $\nabla_{\Sigma_k} Q(\theta|\theta^{(t)})$

$\nabla_{\mu_k} Q(\theta|\theta^{(t)}) = 0 \Rightarrow$ Take $\mu_k^* = \frac{\sum_{i=1}^{N} \delta_{ik}^{(t)} x_i}{\sum_{i=1}^{N} \delta_{ik}^{(t)}}$, $\nabla_{\Sigma_k} Q(\theta|\theta^{(t)}) = 0 \Rightarrow$ Take $\Sigma_k^* = \frac{\sum_{i=1}^{N} \delta_{ik}^{(t)} (x_i-\mu_k)(x_i-\mu_k)^T}{\sum_{i=1}^{N} \delta_{ik}^{(t)}}$

Maximum $Q(\theta|\theta^{(t)})$ over $\pi_1, \pi_2 \dots \pi_K$ with constraint $\pi_1 + \dots \pi_K = 1$

$\Rightarrow$ Take $\pi_k^* = \frac{1}{N}\sum_{i=1}^{N} \delta_{ik}^{(t)}$

• $\nabla_{\mu_k} Q(\theta|\theta^{(t)}) = \sum_{i=1}^{N} \delta_{ik}^{(t)} \Sigma_k^{-1}(x_i-\mu_k) = \Sigma_k^{-1} \sum_{i=1}^{N} \delta_{ik}^{(t)} (x_i-\mu_k)$

↳ 因 $\delta_{ik}^{(t)}$ 和 θ 無關故也和 μ_k 無關

代入之後 $\nabla_{\mu_k} Q(\theta|\theta^{(t)}) = 0$

代入後使 $\frac{\partial}{\partial a_{ij}^k} Q(\theta|\theta^{(t)}) = 0$

• Let $\Sigma_k^{-1} = \begin{bmatrix} a_{11}^k & a_{12}^k & \cdots & a_{1m}^k \\ a_{21}^k & a_{22}^k & & \\ a_{31}^k & \vdots & \ddots & \\ a_{m1}^k & a_{m2}^k & & a_{mm}^k \end{bmatrix}$ (要 optimize Σ_k 就是 optimize Σ_k^{-1})

$$\frac{\partial}{\partial a_{ij}^k} Q(\theta|\theta^{(t)}) = \frac{1}{2}\sum_{l=1}^{N} \delta_{lk}^{(t)} e_j^T \left[\Sigma_k - (x_l-\mu_k)(x_l-\mu_k)^T\right] e_i = 0 \quad \forall\, i,j$$

$$\frac{\partial}{\partial \pi_k}\left(Q(\theta|\theta^{(t)}) - \lambda \sum_{j=1}^{K} \pi_j\right) = \sum_{i=1}^{N} \delta_{ik}^{(t)} \cdot \frac{1}{\pi_k} - \lambda$$ (有 constraint 的條件算極值用 Lagrange Multipler)

Take $\pi_k^* = \frac{1}{\lambda}\sum_{i=1}^{N} \delta_{ik}^{(t)}$ $\because \pi_1^* + \dots \pi_K^* = 1$ $\therefore 1 = \frac{1}{\lambda}\sum_{k=1}^{K}\sum_{i=1}^{N} \delta_{ik}^{(t)} = \frac{N}{\lambda}$ $\therefore \lambda = N$

$$\nabla_{\mu_k} Q(\theta|\theta^{(t)}) = \frac{\partial}{\partial \mu_k} \sum_{i=1}^{N} \sum_{k=1}^{K} \delta_{ik}^{(t)} \left[\log \frac{\pi_k^{(t)}}{\sqrt{(2\pi)^m |\Sigma_k|}} - \frac{1}{2}(x_i - \mu_k)^T \Sigma_k^{-1} (x_i - \mu_k) \right]$$

$\delta_{ik}^{(t)}$: const; $\log \frac{\pi_k^{(t)}}{\sqrt{(2\pi)^m |\Sigma_k|}}$: 無關

$$= \sum_{i=1}^{N} \delta_{ik}^{(t)} \frac{\partial}{\partial \mu_k} -\frac{1}{2} \left(x_i^T \Sigma_k^{-1} x_i - \mu_k^T \Sigma_k^{-1} x_i - x_i^T \Sigma_k^{-1} \mu_k + \mu_k^T \Sigma_k^{-1} \mu_k \right)$$

$$= \sum_{i=1}^{N} \delta_{ik}^{(t)} -\frac{1}{2} \left(-\Sigma_k^{-1} x_i - (x_i^T \Sigma_k^{-1})^T + \left[\Sigma_k^{-1} + (\Sigma_k^{-1})^T \right] \mu_k \right)$$

$$= \sum_{i=1}^{N} \delta_{ik}^{(t)} -\frac{1}{2} \left(-\Sigma_k^{-1} x_i - \Sigma_k^{-1} x_i + 2\Sigma_k^{-1} \mu_k \right) = \sum_{i=1}^{N} \delta_{ik}^{(t)} \left(\Sigma_k^{-1} x_i - \Sigma_k^{-1} \mu_k \right)$$

$$= \sum_{i=1}^{N} \delta_{ik}^{(t)} \Sigma_k^{-1} (x_i - \mu_k)$$ Q.E.D.

$$\nabla_{\pi_k} \left(Q(\theta|\theta^{(t)}) - \lambda \sum_{j=1}^{k} \pi_j \right) = \frac{\partial}{\partial \pi_k} \sum_{i=1}^{N} \sum_{k=1}^{K} \delta_{ik}^{(t)} \left[\log \pi_k^{(t)} - \log \sqrt{(2\pi)^m |\Sigma_k|} - \frac{1}{2}(x_i - \mu_k)^T \Sigma_k^{-1} (x_i - \mu_k) \right] - \lambda \sum_{j=1}^{k} \pi_j$$

$$= \sum_{i=1}^{N} \delta_{ik}^{(t)} \frac{1}{\pi_k} - \lambda \cdot 1$$