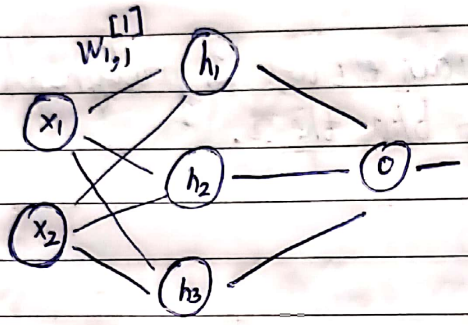
PS-3

(d) $l = \frac{1}{m} \sum_{i=1}^{m} (o^{(i)} - y^{(i)})^2$

(a) let, $l^{(i)} = (o^{(i)} - y^{(i)})^2$



$$m^{(i)} = \begin{pmatrix} x_1^{(i)} \\ x_2^{(i)} \end{pmatrix}_{2\times 1}$$

let $w^{[1]} = \begin{pmatrix} w_{1,1}^{[1]} & w_{2,1}^{[1]} \\ w_{1,2}^{[1]} & w_{2,2}^{[1]} \\ w_{1,3}^{[1]} & w_{2,3}^{[1]} \end{pmatrix}_{3\times 2}$

& $b^{[1]} = \begin{pmatrix} w_{0,1}^{[1]} \\ w_{0,2}^{[1]} \\ w_{0,3}^{[1]} \end{pmatrix}_{3\times 1}$

$$\Rightarrow z^{[1](i)} = w^{[1]} m^{(i)} + b^{[1]}$$

$$a^{[1](i)} = \sigma(z^{[1](i)})$$

let $w^{[2]} = (w_1^{[2]} \quad w_2^{[2]} \quad w_3^{[2]})_{1\times 3}$

& $b^{[2]} = (w_0^{[2]})$

$$\Rightarrow o^{(i)} = a^{[2](i)} = \sigma(z^{[2](i)})$$

$$z^{[2](i)} = W^{[2]} a^{[1](i)} + b^{[2]}$$

$$\frac{\partial l^{(i)}}{\partial o^{(i)}} = 2(o^{(i)} - y^{(i)})$$

$$\frac{\partial l^{(i)}}{\partial a^{[2](i)}} = 2(a^{[2](i)} - y^{(i)})$$

$$\frac{\partial l^{(i)}}{\partial z^{[2](i)}} = \frac{\partial l^{(i)}}{\partial a^{[2](i)}} \times \frac{\partial a^{[2](i)}}{\partial z^{[2](i)}}$$

$$\frac{\partial l^{(i)}}{\partial z^{[2](i)}} = 2(a^{[2](i)} - y^{(i)}) * a^{[2](i)}(1 - a^{[2](i)})$$

↝ element-wise product

$$\frac{\partial l^{(i)}}{\partial a^{[1](i)}} = \frac{\partial l^{(i)}}{\partial z^{[2](i)}} \times \frac{\partial z^{[2](i)}}{\partial a^{[1](i)}}$$

$$= \underbrace{\frac{\partial l^{(i)}}{\partial z^{[2](i)}}}_{1\times 1} \times \underbrace{(W^{[2]})^T}_{3\times 1} = (W^{[2]})^T \times \frac{\partial l^{(i)}}{\partial z^{[2](i)}}$$

$$\frac{\partial l^{(i)}}{\partial z^{[1](i)}} = \frac{\partial l^{(i)}}{\partial a^{[1](i)}} \times \frac{\partial a^{[1](i)}}{\partial z^{[1](i)}}$$

$$= \underbrace{\frac{\partial l^{(i)}}{\partial a^{[1](i)}}}_{3\times 1} * \underbrace{a^{[1](i)}(1 - a^{[1](i)})}_{3\times 1}$$

↳ this is okay!

$$\frac{\partial l^{(i)}}{\partial W^{[1]}} = \frac{\partial l^{(i)}}{\partial z^{[1](i)}} \times \frac{\partial z^{[1](i)}}{\partial W^{[1]}}$$

$$= \underbrace{\frac{\partial l^{(i)}}{\partial z^{[1](i)}}}_{3\times 1} \times \underbrace{(x^{(i)})^T}_{1\times 2}$$

↳ done!

$$\therefore \frac{\partial l^{(i)}}{\partial w^{[1]}} = \underbrace{(w^{[2]})^T}_{3\times 1} \times \Big(2(a^{[2](i)} - y^{(i)}) * a^{[2](i)}(1-a^{[2](i)}) \Big) * \underbrace{a^{[1](i)}(1-a^{[1](i)})}_{3\times 1} \times \underbrace{(M^{(i)})^T}_{1\times 2}$$

$$= \begin{pmatrix} w_1^{[2]} \\ w_2^{[2]} \\ w_3^{[2]} \end{pmatrix} * \begin{pmatrix} w_{0,1}^{[1]} + w_{1,1}^{[1]} M_1^{(i)} + w_{2,1}^{[1]} M_2^{(i)} \\ w_{0,2}^{[1]} + w_{1,2}^{[1]} M_1^{(i)} + w_{2,2}^{[1]} M_2^{(i)} \\ w_{0,3}^{[1]} + w_{1,3}^{[1]} M_1^{(i)} + w_{2,3}^{[1]} M_2^{(i)} \end{pmatrix} \times (1\times 1) \times \begin{pmatrix} M_1^{(i)} & M_2^{(i)} \end{pmatrix}$$

finally we need $\frac{\partial l^{(i)}}{\partial w_{1,2}^{[1]}}$ =

again rearranging,

$$\frac{\partial l^{(i)}}{\partial w^{[1]}} = (w^{[2]T}) \times 2 (a^{[2](i)} - y^{(i)}) * (a^{[1](i)}(1-a^{[1](i)}$$

$$\frac{\partial l^{(i)}}{\partial w^{[1]}} = (w^{[2].T}) \times \Big(2 (o^{(i)} - y^{(i)}) * o^{(i)}(1-o^{(i)}) \Big) * a^{[1](i)} (1-a^{[1](i)}) \times (M^{(i)})^T$$

↳ 2nd row of this ← ↳ 1st col^m of this

$$\Rightarrow \frac{\partial l^{(i)}}{\partial w_{1,2}^{[1]}} = 2 w_2^{[2]} (o^{(i)} - y^{(i)}) o^{(i)}(1-o^{(i)}) \times (w_{0,2}^{[1]} + w_{1,2}^{[1]} M_1^{(i)} + w_{2,2}^{[1]} M_2^{(i)}) * M_2^{(i)}$$

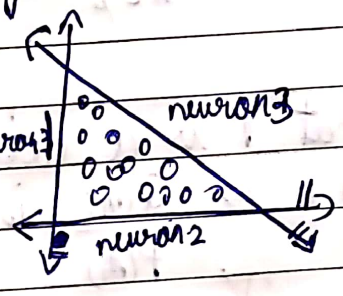
to stop the formula from becoming too long, keep $(a^{[1](i)}_2) = \sigma(W^{[1]}_{0,2} + W^{[1]}_{1,2} x^{(i)}_1 + W^{[1]}_{2,2} x^{(i)}_2)$

$$\therefore \frac{\partial \mathcal{L}^{(i)}}{\partial W^{[1]}_{1,2}} = 2 W^{[2]}_{2} \times (a^{(i)} - y^{(i)})(a^{(i)})(1-a^{(i)}) \times (a^{[1](i)}_2)(1-a^{[1](i)}_2) \times x^{(i)}_1$$

$\therefore$ over all m eg.,

$$W^{[1]}_{1,2} := W^{[1]}_{1,2} - \alpha \times \frac{2 W^{[2]}_2}{m} \sum_{i=1}^{m} (\quad)$$

(b) Yes, it is possible to get 100% accuracy on the data.
possible ∵ as given in the hint, the 3 neurons (hidden) can learn a Δ boundary s.t. the region of os overlap with os in the data set.
something like,



neuron 1 can learn the line - $x_1 = 0.5$
neuron 2 " " " " - $x_2 = 0.5$
∴ " 3 " " " " - $x_1 + x_2 = 4$

then o/p layer weights will need to invert o/p from neuron 1 & 2. This way we will have a classifier detecting pts. inside vs outside the Δ.

(c) Not possible, the whole network becomes a single linear unit & we can learn only one line & dataset is not linearly separable.

$$D_{KL}(P\|Q) = \sum_{n \in \mathcal{X}} P(n) \log \frac{P(n)}{Q(n)}$$

$\hookrightarrow \equiv D_{KL}(P(x)\|Q(x))$

$P(n) > 0 \; \forall n$
& $0 \log 0 = 0$.

(a)

$$D_{KL}(P\|Q) = \sum_{n \in \mathcal{X}} P(n) \log P(n)$$

$$= \sum_{n \in \mathcal{X}} P(n) \times (+\log(Q(n)))$$

$$\bullet\; D_{KL}(P\|Q) = -\sum_{n \in \mathcal{X}} P(n) \log \frac{Q(n)}{P(n)}$$

note that P is a deterministic f^n for an $n \Rightarrow \frac{Q(n)}{P(n)}$ is rv
& so if $X \sim P$ i.e. frequency of $n \in \mathcal{X} = P(n) \; \forall n \in \mathcal{X}$

$$D_{KL}(P\|Q) = E[-\log(Q(n)/P(n))]$$

$$\geq -\log E[Q(n)/P(n)]$$

$$= -\log\left(\sum_{n \in \mathcal{X}} P(n) \times \frac{Q(n)}{P(n)}\right)$$

$$D_{KL}(P\|Q) \geq -\log\left(\sum_{x\in\mathcal{X}} Q(x)\right)$$

$$\geq -\log(1)$$

$$D_{KL}(P\|Q) \geq 0, \quad P=Q \text{ iff } D_{KL}(P\|Q)=0$$

$\because$ P & Q are pmfs of $\mathcal{X}$.

$\because$ then $E[f(x)] = f(E[x]) = 0$

(a)

$$D_{KL}(P(X|Y)\|Q(X|Y)) = \sum_y \sum_x P(y) P(x|y) \log \frac{P(x|y)}{Q(x|y)}$$

(b) $$D_{KL}(P(X,Y)\|Q(X,Y)) = \sum_{x,y} P(x,y) \log \frac{P(x,y)}{Q(x,y)}$$

$$= \sum_x P(x) \sum_y P(y|x) \log \frac{P(x)P(y|x)}{Q(x)Q(y|x)}$$

$$= \sum_x \left(P(x) \sum_y P(y|x) \log \frac{P(x)}{Q(x)} \right) + D_{KL}(P(Y|x)\|Q(Y|x))$$

$$= \sum_x P(x) \log \frac{P(x)}{Q(x)} \times \sum_y P(y|x) \quad (\to 1) \quad + D_{KL}(\quad)$$

$$= D_{KL}(P(X)\|Q(X)) + D_{KL}(P(Y|x)\|Q(Y|x))$$

(c) ~~clearly,~~ $\arg\max_\theta \sum_{i=1}^{m} P_\theta(M^{(i)})$

(c) clearly, $\arg\max_\theta \sum_{i=1}^{m} \log P_\theta(M^{(i)}) = \arg\max_\theta \sum_{i=1}^{m} \frac{1}{m} \log\left(\frac{P_\theta(M^{(i)})}{1/m}\right)$

∵ we are dividing & subtracting constants.

$$\Rightarrow \arg\max_\theta \sum_{i=1}^{m} \log P_\theta(M^{(i)}) = \arg\max_\theta \sum_{i=1}^{m} \hat{P}(M^{(i)}) \log \frac{P_\theta(M^{(i)})}{\hat{P}(M^{(i)})}$$

$$= \arg\max_\theta - D_{KL}(\hat{P} \| P_\theta)$$

$$= \arg\min_\theta D_{KL}(\hat{P} \| P_\theta)$$

~~the remark is also~~

v good remark.

Q5 (a) Code

(b) 16 colors ⇒ 4 bits to completely describe a color.
So $\frac{512 \times 512 \times 1}{2}$ size from $512 \times 512 \times 3$

by a factor of 6.

PS-3 Contd.

Ans (a) Score fⁿ - $\nabla_\theta \log\{p(y;\theta)\}$

$$E[\text{score } f^n] = E_{y\sim p(y;\theta)}\left[\nabla_{\theta'} \log\{p(y;\theta)\}\Big|_{\theta'=\theta}\right]$$

$$= \int_{-\infty}^{\infty} \left(p(y;\theta)\right) \nabla_{\theta'} \log p(y;\theta')\Big|_{\theta'=\theta} dy$$

$$= \int_{-\infty}^{\infty} \frac{p(y;\theta)}{p(y;\theta)} \times 1 \; \nabla_{\theta'} p(y;\theta')\Big|_{\theta'=\theta} dy$$

$$= \nabla_{\theta'} \left(\int_{-\infty}^{\infty} p(y;\theta')\, dy \right)\Big|_{\theta'=\theta}$$

$$= \nabla_{\theta'} (1)\Big|_{\theta'=\theta}$$

$$= 0$$

(b) Fisher information - $I(\theta) = Cov_{y\sim p(y;\theta)}\left[\nabla_{\theta'} \log p(y;\theta')\Big|_{\theta'=\theta}\right]$

↳ covariance matrix of the score fⁿ.

let • $\nabla_{\theta'} \log p(y;\theta')\Big|_{\theta'=\theta} = S$

$$I(\theta) = Cov_{y\sim p(y;\theta)}[S]$$

$$= E_{y\sim p(y;\theta)}\left[(S - E_{y\sim p(y;\theta)}[S])(S - E_{y\sim p(y;\theta)}[S])^T \right]$$

$$I(\theta) = E_{y\sim p(y;\theta)}[(s)(s)^T] \quad (\because E[s]=0)$$

$$\Rightarrow I(\theta) = E_{y\sim p(y;\theta)}\left[\nabla_{\theta'} \log p(y;\theta') \times \nabla_{\theta'} \log p(y;\theta')^T \Big|_{\theta'=\theta}\right]$$

(c) $$\left[\nabla_{\theta'} \log p(y;\theta') \nabla_{\theta'} \log p(y;\theta')^T \Big|_{\theta'=\theta}\right]_{ij} = \frac{\partial \log p(y;\theta')}{\partial \theta_i'} \times \frac{\partial \log p(y;\theta')}{\partial \theta_j'}\Big|_{\theta'=\theta}$$

$$= \frac{1}{(p(y;\theta))^2} \times \left(\frac{\partial p(y;\theta')}{\partial \theta_i} \times \frac{\partial p(y;\theta')}{\partial \theta_j}\right)\Big|_{\theta'=\theta}$$

$$= \frac{1}{p(y;\theta)^2} \times \frac{\partial^2 p(y;\theta')}{\partial \theta_i' \partial \theta_j'}\Big|_{\theta'=\theta}$$

by Clairaut's thm.

now,

note that we cannot split this because log is not twice differentiable.

$$\left[-\nabla_{\theta'}^2 \log p(y;\theta')\Big|_{\theta=\theta'}\right]_{ij} = -\frac{\partial^2 \{\log p(y;\theta')\}}{\partial \theta_i' \partial \theta_j'}\Big|_{\theta'=\theta}$$

$$= -\frac{\partial}{\partial \theta_i'}\left(\frac{1}{p(y;\theta')} \times \frac{\partial p(y;\theta')}{\partial \theta_j'}\right)\Big|_{\theta'=\theta}$$

$$\frac{1\times -1}{(p(y;\theta'))^2} \times$$

$$= \left(\frac{-1\times -1}{p(y;\theta')^2} \times \frac{\partial^2 p(y;\theta')}{\partial \theta_i' \partial \theta_j'} - \frac{1}{p(y;\theta')} \times \frac{\partial^2 p(y;\theta')}{\partial \theta_i' \partial \theta_j'}\Big|_{\theta'=\theta}\right.$$

$$\Rightarrow \left[-\nabla^2_{\theta'} \log p(y;\theta')\big|_{\theta'=\theta}\right]_{ij} = \frac{1}{p(y;\theta)^2}\frac{\partial^2 p(y;\theta)}{\partial\theta_i \partial\theta_j} - \frac{1}{p(y;\theta)}\frac{\partial^2 p(y;\theta')}{\partial\theta_i \partial\theta_j}\Big|_{\theta'=\theta}$$

$$E_{y\sim p(y;\theta)}\left[-\nabla^2_{\theta'} \log p(y;\theta')\big|_{\theta'=\theta}\right]_{ij} = E_{y\sim p(y;\theta)}\left[\frac{1}{p(y;\theta)^2}\frac{\partial^2(\cdots)}{\partial\theta_i\partial\theta_j}\Big|_{\theta'=\theta}\right]_{ij}$$

$$- \int_{-\infty}^{\infty} \frac{p(y;\theta)}{p(y;\theta)}\frac{\partial^2 p(y;\theta')}{\partial\theta_i\partial\theta_j}\Big|_{\theta'=\theta} dy$$

↳ this is $= 0$.

$$= E_{y\sim p(y;\theta)}\left[\frac{1}{p(y;\theta)^2}\frac{\partial^2 p(y;\theta)}{\partial\theta_i\partial\theta_j}\Big|_{\theta'=\theta}\right]_{ij}$$

∴ proved.

(d) $$D_{KL}(p_\theta \| p_{\theta+d}) = \int_{-\infty}^{\infty} p(y;\theta) \log \frac{p(y;\theta)}{p(y;\theta+d)}\, dy$$

if $\tilde{\theta} = \theta + d$, then Taylor Series helps us approximate it-

$$f(\tilde{\theta}) = f(\theta) + (\tilde{\theta}-\theta)^T \nabla_{\theta'} f(\theta')\big|_{\theta'=\theta} + \frac{1}{2}(\tilde{\theta}-\theta)^T \nabla^2_{\theta'} f(\theta')\big|_{\theta'=\theta}(\tilde{\theta}-\theta)$$

now,

$$D_{KL}(p_\theta \| p_{\theta+d}) = E_{y\sim p(y;\theta)}[p(y;\theta)] - E_{y\sim p(y;\theta)}[p(y;\tilde{\theta})]$$

$= E_{y\sim p(y;\theta)}[p(y;d]$

$- \left(E_{y\sim p(y;\theta)}[p(y;\theta)] + d^T E_{y\sim p(y;\theta)}[\nabla_{\theta'} \right.$

$\approx E_{y\sim p(y;\theta)}[p(y;\theta)] - E_{y\sim p(y;\theta)}[p(y;\theta)]$

Applying the Taylor series expansion $\hat{\theta} - \theta = d$

$- E_{y\sim p(y;\theta)}[d^T \nabla_{\theta'} p(y;\theta')|_{\theta'=\theta}]$

$- \frac{1}{2} E_{y\sim p(y;\theta)}[d^T \nabla^2_{\theta'} p(y;\theta')|_{\theta'=\theta} d]$

→ is 0 ! part (a).

$\approx -d^T E_{y\sim p(y;\theta)}[\nabla_{\theta'} p(y;\theta')|_{\theta'=\theta}]$

$- \frac{1}{2} d^T E_{y\sim p(y;\theta)}[\nabla^2_{\theta'} p(y;\theta')|_{\theta'=\theta}] d$

$\approx \frac{1}{2} d^T I(\theta) d$

(e) $d^* = \arg\max_d l(\theta + d)$

subject to $D_{KL}(p_\theta || p_{\theta+d}) = c$

Langrangian -

$\mathcal{L}(d, \lambda) = l(\theta) + d^T \nabla_{\theta'} l(\theta)|_{\theta'=\theta} - \lambda[\frac{1}{2} d^T I(\theta) d - c]$

$$\Rightarrow \mathcal{L}(d,\lambda) \approx \log p(y;\theta) + d^T \nabla_{\theta'} \log p(y;\theta')\big|_{\theta'=\theta} - \lambda\left(\frac{1}{2} d^T \mathcal{I}(\theta) d - c\right)$$

$$= \log p(y;\theta) + d^T \nabla_{\theta'} \frac{p(y;\theta')\big|_{\theta'=\theta}}{p(y;\theta)} - \lambda\left(\frac{1}{2} d^T \mathcal{I}(\theta) d - c\right)$$

Linear Systems -

$$\nabla_d \mathcal{L}(d,\lambda) = 0 + \nabla_{\theta'} \frac{p(y;\theta')\big|_{\theta'=\theta}}{p(y;\theta)} - \frac{\lambda}{2} \mathcal{I}(\theta) d \times 2$$

$$= 0$$

$$\Rightarrow \nabla_{\theta'} p(y;\theta')\big|_{\theta'=\theta} = p(y;\theta) \times \lambda \mathcal{I}(\theta) d \quad \hookrightarrow (a)$$

$$\nabla_\lambda \mathcal{L}(d,\lambda) = 0 + 0 - \left(\frac{1}{2} d^T \mathcal{I}(\theta) d - c\right) = 0$$

$$\Rightarrow c = \frac{1}{2} d^T \mathcal{I}(\theta) d \rightarrow (b)$$

from (a) -

$$\tilde{d} = \frac{1}{p(y;\theta)\lambda} \mathcal{I}(\theta)^{-1} \nabla_{\theta'} p(y;\theta')\big|_{\theta'=\theta}$$

from (b) & $\hat{d}$ -

$$\frac{1}{2}\times\left(\frac{1}{p(y;\theta)\lambda}\right)^2\times\left(\mathcal{I}(\theta)^{-1}\nabla_{\theta'}p(y;\theta')|_{\theta'=\theta}\right)^T\mathcal{I}(\theta)\times\left(\mathcal{I}(\theta)^{-1}\nabla_{\theta'}p(y;\theta')|_{\theta'=\theta}\right)=c$$

$$\frac{1}{2\lambda^2 p(y;\theta)^2}\times\left(\nabla_{\theta'}p(y;\theta')|_{\theta'=\theta}\right)^T\mathcal{I}(\theta)^{-1}\nabla_{\theta'}p(y;\theta')|_{\theta'=\theta}=c$$

$$\lambda=\sqrt{\frac{1}{2cp(y;\theta)^2}\nabla_{\theta'}p(y;\theta')|_{\theta'=\theta}^T\mathcal{I}(\theta)^{-1}\nabla_{\theta'}p(y;\theta')|_{\theta'=\theta}}$$

Calculating d^* -

$$d^*=\frac{\sqrt{2c}}{\sqrt{\nabla_{\theta'}p(y;\theta')|_{\theta'=\theta}^T\mathcal{I}(\theta)^{-1}\nabla_{\theta'}p(y;\theta')|_{\theta'=\theta}}}$$

$$d^*=\sqrt{\frac{2c}{\nabla_{\theta'}p(y;\theta')|_{\theta'=\theta}^T\mathcal{I}(\theta)^{-1}\nabla_{\theta'}p(y;\theta)|_{\theta'=\theta}}\mathcal{I}(\theta)^{-1}\nabla_{\theta'}p(y;\theta')|_{\theta'=\theta}}$$

(q) for GLMs,

~~$p(y;\theta)$~~

$$\theta^{(t+1)} = \theta^{(t)} - H^{-1} \nabla_\theta \ell(\theta)$$

Newton's method.

$$I(\theta) = -E_{y\sim p(y;\theta)}\left[+\nabla^2_{\theta'} p(y;\theta')\big|_{\theta'=\theta}\right]$$

$$= -E_{y\sim p(y;\theta)}\left[+\nabla^2_\theta \ell(\theta)\right]$$

$$= -E_{y\sim p(y;\theta)}[H]$$

$\Rightarrow$ using $\hat{d} = \frac{1}{\lambda} I(\theta)^{-1} \nabla_{\theta'} \log p(y;\theta')\big|_{\theta'=\theta}$

$$\tilde{\theta} = \theta + \hat{d}$$

$$= \theta + \frac{1}{\lambda} \times (-1)\, E_{y\sim p(y;\theta)}[H]^{-1} \nabla_\theta \ell(\theta)$$

$$\boxed{\tilde{\theta} = \theta - \frac{1}{\lambda} E_{y\sim p(y;\theta)}[H]^{-1} \nabla_\theta \ell(\theta)}$$

Same!

PS-3 Contd

Q4 $\ell_{sup}(\theta) = \sum_{i=1}^{\tilde{m}} \log p(\tilde{x}^{(i)}, \tilde{z}^{(i)}; \theta)$

$\ell_{semi-sup}(\theta) = \ell_{unsup}(\theta) + \alpha \ell_{sup}(\theta)$

$\ell_{unsup}(\theta) = \sum_{i=1}^{m} \log p(x^{(i)}; \theta)$

$= \sum_{i=1}^{m} \log \sum_{z^{(i)}} p(x^{(i)}, z^{(i)}; \theta)$

for Estep –

~~$\ell_{semi-sup}(\theta) = \sum_{i=1}^{m} \log Q_i(z^{(i)})$~~

$\ell_{semi-sup}(\theta) = \sum_{i=1}^{m} \log \sum_{z^{(i)}} Q_i(z^{(i)}) \cdot \frac{p(x^{(i)}, z^{(i)}; \theta)}{Q_i(z^{(i)})} + \ell_{sup}(\theta)$

now with the same steps as in the notes, we would apply the Jensen's inequality & try to make the lower bound tight at current θ. Note that to ensure "tightness" it only depends on the first term so we would derive

$$Q_i(z^{(i)}) = p(z^{(i)} | x^{(i)}; \theta)$$

same as in the lecture notes.

now with Jensen's inequality,

$$\ell_{semi-sup}(\theta) \geq \sum_{i=1}^{m} \sum_{z^{(i)}} Q_i(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)}; \theta)}{Q_i(z^{(i)})} + \alpha \ell_{sup}(\theta)$$

⇒ for the M-step we will update

$$\theta := \arg\max_{\theta} \left(\sum_{i=1}^{m} \sum_{z^{(i)}} Q_i(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)}; \theta)}{Q_i(z^{(i)})} + \alpha \sum_{i=1}^{\tilde{m}} \log p(\tilde{x}^{(i)}, \tilde{z}^{(i)}; \theta) \right)$$

(a) $$l_{\text{semi-sup}}(\theta^{(t+1)}) = \sum_{i=1}^{m} \log \sum_{z^{(i)}} Q_i(z^{(i)}) \times \frac{p(x^{(i)}, z^{(i)}; \theta^{(t+1)})}{Q_i(z^{(i)})} + \alpha \sum_{i=1}^{\tilde{m}} \log p(\tilde{x}^{(i)}, \tilde{z}^{(i)}; \theta^{(t+1)})$$

$$\geq \sum_{i=1}^{m} \sum_{z^{(i)}} Q_i(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)}; \theta^{(t+1)})}{Q_i(z^{(i)})} + \alpha \sum_{i=1}^{\tilde{m}} \log p(\tilde{x}^{(i)}, \tilde{z}^{(i)}; \theta^{(t+1)})$$

(Jensen's inequality)

now, $\because \theta^{(t+1)} := \arg\max_{\theta} \left(l_{\text{semi-sup}}(\theta^{(t)}) \right)$

$$l_{\text{semi-sup}}(\theta^{(t+1)}) \geq \sum_{i=1}^{m} \sum_{z^{(i)}} Q_i(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)}; \theta^{(t)})}{Q_i(z^{(i)})} + \alpha \sum_{i=1}^{\tilde{m}} \log p(\tilde{x}^{(i)}, \tilde{z}^{(i)}; \theta^{(t)})$$

$$l_{\text{semi-sup}}(\theta^{(t+1)}) \geq l_{\text{semi-sup}}(\theta^{(t)})$$

note that the first term in the last eqn. is exactly equal to $l_{unsup}(\theta^{(t)})$ ∵ of our choice of $Q_i(z^{(i)})$ to make the lowerbound tight.

(b) $Q_i(z^{(i)}) = p(z^{(i)} | x^{(i)}; \theta)$ ← re-estimation of the latent z variables

from given info ↑ , i.e., E-step only involves

here $\theta = \{\mu, \Sigma, \phi\}$

$$\Rightarrow Q_i(z^{(i)}=j) = p(z^{(i)}=j | x^{(i)}; \mu, \Sigma, \phi)$$

$$= \frac{p(x^{(i)} | z^{(i)}=j; \mu, \Sigma) \times p(z^{(i)}=j | \phi)}{\sum_{j=1}^{k} \{p(x^{(i)} | z^{(i)}=j; \mu, \Sigma) \times p(z^{(i)}=j | \phi)\}}$$

let $Q_i(z^{(i)}=j) = w_j^{(i)}$.

(c) M step - (parameters re-estimated include Σ, μ, ϕ)

$$l_{semi\text{-}sup}(\theta^{(t)}) = \sum_{i=1}^{m} \sum_{j=1}^{k} w_j^{(i)} \log \frac{p(x^{(i)} | z^{(i)}=j; \mu, \Sigma) \times p(z^{(i)}=j | \phi)}{w_j^{(i)}}$$

$$+ \alpha \sum_{i=1}^{\tilde{m}} \log \{p(\tilde{x}^{(i)} | \tilde{z}^{(i)}=j; \mu, \Sigma) \times p(\tilde{z}^{(i)}=j | \phi)\}$$

terms involving ϕ,

$$\mathcal{L}(\phi) = \sum_{i=1}^{m} \sum_{j=1}^{k} w_j^{(i)} \log p(z^{(i)}=j | \phi) + \alpha \sum_{i=1}^{\hat{m}} \log p(\hat{z}^{(i)}=j | \phi)$$

$$+ \beta \left(\sum_{i=1}^{k} \phi_i - 1 \right)$$

note that α is not a Lagrange multiplier but β is. ($\because \sum_j \phi_j = 1$)

~~$\nabla_\phi L(\phi) =$~~

$$\frac{\partial L(\phi)}{\partial \phi_j} = \sum_{i=1}^{m} w_j^{(i)} \times \frac{1}{\phi_j} + \alpha \times \sum_{i=1}^{\tilde{m}} \frac{I\{\tilde{z}^{(i)} = j\}}{\phi_j} + \beta$$

$$= 0$$

$$\frac{1}{\phi_j}\left(\sum_{i=1}^{m} w_j^{(i)} + \alpha \sum_{i=1}^{\tilde{m}} I\{\tilde{z}^{(i)} = j\}\right) = -\beta$$

$$\phi_j = \left(\sum_{i=1}^{m} w_j^{(i)} + \alpha \sum_{i=1}^{\tilde{m}} I\{\tilde{z}^{(i)} = j\}\right) / (-\beta)$$

applying $\sum_{j=1}^{k}(\cdot)$ on both sides,

$$1 = \left(\sum_{i=1}^{m}(1) + \alpha \sum_{i=1}^{\tilde{m}}(1)\right) / (-\beta)$$

$$\boxed{-\beta = m + \alpha \tilde{m}}$$

$$\Rightarrow \phi_j = \frac{\sum_{i=1}^{m} w_j^{(i)} + \alpha \sum_{i=1}^{\tilde{m}} I\{\tilde{z}^{(i)} = j\}}{m + \alpha \tilde{m}}$$

terms involving μ_j,

$$\mathcal{L}(\mu) = \sum_{i=1}^{m}\sum_{j=1}^{k} w_j^{(i)} \left(\log p(x^{(i)} | z^{(i)}=j; \mu, \Sigma) - \log w_j^{(i)} \right) + \alpha \sum_{i=1}^{\tilde{m}} \log \left(p(\tilde{x}^{(i)} | \tilde{z}^{(i)}; \mu, \Sigma) \right)$$

$$= \sum_{i=1}^{m}\sum_{j=1}^{k} \{ w_j^{(i)} \log p(x^{(i)} | z^{(i)}=j; \mu, \Sigma) \} + \alpha \sum_{i=1}^{\tilde{m}} \log p(\tilde{x}^{(i)} | \tilde{z}^{(i)}; \mu, \Sigma)$$

$$\frac{\partial \mathcal{L}(\mu)}{\partial \mu_j} = \Sigma_j^{-1} \left[\sum_{i=1}^{m} w_j^{(i)} x^{(i)} - \mu_j \sum_{i=1}^{m} w_j^{(i)} + \alpha \left(\sum_{i=1}^{\tilde{m}} I\{z^{(i)}=j\} \tilde{x}^{(i)} - \mu_j \sum_{i=1}^{\tilde{m}} I\{z^{(i)}=j\} \right) \right]$$

$$= 0$$

$$\Rightarrow \mu_j = \frac{\sum_{i=1}^{m} w_j^{(i)} x^{(i)} + \alpha \sum_{i=1}^{\tilde{m}} I\{z^{(i)}=j\} \tilde{x}^{(i)}}{\sum_{i=1}^{m} w_j^{(i)} + \alpha \sum_{i=1}^{m} I\{z^{(i)}=j\}}$$

similarly, $\Sigma_j = \dfrac{\sum_{i=1}^{m} w_j^{(i)} (x^{(i)} - \mu_j)(x^{(i)} - \mu_j)^T + \alpha \sum_{i=1}^{\tilde{m}} I\{z^{(i)}=j\} (\cdots)}{\sum_{i=1}^{m} w_j^{(i)} + \alpha \sum_{i=1}^{\tilde{m}} I\{\tilde{z}^{(i)}=j\}}$

Code

Code

(SSEM)

(i) Semi-Supervised EM took much less iterations ~50
Un- " EM ~ 1000
(USEM)

(ii) SSEM was very stable, even with different initializati
I got the same plots.
USEM was not very stable, although some pts. were
classified to the same class, the class itself changed

(iii) SSEM ~~was~~ is ~~correct~~ almost correct if not
completely correct, acc. to given info..
USEM is not very accurate to the give info.