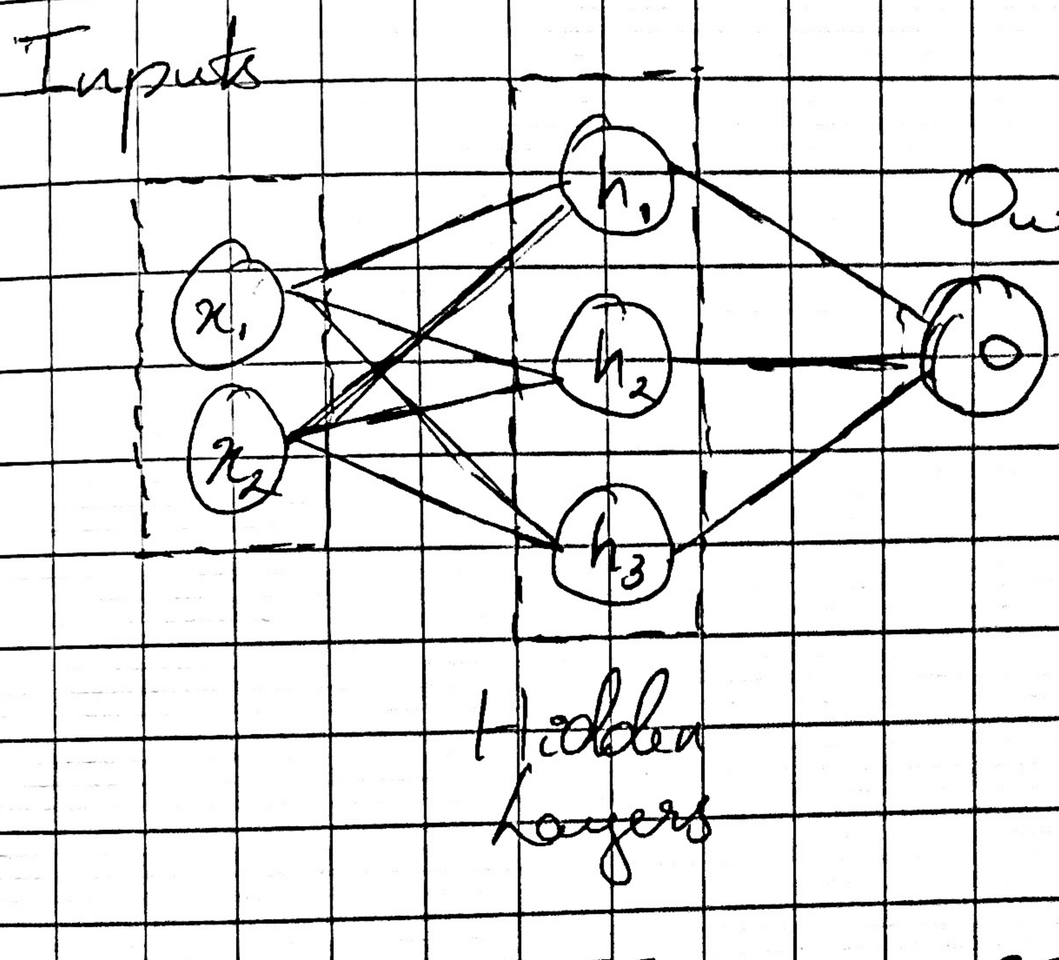
CS 229 Fall 2018

Problem Set #3: Deep learning & Unsupervised learning



$$h_j^{(i)} = g(w_{1,j}^{[1]} x_1^{(i)} + w_{2,j}^{[1]} x_2^{(i)} + w_{0,j}^{[1]})$$

$$O^{(i)} = g(w_1^{[2]} h_1^{(i)} + w_2^{[2]} h_2^{(i)} + w_3^{[2]} h_3^{(i)} + w_0^{[2]})$$

where g is the sigmoid function

a) $$\frac{\partial l}{\partial w_{1,2}^{[1]}} = \frac{\partial h_2}{\partial w_{1,2}^{[1]}} \cdot \frac{\partial O}{\partial h_2} \cdot \frac{\partial l}{\partial O}$$

$$= \sum_{i=0}^{m} x_1^{(i)} h_2^{(i)} (1 - h_2^{(i)}) \cdot w_2^{[2]} O^{(i)} (1 - O^{(i)}) \cdot \frac{2}{m} (O^{(i)} - y^{(i)})$$

where $h_2^{(i)} = g(w_{1,2}^{[1]} x_1^{(i)} + w_{2,2}^{[1]} x_2^{(i)} + w_{0,2}^{[1]})$

With the update rule of gradient descent we get

$$w_{1,2}^{[1]} := w_{1,2}^{[1]} - \alpha \frac{\partial l}{\partial w_{1,2}^{[1]}}$$

where α is the learning rate.

b) Each neuron in the hidden layer will check the position of the point $x^{(i)}$ with respect to each edge of the triangle.

$w_{0,1}^{[1]}$ A

$w_{1,1}^{[1]} = -1$

$w_{2,1}^{[1]} = -1$

$w_{0,2}^{[1]} = -0,7$

$w_{1,2}^{[1]} = 1$

$w_{2,2}^{[1]} = 0$

$w_{0,3}^{[1]} = -0,5$

$w_{1,3}^{[1]} = 0$

$w_{2,3}^{[1]} = 1$

$w_0^{[2]} = -3$

$w_1^{[2]} = 1$

$w_2^{[2]} = 1$

$w_3^{[2]} = 1$

With the output layer we can verify if all the three conditions were met.

c) With the hidden layer we don't have any information gain.

2 KL divergence and Maximum likelihood

$$D_{KL}(P \| Q) = \sum_{x \in X} P(x) \log\left(\frac{P(x)}{Q(x)}\right)$$

a) Prove that $\forall P, Q$, $D_{KL}(P \| Q) \geq 0$ and

$D_{KL}(P \| Q) = 0$ if and only if $P = Q$

$$\Rightarrow D_{KL}(P \| Q) = \sum_{x \in X} P(x) \log\left(\frac{P(x)}{Q(x)}\right)$$

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

Jenson's Inequality

$$\geq -\sum_{x \in X} \log(Q(x))$$

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

$$\geq 0$$

$$D_{KL}(P \| Q) = 0$$

According to the inequality of Jensen

$$-\sum_{x \in X} P(x) \log\left(\frac{P(x)}{Q(x)}\right) = \sum \log\left(\frac{P^2(x)}{Q(x)}\right)$$

If and only if $\frac{P(x)}{Q(x)}$ is constant

and since $P(x) > 0$ $D_{KL}(P \| Q) = 0$ if

and only if $P(x) = Q(x)$.

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

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

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

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

$$\sum_y \sum_x p(x,y) \log\left(\frac{P(y,x)}{Q(y,x)}\right)$$

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

c) Prove that $\operatorname{argmin}_\theta D_{KL}(\hat{P} \| P_\theta) = \operatorname{argmax}_\theta \sum_{i=1}^{m} \log P_\theta(x^{(i)})$

$$\operatorname{argmin}_\theta D_{KL}(\hat{P} \| P_\theta) = \sum_{x \in X} \hat{P}(x) \log\left(\frac{\hat{P}(x)}{P_\theta(x)}\right)$$

$$= \operatorname{argmin}_\theta \sum_{x \in X} \hat{P}(x) \log(\hat{P}(x)) - \sum_{x \in X} \hat{P}(x) \log(P_\theta(x))$$

$$= \operatorname{argmin}_\theta - \sum_{x \in X} \hat{P}(x) \log(P_\theta(x))$$

$$= \operatorname{argmax}_\theta \sum_{x \in X} \frac{1}{m} \sum_{i=1}^{m} 1\{x^{(i)} = x\} \log(P_\theta(x))$$

$$= \operatorname{argmax}_\theta \sum_{i=1}^{m} \log P_\theta(x^{(i)})$$

3) KL Divergence, Fisher Information and the Natural Gradient

$$D_{KL}(p||q) = \int_{-\infty}^{\infty} p(x) \log\left(\frac{p(x)}{q(x)}\right) dx$$

$$= E_{x\sim p(x)}[\log(p(x))] - E_{x\sim p(x)}[\log(q(x))]$$

a) The score function of $p(y;\theta)$ is $\nabla_\theta \log p(y;\theta)$

Show that the expected value of the score function is 0

$$E_{y\sim p(y;\theta)}[\nabla_{\theta'} \log p(y;\theta') | \theta' = \theta] = 0$$

$$* E_{y\sim p(y;\theta)}[\nabla_{\theta'} \log p(y;\theta') | \theta' = \theta] = \int_{-\infty}^{+\infty} p(y;\theta) \nabla_{\theta'} \log p(y;\theta) \, dy$$

$$= \int \frac{p(y;\theta) \times 1}{p(y;\theta)} \times \nabla_\theta p(y;\theta) \, dy$$

$$= \int \nabla_\theta p(y;\theta) dy = \nabla_\theta \int p(y,\theta) dy = \nabla_\theta (1) = 0$$

b) Fisher Information

$$I(\theta) = Cov_{y\sim p(y;\theta)}\left[\nabla_{\theta'} \log p(y;\theta') \mid \theta' = \theta\right]$$

Show that the Fisher Information can be written as:

$$I(\theta) = E_{y\sim p(y;\theta)}\left[\nabla_{\theta'} \log (p(y;\theta)) \nabla_{\theta'} \log (p(y;\theta'))^T \mid \theta' = \theta\right]$$

$$* \; Cov_{y\sim p(y;\theta)}\left[\nabla_{\theta'} \log p(y;\theta') \mid \theta' = \theta\right]$$

$$= E\left(\left[\nabla_\theta \log p(y;\theta) - E(\nabla_\theta \log p(y;\theta))\right]\right[$$

$$\nabla_\theta \log p(y;\theta) - E(\nabla_\theta \log p(y;\theta))^T\left]\right]$$

$$= E\left(\nabla_\theta \log (p(y;\theta))\, \nabla_\theta \log (p(y;\theta))^T\right.$$

c) $E_{y\sim p(y;\theta)}\left[-\nabla_\theta^2 \log p(y;\theta)\right]_{ij}$

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

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

Approximation of D_{KL} with Fisher Information

$$D_{KL}(p_\theta \| p_{\theta+d}) = \mathbb{E}_{y\sim p(y;\theta)}[\log p(y;\theta)] - \mathbb{E}_{y\sim p(y;\theta)}[\log p(y;\bar{\theta})]$$

Taylor Series of $\log(p(y;\theta))$

$$\Rightarrow \log(p(y;\bar{\theta})) = \log p(y;\theta) + d^T \nabla_\theta \log p(y;\theta) + \frac{1}{2} d^T \nabla_\theta^2 \log(p(y;\theta))\, d$$

$$= \log p(y;\theta) + d^T \frac{\nabla_\theta p(y;\theta)}{p(y;\theta)} + \frac{1}{2} d^T \nabla_\theta^2 \log p(y;\theta)\, d$$

$$\Rightarrow \mathbb{E}_{y\sim p(y;\theta)}[\log p(y;\bar{\theta})] = -\mathbb{E}_{y\sim p(y;\theta)}[\log p(y;\theta)] + \mathbb{E}_{y\sim p(y;\theta)}\Big[-\frac{1}{2} d^T \nabla_\theta^2 \log p(y;\theta)\, d\Big]$$

$$\Rightarrow D_{KL} = \mathbb{E}_{y\sim p(y;\theta)}\Big[-\frac{1}{2} d^T \nabla_\theta^2 \log p(y;\theta)\, d\Big]$$

$$= \frac{1}{2} d^T \mathbb{E}_{y\sim p(y;\theta)}[-\nabla_\theta^2 \log p(y;\theta)]\, d = \frac{1}{2} d^T I(\theta) d$$

4 Semi-supervised EM

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

* $l_{sup}(\theta) = \sum_{i=1}^{\bar{m}} \log \left(p(\bar{x}^{(i)}, \bar{z}^{(i)}; \theta) \right)$

* E-step : $Q_i^{(t)}(z^{(i)}) := p(z^{(i)} | x^{(i)}; \theta^{(t)})$

* M-step : $\theta^{(t+1)} := \arg\max_{\theta} \left[\sum_{i=1}^{m} \left(\sum_{z^{(i)}} Q_i^{(t)}(z^{(i)}) \log \left(\frac{p(x^{(i)}, z^{(i)}; \theta)}{Q_i^{(t)}(z^{(i)})} \right) \right) + \alpha \left(\sum_{i=1}^{\bar{m}} \log p(\bar{x}^{(i)}, \bar{z}^{(i)}; \theta) \right) \right]$

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

$$l_{semi\text{-}sup}(\theta^{t+1}) = \sum_{i=1}^{m} \log \sum_{z^{(i)}} p(x^{(i)}, z^{(i)}; \theta^{(t+1)}) + \alpha \sum_{i=1}^{\bar{m}} \log \left(p(\bar{x}^{(i)}, \bar{z}^{(i)}, \theta^{(t+1)}) \right)$$

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

Jensen's inequality $\geq \sum_{i=1}^{m} \left(\sum_{z^{(i)}} Q_i^{(t+1)}(z^{(i)}) \log \left(\frac{p(x^{(i)}, z^{(i)}, \theta^{t+1})}{Q_i^{t+1}(z^{(i)})} \right) \right) + \alpha \sum_{i=1}^{\bar{m}} \log p(\bar{x}^{(i)}, \bar{z}^{(i)}, \theta^{(t+1)})$

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

$$\geq \sum_{i=1}^{m} \left(\sum_{z^{(i)}} Q_i^{(t)}(z^{(i)}) \log \left(p(x^{(i)}, z^{(i)}, \theta^{(t)}) \right) \right) + \alpha \sum_{i=1}^{\bar{m}} \log \left(p(\bar{x}^{(i)}, \bar{z}^{(i)}, \theta^{(t)}) \right)$$

$$= l_{unsup}(\theta^{(t)}) + \alpha l_{sup}(\theta^{(t)})$$

$$= l_{semi\text{-}sup}(\theta^{(t)})$$

b) Semi-supervised E-Step:

$$Q_j^{(i)} = p(z_j^{(i)} | x^{(i)}; \phi, \mu, \Sigma)$$

$$= \frac{p(x^{(i)} | z_j^{(i)}; \mu, \Sigma)\, p(z_j^{(i)}; \phi)}{\sum_{l=1}^{k} p(x^{(i)} | z_l^{(i)}, \mu, \Sigma)\, p(z_l^{(i)}; \phi)}$$

$$= \frac{\frac{1}{(2\pi)^{n/2} |\Sigma_j|^{1/2}} \exp\left(-\frac{1}{2}(x^{(i)} - \mu_j)^T \Sigma_j^{-1} (x^{(i)} - \mu_j)\right) \phi_j}{\sum_{l=1}^{k} \frac{1}{(2\pi)^{n/2} |\Sigma_l|^{1/2}} \exp\left(-\frac{1}{2}(x^{(i)} - \mu_l)^T \Sigma_l^{-1} (x^{(i)} - \mu_l)\right) \phi_l}$$

Semi-Supervised learning

M-Step:

$$L = L_{unsup} + L_{sup}$$

$$= \sum_{i=1}^{m} \sum_{j=1}^{k} w_j^{(i)} \log\left(p(x^{(i)} | z_j; \theta) p(z_j)\right) +$$

$$\alpha \sum_{i=1}^{\bar{m}} \sum_{j=1}^{k} 1\{\bar{z}^{(i)} = j\} \log(p(\bar{x}^{(i)} | \bar{z}_j) p(\bar{z}_j)) + \beta\left(1 - \sum_{j=1}^{k} \phi_j\right)$$

$$\frac{\partial L}{\partial \phi_j} =$$

$$\Rightarrow \frac{\partial L}{\partial \phi_j} = \sum_{i=1}^{m} \frac{w_j^{(i)}}{\phi_j} + \alpha \sum_{i=1}^{\bar{m}} \frac{1\{\bar{z}^{(i)} = \}}{\phi_j} - \beta = 0$$

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

$$\sum_{j=1}^{k} \phi_j = 1 = \sum_{i=1}^{m} \sum_{j=1}^{k} w_j^{(i)} + \alpha \sum_{i=1}^{\bar{m}} \sum_{j=1}^{k} 1\{\bar{z}^{(i)} = j\}$$

$$\Rightarrow D\beta = m + \alpha \bar{m}$$

* μ_j

$$L_{unsup} = \sum_{i=1}^{m} \sum_{j=1}^{k} w_j^{(i)} \log\left(p(z_j)\, p(x^{(i)} | z_j; \theta)\right)$$

$$= \sum_{i=1}^{m} \sum_{j=1}^{k} w_j^{(i)} \log(\phi_j) + \log\left(p(x^{(i)} | z_j, \theta)\right)$$

$$\frac{\partial L_{unsup}}{\partial \mu_j} = \sum_{i=1}^{m} w_j^{(i)} \Sigma_j^{-1} (x^{(i)} - \mu_j)$$

$$\Rightarrow D \frac{\partial L_{sup}}{\partial \mu_j} = \sum_{i=1}^{\bar{m}} 1\{\bar{z}^{(i)} = j\} \Sigma_j^{-1} (\bar{x}^{(i)} - \mu_j)$$

$$\Rightarrow D \frac{\partial L_{semi-sup}}{\partial \mu_j} = \frac{\partial L_{sup}}{\partial \mu_j} + \frac{\partial L_{unsup}}{\partial \mu_j}$$

$$= \Sigma^{-1} \left[\sum_{i=1}^{m} w_j^{(i)} x^{(i)} + \alpha \sum_{i=1}^{\bar{m}} 1\{\bar{z}^{(i)} = j\} \bar{x}^{(i)} - \mu_j \left(\sum_{i=1}^{m} w_j^{(i)} + \alpha \sum_{i=1}^{\bar{m}} 1\{\bar{z}^{(i)} = j\} \right) \right]$$

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

Likewise:

$$* \frac{\partial L_{unsup}}{\partial \Sigma_j} = -\frac{1}{2} \sum_{i=1}^{m} w_j^{(i)} \Sigma_j^{-1} + \frac{1}{2} \Sigma_j^{-1} \left(\sum_{i=1}^{m} w_j^{(i)} (x^{(i)} - \mu_j)(x^{(i)} - \mu_j)^T \right) \Sigma_j^{-1}$$

$$* \frac{\partial L_{sup}}{\partial \Sigma_j} = -\frac{1}{2} \sum_{i=1}^{\bar{m}} 1\{\bar{z}^{(i)} = j\} \Sigma_j^{-1} + \frac{1}{2} \Sigma_j^{-1} \left(\sum_{i=1}^{\bar{m}} 1\{\bar{z}^{(i)} = j\} (\bar{x}^{(i)} - \mu_j)(\bar{x}^{(i)} - \mu_j)^T \right) \Sigma_j^{-1}$$

$$\Rightarrow \Sigma_j = \frac{\sum_{i=1}^{m} w_j^{(i)} (x^{(i)} - \mu_j)(x^{(i)} - \mu_j)^T + \alpha \sum_{i=1}^{\bar{m}} 1\{\bar{z}^{(i)} = j\} (\bar{x}^{(i)} - \mu_j)(\bar{x}^{(i)} - \mu_j)^T}{\sum_{i=1}^{m} w_j^{(i)} + \alpha \sum_{i=1}^{\bar{m}} 1\{\bar{z}^{(i)} = j\}}$$