CS229 Fall 2017
Problem Set #3 Solutions: Deep Learning & Unsupervised Learning

Author: LFhase   rimemosa@163.com

A Simple Neural Network

(a) Using Chain Rule, we know that

$$\frac{\partial loss}{\partial w_{1,2}^{[1]}} = \frac{\partial loss}{\partial o}\frac{\partial o}{\partial h_2}\frac{\partial h_2}{\partial w_{1,2}^{[1]}}$$

let $g(x)$ denote the sigmoid function, then we have

$$g'(x) = g(x)(1 - g(x))$$

so

$$\frac{\partial loss}{\partial w_{1,2}^{[1]}} = \frac{2}{m}\sum_{i=1}^{m}(o^{(i)} - y^{(i)})o^{(i)}(1 - o^{(i)})w_2^{[2]}h_2^{(i)}(1 - h_2^{(i)})x_1^{(i)}$$

where

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

(b) let $(0.5, 0.5)$, $(3.5, 0.5)$, $(0.5, 3.5)$ be the three poinst of the triangle. The forward transport in the neural network can be written in matrix form.

$$\begin{bmatrix} -1.5 & 3 & 0 \\ -1.5 & 0 & 3 \\ 9 & -3 & 3 \end{bmatrix} \times \begin{bmatrix} 1 \\ x_1 \\ x_2 \end{bmatrix}$$

and

$$\begin{bmatrix} -1 & -1 & -1 & 2.33 \end{bmatrix} \times \begin{bmatrix} 1 \\ h_1 \\ h_2 \\ h_3 \end{bmatrix}$$

Once the point is in the triangle, the first product will be

$$\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$$

So the second product will be -0.67 and the final result will be 0. Otherwise, the second product will be larger or equal to 0.33 and the final result will be 1.

(c) Using $f(x) = x$ as hidden layer activation function, we can see the neural network as **a simple neural network without hidden layer**, who only has the **convex boundary** and can't deal with the problem described in statement.

EM for MAP estimation

The whole process is the same like what discussed in lecture notes.
Firstly, we have log-likelihood:

$$\begin{aligned} l(\theta) &= \sum_{i=1}^{m} log[\sum_{z^{(i)}} Q_i(z^{(i)}) \frac{p(x^{(i)}, z^{(i)}|\theta)}{Q_i(z^{(i)})}] + logp(\theta) \\ &\geq \sum_{i=1}^{m} \sum_{z^{(i)}} Q_i(z^{(i)}) log \frac{p(x^{(i)}, z^{(i)}|\theta)}{Q_i(z^{(i)})} + logp(\theta) \end{aligned}$$

So if we set

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

According to Jensen's Inequality, we have

$$l(\theta) = \sum_{i=1}^{m} \sum_{z^{(i)}} Q_i(z^{(i)}) log \frac{p(x^{(i)}, z^{(i)}|\theta)}{Q_i(z^{(i)})} + logp(\theta)$$

Then we get EM-step as below:
E-step:

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

M-step:

$$\theta = argmax_\theta [\sum_{i=1}^{m} \sum_{z^{(i)}} Q_i(z^{(i)}) log \frac{p(x^{(i)}, z^{(i)}|\theta)}{Q_i(z^{(i)})} + logp(\theta)]$$

In our assumption, the M-step is tractable. Then we have

$$\begin{aligned} l(\theta^{(t+1)}) &\geq \sum_{i=1}^{m} \sum_{z^{(i)}} Q_i(z^{(i)}) log \frac{p(x^{(i)}, z^{(i)}|\theta^{(t+1)})}{Q_i(z^{(i)})} + logp(\theta^{(t+1)}) \\ &\geq \sum_{i=1}^{m} \sum_{z^{(i)}} Q_i(z^{(i)}) log \frac{p(x^{(i)}, z^{(i)}|\theta^{(t)})}{Q_i(z^{(i)})} + logp(\theta^{(t)}) \\ &= l(\theta^{(t)}) \end{aligned}$$

The likelihood will increase monotonically with each iteration of the algorithm.

EM application

(a) (i) Since we have $x^{(pr)} = y^{(pr)} + z^{(pr)} + \epsilon^{(pr)}$, then $X \sim N(\mu_p + \nu_r, \sigma^2 + \sigma_p^2 + \tau_r^2)$ So the joint distribution have the mean vector and covariance matrix as below:

$$\begin{bmatrix} \mu_p \\ \nu_r \\ \mu_p + \nu_r \end{bmatrix}$$

and

$$\begin{bmatrix} \sigma_p^2 & 0 & \sigma_p^2 \\ 0 & \tau_r^2 & \tau_r^2 \\ \sigma_p^2 & \tau_r^2 & \sigma^2 + \sigma_p^2 + \tau_r^2 \end{bmatrix}$$

(ii) Using the formula in the notes, we have the mean vector and covariance matrix as below:

$$\mu_Q = \begin{bmatrix} \mu_p \\ \nu_r \end{bmatrix} + \begin{bmatrix} \sigma_p^2 \\ \tau_r^2 \end{bmatrix} \frac{x^{(pr)} - (\mu_p + \nu_r)}{\sigma^2 + \sigma_p^2 + \tau_r^2}$$

and

$$\Sigma_Q = \begin{bmatrix} \sigma_p^2 & 0 \\ 0 & \tau_r^2 \end{bmatrix} - \begin{bmatrix} \sigma_p^2 \\ \tau_r^2 \end{bmatrix} \frac{\begin{bmatrix} \sigma_p^2 & \tau_r^2 \end{bmatrix}}{\sigma^2 + \sigma_p^2 + \tau_r^2}$$

The expression is:

$$Q_{pr}(y^{(pr)}, z^{(pr)}) = \frac{1}{\sqrt{2\pi^2 |\Sigma_Q|}} exp(-\frac{1}{2}(\begin{bmatrix} y^{(pr)} \\ z^{(pr)} \end{bmatrix} - \mu_Q)^T \Sigma_Q^{-1}(\begin{bmatrix} y^{(pr)} \\ z^{(pr)} \end{bmatrix} - \mu_Q))$$

(b) We want to maxmize the lower bound of the log-likelihood function:

$$\begin{aligned}
\Theta &= argmax_\Theta \sum_p \sum_r E_{(y^{(pr)}, z^{(pr)}) \sim Q_{pr}} [logp(x^{(pr)}, y^{(pr)}, z^{(pr)})] \\
&= argmax_\Theta \sum_p \sum_r E[log \frac{1}{2\pi^{3/2} \sigma \sigma_p \tau_r} - \frac{1}{2\sigma_p^2}(y^{(pr)} - \mu_p)^2 - \frac{1}{2\tau_r^2}(z^{(pr)} - \nu_r)^2 \\
&- \frac{1}{2\sigma^2}(x^{(pr)} - y^{(pr)} - z^{(pr)})^2] \\
&= argmax_\Theta \sum_p \sum_r E[log \frac{1}{\sigma_p \tau_r} - \frac{1}{2\sigma_p^2}(y^{(pr)} - \mu_p)^2 - \frac{1}{2\tau_r^2}(z^{(pr)} - \nu_r)^2]
\end{aligned}$$

Then we calculate the derivatives of each parameter and set them to zero to get the update value.

$$\mu_p = \frac{1}{PR} \sum_p \sum_r \mu_{Q1}$$

$$\nu_r = \frac{1}{PR} \sum_p \sum_r \mu_{Q2}$$

$$\sigma_p^2 = \frac{1}{PR}\sum_p\sum_r(\Sigma_{Q_{11}} + \mu_{Q_1}^2 - 2\mu_{Q_1}\mu_p + \mu_p^2)$$

$$\tau_r^2 = \frac{1}{PR}\sum_p\sum_r(\Sigma_{Q_{22}} + \mu_{Q_2}^2 - 2\mu_{Q_2}\nu_r + \nu_r^2)$$

KL divergence and Maximum Likelihood

(a) gg