CS299, Machine Learning: Assignment3

Due on December 1, 2018

Andrew Ng

Bryan Zhang

Contents

Problem 1

First, we will look at the forward pass with m single data sample. I will use matrix instead of element inside the matrix because of being succinct.
X is our data matrix with the shape of $m \times 2$.
$W^{[1]}$ is weight for the hidden layer with the shape of 2×3.
$b^{[1]}$ is the intercept term for the hidden layer with the shape of 1×3.
H is our hidden layer with the shape of $m \times 3$.
$W^{[2]}$ is weight for the hidden layer with the shape of 3×1.
$b^{[2]}$ is the intercept term for the hidden layer with the shape of 1.
O is our output prediction of label matrix with the shape of $m \times 1$.
Y is our ground truth. l is our l2-loss.
For both the both the hidden layer and our output layer, the activation function is sigmoid, which perform sigmoid function element wise on the product matrix. Then,

$$H = sigmoid(XW^{[1]} + b^{[1]})$$

$$O = sigmoid(HW^{[2]} + b^{[2]})$$

[1]

$$l = \frac{1}{m} sum((O - Y)^2)$$

Second, let's look at the back propagation using chain rule.

$$\dot{O} = \frac{2}{m} * (O - Y)$$
$$\dot{H} = \dot{O} \circ (O \circ (1 - O)) {W^{[2]}}^T$$
$$\dot{W}^{[1]} = X^T [\dot{H} \circ H \circ (1 - H)]$$

a

Expanding our matrix derivatives above, we can get,

$$\dot{w}^{[1]}_{1,2} = \sum_i^m x_1^{(i)} \frac{2}{m} (o^{(i)} - y^{(i)}) o^{(i)} (1 - o^{[i]}) w_2^{[2]} (x_1^{(i)} w^{[1]}_{1,2} + x_2^{(i)} w^{[1]}_{2,2})(1 - x_1^{(i)} w^{[1]}_{1,2} - x_2^{(i)} w^{[1]}_{2,2}) \tag{1}$$

$$\begin{aligned} w^{[1]}_{1,2} &= w^{[1]}_{1,2} - \alpha * \dot{w}^{[1]}_{1,2} \\ &= w^{[1]}_{1,2} - \alpha * \sum_i^m x_1^{(i)} \frac{2}{m} (o^{(i)} - y^{(i)}) o^{(i)} (1 - o^{[i]}) w_2^{[2]} (x_1^{(i)} w^{[1]}_{1,2} + x_2^{(i)} w^{[1]}_{2,2})(1 - x_1^{(i)} w^{[1]}_{1,2} - x_2^{(i)} w^{[1]}_{2,2}) \end{aligned} \tag{2}$$

b

From the data points plot, we can observe the dividing boudary: $x_2 = -x_1 + 4$. This means when $x_2 > -x_1 - 4$ the sample label must be one and that $x_2 < -x_1 - 4$ the sample label must be zero.
Besides this nice criteria,

$$H = sign(XW^{[1]} + b^{[1]})$$

[1] $\circ$ means Hadamard Product, or element wise product

Problem 1 continued on next page...

$$O = sign(HW^{[2]} + b^{[2]})$$

With these two properties, we can achieve 100% accuracy. First, let's look samples with label 1, which means our output must be 1. Then $HW^{[2]} + b^{[2]}$ must be positive. We can simplify this by making $W^{[2]} = [1, 0, 0]$, $b^{[2]} = 0$ Since this is also the output of another step function, then only $(XW^{[1]} + b^{[1]})_1$ needs to be positive. In other words, $w^{[1]}_{0,2}, w^{[1]}_{1,2}, w^{[1]}_{2,2}, w^{[1]}_{1,3}, w^{[1]}_{2,3}$, can be arbitrary numbers.

$$(XW^{[1]} + b^{[1]})_1 = x^{(i)}_1 w^{[1]}_{1,1} + x^{(i)}_2 w^{[1]}_{2,1} + w^{[1]}_{0,1} > 0 \tag{3}$$

Compared with our dividing line equation, we can easily conclude that $w^{[1]}_{1,1} = 1, w^{[1]}_{2,1} = 1, w^{[1]}_{0,1} = 4$. Then one set of the weights that can perfectly classify with step function as our activation function would be

$$w^{[1]}_{0,1} = 4, w^{[1]}_{1,1} = 1, w^{[1]}_{2,1} = 1$$

$$w^{[1]}_{0,2} = 0, w^{[1]}_{1,2} = 0, w^{[1]}_{2,2} = 0$$

$$w^{[1]}_{0,3} = 0, w^{[1]}_{1,3} = 0, w^{[1]}_{2,3} = 0$$

$$w^{[2]}_0 = 0, w^{[2]}_1 = 1, w^{[2]}_2 = 0, w^{[2]}_3 = 0$$

C

Yes, for the similar reason as the (b).

$$w^{[1]}_{0,1} = 4, w^{[1]}_{1,1} = 1, w^{[1]}_{2,1} = 1$$

$$w^{[1]}_{0,2} = 0, w^{[1]}_{1,2} = 0, w^{[1]}_{2,2} = 0$$

$$w^{[1]}_{0,3} = 0, w^{[1]}_{1,3} = 0, w^{[1]}_{2,3} = 0$$

$$w^{[2]}_0 = 0, w^{[2]}_1 = 1, w^{[2]}_2 = 0, w^{[2]}_3 = 0$$

Problem 2

The log-likelihood:

$$
\begin{aligned}
l(\theta) &= \sum_{i=1}^{m} \log p(x^{(i)}|\theta) + \log p(\theta) \\
&= \sum_{i=1}^{m} \log \sum_{z^{(i)}} p(x^{(i)}, z^{(i)}|\theta) + \log p(\theta) \\
&= \sum_{i=1}^{m} \log \sum_{z^{(i)}} Q_i(z^{(i)}) \frac{p(x^{(i)}, z^{(i)}|\theta)}{Q_i(z^{(i)})} + \log p(\theta) \\
&\geq \sum_{i=1}^{m} \sum_{z^{(i)}} Q_i(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)}|\theta)}{Q_i(z^{(i)})} + \log p(\theta)
\end{aligned}
\tag{4}
$$

The last step is using Jensen's inequality, since $\log x$ is a concave function.

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

is just an expectation of the quantity $\frac{p(x^{(i)}, z^{(i)}|\theta)}{Q_i(z^{(i)})}$ with respect to $z^{(i)}$. Thus by Jensen's inequality, we have

$$
\log \mathop{\mathbb{E}}_{z^{(i)} \sim Q_i} [\frac{p(x^{(i)}, z^{(i)}|\theta)}{Q_i(z^{(i)})}] + \log p(\theta) \geq \mathop{\mathbb{E}}_{z^{(i)} \sim Q_i} [\log \frac{p(x^{(i)}, z^{(i)}|\theta)}{Q_i(z^{(i)})}] + \log p(\theta)
$$

In above inequality, $\log p(\theta)$ is just a constant with respect to $z^{(i)}$.First,for E-step, let's assume that we have known θ and compute for $Q_i(z^{(i)})$. We want to make the lower bound give by Jensen's inequality to hold tight. To achieve this, we have to make the entity for expectation to be constant instead of a random variable. Then,

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

c doesn't depend on $z^{(i)}$. Then $Q_i(z^{(i)}) \propto p(x^{(i)}, z^{(i)}|\theta)$. Since

$$
\sum_{z^{(i)}} Q_i(z^{(i)}) = \sum_{z^{(i)}} c \times p(x^{(i)}, z^{(i)}|\theta) = 1
$$

. Thus,

$$
\begin{aligned}
Q_i(z^{(i)}) &= \frac{p(x^{(i)}, z^{(i)}|\theta)}{\sum_{z^{(i)}} p(x^{(i)}, z^{(i)}|\theta)} \\
&= \frac{p(x^{(i)}, z^{(i)}|\theta)}{p(x^{(i)}|\theta)} \\
&= p(z^{(i)}|\theta, x^{(i)})
\end{aligned}
\tag{5}
$$

Thus, Q_i is set to be the posterior distribution of the $z^{(i)}$ given $x^{(i)}$ and θ. Second, for the M-step, we will maximize the lower bound with respect to θ given the Q_i calculated form the E-step.
M-step: find

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

Finally, repeat E-step and M-step iteratively until convergence.

a: Prove M-step is Tractable

To prove M-step is tractable is to prove that above optimization can be achieved in a polynomial time respect to the sample size. I will use a somewhat self-evident conclusion that all convex optimization problems are tractable. Then, I only need to show the problem belongs to convex optimization.
Since we are maximizing, I only need to show that each log function in that linear combinations is concave for θ. This is a assumption given by the problem.
q.e.d. Thus, we can prove M-step, a MAP estimation with x and z observed, is tractable.

b: Prove the Log likelihood Increase monotonically with each iteration

To prove this statement, we only need to compare log-likelihood at t step and t+1 step. After t step,

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

In step t, $\theta^{(t+1)}$ is already computed
After t+1 step,

$$\begin{aligned} l(\theta^{(t+1)}) &= \sum_{i=1}^{m}\sum_{z^{(i)}} Q_i^{(t+1)}(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)}|\theta^{(t+1)})}{Q_i^{(t+1)}(z^{(i)})} + \log p(\theta^{(t+1)}) \\ &\geq \sum_{i=1}^{m}\sum_{z^{(i)}} Q_i^{(t+1)}(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)}|\theta^{(t)})}{Q_i^{(t+1)}(z^{(i)})} + \log p(\theta^{(t)}) \\ &\geq \sum_{i=1}^{m}\sum_{z^{(i)}} Q_i^{(t)}(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)}|\theta^{(t)})}{Q_i^{(t)}(z^{(i)})} + \log p(\theta^{(t)}) \\ &= l(\theta^{(t)}) \end{aligned} \tag{6}$$

This first inequality comes form the fact that for step t+1,

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

This first second comes form the fact that for step t+1,

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

with respect $\theta^{(t)}$ computed form last step.

Problem 3

a: E-step

$$x^{(pr)} = y^{(pr)} + z^{(pr)} + \epsilon^{(pr)}$$
$$y^{(pr)} \sim \mathcal{N}(\mu_p, \sigma_p^2)$$
$$z^{(pr)} \sim \mathcal{N}(\nu_r, \tau_r^2)$$
$$\epsilon^{(pr)} \sim \mathcal{N}(0, \sigma^2)$$

i:Find joint distribution

I will ignore the superscript (pr) for x, y, z temporally.

$$\begin{bmatrix} y \\ z \\ x \end{bmatrix} \sim \mathcal{N}(\mu_{yzx}, \Sigma)$$

According to the summation rule of Gaussian distributions, we can easily get

$$\mu_{yzx} = \begin{bmatrix} \mu_p \\ \nu_r \\ \mu_p + \nu_r \end{bmatrix}$$

Then we can proceed to find the form of their covariance matrix. To compute it, we need calculate $\Sigma_{yy} = \mathbb{E}[(y-\mathbb{E}[y])(y-\mathbb{E}[y])^T]$, $\Sigma_{yz} = \mathbb{E}[(y-\mathbb{E}[y])(z-\mathbb{E}[z])^T]$, $\Sigma_{yx} = \mathbb{E}[(y-\mathbb{E}[y])(x-\mathbb{E}[x])^T]$, $\Sigma_{zy} = \mathbb{E}[(z-\mathbb{E}[z])(y-\mathbb{E}[y])^T]$, $\Sigma_{zz} = \mathbb{E}[(z-\mathbb{E}[z])(z-\mathbb{E}[z])^T]$, $\Sigma_{zx} = \mathbb{E}[(z-\mathbb{E}[z])(x-\mathbb{E}[x])^T]$, $\Sigma_{xy} = \mathbb{E}[(x-\mathbb{E}[x])(y-\mathbb{E}[y])^T]$, $\Sigma_{xz} = \mathbb{E}[(x-\mathbb{E}[x])(z-\mathbb{E}[z])^T]$, $\Sigma_{xx} = \mathbb{E}[(x-\mathbb{E}[x])(x-\mathbb{E}[x])^T]$.
First, we can get $\Sigma yy = \sigma_p^2, \Sigma zz = \nu_r^2, \Sigma xx = \sigma_p^2 + \tau_r^2 + \sigma^2$. Since $y^{(pr)}$ and $z^{(pr)}$ are independent, $\mathbb{E}[(y-\mathbb{E}[y])(z-\mathbb{E}[z])^T] = \mathbb{E}[y]\,\mathbb{E}[z] - \mathbb{E}[y]\,\mathbb{E}[z]] = 0$. For this reason, $\Sigma yz = 0$, $\Sigma zy = 0$, $\Sigma yz = 0$.

$$\begin{aligned}
\Sigma_{yx} &= \mathbb{E}[(y-\mathbb{E}[y])(x-\mathbb{E}[x])^T] \\
&= \mathbb{E}[yx] - \mathbb{E}[y]\,\mathbb{E}[x] \\
&= \mathbb{E}[y(y+z+\epsilon) - \mu_p(\mu_p+\nu_r) \\
&= \mathbb{E}[y^2 + yz + y\epsilon] - \mu_p(\mu_p+\nu_r) \\
&= \sigma_p^2
\end{aligned} \tag{7}$$

For similar reasons, $\Sigma_{xy} = \sigma_p^2$, $\Sigma_{xz} = \tau_r^2$, $\Sigma_{zx} = \tau_r^2$. Then the covariance matrix is,

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

ii: Derive expression for E-step

Then we can calculate $Q_{pr}(y^{(pr)}, x^{(pr)})$ According the supplementary notes on multivariate Gaussian,

$$x_A | x_B \sim \mathcal{N}(\mu_A + \Sigma_{AB}\Sigma_{BB}^{-1}(x_B - \mu_B), \Sigma_{AA} - \Sigma_{AB}\Sigma_{BB}^{-1}\Sigma_{BA})$$

We will make

$$x_A = \begin{bmatrix} y^{(pr)} \\ z^{(pr)} \end{bmatrix}$$

Then $x_A \sim \mathcal{N}(\mu_A, \Sigma_A)$

$$\mu_A = \begin{bmatrix}\mu_p \\ \nu_r\end{bmatrix}$$

$$\Sigma_A = \begin{bmatrix}\sigma_p^2 & 0 \\ 0 & \tau_r^2\end{bmatrix}$$

$x_B = x$, then $\mu_b = \mu_p + \nu_r$, $\Sigma_B = \sigma_p^2 + \tau_r^2 + \sigma^2$
$\Sigma_{AA} = \Sigma_A$, $\Sigma_{BB} = \Sigma_B$

$$\begin{aligned}\Sigma_{AB} &= \mathbb{E}[(x_A - \mu_A)(x_B - \mu_B)^T] \\ &= \begin{bmatrix}\Sigma_{yx} \\ \Sigma_{zx}\end{bmatrix} \\ &= \begin{bmatrix}\sigma_p^2 \\ \tau_r^2\end{bmatrix}\end{aligned} \tag{8}$$

$$\begin{aligned}\Sigma_{BA} &= \mathbb{E}[(x_B - \mu_B)(x_A - \mu_A)^T] \\ &= \begin{bmatrix}\Sigma_{yx} & \Sigma_{zx}\end{bmatrix} \\ &= \begin{bmatrix}\sigma_p^2 & \tau_r^2\end{bmatrix}\end{aligned} \tag{9}$$

Then we can plug this entities into the multivariate Gaussian conditional formula.

$$\begin{aligned}\mu_{y^{(pr)},z^{(pr)}|x^{(pr)}} &= \begin{bmatrix}\mu_p \\ \nu_r\end{bmatrix} + \begin{bmatrix}\sigma_p^2 \\ \tau_r^2\end{bmatrix} \frac{x^{(pr)} - \mu_p - \nu_r}{\sigma_p^2 + \tau_r^2 + \sigma^2} \\ &= \begin{bmatrix}\mu_p + \frac{\sigma_p^2(x^{(pr)} - \mu_p - \nu_r)}{\sigma_p^2+\tau_r^2+\sigma^2} \\ \nu_r + \frac{\tau_r^2(x^{(pr)} - \mu_p - \nu_r)}{\sigma_p^2+\tau_r^2+\sigma^2}\end{bmatrix}\end{aligned} \tag{10}$$

$$\begin{aligned}\Sigma_{y^{(pr)},z^{(pr)}|x^{(pr)}} &= \begin{bmatrix}\sigma_p^2 & 0 \\ 0 & \tau_r^2\end{bmatrix} - \begin{bmatrix}\sigma_p^2 \\ \tau_r^2\end{bmatrix} \frac{1}{\sigma_p^2 + \tau_r^2 + \sigma^2} \begin{bmatrix}\sigma_p^2 & \tau_r^2\end{bmatrix} \\ &= \begin{bmatrix}\frac{\sigma_p^2(\tau_r^2+\sigma^2)}{\sigma_p^2+\tau_r^2+\sigma^2} & -\frac{\sigma_p^2\tau_r^2}{\sigma_p^2+\tau_r^2+\sigma^2} \\ -\frac{\sigma_p^2\tau_r^2}{\sigma_p^2+\tau_r^2+\sigma^2} & -\frac{\tau_r^2(\sigma_p^2+\sigma^2)}{\sigma_p^2+\tau_r^2+\sigma^2}\end{bmatrix}\end{aligned} \tag{11}$$

Finally,

$$Q_{pr}(y^{(pr)}, z^{(pr)}|x^{(pr)}; \mu_p, \nu_r, \sigma_p, \tau_r) = \frac{1}{(2\pi)^{n/2}|\Sigma_{y^{(pr)},z^{(pr)}|x^{(pr)}}|} \exp(-\frac{1}{2}(\begin{bmatrix}y^{(pr)} \\ z^{(pr)}\end{bmatrix} - \mu_{y^{(pr)},z^{(pr)}|x^{(pr)}})^T$$
$$\Sigma^{-1}_{y^{(pr)},z^{(pr)}|x^{(pr)}})(\begin{bmatrix}y^{(pr)} \\ z^{(pr)}\end{bmatrix} - \mu_{y^{(pr)},z^{(pr)}|x^{(pr)}})) \tag{12}$$

b: M-step

The lower bound given by Jensen's inequality is (p is respect to i and r is respect to $z^{(i)}$)

$$\begin{aligned}J(Q, \mu_p, \nu_r, \sigma_p, \tau_r) &= \sum_{r=1}^{R}\sum_{p=1}^{P}\int_{y^{(pr)}}\int_{z^{(pr)}} Q_{pr}(y^{(pr)}, x^{(pr)}) \log \frac{p(y^{(pr)}, z^{(pr)}, x^{(pr)}; \mu_p, \nu_r, \sigma_p, \tau_r)}{Q_{pr}(y^{(pr)}, x^{(pr)})} dy^{(pr)} dx^{(pr)} \\ &= \sum_{r=1}^{R}\sum_{p=1}^{P} \mathop{\mathbb{E}}_{y^{(pr)},z^{(pr)} \sim Q_{pr}} [\log p(x^{(pr)}, y^{(pr)}, z^{(pr)}) - \log Q_{pr}(y^{(pr)}, x^{(pr)})]\end{aligned}$$

(13)

Form now on, I will drop the double summation, the distribution under the expectation. Since Q_{pr} is fixed at M-step, I will drop it two. Then, we find that we only need to maximize a multivariate Gaussian distribution expectation. I will also constants and in expectation and still abuse the equal sign.

$$\mathbb{E}[\log p(x^{(pr)}, y^{(pr)}, z^{(pr)})] = \mathbb{E}[-\frac{1}{2}\log|\Sigma| + \frac{1}{2}(\begin{bmatrix} y^{(pr)} \\ z^{(pr)} \\ x^{(pr)} \end{bmatrix} - \begin{bmatrix} \mu_p \\ \nu_r \\ \mu_p + \nu_r \end{bmatrix})^T \Sigma^{-1} (\begin{bmatrix} y^{(pr)} \\ z^{(pr)} \\ x^{(pr)} \end{bmatrix} - \begin{bmatrix} \mu_p \\ \nu_r \\ \mu_p + \nu_r \end{bmatrix})] \tag{14}$$

with Σ caculated in a:i,

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

$$|\Sigma| = \sigma_p^2 \tau_r^2 \sigma^2$$

$$\Sigma^{-1} = \begin{bmatrix} \frac{1}{\sigma_p^2} + \frac{1}{\sigma^2} & \frac{1}{\sigma^2} & -\frac{1}{\sigma^2} \\ \frac{1}{\sigma^2} & \frac{1}{\sigma^2} + \frac{1}{\tau_r^2} & -\frac{1}{\sigma^2} \\ -\frac{1}{\sigma^2} & -\frac{1}{\sigma^2} & \frac{1}{\sigma^2} \end{bmatrix}$$

We continue to ignore the constants($x^{(pr)}$ and σ is constant too) and move pure parameters out of the expectation symbol. I will define three symbols to simplify calculations.

$$[y] = y^{(pr)} - \mu_p$$

$$[z] = z^{(pr)} - \nu_r$$

$$[x] = x^{(pr)} - \mu_p - \nu_r$$

$$
\begin{aligned}
&\mathbb{E}[\log p(x^{(pr)}, y^{(pr)}, z^{(pr)})] \\
&= -\log \sigma_p - \log \tau_r + \mathbb{E}[\begin{bmatrix}[y] & [z] & [x]\end{bmatrix} \begin{bmatrix} \frac{1}{\sigma_p^2} + \frac{1}{\sigma^2} & \frac{1}{\sigma^2} & -\frac{1}{\sigma^2} \\ \frac{1}{\sigma^2} & \frac{1}{\sigma^2} + \frac{1}{\tau_r^2} & -\frac{1}{\sigma^2} \\ -\frac{1}{\sigma^2} & -\frac{1}{\sigma^2} & \frac{1}{\sigma^2} \end{bmatrix} \begin{bmatrix}[y] \\ [z] \\ [x]\end{bmatrix}] \\
&= -\log \sigma_p - \log \tau_r + \mathbb{E}[(\frac{1}{\sigma_p^2} + \frac{1}{\sigma^2})[y]^2 + \frac{1}{\sigma^2}[z][y] - \frac{1}{\sigma^2}[x][y] + \frac{1}{\sigma^2}[y][z] \\
&\quad + (\frac{1}{\sigma^2} + \frac{1}{\tau_r^2})[z]^2 - \frac{1}{\sigma^2}[x][z] - \frac{1}{\sigma^2}[y][x] - \frac{1}{\sigma^2}[z][x] + \frac{1}{\sigma^2}[x]^2] \\
&= -\log \sigma_p - \log \tau_r + (\frac{1}{\sigma_p^2} + \frac{1}{\sigma^2})(\Sigma_Q^{yy} + {\mu_Q^y}^2 - 2\mu_p \mu_Q^y + 2\mu_p^2) + (\frac{1}{\sigma^2} + \frac{1}{\tau_r^2})(\Sigma_Q^{zz} + \\
&\quad {\mu_Q^z}^2 - 2\nu_r \mu_Q^z + 2\nu_r^2) + \frac{2}{\sigma^2}\Sigma_Q^{yz} - \frac{2}{\sigma^2}[x]\mu_Q^y - \frac{2}{\sigma^2}[x]\mu_Q^z \\
&= -\log \sigma_p - \log \tau_r + (\frac{1}{\sigma_p^2} + \frac{1}{\sigma^2})(\Sigma_Q^{yy} + {\mu_Q^y}^2 - 2\mu_p \mu_Q^y + \mu_p^2) + (\frac{1}{\sigma^2} + \frac{1}{\tau_r^2})(\Sigma_Q^{zz} + \\
&\quad {\mu_Q^z}^2 - 2\nu_r \mu_Q^z + \nu_r^2) - \frac{2}{\sigma^2}(x^{(pr)} - \mu_p - \nu_r)\mu_Q^y - \frac{2}{\sigma^2}(x^{(pr)} - \mu_p - \nu_r)\mu_Q^z
\end{aligned}
$$

Notice that I have dropped any term that is only related to x or σ. Now we can plug into our mean vector and covariance matrix of the posterior distribution $Q^{(pr)}$. both of which are known.
NB:: $\mathbb{E}[(y - \mu_p)^2] \neq {\mu_Q^y}^2$
Taking derivative to σ_P, we can get

$$
\nabla_{\mu_p} \mathbb{E}[\log p(x^{(pr)}, y^{(pr)}, z^{(pr)})] = \frac{2}{\sigma^2}(\mu_Q^y + \mu_Q^z) - 2\mu_Q^y(\frac{1}{\sigma_p^2} + \frac{1}{\sigma^2}) + 2\mu_p(\frac{1}{\sigma_p^2} + \frac{1}{\sigma^2}) \tag{15}
$$

Setting above equation to zero, for $p = 1, 2 \ldots P$, we can get

$$
\mu_p = \frac{1}{R}\sum_{r=1}^{R}(\frac{\sigma^2}{\sigma_p^2 + \sigma^2}\mu_{Q_{pr}}^y - \frac{\sigma_p^2}{\sigma_p^2 + \sigma^2}\mu_{Q_{pr}}^z)
$$

Similarly, we can have

$$
\nu_r = \frac{1}{P}\sum_{p=1}^{P}(\frac{\sigma^2}{\tau_r^2 + \sigma^2}\mu_{Q_{pr}}^z - \frac{\tau_r^2}{\tau_r^2 + \sigma^2}\mu_{Q_{pr}}^y)
$$

$$
\sigma_p^2 = \frac{2}{R}\sum_{r=1}^{r}(\Sigma_Q^{yy} + {\mu_Q^y}^2 - 2\mu_p \mu_Q^y + \mu_p^2)
$$

$$
\tau_r^2 = \frac{2}{P}\sum_{p=1}^{P}(\Sigma_Q^{zz} + {\mu_Q^z}^2 - 2\nu_r \mu_Q^z + \nu_r^2)
$$

All the parameters on the right hand side of the equation is from the last iteration.

Interpretation

I adopted a different approach than the answers key provided here. The expectation about $x^{(pr)}$ should be dropped since they are observed and thus fixed. It seems more intuitive as well, since we need to use the posterior mean of Z to revise the mean vector for $y^{(pr)}$.

Problem 4

a: Non-negativity

$$
\begin{aligned}
KL(P||Q) &= \sum_x P(x)\log\frac{P(x)}{Q(x)} \\
&= \sum_x P(x)(-\log\frac{Q(x)}{P(x)}) \\
&= \mathbb{E}[-\log\frac{Q(x)}{P(x)}] \\
&\geq -\log\mathbb{E}[\frac{Q(x)}{P(x)}] \\
&= -\log\sum_x P(x)\frac{Q(x)}{P(x)} \\
&= -\log 1 \\
&= 0
\end{aligned}
\tag{16}
$$

Then we have proved that $\forall P, Q, KL(P||Q) \geq 0$ According to the boundary condition of Jensen's inequality, $KL(P||Q) = 0$ only when $\frac{Q(x)}{P(x)} = \mathbb{E}[\frac{Q(x)}{P(x)}] = 1$. Thus, $KL(P||Q) = 0, iffP = Q$.

b: Chain Rule for KL divergence

$$
\begin{aligned}
RHL &= KL(P(X,Y)||Q(X,Y)) \\
&= \sum_y\sum_x P(x,y)\log\frac{P(x,y)}{Q(x,y)} \\
LHS &= \sum_x P(x)\log\frac{P(x)}{Q(x)} + \sum_x P(x)(\sum_y P(y|x)\log\frac{P(y|x)}{Q(y|x)}) \\
&= \sum_x P(x)(\log\frac{P(x)}{Q(x)} + \sum_y P(y|x)\log\frac{P(y|x)}{Q(y|x)}) \\
&= \sum_x P(x)\sum_y P(y|x)(\log\frac{P(x)}{Q(x)} + \log\frac{P(y|x)}{Q(y|x)}) \\
&= \sum_x\sum_y P(x)P(y|x)\log\frac{P(x)P(y|x)}{Q(x)Q(y|x)} \\
&= \sum_x\sum_y P(x,y)\log\frac{P(x,y)}{Q(x,y)} \\
&= RHL
\end{aligned}
\tag{17}
$$

q.e.d.

Problem 4 continued on next page...

c: KL and maximum likelihood

$$\begin{aligned}
\arg\min_\theta KL(\hat{P}||P_\theta) &= \arg\min_\theta \sum_x \hat{P}(x)\log\frac{\hat{P}(x)}{P_\theta(x)} \\
&= \arg\min_\theta [\sum_x \hat{P}(x)\log\hat{P}(x) - \sum_x \hat{P}(x)\log P_\theta(x)] \\
&= \arg\min_\theta - \sum_x \hat{P}(x)\log P_\theta(x) \\
&= \arg\max_\theta \sum_x \frac{1}{m}1\{x^{(i)} = x\}\log P_\theta(x) \\
&= \arg\max_\theta \log P_\theta(x^{(i)})
\end{aligned} \tag{18}$$

Problem 5

a, b, c, d: K-means for compression

I use the gpuarray function from MATLAB to accelerate the computation.

```
%% Function: KMeansCompressions
%% ----------------------------
%% original image with size M< N, C
%% return and show compressed image side with the original image
function result = KMeansCompression(~, ~)
%% Imenplementation Notes: This is a vetorized version of K-means.
%% I didn't compare the running time with non-vectorized version.
clear; clc;
K = 16;
image = gpuArray(double(imread('mandrill-large.tiff')));
[M, N, C] = size(image);
result = gpuArray(zeros([M, N, C]));
labels = gpuArray(zeros([M, N]));

imageLarge = repmat(image, 1, 1, 1, K); % with same pixel piling up at the fourth demension
centeroids = datasample(reshape(image, [M*N, C]), K); %K, C shape random centeroids.
iterationCount = 0;
centeroidsUpdate = 0;

while((iterationCount < 31)|| (centeroidsUpdate > 1e-4))
    %increase iteration count
    iterationCount =  iterationCount +1;

    %assign labels
    centeroidsLarge = permute(repmat(centeroids, 1, 1, N, M), [4 3 2 1]); %M, N, C, K matrixs.
    distance = reshape(sum((centeroidsLarge - imageLarge).^2, 3), [M, N, K]);
    [~, labels] = min(distance, [], 3);

    preCenteroids = centeroids;
    %update centeroids
```

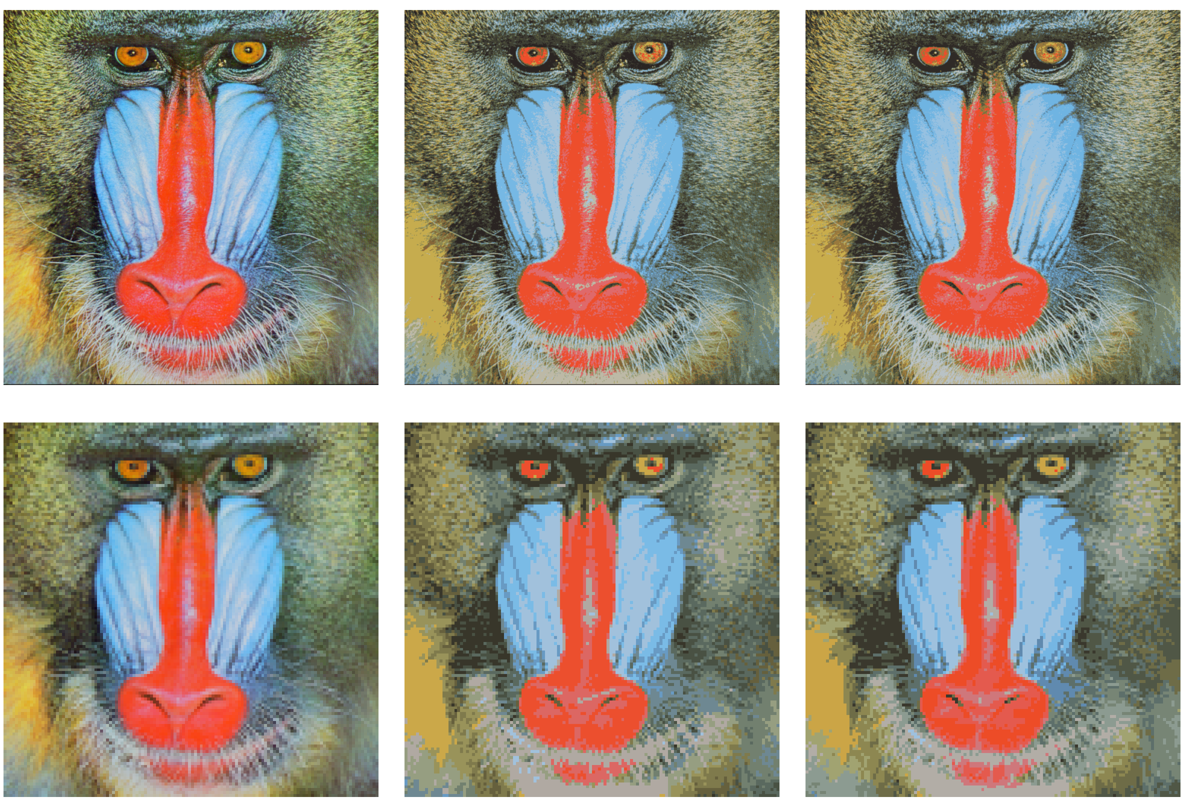
```
    for label = 1:K
        mask2d = labels==label;
        mask3d = repmat(mask2d, 1, 1, C);
        centeroids(label, :) = reshape(sum(sum(mask3d.*image, 1), 2), [1, 3]) ./ sum(
            sum(mask2d, 1), 2);
    end

    %calculate the delta for testing convergence
    centeroidsUpdate = sum(sum((preCenteroids - centeroids).^2, 1), 2);

     %for debug
    disp(['Iterations ', num2str(iterationCount), ': delta, ', num2str(
        centeroidsUpdate)]);
end

%update the compressed image
result = reshape(centeroids(labels, :), [M, N, C]);
imshow(uint8(round(result)));
result_cpu = uint8(round(gather(result)));
imwrite(result_cpu, 'compressed-large-10.png');
end
```

Compression result Comparison



From left to right, they are original image, image after 10 iterations, and the converged image. The upper row is the mandrill-large. The lower row is mandrill-small.

In theory, file size should be smaller in the times of 6, since originally one pixel needs 24 bit but now only requires 4 bit for the encoding from the 16 colors.