

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 point 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:

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

So if we set

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

According to Jensen's Inequality, we have

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

Then we get EM-step as below:

E-step:

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

M-step:

$$\theta = \operatorname{argmax}_{\theta} \left[\sum_{i=1}^m \sum_{z^{(i)}} Q_i(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)} | \theta)}{Q_i(z^{(i)})} + \log p(\theta) \right]$$

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

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

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:

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

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

$$\begin{aligned} \Theta &= \operatorname{argmax}_{\Theta} \sum_p \sum_r E_{(y^{(pr)}, z^{(pr)}) \sim Q_{pr}} [\log p(x^{(pr)}, y^{(pr)}, z^{(pr)})] \\ &= \operatorname{argmax}_{\Theta} \sum_p \sum_r E \left[\log \frac{1}{2\pi^{3/2} \sigma \sigma_p \tau_r} - \frac{1}{2\sigma_p^2} (y^{(pr)} - \mu_p)^2 - \frac{1}{2\tau_r^2} (z^{(pr)} - \nu_r)^2 \right. \\ &\quad \left. - \frac{1}{2\sigma^2} (x^{(pr)} - y^{(pr)} - z^{(pr)})^2 \right] \\ &= \operatorname{argmax}_{\Theta} \sum_p \sum_r E \left[\log \frac{1}{\sigma_p \tau_r} - \frac{1}{2\sigma_p^2} (y^{(pr)} - \mu_p)^2 - \frac{1}{2\tau_r^2} (z^{(pr)} - \nu_r)^2 \right] \end{aligned}$$

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

$$\begin{aligned} \mu_p &= \frac{1}{PR} \sum_p \sum_r \mu_{Q1} \\ \nu_r &= \frac{1}{PR} \sum_p \sum_r \mu_{Q2} \end{aligned}$$

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

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

KL divergence and Maximum Likelihood

(a)

$$\begin{aligned} KL(P||Q) &= \sum_x P(x) - \log \frac{Q(X)}{P(X)} \\ &\geq -\log \sum_x (P(x) \frac{Q(X)}{P(X)}) \\ &= -\log \sum_x Q(x) \end{aligned}$$

Since $\sum_x Q(x) = 1$, so $KL(P||Q) \geq 0$.

If $P = Q$, it's easily to see that $KL(P||Q) = 0$.

Because $-\log(x)$ is a strictly convex function, so we have the $=$ when $X = E[X]$ with probability 1 where $X = \frac{Q}{P}$.

(b)

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

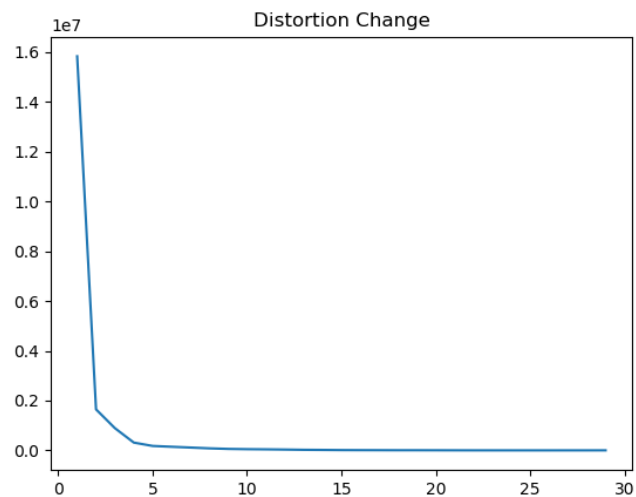
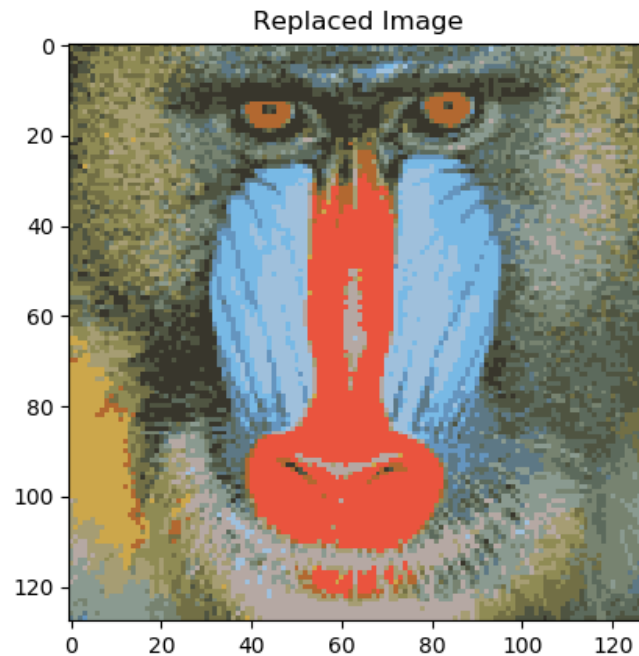
(c)

$$\begin{aligned} KL(\hat{P}||P_\theta) &= \sum_x \hat{P} \log \frac{\hat{P}(x)}{P_\theta(x)} \\ &= - \sum_x \frac{1}{m} \sum_{i=1}^m \mathbf{1}\{x^{(i)} = x\} \log \frac{P_\theta(x)}{\sum_{i=1}^m \mathbf{1}\{x^{(i)} = x\}} \\ &= - \frac{1}{m} \sum_{i=1}^m \log P_\theta(x^{(i)}) \end{aligned}$$

So adjust θ to minimize the KL is equivalent to maximize the log-likelihood function.

K-means for compression

Since we have reduced the number of colors to 16, we only need to store the 16 colors and the location information, which decreases the space we need to store the whole image. Here is the result of k-means with mandrill-small.



Here is the result of k-means with mandrill-large.

