Machine Learning 2 - Homework Week 5

Minh Ngo
MSc Artificial Intelligence
University of Amsterdam
nlminhtl@gmail.com

May 12, 2016

Collaborators: Riaan Zoetmulder, Arthur Bražinskas

Problem 1

a. Forever (until computation limit ends):

   1. Sample $x \sim q(x)$
   2. Sample $y \sim Unif(0, cq(x))$
   3. If $Y > p(x)$ then reject the sample, otherwise - accept with an acceptance probability (Bishop 11.14):

$$p(accept) = \int p(x)/cq(z)q(x)dx$$

b. Sampling a new state doesn't dependent on the previous state and consequently they are independent.

c. Weights for important sampling are updated as following (Bishop, p. 533):

$$w_n = \frac{p(x_n)}{q(x_n)} \tag{1}$$

d. Following Bishop 11.33 the Metropolis Hasting accept probability will be:

$$\alpha(x_{t+1}, x_t) = min(1, \frac{q(x_t)p(x_{t+1})}{p(x_t)q(x_{t+1})}) \tag{2}$$

e. In spite of the fact that the proposed new state is independent from the previous state, we still count it for the accept probability. Consequently subsequent samples are in general not independent.

f. With rejecting x_2 and x_5 we simply leave previous element of the sequence as a result. Consequently the sequence of states will be: x_1, x_1, x_3, x_4, x_4

g. For the rejection sampling the acceptance rate will decrease exponentially with dimensionality, therefore this technique is only useful in 1 or 2 dimensions.

   Importance sampling will also have problems with high dimensional data since it becomes harder to develop the envelop distribution in such a way that it will be large enough where the actual distribution has significant probability mass.

   The Independent sampler is only the method that will scales well with the dimensionaly of the sample space, since it uses Markov Chain to decompose complex problem to much simpler subproblems.

Problem 3

$$x \sim \mathcal{N}(x|\mu, \tau^{-1}) \tag{3}$$

$$\mu \sim \mathcal{N}(\mu|\mu_0, s_0) \tag{4}$$

$$\tau \sim Gamma(\tau|a, b) \tag{5}$$

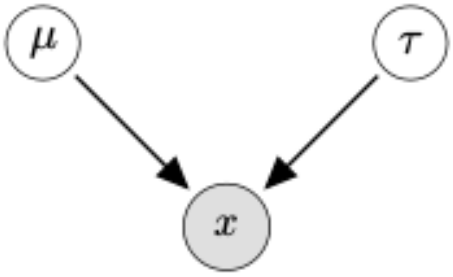

Figure 1: A graph involving an observed Gaussian variable x with prior distributions over its mean μ and precision τ

Update for μ can be computed as following:

$$p(\mu, \tau|x) \propto p(x|\mu, \tau)p(\mu) \propto \mathcal{N}(x|\mu, \tau^{-1})\mathcal{N}(\mu|\mu_0, s_0) \tag{6}$$

Using complete to square we get:

$$\begin{aligned} p(\mu, \tau|x) &\propto \exp\left[-\frac{1}{2}\frac{(x-\mu)^2}{\tau^{-1}} - \frac{1}{2}\frac{(\mu-\mu_0)^2}{s_0}\right] \\ &\propto \exp\left[-\frac{1}{2}\Big(\mu^2(\tau + s_0^{-1}) - 2\mu(x\tau + \mu_0 s_0^{-1})\Big)\right] \\ &\propto \exp\left[-\frac{1}{2}(\tau + s_0^{-1})(\mu - \frac{x\tau + \mu_0 s_0^{-1}}{\tau + s_0^{-1}})^2\right] \\ &\propto \mathcal{N}(\mu|\frac{x\tau + \mu_0 s_0^{-1}}{\tau + s_0^{-1}}, (\tau + s_0^{-1})^{-1}) \end{aligned} \tag{7}$$

Update for τ can be computed as following:

$$\begin{aligned} p(\tau|x, \mu, a, b) &\propto p(x|\mu, \tau)p(\tau, a, b) \propto \mathcal{N}(x|\mu, \tau^{-1})Gamma(\tau|a, b) \\ &\propto \sqrt{\tau}\exp\Big(-\frac{1}{2}(x-\mu)^2\tau\Big)\tau^{a-1}\exp(-b\tau) \\ &\propto \tau^{a-\frac{1}{2}}\exp\Big(-(\frac{1}{2}(x-\mu)^2 + b)\tau\Big) \propto Gamma(\tau|a + \frac{1}{2}, \frac{1}{2}(x-\mu)^2 + b) \end{aligned} \tag{8}$$

Problem 5

a.

$$p(\mathbf{x}|\mu) = \prod_{i=1}^{D} \mu_i^{x_i}(1-\mu_i)^{1-x_i}$$

Mean of x under this distribution will be (Bishop 9.45):

$$E[x_i] = \sum_{i=1}^{N} x_i p_i = 1\mu_i + 0(1-\mu_i) = \mu_i \tag{9}$$

Consequently $E[\mathbf{x}] = \mu$.

b.

$$Var[x_i] = E[x_i^2] - E[x_i]^2 = \mu_i - \mu_i^2 = \mu_i(1 - \mu_i) \tag{10}$$

Consequently covariance will be (Bishop 9.46):

$$Cov[\mathbf{x}] = diag(\mu_i(1 - \mu_i)) \tag{11}$$

c. Mean of x under the mixture distribution will be (Bishop 9.49):

$$E[\mathbf{x}] = \sum_{k=1}^{K} \pi_k \mu_{\mathbf{k}} \tag{12}$$

d. Log likelihood for the mixture distribution will be (Bishop 9.51):

$$\ln p(\mathbf{X}|\mu, \pi) = \sum_{n=1}^{N} \ln \left[\sum_{k=1}^{K} \pi_k p(x_n|\mu_{\mathbf{k}}) \right] \tag{13}$$

e. The standard maximum-likelihood doesn't work in this case since there is no closed form solution for log-likelihood with summation inside the logarithm.

f.

$$p(\mathbf{x}_n, \mathbf{z}_n|\mu, \pi) = p(\mathbf{z}_n|\pi)p(\mathbf{x}_n|\mathbf{z}_n, \mu) = \prod_{k=1}^{K} \pi_k^{z_{nk}} p(\mathbf{x}_n|\mu_k)^{z_{nk}}$$

Log-likelihood consequently will be:

$$\ln p(\mathbf{X}, \mathbf{Z}|\mu, \pi) = \sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk} \left(\ln \pi_k + \sum_{d=1}^{D} (x_{nd} \ln \mu_{kd} + (1 - x_{nd}) \ln(1 - \mu_{kd})) \right) \tag{14}$$

g. The corresponding graphical model will be:

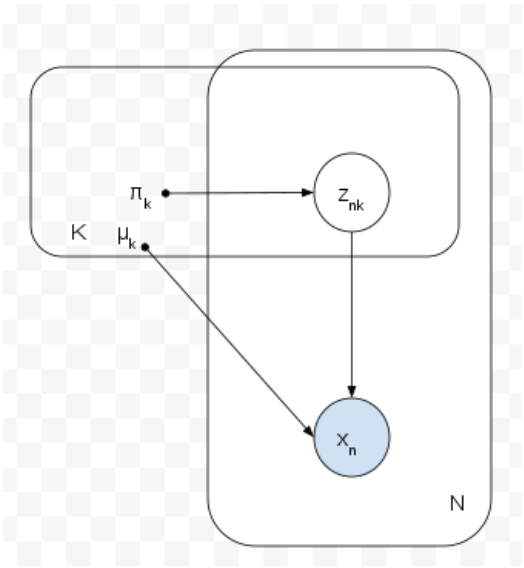

Figure 2: Graphical model

h. Using (Bishop 9.71):

$$\begin{aligned}\mathcal{L}(\mu,\pi) &= \sum_{n=1}^{N}\ln\sum_{z_n}\sum_{k=1}^{K}\frac{q(z_{nk})}{q(z_{nk})}p(x_n,z_{nk}|\mu,\pi)\\ &= \sum_{n=1}^{N}\sum_{z_n}\sum_{k=1}^{K}q(z_{nk})\ln\frac{p(x_n,z_{nk}|\mu,\pi)}{q(z_{nk})}\end{aligned} \tag{15}$$

Objective function for VEM will be:

$$\begin{aligned}\mathcal{B}(q(z_{nk},\mu,\pi) &= \sum_{n=1}^{N}\sum_{z_n}\sum_{k=1}^{K}q(z_{nk)})\ln p(x_n,z_{nk}|\mu,\pi) - \sum_{n=1}^{N}\sum_{z_n}\sum_{k=1}^{K}q(z_{nk})\ln q(z_{nk})\\ &= \sum_{n=1}^{N}\sum_{z_n}\sum_{k=1}^{K}q(z_{nk})\Bigg(\ln\pi_{z_{nk}} + \sum_{d=1}^{D}(x_{nd}\ln\mu_{z_{nkd}} + (1-x_{nd})\ln(1-\mu_{z_{nkd}}))-\\ &- \ln q(z_{nk})\Bigg)\end{aligned} \tag{16}$$

i. Including Lagrangian multipliers we get:

$$\mathcal{B}^{'}(q(z_n), \mu, \pi, \lambda, \lambda_n) = \mathcal{B}(q(z_n), \mu, \pi) + \lambda(\sum_{k=1}^{K} \pi_k - 1) + \sum_{n=1}^{N} \lambda_n (\sum_{z_n} \sum_{k=1}^{K} q(z_{nk}) - 1) \tag{17}$$

j. E step:

$$\frac{\partial \mathcal{B}^{'}(...)}{\partial q(z_{nk})} = 0 \tag{18}$$

$$\ln \pi_{z_{nk}} + \sum_{d=1}^{D} (x_{nd} \ln \mu_{z_{nk}d}) + (1 - x_{nd}) \ln(1 - \mu_{z_{nk}d})) - \ln q(z_{nk}) + \lambda_n - 1 = 0 \tag{19}$$

$$\ln \pi_{z_{nk}} + \sum_{d=1}^{D} (x_{nd} \ln \mu_{z_{nk}d} + (1 - x_{nd}) \ln(1 - \mu_{z_{nk}d})) + (\lambda_n - 1) = \ln q(z_{nk}) \tag{20}$$

$$q(z_{nk}) = \pi_{z_{nk}} \prod_{d=1}^{D} (\mu_{z_{nk}d}^{x_{nd}} (1 - \mu_{z_{nk}d}^{1-x_{nd}}) \exp(\lambda_n - 1) \tag{21}$$

k. M-step:

$$\frac{\partial \mathcal{B}^{'}(...)}{\partial \pi_k} = 0 \tag{22}$$

$$\sum_{n=1}^{N} \sum_{z_n} \frac{q(z_{nk})}{\pi_k} + \lambda = 0 \tag{23}$$

$$\pi_k = \frac{\sum_{n=1}^{N} \sum_{z_n} q(z_{nk})}{N} \tag{24}$$