ML-2

Homework Assignment - 5

Shantanu Chandra
(12048461)

Q1.

(1) The given complete-data log-likelihood has the constraint $\sum_{k=1}^{K} \pi_k = 1$
Hence, we introduce the Lagrange multiplier such as:

$$L = \mathbb{E}_{posterior}\left[\ln p(x, z \mid \mu, \Sigma, \pi)\right] + \lambda\left(\sum_{k=1}^{K} \pi_k - 1\right)$$

Also note that the updates are done in the M-step & hence $\gamma(z_{nk})$ = constant.

a) update rule for π_k:

$$\frac{\partial L}{\partial \pi_k} = \frac{\partial}{\partial \pi_k}\left(\sum_{n=1}^{N}\sum_{k=1}^{K} \gamma(z_{nk})\left[\ln \pi_k - \frac{D}{2}\ln 2\pi - \frac{1}{2}\ln|\Sigma_k| - \frac{1}{2}(x_n-\mu_k)\Sigma_k^{-1}(x_n-\mu_k)^T\right] + \lambda\left(\sum_{k=1}^{K} \pi_k - 1\right)\right)$$

$$0 = \underbrace{\sum_{n=1}^{N} \gamma(z_{nk})}_{= N_k}\left[\frac{1}{\pi_k} - 0 - 0 - 0\right] + \lambda$$

$$\pi_k \lambda = -N_k \quad \text{—— (1)}$$

$$\lambda \underbrace{\sum_k \pi_k}_{=1} = -\underbrace{\sum_k N_k}_{= N}$$ (multiplying by Σ_k on both sides)
↳ assuming it to be non-singular

$$\boxed{\lambda = -N}$$

Substituting λ in (1),

$$-N\pi_k = -N_k \implies \boxed{\pi_k = \frac{N_k}{N}}$$ // Ans.

b) update rule for μ_k:

$$\frac{\partial L}{\partial \mu_k} = \frac{\partial}{\partial \mu_k} \sum_{n=1}\sum_{k=1} \gamma(z_{nk})\left[\ln \pi_k - \frac{D}{2}\ln 2\pi - \frac{1}{2}\ln|\Sigma_k| - \frac{1}{2}(x_n-\mu_k)\Sigma_k(x_n-\mu_k)^T\right] + \lambda\left(\sum_{k=1} \pi_k - 1\right)$$

Using 86 from Matrix cook book (derivation of a gaussian w.r.t its mean):

$$0 = \sum_{n=1}^{N} \gamma(z_{nk})\left(\frac{+2}{2}\cdot \Sigma_k^{-1}(x_n - \mu_k)\right)$$

$$\cancel{\Sigma_k^{-1}} \sum_{n=1}^{N} \gamma(z_{nk}) x_n = \cancel{\Sigma_k^{-1}} \underbrace{\sum_{n=1}^{N} \gamma(z_{nk})}_{= N_k} \mu_k$$

$$\therefore \boxed{\mu_k = \frac{\sum_{n=1}^{N} \gamma(z_{nk}) x_n}{N_k}}$$ // Ans.

c) Update rule for Σ_k :

for clarity I use $z = (x_n - \mu_k)$ -

$$\frac{\partial L}{\partial \Sigma_k} = \frac{\partial}{\partial \Sigma_k} \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma(z_{nk}) \left[\ln \pi_k - \frac{D}{2} \ln 2\pi - \frac{1}{2} \ln |\Sigma_k| - \frac{1}{2} \underbrace{z \Sigma_k^{-1} z^T}_{= \text{scalar}} \right] + \lambda \left(\sum_{k=1}^{K} \pi_k - 1 \right)$$

$\therefore$ can write as $Tr[z \Sigma_k^{-1} z^T]$

$$0 = \sum_{n=1}^{N} \gamma(z_{nk}) \left[\frac{-1}{2} \frac{\partial}{\partial \Sigma_k} \ln |\Sigma_k^{-1}| - \frac{1}{2} \frac{\partial}{\partial \Sigma_k} Tr\left[z z^T \Sigma_k^{-1} \right] \right] + 0$$

(using $Tr[ABC] = Tr[ACB]$)

$$0 = \sum_{n=1}^{N} \gamma(z_{nk}) \left[\frac{\partial}{\partial \Sigma_k} \ln |\Sigma_k^{-1}| - \underbrace{\left(\Sigma_k^{-1} z z^T \Sigma_k^{-1} \right)^T}_{\Sigma_k \text{ is symmetric}} \right]$$

(using 63 from matrix cook book : $\frac{\partial}{\partial X} Tr(A B X^{-1}) = -(X^{-1} A B X^{-1})^T$)

$$0 = \sum_{n=1}^{N} \gamma(z_{nk}) \left[\Sigma_k^{-1} - \Sigma_k^{-1} z z^T \Sigma_k^{-1} \right]$$

(using 57 of cookbook : $\frac{\partial}{\partial X} \ln |\det(X)| = X^{-T}$)

$$\underbrace{\sum_{n=1}^{N} \gamma(z_{nk})}_{= N_k} \cancel{\Sigma_k^{-1}} = \Sigma_k^{-1} \sum_{n=1}^{N} \gamma(z_{nk}) z z^T \cancel{\Sigma_k^{-1}}$$

$$\Sigma_k^{-1} = \frac{N_k}{\sum_{n=1}^{N} \gamma(z_{nk}) (x_n - \mu_k)(x_n - \mu_k)^T}$$

or, $\boxed{\Sigma_k = \frac{\sum_{n=1}^{N} \gamma(z_{nk}) (x_n - \mu_k)(x_n - \mu_k)^T}{N_k}}$

// Ans.

(NOTE: the 'μ_k' used here are the 'μ_k^{new}' we get from updating the μ_k^{old} using update value above.)

(2) In the case where we have a common Σ, we will follow similar steps as above, however the $\sum_{k=1}^{K}$ will still be present since we don't derive w.r.t a specific 'k' :

$$\frac{\partial L}{\partial \Sigma} = \sum_{k=1}^{K} \sum_{n=1}^{N} \gamma(z_{nk}) \left[\Sigma^{-1} - \Sigma^{-1} z z^T \Sigma^{-1} \right]$$

$$\underbrace{\sum_{k=1}^{K} \sum_{n=1}^{N} \gamma(z_{nk})}_{= N} \cancel{\Sigma_k^{-1}} = \Sigma_k^{-1} \left(\sum_{k=1}^{K} \sum_{n=1}^{N} \gamma(z_{nk}) z z^T \cancel{\Sigma_k^{-1}} \right)$$

$$\Sigma^{-1} = \frac{N}{\sum_{k=1}^{K} \sum_{n=1}^{N} \gamma(z_{nk}) (x_n - \mu_k)(x_n - \mu_k)^T}$$

or, $\boxed{\Sigma = \frac{\sum_{k=1}^{K} \sum_{n=1}^{N} \gamma(z_{nk}) (x_n - \mu_k)(x_n - \mu_k)^T}{N}}$

// Ans.

Since π_k & μ_k do not depend on Σ, their update steps remain same as before.

// Ans.

Q2.

Instead of data likelihood ($p(x|\theta)$) we are now maximizing the posterior given by :

$$p(\theta|x) = \frac{p(x|\theta)\cdot p(\theta)}{p(x)}$$

or, $$\ln p(\theta|x) = \ln p(x|\theta) + \ln p(\theta) - \ln p(x)$$

$$= L(q,\theta) + KL(q||p) + \ln p(\theta) - \ln p(x) \qquad \text{(Using Bishop 9.70)}$$

$$\geq \underbrace{L(q,\theta) + \ln p(\theta) - \ln p(x)}_{\text{ELBO}} \qquad (\text{using } KL(q||p) \geq 0)$$

where, $$L(q,\theta) = \sum_z q(z) \ln\left[\frac{p(x,z|\theta)}{q(z)}\right]$$

$$KL(q||p) = -\sum_z q(z) \ln\left[\frac{p(z|x,\theta)}{q(z)}\right]$$

E-step :

In E-step we maximize the ELBO expression w.r.t $q(z)$ while holding θ fixed. In this case, the ELBO is the same as with the case of likelihood maximization. since the new terms (prior & evidence) are constants w.r.t $q(z)$.
Hence, this step is essentially the same for both (and it involves setting $q(z)$ to the posterior $p(z|x,\theta)$).

Ans.

M-step :

In M-step we maximize ELBO wrt the parameters θ. Here we expand the ELBO expression using definitions of $L(q,\theta)$ & $KL(q||p)$ and using the result of E-step ($q(z) = p(z|x,\theta)$).

$$\text{ELBO} = L(q,\theta) + \ln p(\theta) + \text{const}.$$

$$= \sum_z p(z|x,\theta) \ln p(x,z|\theta) + \underbrace{\sum_z p(z|x,\theta) \ln p(z|x,\theta)}_{= H[p(z|x,\theta)] \Rightarrow \text{const. w.r.t } \theta} + \ln p(\theta) + \text{const}.$$

$$\therefore \nabla_\theta \text{ELBO} = \nabla_\theta \left[\sum_z p(z|x,\theta) \ln p(x,z|\theta) + \ln p(\theta)\right] + 0$$

Thus, the quantity that is maximized in M-step is :

$$\sum_z p(z|x,\theta) \ln p(x,z|\theta) + \ln p(\theta)$$

Proved

Q3.

Log-posterior is given by :

$$\ln p(\mu,\pi|x) = \sum_{n=1}^{N} L(\mu,\pi,z_n) + \ln p(\mu,\pi) - \underbrace{\ln p(x_n)}_{\text{constant.}}$$

$$= \sum_{n=1}^{N} L(\mu,\pi,z_n) + \ln p(\mu) + \ln p(\pi)$$

$$= \sum_{n=1}^{N}\sum_{k=1}^{K} \gamma(z_{nk}) \left[\ln \pi_k + \sum_{i=1}^{D} x_{ni} \ln \mu_{ki} + (1-x_{ni}) \ln(1-\mu_{ki})\right] + \sum_{k=1}^{K} \ln \text{Beta}(\mu_k|a_k,b_k) + \ln \text{Dir}(\pi_k|\alpha_k)$$

$$L = \sum_{n=1}^{N}\sum_{k=1}^{K} \gamma(z_{nk}) \left[\ln \pi_k + \sum_{i=1}^{D} x_{ni} \ln \mu_{ki} + (1-x_{ni}) \ln(1-\mu_{ki})\right] + \sum_{k=1}^{K}\sum_{i=1}^{D} (a_k-1) \ln \mu_{ki} + (b_k-1) \ln(1-\mu_{ki})$$

$$+ \sum_{k=1}^{K} (\alpha_k - 1) \ln \pi_k$$

maximize π_k :

To maximize L w.r.t π_k we need to add the Lagrangian multiplier to it $\left[\lambda\left(\sum_{k=1}^{K} \pi_k - 1\right)\right]$ and set the derivative to 0 :

$$\frac{\partial L}{\partial \pi_k} = \frac{\sum_{n=1}^{N} \gamma(z_{nk})}{\pi_k} + \frac{\alpha_k - 1}{\pi_k} - \lambda = 0$$

$$\sum_{n=1}^{N} \gamma(z_{nk}) + (\alpha_k - 1) = \lambda \pi_k \quad \text{—— (1)}$$

$$\underbrace{\sum_{k=1}^{K} \sum_{n=1}^{N} \gamma(z_{nk})}_{=N} + \sum_{k=1}^{K} \alpha_k - K = \lambda \underbrace{\sum_{k=1}^{K} \pi_k}_{=1} \qquad \left(\text{multiplying both sides by } \sum_{k=1}^{K}\right)$$

$$N + \sum_{k=1}^{K} \alpha_k - K = \lambda$$

Substiting λ in (1) :

$$\boxed{\pi_k = \frac{N_k + \alpha_k - 1}{N + \sum_k \alpha_k - K}}$$

// Ans.

maximize μ_k :

$$\frac{\partial L}{\partial \mu_k} = \sum_{n=1}^{N} \gamma(z_{nk}) \left[\frac{x_{ni}}{\mu_{ki}} - \frac{1 - x_{ni}}{1 - \mu_{ki}}\right] + \frac{a_k - 1}{\mu_{ki}} - \frac{b_k - 1}{1 - \mu_{ki}} = 0$$

$$= \sum_{n=1}^{N} \gamma(z_{nk}) \left[\frac{x_{ni}(1 - \mu_{ki}) - \mu_{ki}(1 - x_{ni})}{\mu_{ki}(1 - \mu_{ki})}\right] + \frac{(a_k - 1)(1 - \mu_{ki}) - \mu_{ki}(b_k - 1)}{\mu_{ki}(1 - \mu_{ki})} = 0$$

$$0 = \sum_{n=1}^{N} \gamma(z_{nk})(x_{ni} - \mu_{ki}) + a_k - 1 - \mu_{ki}(a_k + b_k - 2)$$

$$\sum_{n=1}^{N} \gamma(z_{nk}) \mu_{ki} + \mu_{ki}(a_k + b_k - 2) = \sum_{n=1}^{N} \gamma(z_{nk}) x_{ni} + a_k - 1$$

$$\therefore \boxed{\mu_{ki} = \frac{\sum_{n=1}^{N} \gamma(z_{nk}) x_{ni} + a_k - 1}{\sum_{n=1}^{N} \gamma(z_{nk}) + (a_k - 1) + (b_k - 1)}}$$

// Ans.