COMP 5212 Machine Learning (2024 Fall)
Homework 2:
Hand out: October 15, 2024
Due: October 29, 2024, 11:59 PM
Total Points: 50

Name: Tran Chun Yui
SID: 20963715
Your solution should contain below information at the top of its first page.

1. Your name
2. Your student id number

Some Notes:

- Homeworks will not be easy, please start early.
- For this homework, please submit a single PDF file to Canvas. Make sure to prepare your solution to each problem on a separate page.
- The total points of this homework is 100, and a score of 100 already gives you full grades on this homework. There are possibly bonus questions, you can optionally work on them if you are interested and have time.
- You can choose either using LaTeX by inserting your solutions to the problem pdf, or manually write your solutions on clean white papers and scan it as a pdf file – in the case of handwriting, please write clearly and briefly, we are not responsible to extract information from unclear handwriting. **We highly recommend you use LaTeX for the sake of any misunderstandings about the handwriting.** If your submission is a scan of a handwritten solution, make sure that it is of high enough resolution to be easily read. At least 300dpi and possibly denser.
- We encourage students to work in groups for homeworks, but the students need to write down the homework solutions or the code independently. In case that you work with others on the homework, please write down the names of people with whom you've discussed the homework. You are not allowed to copy, refer to, or look at the exact solutions from previous years, online, or other resources.
- **Late Policy:** 3 free late days in total across the semester, for additional late days, 20% penalization applied for each day late. **No assignment will be accepted more than 3 days late**.
- Please refer to the Course Logistics page for more details on the honor code and logisitcs. We have zero tolerance — in the case of honor code violation for a single time, you will fail this course directly.

Expectation Maximization (50 pts)

The EM algorithm is a general technique for finding maximum likelihood solutions for probabilistic models having latent variables. Take a probabilistic model in which we denote all of the *observed* variables as $\mathbf{X}$ and all of the *hidden* variables as $\mathbf{Z}$ (here we assume $\mathbf{Z}$ is discrete, for the sake of simplicity). Let us assume that the joint distribution is $p(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is the set of all parameters describing this distribution (e.g. for a Gaussian distribution, $\boldsymbol{\theta} = (\mu, \Sigma)$). The goal is to maximize the likelihood function

$$p(\mathbf{X} \mid \boldsymbol{\theta}) = \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\theta})$$

1. [**5 points**] For an arbitrary distribution $q(\mathbf{Z})$ over the latent variables, show that the following decomposition holds:

$$\ln p(\mathbf{X} \mid \boldsymbol{\theta}) = \mathcal{L}(q, \boldsymbol{\theta}) + \mathrm{D}_{\mathrm{KL}}\left(q \| p_{\text{post}}\right) \tag{1}$$

where $p_{\text{post}} = p(\mathbf{Z} \mid \mathbf{X}, \boldsymbol{\theta})$ is the posterior distribution. Also find the formulation of $\mathcal{L}(q, \boldsymbol{\theta})$.

1.

$$lnP(X|\theta) = ln\sum_z P(X,Z|\theta) = ln\sum_z P(X|Z,\theta)P(Z|\theta)$$

$$D_{KL}(q||p_{post}) = E_{q(z)}ln(\frac{q(z)}{P(Z|X,\theta)}) = \sum_z q(z)ln(\frac{q(z)}{P(Z|X,\theta)})$$

$$D_{KL}(q||p_{post}) = -\sum_z q(z)ln(\frac{P(X|Z,\theta)P(Z|\theta)}{q(z)\sum_z P(X|Z,\theta)P(Z|\theta)})$$

$$= -\sum_z q(z)ln(\frac{P(X|Z,\theta)P(Z|\theta)}{q(z)}) + \sum_z q(z)ln\sum_z P(X|Z,\theta)P(Z|\theta)$$

where $ln\sum_z P(X|Z,\theta)P(Z|\theta) = lnP(X|\theta)$

$$D_{KL}(q||p_{post}) = lnP(X|\theta)\sum_z q(z) - \sum_z q(z)ln(\frac{P(X|Z,\theta)P(Z|\theta)}{q(z)})$$

where $\sum_z q(z) = 1$
Therefore, $lnP(X|\theta)$ is:

$$lnP(X|\theta) = \sum_z q(z)ln(\frac{P(X|Z,\theta)P(Z|\theta)}{q(z)}) + D_{KL}(q||p_{post})$$

$$\mathcal{L}(q, \boldsymbol{\theta}) = \sum_z q(z)ln(\frac{P(X|Z,\theta)P(Z|\theta)}{q(z)})$$

2. [**5 points**] Prove that $\mathcal{L}(q, \boldsymbol{\theta}) \leq \ln p(\mathbf{X} \mid \boldsymbol{\theta})$, and that equality holds if and only if $q(\mathbf{Z}) = p(\mathbf{Z} \mid \mathbf{X}, \boldsymbol{\theta})$. (You are not allowed to directly use the fact that $\mathrm{D}_{\mathrm{KL}} \geq 0$)

2.

$$\mathcal{L}(q, \boldsymbol{\theta}) = \ln p(\mathbf{X} \mid \boldsymbol{\theta}) - \mathrm{D}_{\mathrm{KL}}\left(q \| p_{\text{post}}\right)$$

while

$$\ln p(\mathbf{X} \mid \boldsymbol{\theta}) = \ln \sum_{z} \frac{q(z) P(X|Z, \theta) P(Z)}{q(z)}$$

By Jensen inequality,

$$\ln p(\mathbf{X} \mid \boldsymbol{\theta}) \geq \sum_{z} q(z) \ln \frac{P(X|Z, \theta) P(Z)}{q(z)}$$

$$\mathcal{L}(q, \boldsymbol{\theta}) \leq \ln p(\mathbf{X} \mid \boldsymbol{\theta})$$

Equality in Jensen's inequality holds if and only if $\frac{P(\mathbf{X}|\mathbf{Z}, \boldsymbol{\theta}) P(\mathbf{Z})}{q(z)}$ is constant with respect to z, which implies:

$$q(z) = \frac{P(\mathbf{X} \mid \mathbf{Z}, \boldsymbol{\theta}) P(\mathbf{Z})}{\sum_z P(\mathbf{X} \mid \mathbf{Z}, \boldsymbol{\theta}) P(\mathbf{Z})} = P(\mathbf{Z} \mid \mathbf{X}, \boldsymbol{\theta})$$

3. [**5 points**] Prove that in the E-step, the lower bound $\mathcal{L}\left(q, \boldsymbol{\theta}_{\text{curr}}\right)$ is maximized with respect to the distribution $q(\mathbf{Z})$

3. During the E-step, we aim to maximize $\mathcal{L}(q, \boldsymbol{\theta}_{\text{curr}})$ with respect to $q(\mathbf{Z})$.

Since we know:

$$q(z) = P(\mathbf{Z} \mid \mathbf{X}, \boldsymbol{\theta}_{\text{curr}})$$

and that:

$$\mathcal{L}(q, \boldsymbol{\theta}_{\text{curr}}) \leq \ln p(\mathbf{X} \mid \boldsymbol{\theta}_{\text{curr}})$$

we see that when $q(z) = P(\mathbf{Z} \mid \mathbf{X}, \boldsymbol{\theta}_{\text{curr}})$, the bound reaches its maximum value:

$$\mathcal{L}(q, \boldsymbol{\theta}_{\text{curr}}) = \ln p(\mathbf{X} \mid \boldsymbol{\theta}_{\text{curr}}) = \max \mathcal{L}(q, \boldsymbol{\theta}_{\text{curr}})$$

4. [**5 points**] In the M-step, the lower bound $\mathcal{L}(q, \boldsymbol{\theta})$ is maximized with respect to $\boldsymbol{\theta}$ while keeping $q(\mathbf{Z})$ fixed, resulting in a new value of parameters $\boldsymbol{\theta}_{\text{new}}$. Prove that this step will result in an increase in left-hand-side of (1) (if it is not already in a local maximum).

4. M-Step:

$$\theta^{t+1} = \underset{\theta^t}{\operatorname{argmax}} \sum_{i} \mathcal{L}(q, \boldsymbol{\theta}^{\boldsymbol{t}})$$

So,

$$\mathcal{L}(q, \boldsymbol{\theta}^{\boldsymbol{t+1}}) \geq \mathcal{L}(q, \boldsymbol{\theta}^{\boldsymbol{t}})$$

$$\ln p(\mathbf{X} \mid \boldsymbol{\theta}^{\boldsymbol{t}}) = \mathcal{L}(q, \boldsymbol{\theta}^{\boldsymbol{t}}) + \mathrm{D}_{\mathrm{KL}}\left(q \| p_{\text{post}}\right)$$

$$\ln p(\mathbf{X} \mid \boldsymbol{\theta}^{\boldsymbol{t}}) \leq \mathcal{L}(q, \boldsymbol{\theta}^{\boldsymbol{t+1}}) + \mathrm{D}_{\mathrm{KL}}\left(q \| p_{\text{post}}\right)$$

$$\ln p(\mathbf{X} \mid \boldsymbol{\theta}^{\boldsymbol{t}}) \leq \ln p(\mathbf{X} \mid \boldsymbol{\theta}^{\boldsymbol{t+1}})$$

so $\ln p(\mathbf{X} \mid \boldsymbol{\theta})$ will increase monotonically until it is local maximum.

5. [**6 points**] Substitute $q(\mathbf{Z}) = p\left(\mathbf{Z} \mid \mathbf{X}, \boldsymbol{\theta}_{\text{curr}}\right)$ in (1), and show that

$$\mathcal{L}(q, \boldsymbol{\theta}) = \mathbb{E}_q[\text{ complete-data log likelihood }] + H(q).$$

In other words, in the M-step we are maximizing the expectation of the complete-data log likelihood $p(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\theta})$, since the entropy term is independent of $\boldsymbol{\theta}$.
5. Again

$$\ln p(\mathbf{X} \mid \boldsymbol{\theta}) = \mathcal{L}(q, \boldsymbol{\theta}) + \mathrm{D}_{\mathrm{KL}}\left(q \| p_{\text{post}}\right)$$

And,

$$\mathcal{L}(q, \boldsymbol{\theta}) = \sum_z P(\mathbf{Z} \mid \mathbf{X}, \boldsymbol{\theta}_{\text{curr}}) \ln \frac{P(\mathbf{X} \mid \mathbf{Z}, \boldsymbol{\theta}) P(\mathbf{Z} \mid \boldsymbol{\theta})}{P(\mathbf{Z} \mid \mathbf{X}, \boldsymbol{\theta}_{\text{curr}})}$$

$$= \sum_z P(\mathbf{Z} \mid \mathbf{X}, \boldsymbol{\theta}_{\text{curr}}) \ln P(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\theta}) - \sum_z P(\mathbf{Z} \mid \mathbf{X}, \boldsymbol{\theta}_{\text{curr}}) \ln P(\mathbf{Z} \mid \mathbf{X}, \boldsymbol{\theta}_{\text{curr}})$$

$$= \mathbb{E}_{P(\mathbf{Z} \mid \mathbf{X}, \boldsymbol{\theta}_{\text{curr}})}\left[\ln P(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\theta})\right] + H(P(\mathbf{Z} \mid \mathbf{X}, \boldsymbol{\theta}_{\text{curr}}))$$

$$= \mathbb{E}_q\left[\ln P(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\theta})\right] + H(q)$$

Where $\ln P(X, Z|\theta)$ is the complete-data log likelihood.

6. [**7 points**] Show that the lower bound $\mathcal{L}(q, \boldsymbol{\theta})$, where $q(\mathbf{Z}) = q^{\star}(\mathbf{Z}) = p\left(\mathbf{Z} \mid \mathbf{X}, \boldsymbol{\theta}_{\text{curr}}\right)$, has the same gradient w.r.t. $\boldsymbol{\theta}$ as the log likelihood function $p(\mathbf{X} \mid \boldsymbol{\theta})$ at the point $\boldsymbol{\theta} = \boldsymbol{\theta}_{\text{curr}}$.

6.

$$q(Z) = P(Z|X, \theta_{curr})$$

From (2), we proved that when $q(Z) = P(Z|X, \theta_{curr})$,

$$\mathcal{L}(q, \boldsymbol{\theta}) \leq \ln p(\mathbf{X} \mid \boldsymbol{\theta})$$

Therefore, the gradient of $\mathcal{L}(q, \boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ is:

$$\nabla_\theta \mathcal{L}(q, \boldsymbol{\theta}) = \nabla_\theta \ln p(\mathbf{X} \mid \boldsymbol{\theta}) = \nabla_\theta \ln \sum_{\mathbf{Z}} P(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\theta}).$$

Expanding this expression, we get:

$$\nabla_\theta \ln p(\mathbf{X} \mid \boldsymbol{\theta}) = \frac{\sum_{\mathbf{Z}} \nabla_\theta P(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\theta})}{\sum_{\mathbf{Z}} P(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\theta})}.$$

Thus, the lower bound $\mathcal{L}(q,\boldsymbol{\theta})$ and the log likelihood function $\ln p(\mathbf{X} \mid \boldsymbol{\theta})$ share the same gradient with respect to $\boldsymbol{\theta}$ when $\boldsymbol{\theta} = \boldsymbol{\theta}_{\text{curr}}$.

7. [**2 points**] Have you found what implies the convergence properties of EM algorithm?

7. The convergence properties of the EM algorithm is from the fact that each step guarantees a non-decreasing likelihood function $\ln p(\mathbf{X} \mid \boldsymbol{\theta})$.
In E-Step, It updates $q(\mathbf{Z})$ to $p(\mathbf{Z} \mid \mathbf{X}, \boldsymbol{\theta}_{\text{curr}})$, which maximizes the lower bound $\mathcal{L}(q, \boldsymbol{\theta}_{\text{curr}})$.
In M-Step, it updates $\mathcal{L}(q,\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$, given a fixed $q(\mathbf{Z})$. These steps ensure that the likelihood $\ln p(\mathbf{X} \mid \boldsymbol{\theta})$ is either strictly increasing or remains constant, leading to convergence towards a local maximum of the likelihood function.

Now we get into practice, consider a data set $D = \{x_1, \ldots, x_M\}$, where $x_j \in (0,1)$ and $M = 1{,}000{,}000$. The observed x values independently come from a mixture of the Uniform distribution on (0,1) (denoted as component 0) and a distribution with density function $\alpha x^{\alpha-1}$ with unknown parameter α (denoted as component 1). Let $z_j \in \{0,1\}$ be the latent variable indicating whether x_j is from component 0 ($z_j = 0$) or component 1 ($z_j = 1$). Then the probabilistic model can be written as:

$$\pi_0 = \Pr(z_j = 0) : x_j \sim \mathcal{U}[0,1], \quad \text{if } z_j = 0,$$

$$\pi_1 = \Pr(z_j = 1) : x_j \sim \alpha x^{\alpha-1}, \quad \text{if } z_j = 1.$$

8. [**5 points**] Let $\Theta = \{\pi_0, \pi_1, \alpha\}$ be the set of unknown parameters to be estimated. Write down the complete data log-likelihood function $L(\Theta) = \sum_i \log p(x_i, z_i)$ for this problem.

8.

$$L(\Theta) = \sum_{i=1}^{M} \log p(x_i, z_i \mid \Theta)$$

$$L(\Theta) = \sum_{i=1}^{M} \left(z_i \log\left(\pi_1 \alpha x_i^{\alpha-1}\right) + (1 - z_i) \log\left(\pi_0 \cdot 1\right) \right)$$

$$L(\Theta) = \sum_{i=1}^{M} \left(z_i \left(\log \pi_1 + \log \alpha + (\alpha - 1) \log x_i\right) + (1 - z_i) \log \pi_0 \right)$$

9. [**10 points**] Derive an EM algorithm for parameter estimation, where $\Theta = \{\pi_0, \pi_1, \alpha\}$ is the parameter set to be estimated and $\{z_1, \ldots, z_M\}$ are considered as missing data.

9. Consider $\gamma = P(Z \mid X) = \frac{P(X|Z)P(Z)}{\sum_i P(X|Z)P(Z)}$.

$$\gamma_0 = \frac{\pi_0}{\pi_0 + \pi_1 \alpha x^{\alpha-1}}$$

To derive an EM algorithm for parameter estimation, where $\Theta = \{\pi_0, \pi_1, \alpha\}$ is the parameter set to be estimated and $\{z_1, \ldots, z_M\}$ are considered as missing data, we proceed as follows:
E Step:

$$\gamma_0 = P(Z_i = 0 \mid X) = \frac{\pi_0}{\pi_1 \alpha x^{\alpha-1} + \pi_0}$$

$$\gamma_1 = P(Z_i = 1 \mid X) = \frac{\pi_1 \alpha x^{\alpha-1}}{\pi_1 \alpha x^{\alpha-1} + \pi_0} = 1 - \gamma_0$$

M Step:
Update π_0 and π_1 by maximizing with the constraint $\pi_0 + \pi_1 = 1$:

$$\pi_0^{(t+1)} = \frac{1}{M} \sum_{i=1}^{M} \gamma_0^{(i)}, \quad \pi_1^{(t+1)} = 1 - \pi_0^{(t+1)}$$

To update α, we take the gradient of the likelihood function:

$$L(\Theta) = \sum_{i=1}^{M} \left(z_i \left(\log \pi_1 + \log \alpha + (\alpha - 1) \log x_i \right) + (1 - z_i) \log \pi_0 \right)$$

The gradient with respect to α is:

$$\frac{\partial L(\Theta)}{\partial \alpha} = \sum_{i=1}^{M} \gamma_1^{(i)} \left(\frac{1}{\alpha} + \log x_i \right)$$

Setting $\frac{\partial L(\Theta)}{\partial \alpha} = 0$, we solve for α:

$$\alpha = -\frac{1}{\sum_{i=1}^{M} \gamma_1^{(i)} \log x_i}$$

10. [**Bonus question, 5 points**] Suppose we have some side information collected in two vectors $\mathbf{A} = [A_1, \ldots, A_M]$ and $\mathbf{B} = [B_1, \ldots, B_M]$, where $A_j \in \{0, 1\}$ and $B_j \in \{0, 1\}$. Each of A_j and B_j is observed together with x_j and follows the i.i.d. assumption (you can think them as additional independent features for the data). To incorporate $\mathbf{A}$ and $\mathbf{B}$ to infer the posterior of $\mathbf{Z}$, we assume the conditional dependence $\Pr(\mathbf{A}, \mathbf{B}, x|\mathbf{Z}) = \Pr(\mathbf{A}|\mathbf{Z}) \Pr(\mathbf{B}|\mathbf{Z}) \Pr(x|\mathbf{Z})$. Then we model the relationship between A_j and z_j as $q_{0,A} = \Pr(A_j = 1|z_j = 0)$ and $q_{1,A} = \Pr(A_j = 1|z_j = 1)$, respectively. Similarly, we have $q_{0,B} = \Pr(B_j = 1|z_j = 0)$ and $q_{1,B} = \Pr(B_j = 1|z_j = 1)$. Derive an EM algorithm to estimate all the parameters $\{\pi_0, \pi_1, \alpha, q_{0,A}, q_{1,A}, q_{0,B}, q_{1,B}\}$. Again, $\{z_1, \ldots, z_M\}$ are considered as missing data.

10.

E Step

Calculate the posterior probabilities γ_0 and γ_1 for each Z_i given the observations A_i, B_i, and X_i:

$$\gamma_0 = P(Z_i = 0 \mid A_i, B_i, X_i) = \frac{P(A_i \mid Z_i = 0)P(B_i \mid Z_i = 0)P(X_i \mid Z_i = 0)P(Z_i = 0)}{\sum_{z\in\{0,1\}} P(A_i \mid Z_i = z)P(B_i \mid Z_i = z)P(X_i \mid Z_i = z)P(Z_i = z)}$$

Substituting each probability term, we get:

$$\gamma_0 = \frac{q_{0,A}^{A_i}(1-q_{0,A})^{1-A_i} \cdot q_{0,B}^{B_i}(1-q_{0,B})^{1-B_i} \cdot \pi_0}{q_{0,A}^{A_i}(1-q_{0,A})^{1-A_i} \cdot q_{0,B}^{B_i}(1-q_{0,B})^{1-B_i} \cdot \pi_0 + q_{1,A}^{A_i}(1-q_{1,A})^{1-A_i} \cdot q_{1,B}^{B_i}(1-q_{1,B})^{1-B_i} \cdot \pi_1 \alpha x_i^{\alpha-1}}$$

Similarly,

$$\gamma_1 = P(Z_i = 1 \mid A_i, B_i, X_i) = 1 - \gamma_0$$

$$= \frac{q_{1,A}^{A_i}(1-q_{1,A})^{1-A_i} \cdot q_{1,B}^{B_i}(1-q_{1,B})^{1-B_i} \cdot \pi_1 \alpha x_i^{\alpha-1}}{q_{0,A}^{A_i}(1-q_{0,A})^{1-A_i} \cdot q_{0,B}^{B_i}(1-q_{0,B})^{1-B_i} \cdot \pi_0 + q_{1,A}^{A_i}(1-q_{1,A})^{1-A_i} \cdot q_{1,B}^{B_i}(1-q_{1,B})^{1-B_i} \cdot \pi_1 \alpha x_i^{\alpha-1}}$$

M Step

Maximize the expected complete log-likelihood with respect to each parameter, given γ_0 and γ_1 calculated in the E Step.

1. Update π_0 and π_1:

$$\pi_0^{(t+1)} = \frac{1}{M}\sum_{i=1}^{M}\gamma_0^{(i)}, \quad \pi_1^{(t+1)} = 1 - \pi_0^{(t+1)} = \frac{1}{M}\sum_{i=1}^{M}\gamma_1^{(i)}$$

2. Update α: Similar to question 9:

$$\alpha^{(t+1)} = -\frac{1}{\sum_{i=1}^{M}\gamma_1^{(i)}\log x_i}$$

3. Update $q_{0,A}, q_{1,A}, q_{0,B}, q_{1,B}$:

$$q_{0,A}^{(t+1)} = \frac{\sum_{i=1}^{M}\gamma_0^{(i)}A_i}{\sum_{i=1}^{M}\gamma_0^{(i)}}$$

$$q_{1,A}^{(t+1)} = \frac{\sum_{i=1}^{M}\gamma_1^{(i)}A_i}{\sum_{i=1}^{M}\gamma_1^{(i)}}$$

Similarly, for $q_{0,B}$ and $q_{1,B}$:

$$q_{0,B}^{(t+1)} = \frac{\sum_{i=1}^{M}\gamma_0^{(i)}B_i}{\sum_{i=1}^{M}\gamma_0^{(i)}}$$

$$q_{1,B}^{(t+1)} = \frac{\sum_{i=1}^{M} \gamma_1^{(i)} B_i}{\sum_{i=1}^{M} \gamma_1^{(i)}}$$