COMP 5212 Machine Learning (2024 Spring)
Homework 2 Solution:
Hand out: March 13, 2024
Due: March 28, 2024, 11:59 PM
Total Points: 50

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 $\mathcal{L}(q, \boldsymbol{\theta})$.

$$\begin{aligned}
\ln p(\mathbf{X} \mid \boldsymbol{\theta}) &= \ln \frac{p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})}{p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})} \\
&= \sum_{\mathbf{z}} q(\mathbf{z}) \cdot ln \left(\frac{p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})}{q(\mathbf{z})} \cdot \frac{q(\mathbf{z})}{p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})} \right) \\
&= \sum_{\mathbf{z}} q(\mathbf{z}) \; ln \frac{p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})}{q(\mathbf{z})} + \sum_{\mathbf{z}} q(\mathbf{z}) \; ln \frac{q(\mathbf{z})}{p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})} \\
&= \sum_{\mathbf{z}} q(\mathbf{z}) \; ln \frac{p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})}{q(\mathbf{z})} + D_{KL}(q||p_{\text{post}})
\end{aligned}$$

And the formulation is just the ELBO : $\sum_{\mathbf{z}} q(\mathbf{z}) \; ln \frac{p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})}{q(\mathbf{z})}$

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 $\mathrm{D}_{\mathrm{KL}} \geq 0$)

$$\begin{aligned}
\log p(x; \theta) &= \log \sum_{z} p(x, z; \theta) \\
&= \log \sum_{z} Q(z) \frac{p(x, z; \theta)}{Q(z)} \\
&\geq \sum_{z} Q(z) \log \frac{p(x, z; \theta)}{Q(z)} = \mathrm{ELBO}(x; Q, \theta)
\end{aligned}$$

Since $f(x) = \log x$ is a concave function, by Jensen's inequality, wo have the above inequality holds. The equality holds iff $\frac{p(x,z|\theta)}{Q(z)} = \text{const}$. This can be easily accomplished by choosing

$Q(z) \propto p(x, z|\theta)$, Since we knoe that $\sum_z Q(z) = 1$ which gives:

$$\begin{aligned} Q(z) &= \frac{p(x, z; \theta)}{\sum_z p(x, z; \theta)} \\ &= \frac{p(x, z; \theta)}{p(x; \theta)} \\ &= p(z|x; \theta) \end{aligned}$$

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 $q(\mathbf{Z})$

In E-step, we have

$$Q_i\left(z^{(i)}\right) := p\left(z^{(i)}|x^{(i)}; \theta\right)$$

We claim that $p(z|x; \theta) = \text{argmax}_{Q(z)} \text{ELBO}(x; Q, \theta)$ when $Q(z) = p(z|x; \theta)$

$$\begin{aligned} \text{ELBO}(x; Q, \theta) &= \sum_z Q(z) \log \frac{p(x, z; \theta)}{Q(z)} \\ &= \sum_z p(z|x; \theta) \log \frac{p(x, z; \theta)}{p(z|x; \theta)} \\ &= \sum_z p(z|x; \theta) \log p(x; \theta) \\ &= \log p(x; \theta) \sum_z p(z|x; \theta) \\ &= \log p(x; \theta) \end{aligned}$$

Therefore, in E-step, we maximize $\text{ELBO}(x; Q, \theta)$ with respect to $Q(z)$.

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).

In M-step, we have

$$\begin{aligned} \theta &:= \underset{\theta}{\text{argmax}} \sum_{i=1}^{n} \text{ELBO}\left(x^{(i)}; Q_i, \theta\right) \\ &= \underset{\theta}{\text{argmax}} \sum_{i=1}^{n} \sum_{z^{(i)}} Q_i\left(z^{(i)}\right) \log \frac{p\left(x^{(i)}, z^{(i)}; \theta\right)}{Q_i\left(z^{(i)}\right)} \\ &= \underset{\theta}{\text{argmax}} \sum_{i=1}^{n} \sum_{z^{(i)}} Q_i\left(z^{(i)}\right) \log p\left(x^{(i)}, z^{(i)}; \theta\right) \end{aligned}$$

Therefore, in M-step , we maximize $\text{EBLO}(x; Q, \theta)$ with respect to θ while keep $Q(z)$ un-

changed. Consider step t and $t-1$ before convergence.

$$\begin{aligned}\log p(x|\theta^{(t)}) &\geq \text{ELBO}\left(x|Q\left(z^{(t-1)}\right);\theta^{(t)}\right)\\ &\geq \text{ELBO}\left(x|Q\left(z^{(t-1)}\right);\theta^{(t-1)}\right)\\ &= \log p(x|\theta^{(t-1)})\end{aligned}$$

Therefore, $\log p(x|\theta)$ increases monotonically after M-step.

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 $\boldsymbol{\theta}$.

$$\begin{aligned}\text{ELBO}(x|Q,\theta) &= \sum_z Q(z)\log\frac{p(x,z|\theta)}{Q(z)}\\ &= \sum_z p(z|x,\theta)\log p(x,z|\theta) - \sum_z p(z|x,\theta)\log p(z|x,\theta)\\ &= E_{z\sim Q(z)}[\log p(x,z|\theta)] + H(Q(z))\end{aligned}$$

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 $\boldsymbol{\theta} = \boldsymbol{\theta}_{\text{curr}}$.

$$\begin{aligned}\nabla_\theta\text{ELBO}(x|Q,\theta) &= \nabla_\theta\sum_z Q(z)\log p(x,z|\theta) - \nabla_\theta\sum_z Q(z)\log Q(z)\\ &= \sum_z Q(z)\frac{\nabla_\theta p(x,z|\theta)}{p(x,z|\theta)}\\ &= \sum_z \frac{p(z|x;\theta_{curr})}{p(x,z|\theta)}\nabla_\theta p(x,z|\theta)\end{aligned}$$

Take $\theta = \theta_{curr}$, we have

$$\begin{aligned}\nabla_\theta\text{ELBO}(x|Q,\theta)\big|_{\theta=\theta_{curr}} &= \sum_z \frac{p(z|x;\theta_{curr})}{p(x,z|\theta)}\nabla_\theta p(x,z|\theta)\big|_{\theta=\theta_{curr}}\\ &= \sum_z \frac{1}{p(x|\theta_{curr})}\nabla_\theta p(x,z|\theta)\big|_{\theta=\theta_{curr}}\\ &= \frac{1}{p(x|\theta_{curr})}\nabla_\theta p(x|\theta)\big|_{\theta=\theta_{curr}}\\ &= \nabla_\theta \log p(x|\theta)\big|_{\theta=\theta_{curr}}\end{aligned}$$

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

The convergence of EM algorithm implies $D_{KL}(Q||p_{z|x}) = 0$, which means that the distribution of $Q(z)$ is the same as $p(z|x,\theta)$. As shown in (4), the monotonically increase of the log likelihood implies the convergence of EM algorithm.

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.

$$p(z|\theta) = \begin{cases} \pi_0 & \text{if } z = 0 \\ \pi_1 & \text{if } z = 1 \end{cases}$$

$$p(x|z;\theta) = \begin{cases} 1 & \text{if } z = 0 \\ \alpha x^{\alpha-1} & \text{if } z = 1 \end{cases}$$

$$\begin{aligned} p(x, z|\theta) &= p(x|z;\theta)p(z|\theta) \\ &= \left(\pi_1 \alpha x_i^{\alpha-1}\right)^{z_i} (\pi_0)^{1-z_i} \end{aligned}$$

$$\begin{aligned} p(x|\theta) &= p(x, z=0|\theta) + p(x, z=1|\theta) \\ &= \pi_1 \alpha x_i^{\alpha-1} + \pi_0 \end{aligned}$$

$$\begin{aligned} \ell(\theta) &= \log \prod_{i=1}^{M} p(x_i, z_i|\theta) \\ &= \log \prod_{i=1}^{M} p(x_i|z_i;\theta)p(z_i|\theta) \\ &= \log \prod_{i=1}^{M} \left(\pi_1 \alpha x_i^{\alpha-1}\right)^{z_i} (\pi_0)^{1-z_i} \\ &= \sum_{i=1}^{M} z_i(\log \pi_1 + \log \alpha + (\alpha-1)\log x_i) + (1-z_i)\log \pi_0 \end{aligned}$$

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.

Algorithm:

1. Initialize π_0, π_1, α.

2. Repeat until convergence: **E-step**: for each $x_i \in D$, update posterior distribution $p_i(z_i)$ according to current θ

$$\pi_{1i} := p_i(z_i = 1) =: p_i(z_i = 1|x_i; \theta) = \frac{\pi_1 \alpha x_i^{\alpha-1}}{\pi_0 + \pi_1 \alpha x_i^{\alpha-1}}$$

$$\pi_{0i} := p_i(z_i = 0) =: p_i(z_i = 0|x_i; \theta) = \frac{\pi_0}{\pi_0 + \pi_1 \alpha x_i^{\alpha-1}}$$

M-step: update the current α

$$\begin{aligned}
\alpha &:= \operatorname*{argmax}_{\alpha} \sum_{i=1}^{M} \sum_{z_i=1} p_i(z_i = 1) \log p(x_i, z_i = 1; \theta) \\
&= \operatorname*{argmax}_{\alpha} \sum_{i=1}^{M} \sum_{z_i=1} \pi_1 (\log \pi_1 + \log \alpha + (\alpha - 1) \log x_i) \\
&= \operatorname*{argmax}_{\alpha} \sum_{i=1}^{M} \sum_{z_i=1} \log \alpha + \alpha \log x_i
\end{aligned}$$

$$\frac{\partial}{\partial \alpha} \sum_{i=1}^{M} \sum_{z_i=1} \log \alpha + \alpha \log x_i = \sum_{i=1}^{M} \sum_{z_i=1} \frac{1}{\alpha} + \log x_i$$

We have

$$\begin{aligned}
\pi_1 &:= \frac{1}{M} \sum_{i=1}^{M} \pi_{i1} \\
\pi_0 &:= \frac{1}{M} \sum_{i=1}^{M} \pi_{i0} \\
\alpha &:= -\frac{\sum_{i=1}^{M} \sum_{z_i=1} 1}{\sum_{i=1}^{M} \sum_{z_i=1} \log x_i}
\end{aligned}$$

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.

Algorithm:

1. Initialize $\pi_0, \pi_1, \alpha, q_{0,A}, q_{1,A}, q_{0,B}, q_{1,B}$.

2. Repeat until convergence: **E-step**: for each $x_i \in D$, update posterior distribution $p_i(z_i)$ according to current θ Let

$$\begin{cases} a_{1i} = p_i(A_i|z_i = 1;\theta) = q_{1,A}^{A_i}(1-q_{1,A})^{1-A_i} \\ b_{1i} = p_i(B_i|z_i = 1;\theta) = q_{1,B}^{B_i}(1-q_{1,B})^{1-B_i} \\ a_{0i} = p_i(A_i|z_i = 0;\theta) = q_{0,A}^{A_i}(1-q_{0,A})^{1-A_i} \\ b_{0i} = p_i(B_i|z_i = 0;\theta) = q_{0,B}^{B_i}(1-q_{0,B})^{1-B_i} \end{cases}$$

$$\pi_{1i} := p_i(z_i = 1|A_i, B_i, x_i;\theta) = \frac{a_{1i}b_{1i}\pi_1\alpha x_i^{\alpha-1}}{a_{1i}b_{1i}\pi_1\alpha x_i^{\alpha-1} + a_{0i}b_{0i}\pi_0}$$

$$\pi_{0i} =: p_i(z_i = 0|A_i, B_i, x_i;\theta) = \frac{a_{0i}b_{0i}\pi_0}{a_{1i}b_{1i}\pi_1\alpha x_i^{\alpha-1} + a_{0i}b_{0i}\pi_0}$$

M-step: update the current θ

$$\theta := \underset{\theta}{\operatorname{argmax}} \sum_{i=1}^{M}\sum_{z_i} p_i(z_i)\log p_i(A_i, B_i, x_i, z_i|\theta)$$

Specifically, we have

$$\begin{aligned} \alpha, q_{1,A}, q_{1,B} &:= \underset{\alpha,q_{1,A},q_{1,B}}{\operatorname{argmax}} \sum_{i=1}^{M}\sum_{z_i=1} p_i(z_i = 1)\log p(A_i, B_i, x_i, z_i = 1|\theta) \\ &= \underset{\alpha,q_{1,A},q_{1,B}}{\operatorname{argmax}} \sum_{i=1}^{M}\sum_{z_i=1} p_i(z_i = 1)\log p(A_i, B_i, x_i|z_i = 1;\theta)p(z_i = 1|\theta) \\ &= \underset{\alpha,q_{1,A},q_{1,B}}{\operatorname{argmax}} \sum_{i=1}^{M}\sum_{z_i=1} \pi_{1i}(\log a_{1i} + \log b_{1i} + \log \pi_{1i} + \log\alpha + (\alpha-1)\log x_i) \end{aligned}$$

$$\begin{cases} \alpha := \operatorname{argmax}_\alpha \sum_{i=1}^{M}\sum_{z_i=1}\log\alpha + \alpha\log x_i \to \alpha := -\frac{\sum_{i=1}^{M}\sum_{z_i=1}1}{\sum_{i=1}^{M}\sum_{z_i=1}\log x_i} \\ q_{1,A} := \operatorname{argmax}_{q_{1,A}} \sum_{i=1}^{M}\sum_{z_i=1}\log a_{1i} \to q_{1,A} := \frac{\sum_{i=1}^{M}\sum_{z_i=1}A_i}{\sum_{i=1}^{M}\sum_{z_i=1}1} \\ q_{1,B} := \operatorname{argmax}_{q_{1,B}} \sum_{i=1}^{M}\sum_{z_i=1}\log b_{1i} \to q_{1,B} := \frac{\sum_{i=1}^{M}\sum_{z_i=1}B_i}{\sum_{i=1}^{M}\sum_{z_i=1}1} \end{cases}$$

$$\begin{aligned} q_{0,A}, q_{0,B} &:= \underset{q_{0,A},q_{0,B}}{\operatorname{argmax}} \sum_{i=1}^{M}\sum_{z_i=0} p_i(z_i = 0)\log p(A_i, B_i, x_i, z_i = 0|\theta) \\ &:= \underset{q_{0,A},q_{0,B}}{\operatorname{argmax}} \sum_{i=1}^{M}\sum_{z_i=0} p_i(z_i = 0)\log p(A_i, B_i, x_i|z_i = 0;\theta)p(z_i = 0|\theta) \\ &= \underset{q_{0,A},q_{0,B}}{\operatorname{argmax}} \sum_{i=1}^{M}\sum_{z_i=0} \pi_0(\log a_{0i} + \log b_{0i} + \log\pi_0) \end{aligned}$$

$$\begin{cases} q_{0,A} := \operatorname{argmax}_{q_{0,A}} \sum_{i=1}^{M} \sum_{z_i=0} \log a_{0i} \rightarrow q_{0,A} := \frac{\sum_{i=1}^{M} \sum_{z_i=0} A_i}{\sum_{i=1}^{M} \sum_{z_i=0} 1} \\ q_{0,B} := \operatorname{argmax}_{q_{0,B}} \sum_{i=1}^{M} \sum_{z_i=0} \log b_{0i} \rightarrow q_{0,B} := \frac{\sum_{i=1}^{M} \sum_{z_i=0} B_i}{\sum_{i=1}^{M} \sum_{z_i=0} 1} \end{cases}$$

In summary, we have

$$\begin{aligned} \pi_1 &:= \frac{1}{M} \sum_{i=1}^{M} \pi_{1i} \\ \pi_0 &:= \frac{1}{M} \sum_{i=1}^{M} \pi_{0i} \\ \alpha &:= -\frac{\sum_{i=1}^{M} \sum_{z_i=1} 1}{\sum_{i=1}^{M} \sum_{z_i=1} \log x_i} \\ q_{1,A} &:= \frac{\sum_{i=1}^{M} \sum_{z_i=1} A_i}{\sum_{i=1}^{M} \sum_{z_i=1} 1} \\ q_{1,B} &:= \frac{\sum_{i=1}^{M} \sum_{z_i=1} B_i}{\sum_{i=1}^{M} \sum_{z_i=1} 1} \\ q_{0,A} &:= \frac{\sum_{i=1}^{M} \sum_{z_i=0} A_i}{\sum_{i=1}^{M} \sum_{z_i=0} 1} \\ q_{0,B} &:= \frac{\sum_{i=1}^{M} \sum_{z_i=0} B_i}{\sum_{i=1}^{M} \sum_{z_i=0} 1} \end{aligned}$$