2024 秋《机器学习概论》作业 3 解答

舒英特

满分 100 分，每道题 25 分（尝试解答给 1 分，答案正确每小题给 8/4 分），中文或英文作答均可。手写答案建议下次用电子版，看不清的按错误处理。

1

(a)

$$
\begin{aligned}
D_{KL}(P\|Q) &= \mathbb{E}_{X\sim P}\left[-\log\frac{Q(X)}{P(X)}\right] \\
&\geq -\log\mathbb{E}_{X\sim P}\left[\frac{Q(X)}{P(X)}\right] \\
&= -\log\sum_x P(x)\frac{Q(x)}{P(x)} \\
&= -\log\sum_x Q(x) \\
&= -\log(1) \\
&= 0
\end{aligned}
$$

我们在第 2 步使用了 Jensen 不等式。等号成立时，$\frac{Q(X)}{P(X)}$ 是一个常数，因此 $P=Q$。

(b)

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

(c)

$$
\begin{aligned}
\arg\min_\theta D_{KL}(\hat{P}||P_\theta) &= \arg\min_\theta \mathbb{E}_{X\sim\hat{P}}\left[\log\frac{\hat{P}(X)}{P_\theta(X)}\right] \\
&= \arg\min_\theta \mathbb{E}_{X\sim\hat{P}}[\log\hat{P}(X)] - \mathbb{E}_{X\sim\hat{P}}[\log P_\theta(X)] \\
&= \arg\min_\theta -\sum_x \hat{P}(x)\log P_\theta(x) \\
&= \arg\max_\theta \sum_x \frac{1}{n}\sum_{i=1}^n 1\{x^{(i)}=x\}\log P_\theta(x) \\
&= \arg\max_\theta \sum_{i=1}^n \sum_x 1\{x^{(i)}=x\}\log P_\theta(x) \\
&= \arg\max_\theta \sum_{i=1}^n \log P_\theta(x^{(i)})
\end{aligned}
$$

2

(a)

$$
\begin{aligned}
&\ell_{\text{semi-sup}}(\theta^{(t+1)}) \\
&= \sum_{i=1}^n \log\sum_{z^{(i)}} Q_i^{(t)}(z^{(i)})\frac{p(x^{(i)},z^{(i)};\theta^{(t+1)})}{Q_i^{(t)}(z^{(i)})} + \alpha\ell_{\text{sup}}(\theta^{(t+1)}) && \text{(等价变换)} \\
&\geq \sum_{i=1}^n \sum_{z^{(i)}} Q_i^{(t)}(z^{(i)})\log\frac{p(x^{(i)},z^{(i)};\theta^{(t+1)})}{Q_i^{(t)}(z^{(i)})} + \alpha\ell_{\text{sup}}(\theta^{(t+1)}) && \text{(Jensen 不等式)} \\
&\geq \sum_{i=1}^n \sum_{z^{(i)}} Q_i^{(t)}(z^{(i)})\log\frac{p(x^{(i)},z^{(i)};\theta^{(t)})}{Q_i^{(t)}(z^{(i)})} + \alpha\ell_{\text{sup}}(\theta^{(t)}) && \text{(M-step 中 argmax 决定)} \\
&= \sum_{i=1}^n \log p(x^{(i)};\theta^{(t)})\sum_{z^{(i)}} p(z^{(i)}|x^{(i)};\theta^{(t)}) + \alpha\ell_{\text{sup}}(\theta^{(t)}) && \text{(等价变换)} \\
&= \ell_{\text{unsup}}(\theta^{(t)}) + \alpha\ell_{\text{sup}}(\theta^{(t)}) && \text{(全概率公式)} \\
&= \ell_{\text{semi-sup}}(\theta^{(t)})
\end{aligned}
$$

参考：https://blog.csdn.net/qq_41554005/article/details/100591525。课件中也有讲到。

(b)

在 E-step 中，我们需要重新估计隐变量 $z^{(i)}$'s。

$$
\begin{aligned}
w_j^{(i)} &= p(z^{(i)}=j|x^{(i)};\theta) \\
&= \frac{p(z^{(i)}=j;\theta)p(x^{(i)}|z^{(i)}=j;\theta)}{\sum_{l=1}^k p(z^{(i)}=l;\theta)p(x^{(i)}|z^{(i)}=l;\theta)} \\
&= \frac{\frac{1}{(2\pi)^{d/2}|\Sigma_j|^{1/2}}\exp\left\{-\frac{1}{2}(x^{(i)}-\mu_j)^T\Sigma_j^{-1}(x^{(i)}-\mu_j)\right\}\phi_j}{\sum_{l=1}^k \frac{1}{(2\pi)^{d/2}|\Sigma_l|^{1/2}}\exp\left\{-\frac{1}{2}(x^{(i)}-\mu_l)^T\Sigma_l^{-1}(x^{(i)}-\mu_l)\right\}\phi_l}
\end{aligned}
$$

(c)

在 M-step 中，我们需要重新估计模型的参数 μ_j's, Σ_j's 和 ϕ_j's，$j \in \{1, \ldots, k\}$ 从而最大化对数似然函数：

$$\sum_{i=1}^{n}\sum_{j=1}^{k} w_j^{(i)} \log \frac{p(x^{(i)}, z^{(i)} = j; \theta)}{w_j^{(i)}} + \alpha \sum_{i=1}^{\tilde{n}} \log p(\tilde{x}^{(i)}, \tilde{z}^{(i)}; \theta)$$

（助教注：此式推导类似于 (a) 中倒数第 3 步）

去掉一些常数项后，此式与下式等价：

$$\sum_{i=1}^{n}\sum_{j=1}^{k} w_j^{(i)} \log p(x^{(i)}, z^{(i)} = j; \theta) + \alpha \sum_{i=1}^{\tilde{n}}\sum_{j=1}^{k} 1\{\tilde{z}^{(i)} = j\} \log p(\tilde{x}^{(i)}, \tilde{z}^{(i)}; \theta)$$

将有标签的数据附在无标签的数据之后，我们就得到了大小为 $n + \tilde{n}$ 的训练集，前 n 个无标签，后 $\tilde{n}$ 个有标签，其编号为 $\{n+1, \ldots, n+\tilde{n}\}$ 此时，我们令 $w_j^{(i)} = \alpha \cdot 1\{z^{(i)} = j\}, i \in \{n+1, \ldots, n+\tilde{n}\}$，则有：

$$\begin{aligned}
&\sum_{i=1}^{n}\sum_{j=1}^{k} w_j^{(i)} \log p(x^{(i)}, z^{(i)} = j; \theta) + \sum_{i=n+1}^{n+\tilde{n}}\sum_{j=1}^{k} w_j^{(i)} \log p(x^{(i)}, z^{(i)} = j; \theta) \\
&= \sum_{i=1}^{n+\tilde{n}}\sum_{j=1}^{k} w_j^{(i)} \log p(x^{(i)}, z^{(i)} = j; \theta)
\end{aligned}$$

此时目标已经与普通的 GMM 模型相同，因此引用 Lecture Notes 中的结论得到（助教注：150 页最上端，已经发给大家）：

$$\begin{aligned}
\phi_j &= \frac{1}{n + \alpha\tilde{n}} \sum_{i=1}^{n+\tilde{n}} w_j^{(i)}, \\
\mu_j &= \frac{\sum_{i=1}^{n+\tilde{n}} w_j^{(i)} x^i}{\sum_{i=1}^{n+\tilde{n}} w_j^{(i)}}, \\
\sum_j &= \frac{\sum_{i=1}^{n+\tilde{n}} w_j^{(i)} (x^{(i)} - \mu_j)(x^{(i)} - \mu_j)^T}{\sum_{i=1}^{n+\tilde{n}} w_j^{(i)}}
\end{aligned}$$

（助教注：第一行分母中的 α 是必要的，试想 $\alpha = 0$ 的情况，如果没有 α 会导致答案偏小。）

3

(a)

$L(\mu) = 30$

(b)

坐标为 (-4,3) 和 (4,0)，$L(\mu) = 4$

(c)

$L_C(\mu) = \sum_{i=1}^{r} c^{(i)} \min_j \left\| x^{(i)} - \mu^{(j)} \right\|^2$

(d)

$L_C(\mu) = 14\mu^2 + 12\mu + 14$，$\mu = -3/7$

(e)

$$L_S(\mu) = L_C(\mu) + \sum_{j=1}^{k} N_C^{(j)} \left\| \mu^{(j)} - x_{DC} \right\|^2$$

其中 L_C 定义见 (c)，此外

$$N_C^{(j)} = \sum_{i=1}^{r} c^{(i)} \cdot 1(y^{(i)} = j)$$

其中

$$y^{(i)} = \arg\min_j \left\| x^{(i)} - \mu^{(j)} \right\|^2$$

(f)

$L_C(\mu) = 28\mu^2 - 268\mu + 1414$，$\mu = 67/14$

4 扩散模型：一个变分推断的例子

(a)

$$\begin{aligned}
\boldsymbol{x}_t &= \sqrt{\alpha_t}\boldsymbol{x}_{t-1} + \sqrt{1-\alpha_t}\boldsymbol{\epsilon}_{t-1} \\
&= \sqrt{\alpha_t}(\sqrt{\alpha_{t-1}}\boldsymbol{x}_{t-2} + \sqrt{1-\alpha_{t-1}}\boldsymbol{\epsilon}_{t-2}) + \sqrt{1-\alpha_t}\boldsymbol{\epsilon}_{t-1} \\
&= \sqrt{\alpha_t\alpha_{t-1}}\boldsymbol{x}_{t-2} + \sqrt{\sqrt{\alpha_t - \alpha_t\alpha_{t-1}}^2 + \sqrt{1-\alpha_t}^2}\boldsymbol{\epsilon}_{t-2} \\
&= \sqrt{\alpha_t\alpha_{t-1}}\boldsymbol{x}_{t-2} + \sqrt{\alpha_t - \alpha_t\alpha_{t-1} + 1 - \alpha_t}\boldsymbol{\epsilon}_{t-2} \\
&= \cdots \\
&= \sqrt{\prod_{i=1}^{t}\alpha_i}\boldsymbol{x}_0 + \sqrt{1 - \prod_{i=1}^{t}\alpha_i}\boldsymbol{\epsilon}_0 \\
&= \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_0 \\
&\sim \mathcal{N}(\boldsymbol{x}_t; \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0, (1-\bar{\alpha}_t)\boldsymbol{I})
\end{aligned}$$

(b)

$$
\begin{aligned}
\log p(\boldsymbol{x}_0) &= \log p(\boldsymbol{x}_0) \int q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0) d\boldsymbol{x}_{1:T} \\
&= \int \log p(\boldsymbol{x}_0) q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0) d\boldsymbol{x}_{1:T} \\
&= \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\log p(\boldsymbol{x}_0)\right] \\
&= \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\log \frac{p(\boldsymbol{x}_{0:T})}{p(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\right] \\
&= \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\log \frac{p(\boldsymbol{x}_{0:T}) q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}{p(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0) q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\right] \\
&= \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\log \frac{p(\boldsymbol{x}_{0:T})}{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\right] + \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\log \frac{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}{p(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\right] \\
&= ELBO + D_{\mathrm{KL}}(q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0) \| p(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0))
\end{aligned}
$$

(c)

$$
\begin{aligned}
& D_{\mathrm{KL}}(q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0) \| p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)) \\
&= D_{\mathrm{KL}}(\mathcal{N}(\boldsymbol{x}_{t-1}; \boldsymbol{\mu}_q, \boldsymbol{\Sigma}_q(t)) \| \mathcal{N}(\boldsymbol{x}_{t-1}; \boldsymbol{\mu}_{\boldsymbol{\theta}}, \boldsymbol{\Sigma}_q(t))) \\
&= \frac{1}{2}\left[\log \frac{|\boldsymbol{\Sigma}_q(t)|}{|\boldsymbol{\Sigma}_q(t)|} - d + tr(\boldsymbol{\Sigma}_q(t)^{-1} \boldsymbol{\Sigma}_q(t)) + (\boldsymbol{\mu}_{\boldsymbol{\theta}} - \boldsymbol{\mu}_q)^T \boldsymbol{\Sigma}_q(t)^{-1} (\boldsymbol{\mu}_{\boldsymbol{\theta}} - \boldsymbol{\mu}_q)\right] \\
&= \frac{1}{2}\left[\log 1 - d + d + (\boldsymbol{\mu}_\theta - \boldsymbol{\mu}_q)^T \boldsymbol{\Sigma}_q(t)^{-1} (\boldsymbol{\mu}_\theta - \boldsymbol{\mu}_q)\right] \\
&= \frac{1}{2}\left[(\boldsymbol{\mu}_\theta - \boldsymbol{\mu}_q)^T \boldsymbol{\Sigma}_q(t)^{-1} (\boldsymbol{\mu}_\theta - \boldsymbol{\mu}_q)\right] \\
&= \frac{1}{2}\left[(\boldsymbol{\mu}_{\boldsymbol{\theta}} - \boldsymbol{\mu}_q)^T (\sigma_q^2(t) \boldsymbol{I})^{-1} (\boldsymbol{\mu}_{\boldsymbol{\theta}} - \boldsymbol{\mu}_q)\right] \\
&= \frac{1}{2\sigma_q^2(t)}\left[\|\boldsymbol{\mu}_{\boldsymbol{\theta}} - \boldsymbol{\mu}_q\|_2^2\right]
\end{aligned}
$$

$$
\begin{aligned}
& \frac{1}{2\sigma_q^2(t)}\left[\|\boldsymbol{\mu}_{\boldsymbol{\theta}} - \boldsymbol{\mu}_q\|_2^2\right] \\
&= \frac{1}{2\sigma_q^2(t)}\left[\left\| \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})\boldsymbol{x}_t + \sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)\hat{\boldsymbol{x}}_0(\boldsymbol{x}_t, t)}{1-\bar{\alpha}_t} - \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})\boldsymbol{x}_t + \sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)\boldsymbol{x}_0}{1-\bar{\alpha}_t} \right\|_2^2\right] \\
&= \frac{1}{2\sigma_q^2(t)}\left[\left\| \frac{\sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)\hat{\boldsymbol{x}}_0(\boldsymbol{x}_t, t)}{1-\bar{\alpha}_t} - \frac{\sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)\boldsymbol{x}_0}{1-\bar{\alpha}_t} \right\|_2^2\right] \\
&= \frac{1}{2\sigma_q^2(t)}\left[\left\| \frac{\sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)}{1-\bar{\alpha}_t}(\hat{\boldsymbol{x}}_0(\boldsymbol{x}_t, t) - \boldsymbol{x}_0) \right\|_2^2\right] \\
&= \frac{\bar{\alpha}_{t-1}(1-\alpha_t)^2}{2\sigma_q^2(t)(1-\bar{\alpha}_t)^2}\left[\|\hat{\boldsymbol{x}}_0(\boldsymbol{x}_t, t) - \boldsymbol{x}_0\|_2^2\right]
\end{aligned}
$$

因此最大化 ELBO 就等价于最小化下面的均方误差:

$$
\mathbb{E}_{t \sim \mathbb{U}[0,T], q(\boldsymbol{x}_t|\boldsymbol{x}_0)} \frac{\bar{\alpha}_{t-1}(1-\alpha_t)^2}{2\sigma_q^2(t)(1-\bar{\alpha}_t)^2}\left[\|\hat{\boldsymbol{x}}_0(\boldsymbol{x}_t, t) - \boldsymbol{x}_0\|_2^2\right]
$$