

从狭义EM到变分自编码器

1. 狭义EM

Reference: [统计学习方法 第二版 第九章]

EM算法是一种迭代算法，1977年由Dempster等人总结提出，用于含有隐变量（hidden variable）的概率模型参数的极大似然估计，或极大后验概率估计。EM算法的每次迭代由两步组成：E步，求期望（expectation）；M步，求极大（maximization）。所以这一算法称为期望极大算法（expectation maximization algorithm），简称EM算法。

1.1 EM算法的引入

概率模型有时既含有观测变量（observable variable），又含有隐变量或潜在变量（latent variable）。如果概率模型的变量都是观测变量，那么给定数据，可以直接用极大似然估计法，或贝叶斯估计法估计模型参数（**此时只有固定的参数未知**）。EM算法就是含有隐变量的概率模型参数的极大似然估计法，或**极大后验概率估计法**。

例1（三硬币模型）：假设有3枚硬币，分别记作A，B，C。这些硬币正面出现的概率分别是 π ， p ， q 。进行如下掷硬币试验：先掷硬币A，根据其结果选出硬币B或硬币C，正面选硬币B，反面选硬币C；然后掷选出的硬币，掷硬币的结果，出现正面记作1，出现反面记作0；独立地重复 n 次试验（这里， $n=10$ ），观测结果如下，如何估计三硬币模型的参数：

1, 1, 0, 1, 0, 0, 1, 0, 1, 1

三硬币模型可以写作

$$P(y|\theta) = \sum_z P(y, z|\theta) = \sum_z P(z|\theta)P(y|z, \theta) = \pi p^y(1-p)^{1-y} + (1-\pi)q^y(1-q)^{1-y} \quad (1)$$

这里随机变量 y 是观测变量，表示一次试验观测到的结果是1或0；随机变量 z 是隐变量，表示未观测到的掷硬币A的结果； $\theta = (\pi, p, q)$ 是模型参数。则所有观测数据的似然函数为

$$P(Y|\theta) = \sum_Z P(Z|\theta)P(Y|Z, \theta) = \prod_{j=1}^n [\pi p^{y_j}(1-p)^{1-y_j} + (1-\pi)q^{y_j}(1-q)^{1-y_j}] \quad (2)$$

考虑求模型参数 $\theta = (\pi, p, q)$ 的极大似然估计，即

$$\hat{\theta} = \arg \max_{\theta} \log P(Y|\theta) \quad (3)$$

这个问题没有解析解，因为 \log 中有相加的两项，只有通过迭代的方法求解。

下列算法为何是EM算法？

E步：计算在模型参数 $\pi^{(i)}, p^{(i)}, q^{(i)}$ 下观测数据 y_j 来自掷硬币B的概率

$$\mu_j^{(i+1)} = \frac{\pi^{(i)}(p^{(i)})^{y_j}(1-p^{(i)})^{1-y_j}}{\pi^{(i)}(p^{(i)})^{y_j}(1-p^{(i)})^{1-y_j} + (1-\pi^{(i)})(q^{(i)})^{y_j}(1-q^{(i)})^{1-y_j}} \quad (4)$$

M步：计算模型参数的新估计值

$$\begin{aligned} \pi^{(i+1)} &= \frac{1}{n} \sum_{j=1}^n \mu_j^{(i+1)} \\ p^{(i+1)} &= \frac{\sum_{j=1}^n \mu_j^{(i+1)} y_j}{\sum_{j=1}^n \mu_j^{(i+1)}} \\ q^{(i+1)} &= \frac{\sum_{j=1}^n (1 - \mu_j^{(i+1)}) y_j}{\sum_{j=1}^n (1 - \mu_j^{(i+1)})} \end{aligned} \quad (5)$$

按照上述迭代步骤直至收敛，若假设模型参数初值为 $\pi^{(0)} = 0.5, p^{(0)} = 0.5, q^{(0)} = 0.5$ ，则模型参数的极大似然估计为 $\hat{\pi} = 0.5, \hat{p} = 0.6, \hat{q} = 0.6$ 。若假设模型参数初值为 $\pi^{(0)} = 0.4, p^{(0)} = 0.6, q^{(0)} = 0.6$ ，则模型参数的极大似然估计为 $\hat{\pi} = 0.4064, \hat{p} = 0.5368, \hat{q} = 0.6432$ 。

一般地，用 Y 表示观测随机变量的数据， Z 表示隐随机变量的数据。 Y 和 Z 连在一起称为完全数据（complete-data），观测数据 Y 又称为不完全数据（incomplete-data）。

首先写出所有观测的似然函数

$$P(Y|\theta) = \prod_{j=1}^n P(y_j|\theta) = \prod_{j=1}^n [\pi p + (1-\pi)q]^{y_j} [\pi(1-p) + (1-\pi)(1-q)]^{(1-y_j)} \quad (6)$$

计算 $P(z_j = 1 | y_j, \theta^{(i)})$ ：

$$\mu_j^{(i+1)} = P(z_j = 1 | y_j, \theta^{(i)}) = \frac{P(y_j | z_j = 1, \theta^{(i)})P(z_j = 1 | \theta^{(i)})}{P(y_j | \theta^{(i)})} = \begin{cases} \frac{\pi^{(i)}p^{(i)}}{\pi^{(i)}p^{(i)} + (1-\pi^{(i)})q^{(i)}} & \text{if } y_j = 1 \\ \frac{\pi^{(i)}(1-p^{(i)})}{\pi^{(i)}(1-p^{(i)}) + (1-\pi^{(i)})(1-q^{(i)})} & \text{if } y_j = 0 \end{cases} \quad (7)$$

计算完全数据的对数似然函数的期望

$$\begin{aligned} Q(\theta | \theta^{(i)}) &= \mathbb{E}_{P(Z|Y, \theta^{(i)})} [\log P(Y, Z | \theta)] \\ &= \mathbb{E}_{P(Z|Y, \theta^{(i)})} \left[\sum_{j=1}^n \log P(y_j, z_j | \theta) \right] \\ &= \sum_{j=1}^n \mathbb{E}_{P(z_j|y_j, \theta^{(i)})} [\log P(y_j, z_j | \theta)] \\ &= \sum_{j=1}^n \sum_{z_j} P(z_j | y_j, \theta^{(i)}) \log P(y_j, z_j | \theta) \\ &= \sum_{j=1}^n \left[\mu_j^{(i+1)} \log \left(\pi p^{y_j} (1-p)^{(1-y_j)} \right) + (1 - \mu_j^{(i+1)}) \log \left((1-\pi) q^{y_j} (1-q)^{(1-y_j)} \right) \right] \end{aligned} \quad (8)$$

对各参数求导，并令其满足一阶条件可得公式 (5)。

算法1 (EM 算法)：

输入：观测变量数据 Y ，隐变量数据 Z ，联合分布 $P(Y, Z | \theta)$ ，条件分布 $P(Z | Y, \theta)$ ；

输出：模型参数 θ 。

(1) 选择参数的初值 $\theta^{(0)}$ ，开始迭代；

(2) E 步：记 $\theta^{(i)}$ 为第 i 次迭代参数的估计值，在第 $i+1$ 次迭代的 E 步，计算

$$Q(\theta, \theta^{(i)}) = E_Z [\log P(Y, Z | \theta) | Y, \theta^{(i)}] = \sum_Z P(Z | Y, \theta^{(i)}) \log P(Y, Z | \theta) \quad (9)$$

(3) M 步：求使 $Q(\theta, \theta^{(i)})$ 极大化的 θ ，确定第 $i+1$ 次迭代的参数的估计值 $\theta^{(i+1)}$

$$\theta^{(i+1)} = \arg \max_{\theta} Q(\theta, \theta^{(i)}) \quad (10)$$

(4) 重复第 (2) 步和第 (3) 步，直到收敛。

注意：参数的初值可以任意选择，但 EM 算法对初值是敏感的。

定义1：Q 函数 (Q function)

完全数据的对数似然函数 $\log P(Y, Z | \theta)$ 关于在给定观测数据 Y 和当前参数 $\theta^{(i)}$ 下对未观测数据 Z 的条件概率分布 $P(Z | Y, \theta^{(i)})$ 的期望称为 Q 函数

$$Q(\theta, \theta^{(i)}) = E_Z [\log P(Y, Z | \theta) | Y, \theta^{(i)}] \quad (11)$$

1.2 EM 算法的导出

1.2.1 方法一（正向推导，只需要有进步即可）

Reference: [统计学习方法 第二版 179页]

面对一个含有隐变量的概率模型，目标是极大化观测数据（不完全数据） Y 关于参数 θ 的对数似然函数，即极大化

$$L(\theta) = \log P(Y | \theta) = \log \sum_Z P(Y, Z | \theta) = \log \left(\sum_Z P(Y | Z, \theta) P(Z | \theta) \right) \quad (12)$$

上式极大化的主要困难在于未观测数据以及对数里的和（或者积分）。

EM 算法是通过迭代逐步近似极大化 $L(\theta)$ 的，假设在第 i 次迭代后 θ 的估计值是 $\theta^{(i)}$ 。我们希望新估计值 θ 能使 $L(\theta)$ 增加，即 $L(\theta) > L(\theta^{(i)})$ ，并逐步达到极大值。

$$L(\theta) - L(\theta^{(i)}) = \log \left(\sum_Z P(Y | Z, \theta) P(Z | \theta) \right) - \log P(Y | \theta^{(i)}) \quad (13)$$

利用 Jensen 不等式得到其下界：

$$\begin{aligned} L(\theta) - L(\theta^{(i)}) &= \log \left(\sum_Z P(Z | Y, \theta^{(i)}) \frac{P(Y | Z, \theta) P(Z | \theta)}{P(Z | Y, \theta^{(i)})} \right) - \log P(Y | \theta^{(i)}) \\ &\geq \sum_Z P(Z | Y, \theta^{(i)}) \log \frac{P(Y | Z, \theta) P(Z | \theta)}{P(Z | Y, \theta^{(i)})} - \log P(Y | \theta^{(i)}) \\ &= \sum_Z P(Z | Y, \theta^{(i)}) \log \frac{P(Y | Z, \theta) P(Z | \theta)}{P(Z | Y, \theta^{(i)})} - \sum_Z P(Z | Y, \theta^{(i)}) \log P(Y | \theta^{(i)}) \\ &= \sum_Z P(Z | Y, \theta^{(i)}) \log \frac{P(Y | Z, \theta) P(Z | \theta)}{P(Z | Y, \theta^{(i)}) P(Y | \theta^{(i)})} \end{aligned} \quad (14)$$

$$B(\theta, \theta^{(i)}) = L(\theta^{(i)}) + \sum_Z P(Z | Y, \theta^{(i)}) \log \frac{P(Y | Z, \theta) P(Z | \theta)}{P(Z | Y, \theta^{(i)}) P(Y | \theta^{(i)})} \quad (15)$$

则 $L(\theta) \geq B(\theta, \theta^{(i)})$, 即函数 $B(\theta, \theta^{(i)})$ 是 $L(\theta)$ 的一个下界, 而且 $B(\theta^{(i)}, \theta^{(i)}) = L(\theta^{(i)})$ 。因此, 任何使 $B(\theta, \theta^{(i)})$ 相较于在 $\theta^{(i)}$ 处增大的 θ 也可以使相应的 $L(\theta)$ 增大, 即 $L(\theta^{(i)}) = B(\theta^{(i)}, \theta^{(i)}) \leq B(\theta^{(i+1)}, \theta^{(i)}) \leq L(\theta^{(i+1)})$ 。

$$\begin{aligned} \theta^{(i+1)} &= \arg \max_{\theta} B(\theta, \theta^{(i)}) \\ &= \arg \max_{\theta} \left(L(\theta^{(i)}) + \sum_Z P(Z | Y, \theta^{(i)}) \log \frac{P(Y | Z, \theta) P(Z | \theta)}{P(Z | Y, \theta^{(i)}) P(Y | \theta^{(i)})} \right) \\ &= \arg \max_{\theta} \left(\sum_Z P(Z | Y, \theta^{(i)}) \log P(Y | Z, \theta) P(Z | \theta) \right) \\ &= \arg \max_{\theta} Q(\theta, \theta^{(i)}) \end{aligned} \quad (16)$$

下图给出 EM 算法的直观解释:

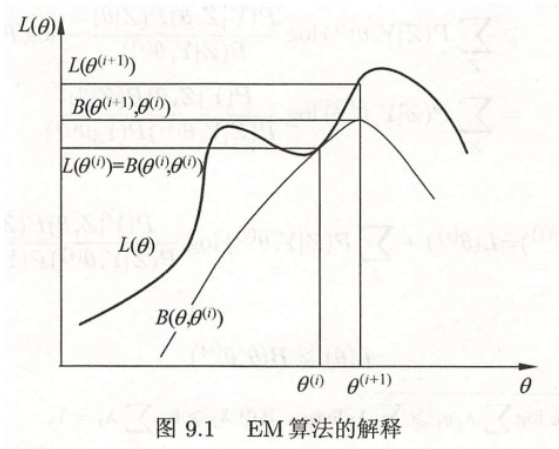


图 9.1 EM 算法的解释

1.2.2 方法二 (通过条件概率公式引入隐变量)

Reference: [变分推断PPT]

对等式两边 $\log P(Y | \theta) = \log P(Y, Z | \theta) - \log P(Z | Y, \theta)$ 分别关于隐变量的后验分布求期望

左边得到

$$\begin{aligned} \text{Left} &= \sum_Z P(Z | Y, \theta^{(i)}) \log P(Y | \theta) \\ &= \log P(Y | \theta) \sum_Z P(Z | Y, \theta^{(i)}) \\ &= \log P(Y | \theta) \end{aligned} \quad (17)$$

右边得到

$$\begin{aligned} \text{Right} &= \sum_Z P(Z | Y, \theta^{(i)}) \log P(Y, Z | \theta) - \sum_Z P(Z | Y, \theta^{(i)}) \log P(Z | Y, \theta) \\ &= Q(\theta, \theta^{(i)}) - H(\theta, \theta^{(i)}) \end{aligned} \quad (18)$$

此处 $Q(\theta, \theta^{(i)})$ 即为 EM 算法中 M 步的优化目标, 因此有 $Q(\theta^{(i+1)}, \theta^{(i)}) \geq Q(\theta^{(i)}, \theta^{(i)})$ 。

而对于 $H(\theta, \theta^{(i)})$, 可以证明

$$\begin{aligned} &H(\theta^{(i+1)}, \theta^{(i)}) - H(\theta^{(i)}, \theta^{(i)}) \\ &= \sum_Z P(Z | Y, \theta^{(i)}) \log P(Z | Y, \theta^{(i+1)}) - \sum_Z P(Z | Y, \theta^{(i)}) \log P(Z | Y, \theta^{(i)}) \\ &= \sum_Z P(Z | Y, \theta^{(i)}) \log \frac{P(Z | Y, \theta^{(i+1)})}{P(Z | Y, \theta^{(i)})} \\ &\leq \log \sum_Z P(Z | Y, \theta^{(i)}) \cdot \frac{P(Z | Y, \theta^{(i+1)})}{P(Z | Y, \theta^{(i)})} \\ &= 0 \end{aligned} \quad (19)$$

从而得到

$$\begin{aligned} &\log P(Y | \theta^{(i+1)}) - \log P(Y | \theta^{(i)}) \\ &= [Q(\theta^{(i+1)}, \theta^{(i)}) - H(\theta^{(i+1)}, \theta^{(i)})] - [Q(\theta^{(i)}, \theta^{(i)}) - H(\theta^{(i)}, \theta^{(i)})] \\ &= [Q(\theta^{(i+1)}, \theta^{(i)}) - Q(\theta^{(i)}, \theta^{(i)})] - [H(\theta^{(i+1)}, \theta^{(i)}) - H(\theta^{(i)}, \theta^{(i)})] \\ &\geq 0 \end{aligned} \quad (20)$$

1.2.3 方法三 (引入隐变量的近似分布, 承接变分推断内容)

Reference: [变分推断PPT]

引入隐变量 Z 的某种分布 $q_\phi(Z)$

$$\begin{aligned}\log P(Y | \theta) &= \log P(Y, Z | \theta) - \log P(Z | Y, \theta) \\ &= \log \frac{P(Y, Z | \theta)}{q(Z)} - \log \frac{P(Z | Y, \theta)}{q(Z)}\end{aligned}\quad (21)$$

对上式两边分别关于分布 $q(Z)$ 求期望，左边得到

$$\begin{aligned}\text{Left} &= \sum_Z q(Z) \log P(Y | \theta) \\ &= \log P(Y | \theta)\end{aligned}\quad (22)$$

右边得到

$$\text{Right} = \sum_Z q(Z) \log \frac{P(Y, Z | \theta)}{q(Z)} - \sum_Z q(Z) \log \frac{P(Z | Y, \theta)}{q(Z)}\quad (23)$$

联立得到

$$\begin{aligned}\underbrace{\log P(Y | \theta)}_{\text{evidence}} &= \sum_Z q(Z) \log \frac{P(Y, Z | \theta)}{q(Z)} - \sum_Z q(Z) \log \frac{P(Z | Y, \theta)}{q(Z)} \\ &= \underbrace{\sum_Z q(Z) \log \frac{P(Y, Z | \theta)}{q(Z)}}_{\text{ELBO}} + \underbrace{\sum_Z q(Z) \log \frac{q(Z)}{P(Z | Y, \theta)}}_{\text{KL}(q(Z) || P(Z | Y, \theta))}\end{aligned}\quad (24)$$

- $\log P(Y | \theta)$ 被称为证据 (evidence)
- $\sum_Z q(Z) \log \frac{P(Y, Z | \theta)}{q(Z)}$ 被称为证据下界 (evidence lower bound, ELBO)
- $\sum_Z q(Z) \log \frac{q(Z)}{P(Z | Y, \theta)} = \text{KL}(q(Z) || P(Z | Y, \theta))$ 是分布 $q(Z)$ 相对于分布 $P(Z | Y, \theta)$ 的 **KL散度** (Kullback-Leibler divergence)

因为 KL 散度非负，从而得到下式，当且仅当 $q(Z) = P(Z | Y, \theta)$ 时取等号

$$\underbrace{\log P(Y | \theta)}_{\text{evidence}} \geq \underbrace{\sum_Z q(Z) \log \frac{P(Y, Z | \theta)}{q(Z)}}_{\text{ELBO}}\quad (25)$$

E 步：固定参数 $\theta^{(i)}$ ，取 $q(Z) = P(Z | Y, \theta^{(i)})$ ，此时有 (**不严谨，为何此时取等号**)

$$\underbrace{\log P(Y | \theta)}_{\text{evidence}} = \underbrace{\sum_Z P(Z | Y, \theta^{(i)}) \log \frac{P(Y, Z | \theta)}{P(Z | Y, \theta^{(i)})}}_{\text{ELBO}}\quad (26)$$

M 步：ELBO 关于参数 θ 求最大，更新参数

$$\begin{aligned}\theta^{(i+1)} &= \arg \max_{\theta} \sum_Z P(Z | Y, \theta^{(i)}) \log \frac{P(Y, Z | \theta)}{P(Z | Y, \theta^{(i)})} \\ &= \arg \max_{\theta} \underbrace{\sum_Z P(Z | Y, \theta^{(i)}) \log P(Y, Z | \theta)}_{Q(\theta, \theta^{(i)})}\end{aligned}\quad (27)$$

应更正为固定参数 $\theta^{(i)}$ ，取 $q(Z) = P(Z | Y, \theta^{(i)})$ ，此时有

$$\log P(Y | \theta) = \underbrace{\sum_Z P(Z | Y, \theta^{(i)}) \log \frac{P(Y, Z | \theta)}{P(Z | Y, \theta^{(i)})}}_A + \underbrace{\sum_Z P(Z | Y, \theta^{(i)}) \log \frac{P(Z | Y, \theta^{(i)})}{P(Z | Y, \theta)}}_B\quad (28)$$

而当 $\theta = \theta^{(i)}$ 时有

$$\log P(Y | \theta^{(i)}) = \underbrace{\sum_Z P(Z | Y, \theta^{(i)}) \log \frac{P(Y, Z | \theta^{(i)})}{P(Z | Y, \theta^{(i)})}}_C + \underbrace{\sum_Z P(Z | Y, \theta^{(i)}) \log \frac{P(Z | Y, \theta^{(i)})}{P(Z | Y, \theta^{(i)})}}_D\quad (29)$$

由 KL 散度性质可知 $B \geq D = 0$ ：

$$\begin{aligned}\theta^{(i+1)} &= \arg \max_{\theta} \underbrace{\sum_Z P(Z | Y, \theta^{(i)}) \log P(Y, Z | \theta)}_{Q(\theta, \theta^{(i)})} \\ &= \arg \max_{\theta} A \\ \Rightarrow A(\theta^{(i+1)}) &\geq C \\ \Rightarrow \log P(Y | \theta^{(i+1)}) &\geq \log P(Y | \theta^{(i)})\end{aligned}\quad (30)$$

1.3 EM 算法的收敛性

定理1 设 $L(\theta) = \log P(Y | \theta)$ 为观测数据的对数似然函数, $\theta^{(i)} (i = 1, 2, \dots)$ 为 EM 算法得到的参数估计序列, $L(\theta^{(i)}) (i = 1, 2, \dots)$ 为对应的对数似然函数序列。

(1) 如果 $P(Y | \theta)$ 有上界, 则 $L(\theta^{(i)}) = \log P(Y | \theta^{(i)})$ 收敛到某一值 L^* ;

(2) **在函数 $Q(\theta, \theta')$ 与 $L(\theta)$ 满足一定条件下**, 由 EM 算法得到的参数估计序列 $\theta^{(i)}$ 的收敛值 θ^* 是 $L(\theta)$ 的稳定点。

证明:

(1) 由 $L(\theta) = \log P(Y | \theta)$ 的单调性及 $P(Y | \theta)$ 的有界性立即得到。

(2) 证明从略, 参阅文献 [1983 On the convergence properties of the EM algorithm]。

1.4 EM 算法在高斯混合模型学习中的应用

定义2 高斯混合模型

高斯混合模型是指具有如下形式的概率分布模型:

$$P(y | \theta) = \sum_{k=1}^K \alpha_k \cdot \phi(y | \theta_k) \quad (31)$$

其中, α_k 是系数, $\alpha_k \geq 0$, $\sum_{k=1}^K \alpha_k = 1$; $\phi(y | \theta_k)$ 是高斯分布密度, $\theta_k = (\mu_k, \sigma_k^2)$,

$$\phi(y | \theta_k) = \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left(-\frac{(y - \mu_k)^2}{2\sigma_k^2}\right) \quad (32)$$

称为第 k 个分模型。一般混合模型可以由任意概率分布密度代替式 (29) 中的高斯分布密度, 此处只介绍最常用的高斯混合模型。

1.4.1 高斯混合模型参数估计的 EM 算法

假设观测数据 y_1, y_2, \dots, y_N 由高斯混合模型生成,

$$P(y | \theta) = \sum_{k=1}^K \alpha_k \cdot \phi(y | \theta_k) \quad (33)$$

其中, $\theta = (\alpha_1, \alpha_2, \dots, \alpha_K; \theta_1, \theta_2, \dots, \theta_K)$ 。

观测数据的产生过程: 首先依概率 $(\alpha_1, \dots, \alpha_K)$ 选择第 k 个高斯分布模型, 然后依第 k 个分模型的概率分布 $\phi(y | \theta_k)$ 生成观测数据 y_j 。这时观测数据 $y_j, j = 1, 2, \dots, N$ 是已知的; 反映观测数据 y_j 来自第 k 个分模型的数据是未知的, $k = 1, 2, \dots, K$, 以隐变量 γ_{jk} 表示, 其定义如下:

$$\gamma_{jk} = \begin{cases} 1, & \text{第 } j \text{ 个观测来自第 } k \text{ 个分模型} \\ 0, & \text{否则} \end{cases} \quad (34)$$

有了观测数据 y_j 及未观测数据 γ_{jk} , 那么完全数据是

$$(y_j, \gamma_{j1}, \gamma_{j2}, \dots, \gamma_{jK}), \quad j = 1, 2, \dots, N \quad (35)$$

于是可以写出完全数据的似然函数

$$\begin{aligned} P(y, \gamma | \theta) &= \prod_{j=1}^N P(y_j, \gamma_{j1}, \gamma_{j2}, \dots, \gamma_{jK} | \theta) \\ &= \prod_{k=1}^K \prod_{j=1}^N [\alpha_k \cdot \phi(y_j | \theta_k)]^{\gamma_{jk}} \\ &= \prod_{k=1}^K \alpha_k^{n_k} \prod_{j=1}^N [\phi(y_j | \theta_k)]^{\gamma_{jk}} \\ &= \prod_{k=1}^K \alpha_k^{n_k} \prod_{j=1}^N \left[\frac{1}{\sqrt{2\pi}\sigma_k} \exp\left(-\frac{(y_j - \mu_k)^2}{2\sigma_k^2}\right) \right]^{\gamma_{jk}} \end{aligned} \quad (36)$$

式中, $n_k = \sum_{j=1}^N \gamma_{jk}$, $\sum_{k=1}^K n_k = N$ 。

那么, 完全数据的对数似然函数为

$$\log P(y, \gamma | \theta) = \sum_{k=1}^K \left\{ n_k \log \alpha_k + \sum_{j=1}^N \gamma_{jk} \left[\log\left(\frac{1}{\sqrt{2\pi}}\right) - \log \sigma_k - \frac{1}{2\sigma_k^2} (y_j - \mu_k)^2 \right] \right\} \quad (37)$$

进一步计算 Q 函数

$$\begin{aligned} Q(\theta, \theta^{(i)}) &= E \left[\log P(y, \gamma | \theta) | y, \theta^{(i)} \right] \\ &= E \left\{ \sum_{k=1}^K \left\{ n_k \log \alpha_k + \sum_{j=1}^N \gamma_{jk} \left[\log\left(\frac{1}{\sqrt{2\pi}}\right) - \log \sigma_k - \frac{1}{2\sigma_k^2} (y_j - \mu_k)^2 \right] \right\} \right\} \\ &= \sum_{k=1}^K \left\{ \sum_{j=1}^N (E[\gamma_{jk}]) \log \alpha_k + \sum_{j=1}^N (E[\gamma_{jk}]) \left[\log\left(\frac{1}{\sqrt{2\pi}}\right) - \log \sigma_k - \frac{1}{2\sigma_k^2} (y_j - \mu_k)^2 \right] \right\} \end{aligned} \quad (38)$$

这里需要计算 $E(\gamma_{jk} | y, \theta)$ ，记为 $\hat{\gamma}_{jk}$ 。

$$\begin{aligned}\hat{\gamma}_{jk} &= E(\gamma_{jk} | y, \theta) = P(\gamma_{jk} = 1 | y, \theta) \\ &= \frac{P(\gamma_{jk} = 1, y_j | \theta)}{\sum_{k=1}^K P(\gamma_{jk} = 1, y_j | \theta)} \quad \text{理解分母} \\ &= \frac{P(y_j | \gamma_{jk} = 1, \theta)P(\gamma_{jk} = 1 | \theta)}{\sum_{k=1}^K P(y_j | \gamma_{jk} = 1, \theta)P(\gamma_{jk} = 1 | \theta)} \\ &= \frac{\alpha_k \phi(y_j | \theta_k)}{\sum_{k=1}^K \alpha_k \phi(y_j | \theta_k)}, \quad j = 1, 2, \dots, N; \quad k = 1, 2, \dots, K\end{aligned}\tag{39}$$

$\hat{\gamma}_{jk}$ 是在当前模型参数下第 j 个观测数据来自第 k 个分模型的概率，称为分模型 k 对观测数据 y_j 的响应度。将 $\hat{\gamma}_{jk} = E[\gamma_{jk}]$ 及 $\hat{n}_k = \sum_{j=1}^N E[\gamma_{jk}]$ 代入式 35，即得

$$Q(\theta, \theta^{(i)}) = \sum_{k=1}^K \left\{ \hat{n}_k \log \alpha_k + \sum_{j=1}^N \hat{\gamma}_{jk} \left[\log \left(\frac{1}{\sqrt{2\pi}} \right) - \log \sigma_k - \frac{1}{2\sigma_k^2} (y_j - \mu_k)^2 \right] \right\}\tag{40}$$

迭代的 M 步是求函数 $Q(\theta, \theta^{(i)})$ 对 θ 的极大值，即求新一轮迭代的模型参数

$$\theta^{(i+1)} = \arg \max_{\theta} Q(\theta, \theta^{(i)})\tag{41}$$

用 $\hat{\mu}_k$, $\hat{\sigma}_k^2$ 及 $\hat{\alpha}_k$, $k = 1, 2, \dots, K$ ，表示 $\theta^{(i+1)}$ 的各参数。求 $\hat{\mu}_k$, $\hat{\sigma}_k^2$ 只需将式 37 分别对 $\hat{\mu}_k$, $\hat{\sigma}_k^2$ 求偏导数并令其为 0，即可得到；求 $\hat{\alpha}_k$ 是在 $\sum_{k=1}^K \alpha_k = 1$ 条件下求偏导数并令其为 0 得到的（**为何不用检验函数的凹凸性**）。结果如下

$$\hat{\mu}_k = \frac{\sum_{j=1}^N \hat{\gamma}_{jk} \cdot y_j}{\sum_{j=1}^N \hat{\gamma}_{jk}}, \quad k = 1, 2, \dots, K\tag{42}$$

$$\hat{\sigma}_k^2 = \frac{\sum_{j=1}^N \hat{\gamma}_{jk} (y_j - \mu_k)^2}{\sum_{j=1}^N \hat{\gamma}_{jk}}, \quad k = 1, 2, \dots, K\tag{43}$$

$$\hat{\alpha}_k = \frac{\hat{n}_k}{N} = \frac{\sum_{j=1}^N \hat{\gamma}_{jk}}{N}, \quad k = 1, 2, \dots, K\tag{44}$$

重复以上计算，直到对数似然函数值不再有明显的变化为止。

1.5 EM 算法的推广（可参考变分推断PPT更简单易理解）

1.5.1 F 函数的极大-极大算法

定义2：F 函数

假设隐变量数据 Z 的概率分布为 $\tilde{P}(Z)$ ，定义分布 \tilde{P} 与参数 θ 的函数 $F(\tilde{P}, \theta)$ 如下

$$F(\tilde{P}, \theta) = E_{\tilde{P}}[\log P(Y, Z | \theta)] + H(\tilde{P})\tag{45}$$

称为 F 函数，式中 $H(\tilde{P}) = -E_{\tilde{P}} \log \tilde{P}(Z)$ 是分布 $\tilde{P}(Z)$ 的熵。

在定义2中，通常假设 $P(Y, Z | \theta)$ 是 θ 的连续函数，因而 $F(\tilde{P}, \theta)$ 是 \tilde{P} 和 θ 的连续函数。函数 $F(\tilde{P}, \theta)$ 还有以下重要性质。

引理1：

对于固定的 θ ，存在唯一的分布 \tilde{P}_θ 极大化 $F(\tilde{P}, \theta)$ ，这时 \tilde{P}_θ 由下式给出

$$\tilde{P}_\theta(Z) = P(Z | Y, \theta)\tag{46}$$

并且 \tilde{P}_θ 随 θ 连续变化。

证明

对于固定的 θ ，可以求得使 $F(\tilde{P}, \theta)$ 达到极大的分布 $\tilde{P}_\theta(Z)$ 。为此，引进拉格朗日乘子 λ ，拉格朗日函数为

$$L = E_{\tilde{P}} \log P(Y, Z | \theta) - E_{\tilde{P}} \log \tilde{P}(Z) + \lambda \left(1 - \sum_Z \tilde{P}(Z) \right)\tag{47}$$

将其对 \tilde{P} 求偏导数（ $\log P(Y, Z | \theta)$ 不是 $\tilde{P}(Z)$ 的函数吗？）

$$\frac{\partial L}{\partial \tilde{P}(Z)} = \log P(Y, Z | \theta) - \log \tilde{P}(Z) - 1 - \lambda\tag{48}$$

令偏导数等于 0，得出

$$\lambda = \log P(Y, Z | \theta) - \log \tilde{P}_\theta(Z) - 1\tag{49}$$

由此推出 $\tilde{P}_\theta(Z)$ 与 $P(Y, Z | \theta)$ 成比例

$$\frac{P(Y, Z | \theta)}{\tilde{P}_\theta(Z)} = \exp(1 + \lambda) \quad (50)$$

再从约束条件 $\sum_Z \tilde{P}_\theta(Z) = 1$ 得到式43。

由假设 $P(Y, Z | \theta)$ 是 θ 的连续函数，得到 \tilde{P}_θ 是 θ 的连续函数。

引理

若 $\tilde{P}_\theta(Z) = P(Z | Y, \theta)$ ，则

$$F(\tilde{P}, \theta) = \log P(Y | \theta) \quad (51)$$

证明

$$\begin{aligned} F(\tilde{P}, \theta) &= E_{\tilde{P}}[\log P(Y, Z | \theta)] + H(\tilde{P}) \\ &= E_{\tilde{P}}[\log \frac{P(Y, Z | \theta)}{P(Z | Y, \theta)}] \\ &= \sum_Z P(Z | Y, \theta) \cdot \log \frac{P(Y, Z | \theta)}{P(Z | Y, \theta)} \\ &= \log P(Y | \theta) \end{aligned} \quad (52)$$

由以上引理，可以得到关于 EM 算法用 F 函数的极大-极大算法的解释。

定理

设 $L(\theta) = \log P(Y | \theta)$ 为观测数据的对数似然函数， $\theta^{(i)}, i = 1, 2, \dots$ ，为 EM 算法得到的参数估计序列，函数 $F(\tilde{P}, \theta)$ 由式42定义。如果 $F(\tilde{P}, \theta)$ 在 \tilde{P}^* 和 θ^* 有局部极大值，那么 $L(\theta)$ 也在 θ^* 有局部极大值。类似地，如果 $F(\tilde{P}, \theta)$ 在 \tilde{P}^* 和 θ^* 达到全局最大值，那么 $L(\theta)$ 也在 θ^* 达到全局最大值。

证明

由引理1和引理2可知， $L(\theta) = \log P(Y | \theta) = F(\tilde{P}_\theta, \theta)$ **对任意 θ 成立**。特别地，对于使 $F(\tilde{P}, \theta)$ 达到极大的参数 θ^* ，有

$$L(\theta^*) = F(\tilde{P}_{\theta^*}, \theta^*) = F(\tilde{P}^*, \theta^*) \quad (53)$$

为了证明 θ^* 是 $L(\theta)$ 的极大点，需要证明不存在接近 θ^* 的点 θ^{**} ，使 $L(\theta^{**}) > L(\theta^*)$ 。假如存在这样的点 θ^{**} ，那么应有 $F(\tilde{P}^{**}, \theta^{**}) > F(\tilde{P}^*, \theta^*)$ ，这里 $\tilde{P}^{**} = \tilde{P}_{\theta^{**}}$ 。但因 \tilde{P}_θ 是随 θ 连续变化的， \tilde{P}^{**} 应接近 \tilde{P}^* ，这与 \tilde{P}^* 和 θ^* 是 $F(\tilde{P}, \theta)$ 的局部极大点的假设矛盾。类似可以证明关于全局最大值的讨论。

定理

EM 算法的一次迭代可由 F 函数的极大-极大算法实现。

设 $\theta^{(i)}$ 为第 i 次迭代参数 θ 的估计， $\tilde{P}^{(i)}$ 为第 i 次迭代函数 \tilde{P} 的估计。在第 $i+1$ 次迭代的两步为：

- (1) 对固定的 $\theta^{(i)}$ ，求 $\tilde{P}^{(i+1)}$ 使 $F(\tilde{P}, \theta^{(i)})$ 极大化；
- (2) 对固定的 $\tilde{P}^{(i+1)}$ ，求 $\theta^{(i+1)}$ 使 $F(\tilde{P}^{(i+1)}, \theta)$ 极大化。

证明 (1) 由引理 9.1，对于固定的 $\theta^{(i)}$ ，

$$\tilde{P}^{(i+1)}(Z) = \tilde{P}_{\theta^{(i)}}(Z) = P(Z | Y, \theta^{(i)}) \quad (54)$$

使 $F(\tilde{P}, \theta^{(i)})$ 极大化。此时，

$$\begin{aligned} F(\tilde{P}^{(i+1)}, \theta) &= E_{\tilde{P}^{(i+1)}}[\log P(Y, Z | \theta)] + H(\tilde{P}^{(i+1)}) \\ &= \sum_Z \log P(Y, Z | \theta) P(Z | Y, \theta^{(i)}) + H(\tilde{P}^{(i+1)}) \end{aligned} \quad (55)$$

由 $Q(\theta, \theta^{(i)})$ 的定义式 (9.11) 有

$$F(\tilde{P}^{(i+1)}, \theta) = Q(\theta, \theta^{(i)}) + H(\tilde{P}^{(i+1)}) \quad (56)$$

(2) 固定 $\tilde{P}^{(i+1)}$ ，求 $\theta^{(i+1)}$ 使 $F(\tilde{P}^{(i+1)}, \theta)$ 极大化。得到

$$\theta^{(i+1)} = \arg \max_{\theta} F(\tilde{P}^{(i+1)}, \theta) = \arg \max_{\theta} Q(\theta, \theta^{(i)}) \quad (57)$$

通过以上两步完成了 EM 算法的一次迭代。由此可知，由 EM 算法与 F 函数的极大-极大算法得到的参数估计序列 $\theta^{(i)}, i = 1, 2, \dots$ ，是一致的。

1.5.2 GEM 算法

算法

输入：观测数据， F 函数；

输出：模型参数。

- (1) 初始化参数 $\theta^{(0)}$ ，开始迭代；
- (2) 第 $i+1$ 次迭代，第 1 步：记 $\theta^{(i)}$ 为参数 θ 的估计值， $\tilde{P}^{(i)}$ 为函数 \tilde{P} 的估计，求 $\tilde{P}^{(i+1)}$ 使 \tilde{P} 极大化 $F(\tilde{P}, \theta^{(i)})$ ；

(3) 第 2 步: 求 $\theta^{(i+1)}$ 使 $F(\hat{P}^{(i+1)}, \theta)$ 极大化;

(4) 重复 (2) 和 (3), 直到收敛。

在 GEM 算法 1 中, 有时求 $Q(\theta, \theta^{(i)})$ 的极大化是很困难的。下面介绍的 GEM 算法 2 和 GEM 算法 3 并不是直接求 $\theta^{(i+1)}$ 使 $Q(\theta, \theta^{(i)})$ 达到极大的 θ , 而是找一个 $\theta^{(i+1)}$ 使得 $Q(\theta^{(i+1)}, \theta^{(i)}) > Q(\theta^{(i)}, \theta^{(i)})$ 。

算法

输入: 观测数据, Q 函数;

输出: 模型参数。

(1) 初始化参数 $\theta^{(0)}$, 开始迭代;

(2) 第 $i+1$ 次迭代, 第 1 步: 记 $\theta^{(i)}$ 为参数 θ 的估计值, 计算

$$\begin{aligned} Q(\theta, \theta^{(i)}) &= E_Z [\log P(Y, Z | \theta) | Y, \theta^{(i)}] \\ &= \sum_Z P(Z | Y, \theta^{(i)}) \log P(Y, Z | \theta) \end{aligned} \quad (58)$$

(3) 第 2 步: 求 $\theta^{(i+1)}$ 使

$$Q(\theta^{(i+1)}, \theta^{(i)}) > Q(\theta^{(i)}, \theta^{(i)}) \quad (59)$$

(4) 重复 (2) 和 (3), 直到收敛。

当参数 θ 的维数为 $d(d \geq 2)$ 时, 可采用一种特殊的 GEM 算法, 它将 EM 算法的 M 步分解为 d 次条件极大化, 每次只改变参数向量的一个分量, 其余分量不改变。

算法

输入: 观测数据, Q 函数;

输出: 模型参数。

(1) 初始化参数 $\theta^{(0)} = (\theta_1^{(0)}, \theta_2^{(0)}, \dots, \theta_d^{(0)})$, 开始迭代;

(2) 第 $i+1$ 次迭代, 第 1 步: 记 $\theta^{(i)} = (\theta_1^{(i)}, \theta_2^{(i)}, \dots, \theta_d^{(i)})$ 为参数 $\theta = (\theta_1, \theta_2, \dots, \theta_d)$ 的估计值, 计算

$$\begin{aligned} Q(\theta, \theta^{(i)}) &= E_Z [\log P(Y, Z | \theta) | Y, \theta^{(i)}] \\ &= \sum_Z P(Z | y, \theta^{(i)}) \log P(Y, Z | \theta) \end{aligned} \quad (60)$$

(3) 第 2 步: 进行 d 次条件极大化:

首先, 在 $\theta_2^{(i)}, \dots, \theta_d^{(i)}$ 保持不变的条件下求使 $Q(\theta, \theta^{(i)})$ 达到极大的 $\theta_1^{(i+1)}$; 然后, 在 $\theta_1 = \theta_1^{(i+1)}, \theta_j = \theta_j^{(i)}, j = 3, 4, \dots, d$ 的条件下求使 $Q(\theta, \theta^{(i)})$ 达到极大的 $\theta_2^{(i+1)}$;

如此继续, 经过 d 次条件极大化, 得到 $\theta^{(i+1)} = (\theta_1^{(i+1)}, \theta_2^{(i+1)}, \dots, \theta_d^{(i+1)})$ 使得

$$Q(\theta^{(i+1)}, \theta^{(i)}) > Q(\theta^{(i)}, \theta^{(i)}) \quad (61)$$

(4) 重复 (2) 和 (3), 直到收敛。

2. 变分推断

Reference: [变分推断PPT]

2.1 变分推断介绍

变分推断 (Variational Inference, VI) 是贝叶斯学习中常用的、含有隐变量模型的学习和推断方法。变分推断和马尔科夫链蒙特卡洛法 (MCMC) 属于不同的技巧:

- MCMC 通过随机抽样的方法近似地计算模型的后验概率 (采样), 适合小数据集以及精确度更重要的场景
- 变分推断通过解析的方法计算模型的后验概率的近似值 (优化), 适合大数据集以及想快速测试多种模型的场景

为什么关心后验概率 $P(\theta | X)$?

1. 推断 (Bayesian Inference): 后验分布 $P(\theta | X)$ 包含了模型的重要信息, 描述了数据样本产生的过程, 例如从用户的观影历史评分信息 Y 中推断用户的偏好模型 θ
2. 决策 (Bayesian Decision Theory): 对于新样本 \tilde{x} , 求 $P(\tilde{x} | X)$

$$\begin{aligned} P(\tilde{x} | X) &= \int_{\theta} P(\tilde{x}, \theta | X) d\theta \\ &= \int_{\theta} P(\tilde{x} | \theta) P(\theta | X) d\theta \\ &= E_{\theta|X} [P(\tilde{x} | \theta)] \end{aligned} \quad (62)$$

被称为后验预测分布 (Posterior predictive distribution), 例如根据用户的历史评分信息 X 预测用户对于新电影 \tilde{x} 的评分

2.2 变分推断推导

贝叶斯参数学习问题的描述：

- X 观测数据
- Z 隐变量+参数
- θ 超参数

注意，这里的符号表示和 EM 算法中的表述有区别，贝叶斯参数学习需要推断的是 Z 中的参数，及学习后验分布 $P(Z | \theta)$

首先是 evidence 的分解

$$\underbrace{\log P(X | \theta)}_{\text{evidence}} = \underbrace{\int_Z q(Z) \log \frac{P(X, Z | \theta)}{q(Z)} dZ}_{\text{ELBO}} + \underbrace{\int_Z q(Z) \log \frac{q(Z)}{P(Z | X, \theta)} dZ}_{\text{KL}(q(Z) || P(Z | X, \theta))} \quad (63)$$

当我们知道超参数 θ 时，上式中 evidence 应是固定的，因为 $\log P(X | \theta) = \log \sum_Z P(X, Z | \theta)$ ，虽然这个值通常求不出来。

变分推断的目标是通过最小化 $\text{KL}(q(Z) || P(Z | X, \theta))$ 来寻找与后验分布 $P(Z | X, \theta)$ 最相似的变分分布 $q(Z)$ 。

$$q(Z)^* = \arg \min_{q(Z)} \text{KL}(q(Z) || P(Z | X, \theta)) \quad (64)$$

后验分布 $P(Z | X, \theta)$ 太复杂，直接估计其参数很苦难，但利用 KL 散度和 ELBO 的和为常数，可以转而求

$$\begin{aligned} q(Z)^* &= \arg \min_{q(Z)} \text{KL}(q(Z) || P(Z | X, \theta)) \\ &= \arg \max_{q(Z)} \text{ELBO} \\ &= \arg \max_{q(Z)} \int_Z q(Z) \log \frac{P(X, Z | \theta)}{q(Z)} dZ \\ &= \arg \max_{q(Z)} \int_Z q(Z) \log P(X, Z | \theta) dZ - \int_Z q(Z) \log q(Z) dZ \\ &= \arg \max_{q(Z)} E_{q(Z)}[\log P(X, Z | \theta)] - E_{q(Z)}[\log q(Z)] \end{aligned} \quad (65)$$

变分分布 $q(Z)$ 有多种参数化方法，要求参数化后的 $q(Z)$ 使得上述优化问题容易求解，一种常用的方法是假设 $q(Z)$ 对 $Z = (Z_1, Z_2, \dots, Z_d)$ 的所有分量 Z_j 都是相互独立的（实际是条件独立于参数），即满足

$$q(Z) = q(Z_1) \cdot q(Z_2) \cdots q(Z_d) \quad (66)$$

这时的变分分布被称为平均场（mean field）。

KL 散度的最小化或证据下界的最大化实际是在平均场的集合，即满足独立假设的分布集合 $Q = \{q(Z) | q(Z) = \prod_{j=1}^d q(Z_j)\}$ 之中进行的

$$q(Z)^* = \arg \max_{q(Z) \in Q} E_{q(Z)}[\log P(X, Z | \theta)] - E_{q(Z)}[\log q(Z)] \quad (67)$$

Reference: [intermediate_vb, PRML chapter 10]

现在我们将目标函数重新写为

$$\begin{aligned} \text{ELBO} &= \int q_\phi(\mathbf{z}) \log \left(\frac{p(\mathbf{x}, \mathbf{z})}{q_\phi(\mathbf{z})} \right) d\mathbf{z} \\ &= \int q_\phi(\mathbf{z}) \log(p(\mathbf{x}, \mathbf{z})) d\mathbf{z} - \int q_\phi(\mathbf{z}) \log(q_\phi(\mathbf{z})) d\mathbf{z} \\ &= \underbrace{\int \prod_{i=1}^M q_i(z_i) \log(p(\mathbf{x}, \mathbf{z})) d\mathbf{z}}_{-H(q, p)} + \underbrace{\left(- \int \prod_{i=1}^M q_i(z_i) \sum_{i=1}^M \log(q_i(z_i)) d\mathbf{z} \right)}_{H(q)} \end{aligned} \quad (68)$$

首先考虑第一项 $-H(q, p)$

$$\begin{aligned} -H(q, p) &= \int \prod_{i=1}^M q_i(z_i) \log(p(\mathbf{x}, \mathbf{z})) d\mathbf{z} \\ &= \int_{Z_1} \int_{Z_2} \cdots \int_{Z_M} \prod_{i=1}^M q_i(z_i) \log(p(\mathbf{x}, \mathbf{z})) d\mathbf{z}_1 d\mathbf{z}_2 \cdots d\mathbf{z}_M \end{aligned} \quad (69)$$

只考虑其中一项 $q_j(z_j)$

$$\begin{aligned} -H(q, p)_j &= \int_{Z_j} q_j(z_j) \left(\int \cdots \int \prod_{i \neq j} q_i(z_i) \log(p(\mathbf{x}, \mathbf{z})) \prod_{i \neq j} dz_i \right) dz_j \\ &= \int_{Z_j} q_j(z_j) \mathbb{E}_{\mathbf{z} \setminus z_j}[\log(p(\mathbf{x}, \mathbf{z}))] dz_j \end{aligned} \quad (70)$$

再考虑第二项 $H(q)$

$$\begin{aligned}
H(q) &= - \int \prod_{i=1}^M q_i(z_i) \sum_{i=1}^M \log(q_i(z_i)) d\mathbf{z} \\
&= \sum_{i=1}^M \left(- \int_{Z_i} q_i(z_i) \log(q_i(z_i)) dz_i \right) \\
&= \sum_{i=1}^M H(q_i(z_i))
\end{aligned} \tag{71}$$

仅考虑其中一项 $q_j(z_j)$

$$\begin{aligned}
H(q)_j &= - \int_{Z_j} q_j(z_j) \log(q_j(z_j)) dz_j + \text{Const.} \\
&= H(q_j(z_j)) + \text{Const.}
\end{aligned} \tag{72}$$

针对 ELBO 只考虑优化 q_j

$$\begin{aligned}
\text{ELBO}(q_j) &= -H(q, p)_j + H(q)_j \\
&= \int_{Z_j} q_j(z_j) \mathbb{E}_{\mathbf{z} \setminus z_j} [\log(p(\mathbf{x}, \mathbf{z}))] dz_j - \int_{Z_j} q_j(z_j) \log(q_j(z_j)) dz_j + \text{Const.} \\
&= \int_{Z_j} q_j(z_j) \log \tilde{p}(\mathbf{x}, z_j) dz_j - \int_{Z_j} q_j(z_j) \log(q_j(z_j)) dz_j + \text{Const.} \\
&= \int_{Z_j} q_j(z_j) \log \left[\frac{\tilde{p}(\mathbf{x}, z_j)}{q_j(z_j)} \right] dz_j + \text{Const.} \\
&= -\mathbb{KL}(q_j(z_j) \parallel \tilde{p}(\mathbf{x}, z_j)) + \text{Const.}
\end{aligned} \tag{73}$$

这里我们定义了一个新分布 $\tilde{p}(\mathbf{x}, z_j)$

$$\log \tilde{p}(\mathbf{x}, z_j) = \mathbb{E}_{\mathbf{z} \setminus z_j} [\log(p(\mathbf{x}, \mathbf{z}))] + \text{Const.} \tag{74}$$

因此我们可以通过最小化下述 KL 散度来最大化 ELBO，而 KL 散度的性质可知其值为零时最小，即 $q_j^* = \tilde{p}(\mathbf{x}, z_j)$

$$\log q_j^*(z_j) = \mathbb{E}_{\mathbf{z} \setminus z_j} [\log(p(\mathbf{x}, \mathbf{z}))] + \text{Const.} \tag{75}$$

注意此处的 $\exp(\mathbb{E}_{\mathbf{z} \setminus z_j} [\log(p(\mathbf{x}, \mathbf{z}))])$ 是伪概率分布 (pseudo distribution)，只能满足概率分布的非负性而不能保证具有归一性，常数项为归一化常数 $\int \exp(\mathbb{E}_{\mathbf{z} \setminus z_j} [\log(p(\mathbf{x}, \mathbf{z}))]) dz_j$ ，保证 \tilde{p} 的归一性和非负性，因此有

$$q_j^*(z_j) = \frac{\exp(\mathbb{E}_{\mathbf{z} \setminus z_j} [\log(p(\mathbf{x}, \mathbf{z}))])}{\int \exp(\mathbb{E}_{\mathbf{z} \setminus z_j} [\log(p(\mathbf{x}, \mathbf{z}))]) dz_j} \tag{76}$$

2.3 Gaussian-Gamma

可观测变量为 $\mathcal{D} = \{x_1, \dots, x_n\}$ ，似然为

$$\begin{aligned}
p(\mathcal{D} \mid \mu, \tau) &= \prod_{i=1}^n \left(\frac{\tau}{2\pi} \right)^{\frac{1}{2}} \exp \left(-\frac{\tau}{2} (x_i - \mu)^2 \right) \\
&= \left(\frac{\tau}{2\pi} \right)^{\frac{n}{2}} \exp \left(-\frac{\tau}{2} \sum_{i=1}^n (x_i - \mu)^2 \right)
\end{aligned} \tag{77}$$

假设先验为

$$\begin{aligned}
p(\mu \mid \tau) &= \mathcal{N}(\mu_0, (\lambda_0 \tau)^{-1}) \propto \exp \left(-\frac{\lambda_0 \tau}{2} (\mu - \mu_0)^2 \right) \\
p(\tau) &= \text{Gamma}(\tau \mid a_0, b_0) \propto \tau^{a_0-1} \exp(-b_0 \tau)
\end{aligned} \tag{78}$$

利用共轭性质可以计算解析后验

$$\begin{aligned}
p(\mu, \tau \mid \mathcal{D}) &\propto p(\mathcal{D} \mid \mu, \tau) p(\mu \mid \tau) p(\tau) \\
&= \mathcal{N}(\mu_n, (\lambda_n \tau)^{-1}) \text{Gamma}(\tau \mid a_n, b_n)
\end{aligned} \tag{79}$$

此处

$$\begin{aligned}
\mu_n &= \frac{\lambda_0 \mu_0 + n \bar{x}}{\lambda_0 + n} \\
\lambda_n &= \lambda_0 + n \\
a_n &= a_0 + \frac{n}{2} \\
b_n &= b_0 + \frac{1}{2} \sum_{i=1}^n (x_i - \bar{x})^2 + \frac{\lambda_0 n (\bar{x} - \mu_0)^2}{2(\lambda_0 + n)}
\end{aligned} \tag{80}$$

但是如果我们不能计算其解析后验，可用变分推断来近似其后验。假设变分分布 $q(\mathbf{z})$ 为

$$q(\mu, \tau) = q_\mu(\mu)q_\tau(\tau) \quad (81)$$

利用式 72 得出的结论

$$\begin{aligned} \log q_\mu^*(\mu) &= \mathbb{E}_{q_\tau(\tau)} [\log p(\mu, \tau, \mathcal{D})] \\ &= \mathbb{E}_{q_\tau(\tau)} [\log p(\mathcal{D} \mid \mu, \tau) + \log p(\mu \mid \tau)] + \text{Const.} \\ &= \mathbb{E}_{q_\tau(\tau)} \left[\frac{n}{2} \log \tau - \frac{\tau}{2} \sum_{i=1}^n (x_i - \mu)^2 - \frac{\lambda_0 \tau}{2} (\mu - \mu_0)^2 \right] + \text{Const.} \\ &= -\frac{1}{2} \mathbb{E}_{q_\tau} [\tau] \underbrace{\left[\sum_{i=1}^n (x_i - \mu)^2 + \lambda_0 (\mu - \mu_0)^2 \right]}_{\text{terms taking out of } \tau} + \text{Const.} \end{aligned} \quad (82)$$

将中括号内式子展开，形成高斯分布 $\mathcal{N}(\mu; \mu^*, \tau^*)$ 的形式（为什么会想到能将其化为高斯分布的形式？因为其为 μ 的 2 次多项式）

$$\begin{aligned} \sum_{i=1}^n (x_i - \mu)^2 + \lambda_0 (\mu - \mu_0)^2 &= n\mu^2 - 2n\mu\bar{x} + \lambda_0\mu^2 - 2\lambda_0\mu_0\mu + \text{Const.} \\ &= (n + \lambda_0)\mu^2 - 2\mu(n\bar{x} + \lambda_0\mu_0) + \text{Const.} \\ &= (n + \lambda_0) \left(\mu^2 - \frac{2\mu(n\bar{x} + \lambda_0\mu_0)}{n + \lambda_0} \right) + \text{Const.} \\ &= (n + \lambda_0) \left(\mu - \frac{n\bar{x} + \lambda_0\mu_0}{n + \lambda_0} \right)^2 + \text{Const.} \end{aligned} \quad (83)$$

因此我们有

$$\begin{aligned} \log q_\mu^*(\mu) &= -\frac{\mathbb{E}_{q_\tau}[\tau]}{2} \left[\sum_{i=1}^n (x_i - \mu)^2 + \lambda_0 (\mu - \mu_0)^2 \right] + \text{Const.} \\ &= -\frac{\mathbb{E}_{q_\tau}[\tau] (n + \lambda_0)}{2} \left(\mu - \frac{n\bar{x} + \lambda_0\mu_0}{n + \lambda_0} \right)^2 + \text{Const.} \\ &= -\frac{1}{2} \underbrace{\mathbb{E}_{q_\tau}[\tau] (n + \lambda_0)}_{\tau^*} \left(\mu - \underbrace{\frac{n\bar{x} + \lambda_0\mu_0}{n + \lambda_0}}_{\mu^*} \right)^2 + \text{Const.} \\ \implies q_\mu^*(\mu) &= \mathcal{N} \left(\frac{n\bar{x} + \lambda_0\mu_0}{n + \lambda_0}, \mathbb{E}_{q_\tau}[\tau] (n + \lambda_0) \right) \quad \because -\frac{\tau}{2} (x - \mu)^2 \end{aligned} \quad (84)$$

利用式 79，去掉期望符号 $\mathbb{E}_{q_\tau}[\cdot]$ ，我们还可以得到 $p(\mu \mid \mathcal{D}, \tau)$ （注意，删掉期望值就是原分布的后验，因为 $p(\mathcal{D}, \tau)$ 在常数项里）

$$\begin{aligned} \log p(\mathcal{D} \mid \mu, \tau) + \log p(\mu \mid \tau) &= -\frac{\tau}{2} \sum_{i=1}^n (x_i - \mu)^2 - \frac{\lambda_0 \tau}{2} (\mu - \mu_0)^2 + \text{Const.} \\ &\quad \underbrace{\log(p(\mathcal{D} \mid \mu, \tau))}_{\log(p(\mathcal{D} \mid \mu, \tau))} \quad \underbrace{\log p(\mu \mid \tau)}_{\log p(\mu \mid \tau)} \\ &= -\frac{\tau}{2} \left[\sum_{i=1}^n (x_i - \mu)^2 + \lambda_0 (\mu - \mu_0)^2 \right] + \text{Const.} \\ &= -\frac{\tau (n + \lambda_0)}{2} \left(\mu - \frac{n\bar{x} + \lambda_0\mu_0}{n + \lambda_0} \right)^2 + \text{Const.} \\ \implies p(\mu \mid \mathcal{D}, \tau) &= \mathcal{N} \left(\frac{n\bar{x} + \lambda_0\mu_0}{n + \lambda_0}, \tau (n + \lambda_0) \right) \end{aligned} \quad (85)$$

同理我们可以计算 $\log q_\tau^*(\tau)$

$$\begin{aligned} \log q_\tau^*(\tau) &= \mathbb{E}_{q_\mu} [\log p(\mu, \tau, \mathcal{D})] \\ &= \mathbb{E}_{q_\mu} [\log p(\mathcal{D} \mid \mu, \tau) + \log p(\mu \mid \tau) + \log p(\tau)] + \text{Const.} \\ &= \mathbb{E}_{q_\mu} \left[\underbrace{\frac{n}{2} \log(\tau)}_{\log p(\mathcal{D} \mid \mu, \tau)} - \underbrace{\frac{\tau}{2} \sum_{i=1}^n (x_i - \mu)^2}_{\log p(\mu \mid \tau)} - \underbrace{\frac{\lambda_0 \tau}{2} (\mu - \mu_0)^2}_{\log p(\tau)} + (a_0 - 1) \log(\tau) - b_0 \tau \right] + \text{Const.} \\ &= \frac{n}{2} \log \tau + (a_0 - 1) \log \tau - b_0 \tau - \frac{\tau}{2} \mathbb{E}_{q_\mu(\mu)} \left[\sum_{i=1}^n (x_i - \mu)^2 + \lambda_0 (\mu - \mu_0)^2 \right] + \text{Const.} \\ &= \left(\underbrace{\frac{n}{2} + a_0 - 1}_{a_n} \right) \log \tau - \tau \left(\underbrace{b_0 + \frac{1}{2} \mathbb{E}_{q_\mu(\mu)} \left[\sum_{i=1}^n (x_i - \mu)^2 + \lambda_0 (\mu - \mu_0)^2 \right]}_{b_n} \right) + \text{Const.} \\ \implies q_\tau^*(\tau) &= \text{Gamma}(a_n, b_n) \end{aligned} \quad (86)$$

可以将 b_n 展开为

$$\begin{aligned}
b_n &= b_0 + \frac{1}{2} \mathbb{E}_{q_\mu} \left[\sum_{i=1}^n (x_i - \mu)^2 + \lambda_0 (\mu - \mu_0)^2 \right] \\
&= b_0 + \frac{1}{2} \mathbb{E}_{q_\mu} [-2\mu n\bar{x} + n\mu^2 + \lambda_0 \mu^2 - 2\lambda_0 \mu_0 \mu] + \sum_{i=1}^n x_i^2 + \lambda_0 \mu_0^2 \\
&= b_0 + \frac{1}{2} \left[(n + \lambda_0) \mathbb{E}_{q_\mu} [\mu^2] - 2(n\bar{x} + \lambda_0 \mu_0) \mathbb{E}_{q_\mu} [\mu] + \sum_{i=1}^n x_i^2 + \lambda_0 \mu_0^2 \right]
\end{aligned} \tag{87}$$

因为前面已经知道 $q_\mu(\mu)$ ，可以计算这里的 $\mathbb{E}_{q_\mu}[\mu^2]$ 和 $\mathbb{E}_{q_\mu}[\mu]$ 。

同样地，也可以轻易地获得原分布的后验 $p(\tau | \mathcal{D}, \mu)$

$$\begin{aligned}
\log p(\tau | \mathcal{D}, \mu) &= \log(p(\mathcal{D} | \mu, \tau)) + \log p(\mu | \tau) + \log p(\tau) + \text{Const.} \\
&= \underbrace{\frac{n}{2} \log(\tau) - \frac{\tau}{2} \sum_{i=1}^n (x_i - \mu)^2}_{\log(p(\mathcal{D} | \mu, \tau))} - \underbrace{\frac{\lambda_0 \tau}{2} (\mu - \mu_0)^2}_{\log p(\mu | \tau)} + \underbrace{(a_0 - 1) \log(\tau) - b_0 \tau}_{\log p(\tau)} + \text{Const.} \\
&= \underbrace{\left(\frac{n}{2} + a_0 - 1 \right) \log(\tau)}_{a_n} - \underbrace{\tau \left(b_0 + \frac{1}{2} \sum_{i=1}^n (x_i - \mu)^2 + \lambda_0 (\mu - \mu_0)^2 \right)}_{b_n} + \text{Const.} \\
\Rightarrow p(\tau | \mathcal{D}, \mu) &= \text{Gamma}(a_n, b_n) \\
&\begin{cases} a_n = \frac{n}{2} + a_0 \\ b_n = b_0 + \frac{1}{2} \sum_{i=1}^n (x_i - \mu)^2 + \lambda_0 (\mu - \mu_0)^2 \end{cases}
\end{aligned} \tag{88}$$

2.4 指数族分布

2.4.1 概览

给定先验和似然都是指数族分布，则他们形成一个共轭对，则变分推断（平均场近似）有下列更新公式

$$\eta_j = \mathbb{E}_{q(\mathbf{z} \setminus z_j | \cdot)} [\eta_{\text{post}}(\mathbf{z} \setminus z_j)] \tag{89}$$

这里的 $\eta_{\text{post}}(\mathbf{z} \setminus z_j)$ 是和后验分布 $p(z_j | \cdot)$ 相关的自然参数。

和通用的更新公式相比

$$\log q_i^*(z_i) = \mathbb{E}_{i \neq j} [\log p(\mathbf{x}, \mathbf{z})] \tag{90}$$

使用指数族更新公式更加直接方便。

2.4.2 指数族

指数族分布通常用自然参数 η 表示为下列形式

$$\begin{aligned}
&h(x) \exp(T(x)^\top \eta - A(\eta)) \\
&= \underbrace{\exp(-A(\eta))}_{\text{normalization}} h(x) \exp(T(x)^\top \eta) \\
\Rightarrow \exp(-A(\eta)) \int_x h(x) \exp(T(x)^\top \eta) &= 1 \\
\Rightarrow \int_x h(x) \exp(T(x)^\top \eta) &= \exp(A(\eta))
\end{aligned} \tag{91}$$

指数族分布具有易求最大似然估计的性质（可用下列高斯分布验证，其求解易于原参数形式）

$$\begin{aligned}
&\arg \max_{\eta} [\log p(X | \eta)] \\
&= \arg \max_{\eta} \left[\log \prod_{i=1}^n p(x_i | \eta) \right] \\
&= \arg \max_{\eta} \left[\log \left\{ \prod_{i=1}^n h(x_i) \exp \left(\sum_{i=1}^n T(x_i)^\top \eta - nA(\eta) \right) \right\} \right] \\
&= \arg \max_{\eta} \underbrace{\left[\sum_{i=1}^n T(x_i)^\top \eta - nA(\eta) \right]}_{\mathcal{L}(\eta)} \\
\Rightarrow \frac{\partial \mathcal{L}(\eta)}{\partial \eta} &= \sum_{i=1}^n T(x_i) - nA'(\eta) = 0 \\
\Rightarrow A'(\eta) &= \sum_{i=1}^n \frac{T(x_i)}{n}
\end{aligned} \tag{92}$$

从另一个角度来看，指数分布族具有性质：对数规范化因子 $A(\eta)$ 对自然参数 η 的导数等于充分统计量 $T(x)$ 的数学期望，这是任何情况都成立的

$$\begin{aligned}
\frac{d}{d\eta} A(\eta) &= \frac{d}{d\eta} \log \int h(x) \exp \{ \eta^T T(x) \} dx \\
&= \frac{\int T(x) \exp \{ \eta^T T(x) \} h(x) dx}{\int h(x) \exp \{ \eta^T T(x) \} dx} \\
&= \int T(x) \exp \{ \eta^T T(x) - A(\eta) \} h(x) dx \\
&= \int T(x) p(x | \eta) dx \\
&= E[T(X)]
\end{aligned} \tag{93}$$

例如一维高斯分布，可以将其写为指数族分布的形式

$$\begin{aligned}
\mathcal{N}(x; \mu, \sigma^2) &= (2\pi\sigma^2)^{-1/2} \exp \left(-\frac{(x - \mu)^2}{2\sigma^2} \right) \\
&= \exp \left(-\frac{x^2 - 2x\mu + \mu^2}{2\sigma^2} - \frac{1}{2} \log(2\pi\sigma^2) \right) \\
&= \exp \left(-\frac{1}{2\sigma^2} x^2 + \frac{\mu}{\sigma^2} x - \frac{\mu^2}{2\sigma^2} - \frac{1}{2} \log(2\pi\sigma^2) \right) \\
&= \exp \left([x \quad x^2] \begin{bmatrix} \frac{\mu}{\sigma^2} & -\frac{1}{2\sigma^2} \end{bmatrix}^\top - \frac{\mu^2}{2\sigma^2} - \frac{1}{2} \log(2\pi\sigma^2) \right)
\end{aligned} \tag{94}$$

其中

$$\begin{aligned}
T(x) &= [x \quad x^2] \\
\boldsymbol{\eta} &= [\eta_1 \quad \eta_2] \\
&= \left[\frac{\mu}{\sigma^2} \quad -\frac{1}{2\sigma^2} \right] \\
\boldsymbol{\theta} &= \begin{bmatrix} \mu \\ \sigma^2 \end{bmatrix} = \begin{bmatrix} \frac{-\eta_1}{2\eta_2} \\ \frac{-1}{2\eta_2} \end{bmatrix}
\end{aligned} \tag{95}$$

现在我们可以移除 μ 和 σ 得到一维高斯分布的指数族分布形式

$$\begin{aligned}
\mathcal{N}_{\text{nat}}(x, \boldsymbol{\eta}) &= \exp \left([x \quad x^2] [\eta_1 \quad \eta_2]^\top - \frac{\mu^2}{2\sigma^2} - \frac{1}{2} \log(2\pi\sigma^2) \right) \\
&= \exp \left([x \quad x^2] [\eta_1 \quad \eta_2]^\top - \frac{\left(\frac{-\eta_1}{2\eta_2} \right)^2}{2 \left(\frac{-1}{2\eta_2} \right)} - \frac{1}{2} \log \left(2\pi \left(\frac{-1}{2\eta_2} \right) \right) \right) \\
&= \exp \left(T(x)^\top \boldsymbol{\eta} + \frac{\eta_1^2}{4\eta_2} - \frac{1}{2} \log \left(\frac{2\pi}{-2\eta_2} \right) \right) \\
&= \exp \left(T(x)^\top \boldsymbol{\eta} + \underbrace{\frac{\eta_1^2}{4\eta_2} + \frac{1}{2} \log(-2\eta_2) - \frac{1}{2} \log(2\pi)}_{A(\boldsymbol{\eta})} \right)
\end{aligned} \tag{96}$$

2.4.3 共轭概率

共轭表示先验和后验是同种形式的概率分布，例如

$$p_{\eta_{\text{post}}}(\boldsymbol{\theta} | \mathbf{x}) \propto p(\mathbf{x} | \boldsymbol{\theta}) p_{\eta_{\text{prior}}}(\boldsymbol{\theta}) \tag{97}$$

使用指数族分布表示时，共轭的含义是先验和后验有相同的充分统计量 $T(\boldsymbol{\theta})$ 和 $h(\boldsymbol{\theta})$ （注意这里的 $\boldsymbol{\theta}$ 是变量），不同的自然参数，即 $\eta_{\text{post}}, \eta_{\text{prior}}$ 以及不同的对数归一化因子。

证明：一个指数族分布的先验必定有对应的似然使其拥有一个共轭的后验

$$\begin{aligned}
p(\boldsymbol{\theta} | x) &\propto p(x | \boldsymbol{\theta}) p(\boldsymbol{\theta}) \\
&= h(x) \exp \{ T(x)\boldsymbol{\theta} - A_I(\boldsymbol{\theta}) \} \times h(\boldsymbol{\theta}) \exp \{ T(\boldsymbol{\theta})^\top \boldsymbol{\alpha} - A(\boldsymbol{\alpha}) \} \quad T(\boldsymbol{\theta}) = \begin{bmatrix} \boldsymbol{\theta} \\ -g(\boldsymbol{\theta}) \end{bmatrix}, \boldsymbol{\alpha} = \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix} \\
&\propto h(\boldsymbol{\theta}) \exp \{ T(x)\boldsymbol{\theta} - A_I(\boldsymbol{\theta}) + \alpha_1 \boldsymbol{\theta} - \alpha_2 g(\boldsymbol{\theta}) \} \\
&= h(\boldsymbol{\theta}) \exp \{ (T(x) + \alpha_1)\boldsymbol{\theta} - (A_I(\boldsymbol{\theta}) + \alpha_2 g(\boldsymbol{\theta})) \} \quad \text{assume } g(\boldsymbol{\theta}) = A_I(\boldsymbol{\theta}) \\
&= h(\boldsymbol{\theta}) \exp \{ (T(x) + \alpha_1)\boldsymbol{\theta} - (1 + \alpha_2)A_I(\boldsymbol{\theta}) \} \\
&= h(\boldsymbol{\theta}) \exp \{ [\hat{\alpha}_1 \quad \hat{\alpha}_2] T(\boldsymbol{\theta}) \}
\end{aligned} \tag{98}$$

即似然函数的对数归一化因子等于先验的充分统计量的第二部分（不止两个参数的情况？），则它们共轭

2.4.4 变分推断

隐变量 β 的后验分布，注意这里的 $h(\beta)$ 和 $T(\beta)$ 是相同的，因为设置变分分布是和原分布同一种分布

$$\begin{aligned}
p(\beta | z, x) &= h(\beta) \exp (T(\beta)^\top \eta(z, x) - A_g(\eta(z, x))) \\
&\approx q(\beta | \lambda) = h(\beta) \exp (T(\beta)^\top \lambda - A_g(\lambda))
\end{aligned} \tag{99}$$

隐变量 z 的后验分布

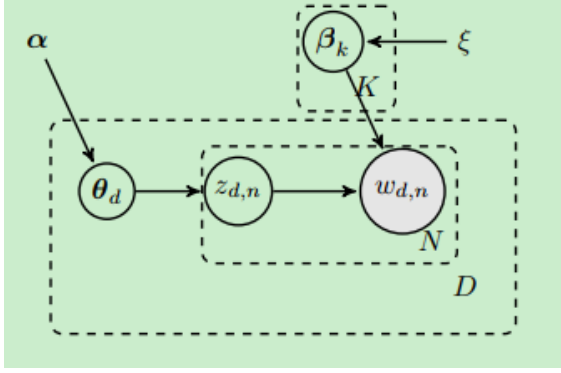
$$\begin{aligned} p(z | \beta, x) &= h(z) \exp(T(z)^\top \eta(\beta, x) - A_I(\eta(\beta, x))) \\ &\approx q(z | \phi) = h(z) \exp(T(z)^\top \phi - A_I(\phi)) \end{aligned} \quad (100)$$

固定 ϕ , 优化 λ , 每一步去除无关项

$$\begin{aligned} \mathcal{L}(\lambda, \phi) &= E_{q(z, \beta)}[\log p(x, z, \beta)] - E_{q(z, \beta)}[\log q(z, \beta)] \\ &= E_{q(z, \beta)}[\log p(\beta | x, z) + \log p(z, x)] - E_{q(z, \beta)}[\log q(\beta)] - E_{q(z, \beta)}[\log q(z)] \\ &= E_{q(z, \beta)}[\log p(\beta | x, z)] - E_{q(z, \beta)}[\log q(\beta)] \\ &= E_{q(z, \beta)}[\log h(\beta)] + E_{q(z, \beta)}[T(\beta)^\top \eta(z, x)] - E_{q(z, \beta)}[A_g(\eta(z, x))] - E_{q(z, \beta)}[\log h(\beta)] - E_{q(z, \beta)}[T(\beta)^\top \lambda] + E_{q(z, \beta)}[A_g(\lambda)] \\ &= E_{q(\beta)}[T(\beta)^\top] E_{q(z)}[\eta(z, x)] - E_{q(z)}[A_g(\eta(z, x))] - E_{q(\beta)}[T(\beta)^\top \lambda] + A_g(\lambda) \quad \text{using } \frac{\partial A_I(\eta)}{\partial \eta} = E_{p(x|\eta)}[T(x)] \\ &= A'_g(\lambda)^\top E_{q(z)}[\eta(z, x)] - \lambda A'_g(\lambda)^\top + A_g(\lambda) \quad \text{taking partial derivative with respect to } \lambda \\ &= A''_g(\lambda)^\top E_{q(z)}[\eta(z, x)] - A'_g(\lambda)^\top - \lambda A'_g(\lambda)^\top + A'_g(\lambda) \\ &= A''_g(\lambda)^\top (E_{q(z)}[\eta(z, x)] - \lambda) = 0 \\ &\implies \lambda = E_{q(z)}[\eta(z, x)] \quad \text{where } q(z) = q(z | \phi) \end{aligned} \quad (101)$$

同理, 固定 λ , 优化 ϕ , 可得到更新公式 $\phi = E_{q(\beta|\lambda)}[\eta(\beta, x)]$

2.5 基于变分推断的LDA参数学习



For each topic k :

$$\beta_k \sim \text{Dir}(\xi, \dots, \xi) \quad \text{for } k \in \{1, \dots, K\} \quad (102)$$

For each document d :

$$\theta_d \sim \text{Dir}(\alpha, \dots, \alpha) \quad (103)$$

For each word $w \in \{1, \dots, N\}$:

$$\begin{aligned} z_{d,n} &\sim \text{Mult}(\theta_d) \\ w_{d,n} &\sim \text{Mult}(\beta_{z_{d,n}}) \end{aligned} \quad (104)$$

因为先验和似然是共轭的, 而选取变分分布时应该选取和后验相同的分布有利于计算, 因此此时选取和先验同样的分布 (如上述通用公式, 作者猜测选取不同的分布也能计算)

$$q(\beta_k) = \text{Dir}(\lambda_k), \quad q(\theta_d) = \text{Dir}(\gamma_d), \quad q(z_{d,n}) = \text{Mult}(\phi_{d,n}) \quad (105)$$

前两个变分分布选取易理解, 最后一个的理由是什么呢? 猜测是写出其后验是多项式分布形式

2.5.1 更新 $\phi_{d,n}$

首先找到后验 $p(z_{d,n} = k | \theta_d, \varphi_k, w_{d,n})$ 的自然参数

$$\begin{aligned} p(z_{d,n} = k | \theta_d, \beta_{1:K}, w_{d,n}) &\propto p(z_{d,n} = k | \theta_d) \cdot p(w_{d,n} | z_{d,n} = k, \beta_{1:K}) \\ &= \theta_{d,k} \cdot \beta_{k,w_{d,n}} \\ &= \exp \left(\underbrace{(\log \theta_{d,k} + \log \beta_{k,w_{d,n}})}_{\eta(\theta_{d,k}, \beta_{1:K}, w_{d,n})} \cdot \underbrace{1}_{T(z_{d,n})} \right) \end{aligned} \quad (106)$$

使用正常的多项式分布可以表达为

$$p(z_{d,n} | \theta_d, \beta_{1:K}, w_{d,n}) = \text{Mult}(\theta_{d,1} \cdot \beta_{1,w_{d,n}}, \dots, \theta_{d,k} \cdot \beta_{k,w_{d,n}}) \quad (107)$$

利用更新公式可知

$$\begin{aligned}
\eta(\phi_{d,n}^k) &= \log(\phi_{d,n}^k) \\
&\propto E_{q(\boldsymbol{\theta}_d, \beta_k)}[\eta_l(\boldsymbol{\theta}_d, \boldsymbol{\beta}_{1:K}, w_{d,n})] \quad \because \text{自然参数是正常参数的对数, 为何正比?} \\
&= E_{q(\boldsymbol{\theta}_d)}[\log(\theta_{d,k})] + E_{q(\beta_k)}[\log(\beta_k, w_{d,n})] \quad \text{迪利克雷分布的性质} \\
&= \Psi(\gamma_{d,k}) - \Psi\left(\sum_{k=1}^K \gamma_{d,k}\right) + \Psi(\lambda_{k,w_{d,n}}) - \Psi\left(\sum_{v=1}^V \lambda_{k,v}\right) \\
&\implies \phi_{d,n}^k \propto \exp\left[\Psi(\gamma_{d,k}) - \Psi\left(\sum_{k=1}^K \gamma_{d,k}\right) + \Psi(\lambda_{k,w_{d,n}}) - \Psi\left(\sum_{v=1}^V \lambda_{k,v}\right)\right] \\
&\propto \exp\left[\Psi(\gamma_{d,k}) + \Psi(\lambda_{k,w_{d,n}}) - \Psi\left(\sum_{v=1}^V \lambda_{k,v}\right)\right]
\end{aligned} \tag{108}$$

2.5.2 更新 γ_d

同样先推导后验 $p(\boldsymbol{\theta}_d \mid \mathbf{z}_d)$ 的表达式

$$\begin{aligned}
p(\boldsymbol{\theta}_d \mid \mathbf{z}_d) &= p(\boldsymbol{\theta}_d \mid \boldsymbol{\alpha}) \prod_{n=1}^N p(z_{d,n} \mid \boldsymbol{\theta}_d) \\
&= \prod_{k=1}^K \left(\theta_{d,k}^{\alpha_k-1} \prod_{n=1}^N \theta_{d,k}^{\mathbb{1}(z_{d,n}=k)} \right) \\
&= \exp \left[\log \left(\prod_{k=1}^K \left(\theta_{d,k}^{\alpha_k-1} \prod_{n=1}^N \theta_{d,k}^{\mathbb{1}(z_{d,n}=k)} \right) \right) \right] \\
&= \exp \left[\sum_{k=1}^K \log \left(\theta_{d,k}^{\alpha_k-1} \prod_{n=1}^N \theta_{d,k}^{\mathbb{1}(z_{d,n}=k)} \right) \right] \\
&= \exp \left[\sum_{k=1}^K \left(\log \theta_{d,k}^{\alpha_k-1} + \sum_{n=1}^N \log \left(\theta_{d,k}^{\mathbb{1}(z_{d,n}=k)} \right) \right) \right] \\
&= \exp \left[\sum_{k=1}^K \left((\alpha_k - 1) \log \theta_{d,k} + \sum_{n=1}^N \mathbb{1}(z_{d,n} = k) \log \theta_{d,k} \right) \right] \\
&= \exp \left[\sum_{k=1}^K \left(\alpha_k - 1 + \sum_{n=1}^N \mathbb{1}(z_{d,n} = k) \right) \log \theta_{d,k} \right] \\
&= \exp \left(\underbrace{\begin{bmatrix} \alpha_1 - 1 + n_1 \\ \vdots \\ \alpha_K - 1 + n_K \end{bmatrix}}_{\eta_l(\boldsymbol{\alpha}, \mathbf{z}_d)}^\top \underbrace{\begin{bmatrix} \log \theta_{d,1} \\ \vdots \\ \log \theta_{d,K} \end{bmatrix}}_{T(\boldsymbol{\theta}_d)} \right) \quad \text{by letting } n_k = \sum_{n=1}^N \mathbb{1}(z_{d,n} = k) \\
&= \text{Dir}(\alpha_1 + n_1, \dots, \alpha_K + n_K)
\end{aligned} \tag{109}$$

接下来用变分分布 $q(\gamma_d) = \text{Dir}(\eta(\gamma_d))$ 来近似 $p(\boldsymbol{\theta}_d \mid \mathbf{z}_d)$, 利用更新公式

$$\begin{aligned}
\eta(\gamma_d) &= E_{q(\mathbf{z}_d \mid \phi_d)}[\eta_l(\boldsymbol{\alpha}, \mathbf{z}_d)] \\
&= E_{q(\mathbf{z}_d \mid \phi_d)}[(\alpha_1 - 1 + n_1) \dots (\alpha_K - 1 + n_K)]
\end{aligned} \tag{110}$$

计算这个期望

$$\begin{aligned}
E_{q(\mathbf{z}_d \mid \phi_d)} \left[\sum_{n=1}^N \mathbb{1}(z_{d,n} = k) \right] &= \sum_{n=1}^N E_{q(\mathbf{z}_d \mid \phi_d)}[\mathbb{1}(z_{d,n} = k)] \\
&= \sum_{n=1}^N q(z_{d,n} = k) \\
&= \sum_{n=1}^N \phi_{d,n}^k
\end{aligned} \tag{111}$$

因此有

$$\begin{aligned}
\eta(\gamma_d) &= \left[\left(\alpha_1 - 1 + \sum_{n=1}^N \phi_{d,n}^1 \right) \dots \left(\alpha_K - 1 + \sum_{n=1}^N \phi_{d,n}^K \right) \right] \\
\implies \boldsymbol{\eta} &= \left[\left(\alpha_1 + \sum_{n=1}^N \phi_{d,n}^1 \right) \dots \left(\alpha_K + \sum_{n=1}^N \phi_{d,n}^K \right) \right] \quad \text{迪利克雷分布的自然参数 } \eta_i = \alpha_i - 1 \\
&= \boldsymbol{\alpha} + \sum_{n=1}^N \boldsymbol{\phi}_{d,n}
\end{aligned} \tag{112}$$

2.5.3 更新 λ_k

与 γ_d 更新公式类似, 有

$$\boldsymbol{\lambda}_k = \boldsymbol{\xi} + \sum_{d=1}^D \sum_{n=1}^N w_{d,n} \cdot \boldsymbol{\phi}_{d,n}^k \tag{113}$$

2.6 随机变分推断

3. 变分自编码器