Mathematical Note On a Novel Collaborative Filtering Paper

马晨

2016 年 10 月 30 日

摘要

最近阅读《A non negative matrix factorization for collaborative filtering recommender systems based on a Bayesian probabilistic model》，这是阅读后的推导细节笔记以及总结。

目录

1 mathematical background

这篇文章很明显借鉴了 David M.Blei 的 <Latent Dirichlet Allocation> 的思路，采用变分推断法推导，这是这篇论文的推导的数学细节。

1.1 Binomial distribution

Binomial distribution(二项分布) 中学就学过，在概率论中，二项分布即重复 n 次独立的伯努利试验。在每次试验中只有两种可能的结果（成功/失败），每次成功的概率为 p，而且两种结果发生与否互相对立，并且相互独立，与其它各次试验结果无关，事件发生与否的概率在每一次独立试验中都保持不变，则这一系列试验总称为 n 重伯努利实验，当试验次数为 1 时，二项分布就是伯努利分布。

在给出二项分布之前，我们来做一个例子，假设你在玩 CS 这个游戏，你拿着狙击枪，敌人出现你打中敌人的概率是 p，打不中敌人的概率是 1-p，那么敌人第一次出现你没打中而第二次出现你打中的概率是 $(1-p)\cdot p$。如果敌人出现了 n 次，而你打中了其中的 k 次，而不确定具体在哪 k 次（第 1 次，还是第 4 次？），这样从 n 次中任取 k 次的次数是 $C_n^k=\binom{n}{k}$ 而这不确定的 k 次打中敌人的概率是：$\binom{n}{k}p^k(1-p)^{n-k}$，通过这个例子我们便得知了二项分布的概率函数。

二项分布的概率密度函数是：

$$\begin{aligned}&f(k;n,p)=P(X=k)=\tbinom{n}{k}p^k(1-p)^{n-k}\\&for\quad k=0,1,...,n,where\\&\tbinom{n}{k}=\tfrac{n!}{k!(n-k)!}\end{aligned}\tag{1.1}$$

1.2 Beta distribution

在概率论中，beta 分布是指一组定义在区间 (0,1) 的连续概率分布，有两个参数 α 和 β，且 $\alpha,\beta>0$。

Beta 分布的概率密度函数是:

$$
\begin{aligned}
f(x;\alpha,\beta) &= \frac{x^{\alpha-1}(1-x)^{\beta-1}}{\int_0^1 u^(\alpha-1)(1-u)^(\beta-1)\,\mathrm{d}u} \\
&= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1} \\
&= \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}
\end{aligned} \tag{1.2}
$$

随机变量 X 服从参数为的 Beta 分布通常写作：$X \sim Beta(\alpha,\beta)$。

公式(1.2)中分母的函数 B 称为 β 函数。

另外一个关于 B 函数的重要的关系是欧拉第一型积分：

$$
B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} = \int_0^1 \mu^{\alpha-1}(1-\mu)^{\beta-1}\,\mathrm{d}\mu \tag{1.3}
$$

1.3 Multinomial distribution

多项分布是二项分布的推广扩展，在 n 次独立试验中每次只输出 k 种结果中的一个，且每种结果都有一个确定的概率 p。多项分布给出了在多种输出状态的情况下，关于成功次数的各种组合的概率。举个例子，投掷 n 次骰子，这个骰子共有 6 种结果输出（k=6），且 1 点出现概率为 p_1，2 点出现概率 p_2，... 多项分布给出了在 n 次试验中，骰子 1 点出现 x_1 次，2 点出现 x_2 次,3 点出现 x_3 次，⋯，6 点出现 x_6 次。这个结果组合的概率为：

$$
\begin{aligned}
f(x_1,\ldots,x_k) &= P(X_1 = x_1 \text{ and } \ldots \text{ and } X_k = x_k) \\
&= \begin{cases} \frac{n!}{x_1!\ldots x_k!} p_1^{x_1} \cdots \cdot p_k^{x_k} & \text{when } \sum_{i=1}^k x_i = n \\ 0 & otherwise \end{cases}
\end{aligned} \tag{1.4}
$$

公式(1.4)是多项分布的概率公式，注意在这个公式中，为第 i 种状态的输出结果的频度，如果 k=2，只有两种情况，此公式将退化为二项分布，所以二项分布是特殊情况下的多项分布。公式(1.4)也可以用 gamma 函数表示（这个写法的形式和 Dirichlet 分布相似）：

$$
f(x_1,\ldots,x_k;n,p_1,\ldots,p_k) = \frac{\Gamma(\sum_i x_i + 1)}{\prod_i \Gamma(x_i+1)} \prod_{i=1}^k p_i^{x_i} \tag{1.5}
$$

1.4 Dirichlet distribution

Dirichlet 分布是 Beta 分布在多项情况下的推广，也是多项分布的共轭先验分布。dirichlet 分布的概率密度函数如下：

$$\begin{aligned} f(p_1,\ldots,p_{k-1};\alpha_1,\ldots,\alpha_k) &= \frac{1}{\Delta(\vec{\alpha})}\prod_{i=1}^{k} p_i^{\alpha_i-1} \\ \text{这里}\Delta(\vec{\alpha}) &= \frac{\prod_{i=1}^{k}\Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^{k}(\alpha_i))} \end{aligned} \tag{1.6}$$

二项分布和多项分布很相似，Beta 分布和 Dirichlet 分布很相似，而至于“Beta 分布是二项式分布的共轭先验概率分布，而狄利克雷分布（Dirichlet 分布）是多项式分布的共轭先验概率分布”这点在下文中说明。

另外，关于 Dirichlet 分布有个重要的性质：

$$\psi(x) = (\ln(\Gamma(x))' = \frac{\Gamma'(x)}{\Gamma(x)} \tag{1.7}$$

1.5 conjugacy prior

所谓的共轭 (conjugacy)，只是我们选取 (choose) 一个函数作为似然函数 (likelihood function) 的 prior probability distribution，使得后验分布函数 (posterior distributions) 和先验分布函数形式一致。比如 Beta 分布是二项式分布的共轭先验概率分布，而狄利克雷分布 (Dirichlet 分布）是多项式分布的共轭先验概率分布。为什么要这样做呢？这得从贝叶斯估计谈起：

根据贝叶斯规则，后验分布 = 似然函数 * 先验分布

$$p(\theta|x) = \frac{\overbrace{p(x|\theta)}^{likelihood}\ \overbrace{p(\theta)}^{priorbelief}}{\underbrace{p(x)}_{evidence}} = \frac{p(x|\theta)p(\theta)}{\int p(x|\theta)p(\theta)\,\mathrm{d}\theta} \propto p(x|\theta)p(\theta) \tag{1.8}$$

参数估计是一个重要的话题。对于典型的离散型随机变量分布：二项式分布，多项式分布；典型的连续型随机变量分布：正态分布。他们都可以看着是参数分布，因为他们的函数形式都被一小部分的参数控制，比如正态分布的均值和方差，二项式分布事件发生的概率等。因此，给定一堆观测数据集（假定数据满足独立同分布），我们需要有一个解决方案来确定这些参数值的大小，以便能够利用分布模型来做密度估计。这就是参数估计！对于参数估计，一直存在两个学派的不同解决方案。一是频率学派解决方

案：通过某些优化准则（比如似然函数）来选择特定参数值；二是贝叶斯学派解决方案：假定参数服从一个先验分布，通过观测到的数据，使用贝叶斯理论计算对应的后验分布。先验和后验的选择满足共轭，这些分布都是指数簇分布的例子。

简而言之，假设参数 θ 也是变量而非常量，而且在做试验前已经服从某个分布 $p(\theta)$（来源于以前做试验数据计算得到，或来自于人们的主观经验），然后现在做新试验去更新这个分布假设。

求证:Beta 分布确实是二项分布的共轭先验分布

证明：

1.Binomial 分布的似然函数：

$$L = \binom{s+f}{s} p^s (1-p)^f$$

2. 先验分布 Beta 分布如下：

$$P(p|\alpha,\beta) = \frac{p^{\alpha-1}(1-p)^{\beta-1}}{B(\alpha,\beta)}, \text{其中} B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

3. 由于 prior distribution * likehood = post distribution

$$\begin{aligned} p(p|s,f,\alpha,\beta) &= \frac{\binom{s+f}{s} \cdot p^s(1-p)^f \cdot p^{\alpha-1}(1-p)^{\beta-1}/B(\alpha,\beta)}{\int_{q=0}^1 \binom{s+f}{s}(q^s(1-q)^f \cdot q^{\alpha-1}(1-q)^{\beta-1}/B(\alpha,\beta))\mathrm{d}q} \\ &= \frac{p^{s+\alpha-1}(1-p)^{f+\beta-1}}{B(s+\alpha,f+\beta)} \end{aligned} \tag{1.9}$$

这就可以看到后验分布 (post distribution) 又变为 beta 分布，也就是和先验 (prior distribution) 一致了，因此我们称之为共轭 (conjugacy)。

2 derivation

2.1 variational Bayes

思想：用一个近似的变分分布 Q 来趋近于真实分布 P，也就是缩小 KL 距离的差距。以下的标准变分贝叶斯内容，在《Pattern Recoginition and Machine Learning》一书第 10 章也有写，但是写的不全。

$$
\begin{aligned}
D_{KL}(Q||P) &= KL(Q(Z)||P(Z|D)) \\
&= \sum_{z} Q(Z) \cdot \log \frac{Q(Z)}{P(Z,D)} + \log P(D)
\end{aligned}
$$

移项：

$$
\underbrace{\log P(D)}_{\text{log-likelihood}} = D_{KL}(Q||P) \underbrace{- \sum_{z} Q(Z) \log \frac{Q(Z)}{P(Z,D)}}_{L(Q(Z))\text{: lower bound}}
$$

来看看 $L(Q(Z))$ 这一项被称为下界：

$$
\begin{aligned}
L(Q(Z)) &= -\sum_{Z} Q(Z) \dot{\log} \frac{Q(Z)}{P(Z,D)} \\
&= \underbrace{\sum_{Z} Q(Z) \log P(Z,D)}_{\text{energy:} E_Q[\log P(Z,D)]} \underbrace{- \sum_{Z} Q(Z) \log Q(Z)}_{\text{entropy:} H(Q)} \\
&= \int (\prod_{i} Q_i(Z_i)) \cdot \ln P(Z,D) \, \mathrm{d}Z - \int (\prod_{k} Q_k(Z_k)) \sum_{i} \ln Q_i(Z_i) \, \mathrm{d}Z
\end{aligned}
\tag{2.1}
$$

公式(2.1)的来源是将其写成积分的形式，再利用 mean-field(平均场) 理论的分解:$Q(Z) = \prod_i Q_i(Z_i)$，现在考虑 $Z = \{Z_i, Z_{-i}\}$，其中 $Z_{-i} = Z \backslash Z_i$，先看 energy 项:

$$
\begin{aligned}
E_Q(Z)[\ln P(Z,D)] &= \int (\prod_{i} Q_i(Z_i) \cdot \ln P(Z,D) \, \mathrm{d}Z \\
&= \int Q_i(Z_i) \, \mathrm{d}Z_i \cdot \int Q_{-i}(Z_{-i}) \cdot \ln P(Z,D) \, \mathrm{d}Z_{-i} \\
&= \int Q_i(Z_i) \cdot \underbrace{\langle \ln P(Z,D) \rangle_{Q_{-i}(Z_{-i})}}_{\int Q_{-i}(Z_{-i}) \cdot \ln P(Z,D) \, \mathrm{d}Z_{-i}} \, \mathrm{d}Z_i \\
&= \int Q_i(Z_i) \cdot \ln \left[\exp\{\langle \ln P(Z,D) \rangle_{Q_{-i}(Z_{-i})}\} \right] \, \mathrm{d}Z_i \\
&= \int Q_i(Z_i) \cdot \ln Q^*(Z_i) \, \mathrm{d}Z_i + \underbrace{\ln C}_{\text{constant}}
\end{aligned}
$$

最后一步中，令 $Q^*(Z_i) = \frac{1}{C}\exp\langle \ln P(Z,D)\rangle_{Q_{-i}(Z_{-i})}$。

再来看 entropy 项：

$$
\begin{aligned}
H(Q(Z)) &= -\sum_Z \underbrace{Q(Z)}_{\prod_k Q_k(Z_k)} \underbrace{\ln Q(Z)}_{\sum_i \ln Q_i(Z_i)} \\
&= -\int (\prod_k Q_k(Z_k) \cdot \sum_i \ln Q_i(Z_i)\,\mathrm{d}Z \\
&= -\sum_i \int (\prod_k Q_k(Z_k)) \cdot \ln Q_i(Z_i) \underbrace{\mathrm{d}Z}_{\text{拆成}\,\mathrm{d}Z_i\,\mathrm{d}Z_{-i}} \\
&= -\sum_i \iint Q_i(Z_i) Q_{-i}(Z_{-i}) \cdot \ln Q_i(Z_i)\,\mathrm{d}Z_i\,\mathrm{d}Z_{-i} \\
&= -\sum_i \left\langle \int Q_i(Z_i) \ln Q_i(Z_i)\,\mathrm{d}Z_i \right\rangle_{Q_{-i}(Z_{-i})} \\
&= -\sum_i \int Q_i(Z_i) \ln Q_i(Z_i)\,\mathrm{d}Z_i
\end{aligned}
$$

上述最后一步的公式是由于 $\int Q_{-i}(Z_{-i})\,\mathrm{d} = 1$ 而得到。

得到泛函:

$$
\begin{aligned}
L(Q(Z)) &= \int Q_i(Z_i) \ln Q_i^*(Z_i)\,\mathrm{d}Z_i - \sum_i \int Q_i(Z_i) \ln Q_i(Z_i)\,\mathrm{d}Z_i + \ln C \\
&= (\int Q_i(Z_i) \ln Q_i^*(Z_i)\,\mathrm{d}Z_i - \int Q_i(Z_i) \ln Q_i(Z_i)\,\mathrm{d}Z_i) \\
&\quad - \sum_{k \neq i} \int Q_k(Z_k) \ln Q_k(Z_k)\,\mathrm{d}Z_k + \ln C \\
&= \int Q_i \ln \frac{Q^*(Z_i)}{Q_i(Z_i)}\,\mathrm{d}Z_i - \sum_{k \neq i} \int Q_k(Z_k) \ln Q_k(Z_k)\,\mathrm{d}Z_k + \ln C \\
&= -D_{KL}(Q_i(Z_i) || Q^*(Z_i)) + H[Q_{-i}(Z_{-i})] + \ln C
\end{aligned}
$$

我们一开始试图最小化 Q 和 P 的距离，这很困难，问题转化为 $Q_i(Z_i)$ 和 $Q^*(Z_i)$ 这样的一维分布的 KL 距离, 现在只需要让 $D_{KL}(Q_i(Z_i)||Q^*(Z_i))$ 这个距离等于 0 即可。

$$
Q_i(Z_i) = Q^*(Z_i) = \frac{1}{C}\exp\{< \ln P(\underbrace{Z}_{Z_i, Z_{-i}}, D) >_{Q_{-i}(Z_{-i})}\} \tag{2.2}
$$

$$
\ln Q_i(Z_i) =< \ln P(Z_i, Z_{-i}, D) >_{Q_{-i}(Z_{-i})} + \text{constant} \tag{2.3}
$$

固定其他 Q_{-i}, 只对其中一个 Q_i 更新，称之为 coordinate-ascent。

2.2 paper model notation

首先有必要对模型进行简短的解释，该模型与 Latent Dirichlet Allocation 相同，都属于生成式模型，该模型最终用户对某个电影的评分过程是：$\rho_{u,i} \sim Bin(R, \kappa_{i,z_{u,i}})$，也就是说，假设一个电影网站一共有 R 个评分等级（比如常见的 5 颗星），该用户打分的过程就是不断对其做伯努利试验，一共做 R 次试验，每次打中一颗星的概率是 $\kappa_{i,z_{u,i}}$，一共对这个影片打的评分就是 R 次试验中打中了几次，这里的 $z_{u,i}$ 是用户组的指定 (assign)，从另一个多项分布得到。

我们先来写出论文里假设的变分分布 q:

$$q(\vec{\phi}_u, \kappa_{i,k}, z_{u,i}) = \prod_{u=1}^{N} q_{\vec{\phi}_u}(\vec{\phi}_u) \prod_{i=1}^{M}\prod_{k=1}^{K} q_{\kappa_{i,k}}(\kappa_{i,k}) \prod_{r_{u,i}\neq\bullet} q_{z_{u,i}}(z_{u,i}) \tag{2.4}$$

N 表示 N 个用户，M 表示 M 个物品，K 表示共把用户分成 K 个组，类似 LDA 中的主题个数，这个 K 值也是用户设置的。u 表示第 u 个用户 (user)，i 表示第 i 个物品 (item)，k 表示第 k 个用户组。

其中:

$$q_{\vec{\phi}_u}(\vec{\phi}_u) \sim Dir(\gamma_{u,1}, \ldots, \gamma_{u,K}) \tag{2.5}$$

解释：其中 $\vec{\phi}_u$ 是用户 u 属于各个组 $(\phi_{u,1}, \ldots, \phi_{u,K})$ 的概率向量。

$$q_{\kappa_{i,k}}(\kappa_{i,k}) \sim Beta(\epsilon^+_{i,k}, \epsilon^-_{i,k}) \tag{2.6}$$

解释：其中 $\kappa_{i,k}$ 表示第 k 组中的用户喜欢物品 i 的概率。

$$q_{z_{u,i}}(z_{u,i}) \sim Mult(\lambda_{u,i,1}, \ldots, \lambda_{u,i,K}) \tag{2.7}$$

解释：其中 $z_{u,i}$ 中的 k 值表示用户 u 在给物品 i 打分 (rate) 时，用户就好像 (as if) 属于第 k 组。也就是说 $z_{u,i}$ 表示用户 u 在给物品 i 打分的时候，落在了 $1, \ldots, K$ 中哪个组里了。

这里的 $\lambda_{u,i,k}$ 是用户 u 在对物品 i 打分时，落在第 k 组的概率（服从多项分布）。所以：

$$\lambda_{u,i,k} = q_z(z_{u,i} = k) \tag{2.8}$$

$$\lambda_{u,i,1} + \cdots + \lambda_{u,i,K} = 1 \tag{2.9}$$

后面推导中的其他符号定义：

$$\begin{aligned} r^+_{u,i} &= \rho_{u,i} = R \cdot r^*_{u,i} \\ r^-_{u,i} &= R - \rho_{u,i} = R \cdot (1 - r^*_{u,i}) \end{aligned} \tag{2.10}$$

解释：$r^+_{u,i}$ 表示用户 u 对物品 i 的打的评分是多少分，R 表示一共 R 个等级的评分，比如电影网站常见的 5 个评分等级。$r^-_{u,i}$ 表示用户 u 未对物品 i 打的分数（以 R 来衡量），所以是 $R - r^+_{u,i}$。

2.3 paper model derivation

2.3.1 derivation of $q_{\vec{\phi}_u(\vec{\phi}_u)}$

我们利用公式(2.2)开始推导，我们先来看看公式(2.4)中的 $q_{\vec{\phi}_u(\vec{\phi}_u)}$，此时公式(2.2)中的 Q_{-i} 变为了 q_k, q_z：

$$\begin{aligned} q_{\vec{\phi}_u(\vec{\phi}_u)} &\propto \exp\{\mathbf{E}_{q_k,q_z}[\ln p(\vec{\phi}_u, \kappa, z, \rho)]\} \\ &= \exp\{\mathbf{E}_{q_k,q_z}[\ln\{p(\vec{\phi}_u|\kappa, z, \rho)p(\kappa, z, \rho)\}]\} \\ &= \exp\{\mathbf{E}_{q_k,q_z}[\ln p(\vec{\phi}_u|\kappa, z, \rho)] + \underbrace{\mathbf{E}_{q_k,q_z}[\ln p(\kappa, z, \rho)]}_{q_{\vec{\phi}_u(\vec{\phi}_u)}\text{only w.r.t. } \vec{\phi}_u}\} \\ &\propto \exp\{\mathbf{E}_{q_k,q_z}[\ln p(\vec{\phi}_u|\kappa, z, \rho)]\} + constant \qquad (2.11) \\ &\propto \exp\{\mathbf{E}_{q_z}[\ln p(\vec{\phi}_u|z_{u,i_1}, \ldots, z_{u,i_M})]\} \qquad (2.12) \end{aligned}$$

上文最后一步公式(2.11)到公式(2.12)的中 $\mathbf{E}_{q_k,q_z}$ 变为 $\mathbf{E}_{q_z}$ 是由于 q_k 与 $\vec{\phi}_u$ 没有关系，而 q_z 与 $\vec{\phi}_u$ 有关系，这一点从概率图模型也可以看得出来。

另外，$\vec{\phi}_u|z_{u,i_1}, \ldots, z_{u,i_M}$ 是来源于 Dirichlet 分布和 Multinomial 分布的共轭性质（ϕ_u 来源于 Dirichlet 分布，$z_{u,i}$ 来源于 Multinomial 分布），该贝叶斯共轭如下：

$$\vec{\phi}_u|z_{u,i_1},\ldots,z_{u,i_M} \sim Dir\left(\alpha+\sum_{i,r_{u,i}\neq\bullet} I(z_{u,i}=1),\ldots,\alpha+\sum_{i,r_{u,i}\neq\bullet} I(z_{u,i}=K)\right) \tag{2.13}$$

可以清晰地看到公式(2.13)中的 Dir 共有 K 个参数，而 $\sum_{i,r_{u,i}\neq\bullet} I(z_{u,i}=K)$ 表示用户 u 对所有有评分的物品 $(i, r_{u,i}\neq\bullet)$ 的评分过程中，该用户被归为第 K 组的次数（I 是示性函数，当 $z_u, i=K$ 时返回 1）。回忆起 $\vec{\phi}_u$ 是用户 u 属于各个组 $(\phi_{u,1},\ldots,\phi_{u,K})$ 的概率向量，所以这个 Dirichlet 分布的超参数得到了解释。

所以：

$$\begin{aligned}
q_{\vec{\phi}_u}(\vec{\phi}_u) &\propto \exp\{\mathbf{E}_{q_z}[\ln p(\vec{\phi}_u|z_{u,i_1},\ldots,z_{u,i_M})]\} \\
&= \exp\left\{\mathbf{E}_{q_z}\left[\ln\left(\frac{\Gamma(\sum_{k=1}^K \alpha+\sum_{i,r_{u,i}\neq\bullet} I(z_{u,i}=k))}{\prod_{k=1}^K \Gamma(\alpha+\sum_{i,r_{u,i}\neq\bullet} I(z_{u,i}=k))} \times \prod_{k=1}^K \phi_{u,k}^{\alpha+\sum_{i,r_{u,i}\neq\bullet} I(z_{u,i}=k)-1}\right)\right]\right\} \\
&\propto \exp\left\{\mathbf{E}_{q_z}\left[\ln \prod_{k=1}^K \phi_{u,k}^{\alpha+\sum_{i,r_{u,i}\neq\bullet} I(z_{u,i}=k)-1}\right]\right\} \\
&= \exp\left\{\mathbf{E}_{q_z}\left[\sum_{k=1}^K (\alpha+\sum_{i,r_{u,i}\neq\bullet} I(z_{u,i}=k)-1)\ln\phi_{u,k}\right]\right\} \\
&= \exp\left\{\sum_{k=1}^K\left(\alpha+\sum_{i,r_{u,i}\neq\bullet} \underbrace{\mathbf{E}_{q_z}[I(z_{u,i}=k)]}_{\boxed{\begin{array}{l} q_{z_{u,i}}(z_{u,i}) \sim Mult(\lambda_{u,i,1},\ldots,\lambda_{u,i,K}) \\ \lambda_{u,i,k}=q_z(z_{u,i}=k),\text{so only w.r.t } z_{u,i}\end{array}}} -1\right)\ln\phi_{u,k}\right\} \\
&= \exp\left\{\sum_{k=1}^K\left(\alpha+\sum_{i,r_{u,i}\neq\bullet} q_z(z_{u,i}=k)-1\right)\ln\phi_{u,k}\right\}
\end{aligned}$$

走到了这一步，回忆起 $q_{\vec{\phi}_u}(\vec{\phi}_u) \sim Dir(\gamma_{u,1},\ldots,\gamma_{u,K})$，所以根据 Dirichlet 分布的概率密度公式:$q_{\vec{\phi}_u}(\vec{\phi}_u) \propto \prod_{i=1}^K \phi_{u,k}^{\gamma_{u,k}-1}$，另外，公式(2.8)已

经定义出 $\lambda_{u,i,k} = q_z(z_{u,i} = k)$，所以：

$$
\begin{aligned}
q_{\vec{\phi}_u}(\vec{\phi}_u) &\propto \exp\left\{\sum_{k=1}^{K}\left(\underbrace{\alpha + \sum_{i,r_{u,i}\neq\bullet}\lambda_{u,i,k}}_{\gamma_{u,k}} - 1\right)\ln\phi_{u,k}\right\} \\
&= \exp\left\{\sum_{k=1}^{K}(\gamma_{u,k} - 1)\ln\phi_{u,k}\right\} \\
&= \prod_{k=1}^{K}(\phi_{u,k})^{\gamma_{u,k}-1}
\end{aligned}
$$

所以由此发现只需令：

$$
\gamma_{u,k} = \alpha + \sum_{i,r_{u,i}\neq\bullet}\lambda_{u,i,k} \tag{2.14}
$$

就可以得到 $q_{\vec{\phi}_u}(\vec{\phi}_u) \propto \prod_{i=1}^{K}\phi_{u,k}^{\gamma_{u,k}-1}$，这里的 $q_{\vec{\phi}_u}$ 代表在已知评分矩阵 (rating matrix) 的情况下，$\vec{\phi}_u$ 的条件概率的近似概率 (approximated conditional probability)。

这样我们就得到了参数 $\gamma_{u,k}$ 的更新公式(2.14)，这个公式在程序中用来更新 $\gamma_{u,k}$。

2.3.2 derivation of $q_\kappa(\kappa_{i,k})$

再来看 $q_\kappa(\kappa_{i,k})$，其中 $\kappa_{i,k}$ 是第 k 组的用户喜欢第 i 号物品的概率:

$$
\begin{aligned}
q_\kappa(\kappa_{i,k}) &\propto \exp\{\mathbf{E}_{q_{\vec{\phi}},q_z}[\ln p(\vec{\phi},\kappa_{i,k},z,\rho)]\} \\
&= \exp\{\mathbf{E}_{q_{\vec{\phi}},q_z}[\ln\{p(\kappa_{i,k}|\vec{\phi},z,\rho)\cdot p(\vec{\phi},z,\rho)\}]\} \\
&= \exp\{\mathbf{E}_{q_{\vec{\phi}},q_z}[\ln\{p(\kappa_{i,k}|\vec{\phi},z,\rho) + \underbrace{\mathbf{E}_{q_{\vec{\phi}},q_z}p(\vec{\phi},z,\rho)\}]}_{q_\kappa(\kappa_{i,k})\text{only w.r.t. }\kappa_{i,k}}\} \\
&\propto \exp\{\mathbf{E}_{q_{\vec{\phi}},q_z}[\ln\{p(\kappa_{i,k}|\vec{\phi},z,\rho)]\} + constant \qquad (2.15) \\
&\propto \exp\{\mathbf{E}_{q_z}[\ln p(\kappa_{i,k}|z_{u_1,i},\rho_{u_1,i},\ldots,z_{u_N,i},\rho_{u_N,i})]\} \qquad (2.16)
\end{aligned}
$$

上文最后一步公式(2.15)到公式(2.17)的中 $\mathbf{E}_{q_{\vec{\phi}},q_z}$ 变为 $\mathbf{E}_{q_z}$ 是由于 $q_{\vec{\phi}}$ 与 $\kappa_{i,k}$ 没有关系，而 q_z 与 $\kappa_{i,k}$ 有关系，这一点从概率图模型以及贝叶斯

共轭的性质也可以看得出来。同样：

$$\begin{aligned}\kappa_{i,k}|z_{u_1,i},\rho_{u_1,i},\ldots,z_{u_N,i},\rho_{u_N,i} \sim Beta\Bigg(&\beta+\sum_{u,r_{u,i}\neq\bullet}\rho_{u,i}\cdot I(z_{u,i}=k)\\ &,\beta+\sum_{u,r_{u,i}\neq\bullet}(R-\rho_{u,i})\cdot I(z_{u,i}=k)\Bigg)\\ &=Beta\left(\beta+\sum_{u,r_{u,i}\neq\bullet}r^{+}_{u,i}\cdot I(z_{u,i}=k),\beta+\sum_{u,r_{u,i}\neq\bullet}r^{-}_{u,i}\cdot I(z_{u,i}=k)\right)\end{aligned}$$

上文最后一步使用了公式(2.10)的 $r^{-}_{u,i}=R-\rho_{u,i}=R\cdot(1-r^{*}_{u,i})$ 定义。另外特别注意到上文红笔的 k，而 $\sum_{u,r_{u,i}\neq\bullet}\rho_{u,i}\cdot I(z_{u,i}=k)$ 表示所有用户 u$(u,r_{u,i}\neq\bullet)$ 对特定物品 i 的评分过程中，所有用户中被归为第 k 组用户的人去进行评分的分数之和（I 是示性函数，当 $z_u,i=k$ 时返回 1），回忆起 $\kappa_{i,k}$ 的定义：表示第 k 组中的用户喜欢物品 i 的概率。所以这一贝叶斯共轭的得到了直观的解释。

接着来分析 $q_\kappa(\kappa_{i,k})$:

$$\begin{aligned}q_\kappa(\kappa_{i,k}) &\propto \exp\{\mathbf{E}_{q_z}[\ln p(\kappa_{i,k}|z_{u_1,i},\rho_{u_1,i},\ldots,z_{u_N,i},\rho_{u_N,i})]\}\\
&=\exp\left\{\mathbf{E}_{q_z}\left[\ln\frac{\kappa_{i,k}^{\beta+\sum_{u,r_{u,i}\neq\bullet}r^{+}_{u,i}\cdot I(z_{u,i}=k)-1}(1-\kappa_{i,k})^{\beta+\sum_{u,r_{u,i}\neq\bullet}r^{-}_{u,i}\cdot I(z_{u,i}=k)-1}}{B(\beta+\sum_{u,r_{u,i}\neq\bullet}r^{+}_{u,i}\cdot I(z_{u,i}=k),\beta+\sum_{u,r_{u,i}\neq\bullet}r^{-}_{u,i}\cdot I(z_{u,i}=k))}\right]\right\}\\
&\propto\exp\Bigg\{\mathbf{E}_{q_z}\Bigg[\left(\beta+\sum_{u,r_{u,i}\neq\bullet}r^{+}_{u,i}\cdot I(z_{u,i}=k)-1\right)\ln(\kappa_{i,k})\\
&\quad+\left(\beta+\sum_{u,r_{u,i}\neq\bullet}r^{-}_{u,i}\cdot I(z_{u,i}=k)-1\right)\ln(1-\kappa_{i,k})\Bigg]\Bigg\}\\
&=\exp\Bigg\{\left(\beta+\sum_{u,r_{u,i}\neq\bullet}r^{+}_{u,i}\cdot\underbrace{\mathbf{E}_{q_z}[I(z_{u,i}=k)]}_{q_z(z_{u,i}=k)}-1\right)\ln(\kappa_{i,k})\\
&\quad+\left(\beta+\sum_{u,r_{u,i}\neq\bullet}r^{-}_{u,i}\cdot\mathbf{E}_{q_z}[I(z_{u,i}=k)]-1\right)\ln(1-\kappa_{i,k})\Bigg\}\\
&=\exp\Bigg\{\left(\beta+\sum_{u,r_{u,i}\neq\bullet}r^{+}_{u,i}\cdot q_z(z_{u,i}=k)-1\right)\ln(\kappa_{i,k})\\
&\quad+\left(\beta+\sum_{u,r_{u,i}\neq\bullet}r^{-}_{u,i}\cdot q_z(z_{u,i}=k)-1\right)\ln(1-\kappa_{i,k})\Bigg\}\end{aligned}$$

走到了这一步，回忆起 $q_{\kappa_{i,k}}(\kappa_{i,k}) \sim Beta(\epsilon^+_{i,k}, \epsilon^-_{i,k})$，所以根据 Beta 分布的概率密度公式:$q_{\kappa_{i,k}}(\kappa_{i,k}) \propto (\kappa_{i,k})^{\epsilon^+_{i,k}-1}(1-\kappa_{i,k})^{\epsilon^-_{i,k}-1}$，另外，公式(2.8)已经定义出 $\lambda_{u,i,k} = q_z(z_{u,i} = k)$，所以：

$$
\begin{aligned}
q_{\kappa_{i,k}}(\kappa_{i,k}) \propto \exp\Bigg\{ & \Bigg(\underbrace{\beta + \sum_{u, r_{u,i} \neq \bullet} r^+_{u,i} \cdot \lambda_{u,i,k} - 1}_{\epsilon^+_{i,k}}\Bigg) \ln(\kappa_{i,k}) \\
& + \Bigg(\underbrace{\beta + \sum_{u, r_{u,i} \neq \bullet} r^-_{u,i} \cdot \lambda_{u,i,k} - 1}_{\epsilon^-_{i,k}}\Bigg) \ln(1-\kappa_{i,k}) \Bigg\} \\
= & \exp\{(\epsilon^+_{i,k} - 1)\ln(\kappa_{i,k}) + (\epsilon^-_{i,k} - 1)\ln(1-\kappa_{i,k})\} \\
= & (\kappa_{i,k})^{\epsilon^+_{i,k}-1}(1-\kappa_{i,k})^{\epsilon^-_{i,k}-1}
\end{aligned}
$$

所以由此发现只需令：

$$
\epsilon^+_{i,k} = \beta + \sum_{u, r_{u,i} \neq \bullet} r^+_{u,i} \cdot \lambda_{u,i,k} \tag{2.17}
$$

$$
\epsilon^-_{i,k} = \beta + \sum_{u, r_{u,i} \neq \bullet} r^-_{u,i} \cdot \lambda_{u,i,k} \tag{2.18}
$$

就可以得到 $q_{\kappa_{i,k}}(\kappa_{i,k}) \propto (\kappa_{i,k})^{\epsilon^+_{i,k}-1}(1-\kappa_{i,k})^{\epsilon^-_{i,k}-1}$，这里的 $q_{\kappa_{i,k}}$ 代表在已知评分矩阵 (rating matrix) 的情况下，$\kappa_{i,k}$ 的条件概率的近似概率 (approximated conditional probability)。

2.3.3 derivation of $q_{z_{u,i}}$

因为 z 是服从多项分布 (multinomial distribution)，现在来推导 $q_{z_{u,i}}$:

$$
\begin{aligned}
\lambda_{u,i,k} = q_z(z_{u,i} = k) & \propto \exp\{\mathbf{E}_{q_{\vec{\phi}}, q_\kappa}[\ln p(\vec{\phi}, \kappa, z, \rho)]\} \\
& \propto \exp\{\mathbf{E}_{q_{\vec{\phi}}, q_\kappa}[\ln p(z_{u,i} = k|\vec{\phi}_u) \cdot p(\rho_{u,i}|\kappa_{i,k})]\} \\
& \quad \text{上一步与 LDA 模型类似，此步可以从概率图模型中得到解释} \\
& \propto \exp\{\mathbf{E}_{q_{\vec{\phi}}, q_\kappa}[\ln p(z_{u,i} = k|\vec{\phi}_u) + \ln p(\rho_{u,i}|\kappa_{i,k})]\} \\
& \propto \exp\{\mathbf{E}_{q_{\vec{\phi}}}[\ln p(z_{u,i} = k|\vec{\phi}_u)] + \mathbf{E}_{q_\kappa}[\ln p(\rho_{u,i}|\kappa_{i,k})]\}
\end{aligned}
$$

因为 $z_{u,i}$ 服从参数为 $\vec{\phi}_u$ 的多项分布，我们可以得到：

$$p(z_{u,i}|\vec{\phi}_u) = \phi_{u,k}$$

因为 $\rho_{u,i}$ 服从参数为 R 和 $\kappa_{i,k}$ 的二项分布 (Binomial distribution): $\rho_{u,i} \sim Bin(R, \kappa_{i,z_{u,i}})$，所以我们可以得到：

$$p(\rho_{u,i}|\kappa_{i,k}) \propto (\kappa_{i,k})^{\rho_{u,i}}(1-\kappa_{i,k})^{R-\rho_{u,i}}$$

因此:

$$\begin{aligned}
\lambda_{u,i,k} &\propto \exp\{\mathbf{E}_{q_{\vec{\phi}}}[\ln\phi_{u,k}] + \mathbf{E}_{q_\kappa}[\ln p(\rho_{u,i}|\kappa_{i,k})]\} \\
&\propto \exp\{\mathbf{E}_{q_{\vec{\phi}}}[\ln\phi_{u,k}] + \rho_{u,i}\mathbf{E}_{q_\kappa}[\ln\kappa_{i,k}] + (R-\rho_{u,i})\mathbf{E}_{q_\kappa}[\ln(1-\kappa_{i,k})]\} \\
&= \exp\{\mathbf{E}_{q_{\vec{\phi}}}[\ln\phi_{u,k}] + \underbrace{r^+_{u,i}}_{\rho_{u,i}}\mathbf{E}_{q_\kappa}[\ln\kappa_{i,k}] + \underbrace{r^-_{u,i}}_{R-\rho_{u,i}}\mathbf{E}_{q_\kappa}[\ln(1-\kappa_{i,k})]\}
\end{aligned}$$

因为 $q_{\vec{\phi}_u}(\vec{\phi}_u) \sim Dir(\gamma_{u,1},\ldots,\gamma_{u,K})$，$q_{\kappa_{i,k}}(\kappa_{i,k}) \sim Beta(\epsilon^+_{i,k},\epsilon^-_{i,k})$，而 Beta 分布是 2 个超参数情况下特殊的 Dirichlet 分布，由于根据(1.7)公式：

$$\psi(x) = (\ln(\Gamma(x))' = \frac{\Gamma'(x)}{\Gamma(x)}$$

可以得到:

$$\mathbf{E}_{p(\theta|\alpha)}(\ln(\theta_i)) = \frac{\mathrm{d}}{\mathrm{d}\alpha_i}\left(\sum_{i=1}^{k}\ln(\Gamma(\alpha_i)) - \ln(\Gamma(\sum_{i=1}^{k}\alpha_i))\right) = \psi(\alpha_i) - \psi(\sum_{i=1}^{k}\alpha_i)$$

所以：

$$\mathbf{E}_{q_{\vec{\phi}}}[\ln\phi_{u,k}] = \psi(\gamma_{u,k}) - \psi(\sum_{k=1}^{K}\gamma_{u,k}) \tag{2.19}$$

$$\mathbf{E}_{q_\kappa}[\ln\kappa_{i,k}] = \psi(\epsilon^+_{i,k}) - \psi(\epsilon^+_{i,k}+\epsilon^-_{i,k}) \tag{2.20}$$

$$\mathbf{E}_{q_\kappa}[\ln(1-\kappa_{i,k})] = \psi(\epsilon^-_{i,k}) - \psi(\epsilon^+_{i,k}+\epsilon^-_{i,k}) \tag{2.21}$$

将 $\psi(\sum_{k=1}^{K}\gamma_{u,k})$ 视为常数，所以，最终得到 $\lambda_{u,i,k}$ 的更新公式:

$$\lambda_{u,i,k} \propto \exp(\psi(\gamma_{u,k}) + r^+_{u,i}\cdot\psi(\epsilon^+_{i,k}) + r^-_{u,i}\cdot\psi(\epsilon^-_{i,k}) - \underbrace{R}_{r^+_{u,i}+r^-_{u,i}}\cdot\psi(\epsilon^+_{i,k}+\epsilon^-_{i,k}) \tag{2.22}$$

这样便得到了所有的更新公式。

3 summary

这个论文看了我一整个周末，给了我一个变分贝叶斯的例子，其中有些数学技巧值得学习，这也是论文最大的收获，所以说 Latent Dirichlet Allocation 只是其中的一个推导方法，而不是全部故事。