流行度偏差的变分推断

1. 变分推断的目标

考虑 user–item 交互集合 $\mathcal{D}=\{u_k,i_k,z_k,c_k\}_{k=1}^N$ ，其中 $c_k\in\{0,1\}$ 是交互标签， $u_k,i_k\in\mathbb{R}^p$ 分别是 user 和 item 的特征向量，由因子分解机得到， $z_k\in(0,+\infty)$ 是不可观测的隐变量，代表对应 item 的流行度，我们的目标得到估计概率 $\mathbb{P}(c_k=1|u_k,i_k)$

但在流行度曝光影响下，数据是非随机缺失的，为了去除流行度偏差，需要使用干预公式得到：

$$\mathbb{P}(c_k=1|u_k,i_k)=\int_z\mathbb{P}(c_k=1|u_k,i_k,z_k)\cdot\mathbb{P}(z_k)dz$$

在可观测数据上，我们只能拟合后验 $p_k=\mathbb{P}(c_k=1|u_k,i_k,z_k)$ ，极大化下面的对数似然函数：

$$\ell=\sum_{k=1}^N[c_k\cdot\log p_k+(1-c_k)\cdot\log(1-p_k)]$$

假设隐变量 $z\sim f_z(\Theta_z)$ ，隐变量 z 无法观测，计划使用变分推断估计 f_z （或者说获得关于流行度的后验 $p(z_k|u_k,i_k,c_k)$ ），同时极大化上面的似然函数 ℓ ，以得到推荐系统排序模型

我们希望通过数据获得关于流行度 z 分布的描述 fz ，我们用一个参数化的后验 $q(z;\boldsymbol{\phi})$ 来近似 f_z ，从 $p(c_k|u_k,i_k)$ 出发引入变分分布 $q(z;\boldsymbol{\phi})$ 改写 $\log\mathbb{P}(c_k|u_k,i_k)$ ：

$$\log\mathbb{P}(c_k|u_k,i_k)=\log\mathbb{E}_q\left[\frac{\mathbb{P}(c_k|u_k,i_k,z_k)\cdot\mathbb{P}(z_k)}{q(z_k;\boldsymbol{\phi})}\right]$$

由对数函数的凹性，根据 Jensen 不等式，有：

$$\log\mathbb{P}(c_k|u_k,i_k)=\mathbb{E}_q\left[\log\frac{\mathbb{P}(c_k|u_k,i_k,z_k)\cdot\mathbb{P}(z_k)}{q(z_k;\boldsymbol{\phi})}\right]$$

展开得到：

$$\log \mathbb{P}(c_k|u_k,i_k) = \mathbb{E}_q\left[\log \mathbb{P}(c_k|u_k,i_k,z_k)\right] - \mathrm{KL}(q(z_k;\boldsymbol{\phi})||f_z(z_k))$$

左边 $\log \mathbb{P}(c_k|u_k,i_k)$ 是与 Z 无关的常数，即为 ELBO

以上就是变分推断优化目标，优化目标分为两部分：

- 第一项：重构误差 $\mathbb{E}_q\left[\log \mathbb{P}(c_k|u_k,i_k,z_k)\right]$

  该目标是给定 u_k,i_k,z_k 的条件下估计 c_k ，即最大化对数似然：

  $$\ell = \mathbb{E}_q\left[\log \mathbb{P}(c_k|u_k,i_k,z_k)\right] = \sum_{k=1}^{L}\left[c_k \cdot \log p_k + (1-c_k)\cdot \log(1-p_k)\right]$$

  这里需要考虑的问题包括：

  - 如何建模概率 $p_k = \mathbb{P}(c_k|u_k,i_k,z_k)$ ？

    PDA 方案下的 $p_k = ELU(u_k^T i_k)\cdot z_k^{\gamma}$ 缺少概率含义，但优点是 z_k 被分离，后续推理时对 z 的积分计算与 u_k,i_k 无关

    例如考虑如下的概率形式：

    $$\mathbb{P}(c_k=1|u_k,i_k,,z_k) = \frac{1}{1+z_k^{-1}\exp(-u_k^T i_k+\beta_0)} \in (0,1)$$

    $$\mathbb{P}(c_k=1|u_k,i_k,,z_k) = \frac{1}{1+\exp(-u_k^T i_k+\gamma\cdot z_k+\beta_0)} \in (0,1)$$

    以上形式都无法将 z_k 分离到 $u_k^T i_k$ 之外，推理时计算 $\mathbb{P}(c_k=1|u_k,i_k)$ 时需要处理积分：

    $$\mathbb{P}(c_k=1|u_k,i_k) = \int_z \mathbb{P}(c_k=1|u_k,i_k,z_k)\cdot\mathbb{P}(z_k)dz$$

  - z_k 的值如何确定
    - 从近似后验 $q(z;\boldsymbol{\phi})$ 中进行抽样，然后用重参数化技巧可以维持优化梯度
    - 或者是直接采样，然后使用 Monte Carlo 估计
- 第二项：先验惩罚 $-\mathrm{KL}(q(z_k;\boldsymbol{\phi})||f_z(z_k))$

该目标是让后验 $q(z;\boldsymbol{\phi})$ 尽可能靠近先验 f_z ，视为优化目标的正则化项，为先验 f_z 和后验近似 $q(z)$ 选择不同的分布，计算散度 $\mathrm{KL}(q(z_k;\boldsymbol{\phi})||f_z(z_k))$ 的计算和积分

$$\int_z \mathbb{P}(c_k=1|u_k,i_k,z_k)\cdot\mathbb{P}(z_k)dz$$

的难度不同，选择合适的分布尽可能导出闭式解能减少 Monte Carlo 等数值方法的使用，提高估计的稳定性

2. 流行度分布的选择

2.1. 对数正态分布下的散度

假设取后验近似 $q(z_k;\phi)$ 为对数正态 $\text{LogNormal}(\mu_k,\sigma_k^2)$ ，变分参数为 $\phi_k=\{\mu_k,\sigma_k^2\}$ ，并且假设先验 $f_z(z_k)$ 也服从对数正态分布 $f_z(z_k)=\text{LogNormal}(\tilde{\mu}_k,\tilde{\sigma}_k^2)$ ，其中先验参数 $\tilde{\mu},\tilde{\sigma}^2$ 可以从用观测数据的交互占比的统计值估计，例如：

$$\tilde{\mu}_k=\frac{\sum_{s=1}^{N}\mathbb{I}(i_s=k)}{N}=\frac{D_k}{\sum_{i\in\mathcal{I}}D_i},\quad \tilde{\sigma}^2\equiv\sigma_0$$

下面，计算散度 $\text{KL}(q(z_k;\phi)||f_z(z_k))$ ：

$$\text{KL}(q(z_k;\phi)||f_z(z_k))=\int q(z;\mu_k,\sigma_k^2)\cdot\log\frac{q(z;\mu_k,\sigma_k^2)}{f(z;\tilde{\mu}_k,\tilde{\sigma}^2)}dz$$

对于对数正态分布 $p(z;\mu,\sigma^2)$ ，有概率密度函数：

$$p(z;\mu,\sigma)=\frac{1}{z\sigma\sqrt{2\pi}}\cdot\exp\left[\frac{-(\log z-\mu)^2}{2\sigma^2}\right]$$

散度的积分中做换元 $t=\log z$ ，则 $z=e^t,dz=e^tdt$ ，则：

$$\log\frac{q(z;\mu_k,\sigma_k^2)}{f(z;\tilde{\mu}_k,\tilde{\sigma}^2)}=\log\left[\frac{\tilde{\sigma}_k}{\sigma_k}\times\exp\left(\frac{(t-\tilde{\mu}_k)^2}{2\tilde{\sigma}_k^2}-\frac{(t-\mu_k)^2}{2\sigma_k^2}\right)\right]$$

化简整理得到：

$$\log\frac{q(z;\mu_k,\sigma_k^2)}{f(z;\tilde{\mu}_k,\tilde{\sigma}^2)}=\log\frac{\tilde{\sigma}_k}{\sigma_k}+\frac{(t-\tilde{\mu}_k)^2}{2\tilde{\sigma}_k^2}-\frac{(t-\mu_k)^2}{2\sigma_k^2}$$

代入回到 KL 的表达式中，并利用：

$$\int q(z;\mu_k,\sigma_k^2)dz=1$$

$$\int q(z;\mu_k,\sigma_k^2)(t-\tilde{\mu}_k)^2dz=(\mu_k-\tilde{\mu}_k)^2+\sigma_k^2$$

$$\int q(z;\mu_k,\sigma_k^2)(t-\mu_k)^2dz=\sigma_k^2$$

得到：

$$\text{KL}(q(z_k;\phi)||f_z(z_k))=\log\frac{\tilde{\sigma}_k}{\sigma_k}+\frac{\sigma_k^2+(\mu_k-\tilde{\mu}_k)^2-\tilde{\sigma}_k^2}{2\tilde{\sigma}_k^2}$$

2.2. 伽马分布下的散度

假设后验近似 $q(z_k;\boldsymbol{\phi})$ 为伽马分布 $\text{Gamma}(\alpha_k,\beta_k)$ ，变分参数为 $\boldsymbol{\phi}_k=\{\alpha_k,\beta_k\}$ ，并且假设先验 $f_z(z_k)$ 也服从伽马分布 $f_z(z_k)=\text{Gamma}(\tilde{\alpha}_k,\tilde{\beta}_k)$ ，伽马分布 $p(z;\alpha,\beta)$ 具有密度：

$$p(z;\alpha,\beta)=\frac{\beta^{\alpha}}{\Gamma(\alpha)}z^{\alpha-1}e^{-\beta z}$$

下面展开 $\log\dfrac{q(z_k;\alpha_k,\beta_k)}{f_z(z_k;\tilde{\alpha}_k,\tilde{\beta}_k)}$

$$\log\frac{q(z_k;\alpha_k,\beta_k)}{f_z(z_k;\tilde{\alpha}_k,\tilde{\beta}_k)}=\log\frac{\beta_k^{\alpha_k}}{\Gamma(\alpha_k)}-\log\frac{\tilde{\beta}_k^{\tilde{\alpha}_k}}{\Gamma(\tilde{\alpha}_k)}+(\alpha_k-\tilde{\alpha}_k)\log z-(\beta_k-\tilde{\beta}_k)z$$

利用期望的线性组合性质：

$$\text{KL}(q(z_k;\boldsymbol{\phi})||f_z(z_k))=\log\frac{\beta_k^{\alpha_k}}{\Gamma(\alpha_k)}-\log\frac{\tilde{\beta}_k^{\tilde{\alpha}_k}}{\Gamma(\tilde{\alpha}_k)}+(\alpha_k-\tilde{\alpha}_k)\mathbb{E}_q[\log z]-(\beta_k-\tilde{\beta}_k)\mathbb{E}_q[z]$$

利用伽马分布的性质：

$$\mathbb{E}_q[z]=\frac{\alpha_k}{\beta_k},\quad \mathbb{E}_q[\log z]=\psi(\alpha_k)-\log\beta_k$$

其中， $\psi(x)$ 是双伽马函数，代入到散度，整理得到：

$$\text{KL}(q(z_k;\boldsymbol{\phi})||f_z(z_k))=(\alpha_k-\tilde{\alpha}_k)\psi(\alpha_k)+\log\frac{\Gamma(\tilde{\alpha}_k)}{\Gamma(\alpha_k)}+\tilde{\alpha}_k\log\frac{\beta_k}{\tilde{\beta}_k}+\alpha_k(\frac{\tilde{\beta}_k}{\beta_k}-1)$$

2.3. 指数分布下的散度

当 $\alpha=1$ 时，伽马分布退化为指数分布：

$$p(z;\beta)=p(z;\gamma)=\gamma\cdot e^{-\gamma z}$$

因此散度为（注意 $\gamma=\beta^{-1}$ ）：

$$\text{KL}(q(z_k;\boldsymbol{\phi})||f_z(z_k))=\log\frac{\tilde{\gamma}_k}{\gamma_k}+(\frac{\gamma_k}{\tilde{\gamma}_k}-1)$$

3. 建模概率 $\mathbb{P}(c_k|u_k,i_k,z_k)$

3.1. 分离式的乘积概率模型

PDA 方案下的 $p_k=ELU(u_k^Ti_k)\cdot z_k^{\gamma}$ 缺少概率含义，**但优点是 z_k 被分离**，即概率 $\mathbb{P}(c_k|u_k,i_k,z_k)$ 可以被分解为：

$$\mathbb{P}(c_k|u_k,i_k,z_k)=h(u_k,i_k)\cdot g(z_k)$$

此时在计算 $\mathbb{P}(c_k|u_k,i_k)$ 时，**处理积分会更加容易：**

$$\mathbb{P}(c_k|u_k,i_k)=\int_z h(u_k,i_k)\cdot g(z_k)\cdot p(z)dz=h(u_k,i_k)\cdot\int_z g(z_k)p(z)dz$$

否则，计算 $\mathbb{P}(c_k|u_k,i_k)$ 将只能使用 Monte Carlo 等数值方法处理，从优化的角度我们还是尽可能得到闭式解，因此我们建模的目标明确为：

- 尽可能使用分离式的概率模型 $\mathbb{P}(c_k|u_k,i_k,z_k)=h(u_k,i_k)\cdot g(z_k)$ ，单独处理 $g(z_k)$
- $g(z)$ 的设计应满足两点：
  - 让 $h(u_k,i_k)\cdot g(z_k)\in(0,1)$ 拥有概率含义
  - $g(z_k)>0$ 且 g 应该随着 z_k 单调递增（代表流行度增加，则交互概率上升）
- 针对我们讨论和选择的先验（或近似后验）分布 $p(z)$ ，选择合适的 $g(z)$ 的形式，让积分的计算尽可能简单，或直接有闭式解：

$$\int_z g(z_k)p(z)dz$$

假设我们用一般的 logitstic-link function 建模 $h(u_k,i_k)$ ：

$$h(u_k,i_k)=\frac{1}{1+\exp(-u_k^Ti_k+b_0)}=\sigma(u_k^Ti_k+b_0)$$

b_0 可以作为一个全局偏置常数，则 $h(u_k,i_k)\in(0,1)$ ， $g(z)$ 的选择多样，例如：

$$g(z)=\frac{z}{z+C},\quad g(z)=1-e^{-\lambda z},\quad g(z)=\sigma(-\lambda z)$$

指数生存函数 $g(z)=1-e^{-\lambda z}$ 组合 $p(z)$ 为伽马分布可以得到积分的闭式解，而 $p(z)$ 如果是对数正态分布则很难导出一个具有以上性质的闭式解的积分，其他 $g(z)$ 的函数形式的复杂性，也较难有闭式解

下面我们取 $g(z)=1-e^{-\lambda z}$ ，而取 z 的分布为伽马分布 $p(z)\sim\Gamma(z;\alpha,\beta)$ ，则目标积分为：

$$\int_0^\infty(1-e^{-\lambda z})\frac{\beta^\alpha}{\Gamma(\alpha)}z^{\alpha-1}e^{-\beta z}dz$$

积分可以被拆分为两个部分：

$$\int_0^\infty\frac{\beta^\alpha}{\Gamma(\alpha)}z^{\alpha-1}e^{-\beta z}dz-\int_0^\infty\frac{\beta^\alpha}{\Gamma(\alpha)}z^{\alpha-1}e^{-(\beta+\lambda)z}dz$$

第一个积分是对伽马分布密度函数的积分等于 1 ，第二个积分满足伽马分布的形式，速率参数改变为 $\beta+\lambda$ ，利用伽马函数的定义：

$$\int_0^\infty z^{\alpha-1}e^{-(\beta+\lambda)z}dz=\frac{\Gamma(\alpha)}{(\beta+\lambda)^\alpha}$$

代入得到：

$$\int_z g(z_k)p(z)dz = \int_0^{\infty}(1-e^{-\lambda z})\frac{\beta^{\alpha}}{\Gamma(\alpha)}z^{\alpha-1}e^{-\beta z}dz = 1-\left(\frac{\beta}{\beta+\lambda}\right)^{\alpha}$$

特别地，当先验（近似后验）选择为指数分布时，即 $\alpha=1$ ，积分具有一个更简洁的表达：

$$\int_z g(z_k)p(z)dz = 1-\left(\frac{\beta}{\beta+\lambda}\right) = \frac{\lambda}{\beta+\lambda}$$

3.2. 乘积概率模型的问题

在实践中，**发现乘积概率模型虽然不影响性能，但会丧失统计模型的估计性质**：

现在我们假设统计模拟实验，我们所考虑的概率模型被替换为：

$$\mathbb{P}(c_k=1|u_k,i_k,z_k) = \sigma(u_k^T\beta_u + i_k^T\beta_i + \beta_0)\cdot g(z_k;\boldsymbol{\phi}_k)$$

- 伯努利分布下的优化过程是在线性 logit-odd 空间中进行的，因此当使用交叉熵损失（即最大化伯努利分布的对数似然函数时），理论上我们需要得到 $\sigma(u_k^T\beta_u + i_k^T\beta_i + \beta_0)$ 中 β_u,β_i,β_0 的无偏估计
- 但是现在我们为了可分离性，在概率层上加入了一个乘性因子 $g(z_k;\boldsymbol{\phi}_k)$ ，这**带来了模型的“不可识别”问题，即对于同一目标的多解问题**：
  - 例如对于某个 $c_k=1$ 的样本，优化方向是增大 $\mathbb{P}(c_k=1|u_k,i_k,z_k)$ ，但在乘积空间中，这可以通过调整 β_u,β_i,β_0 或仅调整 $\boldsymbol{\phi}_k$ 实现，在梯度下降过程中，模型只需要任选一个方向、或者同时分配到两个方向实现即可，导致收敛时得到的参数估计不是对真实概率生成模型参数的无偏估计
- 乘积问题带来的不可识别问题只会影响模型统计性质（无法得到生成模型参数的准确刻画），但不会影响模型的性能

但考虑到我们还是**保证模型参数估计的性质，我们还是只能使用加性的经典 logit-odd 模型形式**，把流行度放入到 σ 函数之内作为一个影响交互概率的偏置项：

$$\mathbb{P}(c_k=1|u_k,i_k,,z_k) = \sigma(u_k^T\beta_u + i_k^T\beta_i + \gamma\cdot z_k + \beta_0)\in(0,1)$$

采用这个加性模型后，带来的优缺点分别是：

- **带来的优点**
  - 可以获得生成模型参数的无偏估计了
  - 可以进一步放宽 z_k 分布的限制，现在不要求 $z_k>0$ 了，例如可以取 $z_k\sim N(\mu_k,\sigma_k^2)$ ，因为整个交互概率依然保持了和 z_k 的单调性

- **带来的缺点**
  - 在处理积分边际分布的积分时需要使用数值积分工具，带来额外的计算开销：

$$\mathbb{P}(c_k = 1|u_k, i_k) = \int_z \mathbb{P}(c_k = 1|u_k, i_k, z_k) \cdot \mathbb{P}(z_k) dz$$

但一些特殊的分布可以用近似公式计算求解，减少复杂度，例如当假设 $z_k \sim N(\mu_k, \sigma_k^2)$ 时，有如下的 Logistic–Normal 积分近似：

$$\mathbb{P}(c_k = 1|u_k, i_k) = \sigma\left(\frac{u_k^T \beta_u + i_k^T \beta_i + \gamma \cdot \mu_k + \beta_0}{\sqrt{1 + \gamma^2 \sigma_k^2}}\right)$$

  - 如何从收集到的交互记录中，确定一个关于流行度 z_k 的先验分布的参数