Author：Liu Jian

Time：2020-06-27

机器学习11a-概率图模型理论

1 概率图模型综述

概率图模型就是用图对随机向量 Y 的联合概率 $P(Y)=P(Y_1,\cdots,Y_n)$ 进行建模，注意， Y_i 表示 Y 的第 i 个随机分量 (当然， Y_i 也有可能是一个随机向量)，而不是第 i 个样本，有时也直接称之为随机变量或属性。一般地，我们会对高维随机变量各分量间的相关关系进行合理的简化，而图模型能简洁紧凑地表示这种相关关系，其中，结点表示随机变量，边表示随机变量间的相关关系。概率图模型可以分为：

1. 使用有向无环图 (Directed Acyclic Graph) 表示变量间的**依赖关系**，称为概率有向图模型/贝叶斯网络 (Bayesian network)/信念网 (belief network)。贝叶斯网络又分为：
    - 静态贝叶斯网络
    - 动态贝叶斯网络：马尔可夫链，隐马尔科夫模型；
2. 使用无向图表示变量间的**相关关系**，称为概率无向图模型 (probabilistic undirected graphical model)/马尔可夫网络 (Markov network)，典型的有马尔可夫随机场 (Markov random field)

若变量间存在显示的因果关系，则常使用贝叶斯网络；若变量间存在相关性，但难以获得显示的因果关系，则常使用马尔可夫网络。

两个重要的任务是：(1) 对模型进行**学习 (learning)**，包括网络结构的学习和参数的学习；(2) 基于学得的模型进行**推断 (inference)**，比如计算某些概率，期望等。对于学习问题，网络结构可以采用评分函数进行评估，比如 AIC、BIC、KIC 等，参数估计则可采用最大似然估计和最大后验估计，而遇到存在隐变量的情况时采用 EM 算法进行优化，不再详细展开，**这里我们重点关注推断问题。**

2 概率有向图模型/贝叶斯网络 (Bayesian network)/信念网 (belief network)

概率有向图的有向边直接给出了属性间的**依赖关系**，记随机变量 Y_i 的父结点集为 π_i ，则

$$P(Y)=P(Y_1,\cdots,Y_n)=\prod_{i=1}^{n}P(Y_i|\pi_i)$$

上式就是贝叶斯网络的因子分解，**因子分解的因子为条件概率。**

除了依赖关系，从概率有向图中我们也可以得到属性间的**条件独立性。**分析条件独立性的三种典型**局部**结构有：(1) head to tail (顺序结构)；(2) tail to tail (同父结构)；(3) head to head (V 型结构)，**它们对应的条件独立性结论可由链式法则和因子分解推出，不再赘述。**此外，为了在**全局**上更好地分析概率有向图中属性间的条件独立性，可以先使用"有向分离" (D-separation) 将**有向图转变为一个无向图，即所谓的道德图 (moral graph)**：

1. 找出有向图中的所有 V 型结构 (head to head)，在 V 型结构的两个父结点之间加上一条无向边 (这一过程称为道德化 (moralization))；
2. 将所有有向边改为无向边。

接着，**基于得到的道德图，我们就能直观、迅速地找到变量间的条件独立性**：对属性 Y_i, Y_j 和属性集合 Y_O，若在道德图中，属性 Y_i, Y_j 能被属性集合 Y_O 分开，即将属性集合 Y_O 在道德图中去除后，Y_i 和 Y_j 分属两个连通分支，则称属性 Y_i 和 Y_j 被 Y_O 有向分离，即有条件独立性 $Y_i \perp Y_j | Y_O$ 。

3 概率无向图模型 (probabilistic undirected graphical model)/马尔可夫网络 (Markov network)/马尔可夫随机场 (Markov random field)

概率无向图模型使用无向图 $G = (V, E)$ 对联合概率进行建模，其中 V 表示结点集合，E 表示边的集合。结点 V 对应随机变量，边 E 则给出了随机变量间的相关和独立关系，体现为如下两点：

1. 相关关系：两个结点之间有边相连，则对应的随机变量存在概率相关关系；
2. 条件独立关系：即满足成对马尔可夫性 (pairwise Markov property)/局部马尔可夫性 (local Markov property)/全局马尔可夫性 (global Markov property)，这三者是等价的，我们分别描述如下：

(a) 成对马尔可夫性 (pairwise Markov property)：结点 u, v 没有边连接，其他所有结点记为 O，则给定随机变量集合 Y_O 的条件下，Y_u, Y_v 相互独立：

$$P(Y_u, Y_v | Y_O) = P(Y_u | Y_O) P(Y_v | Y_O)$$

(b) 局部马尔可夫性 (local Markov property)：结点 v，结点集合 W 为与 v 有边连接的所有结点，结点集合 $O = V - W - v$ 是除 v, W 的所有其他结点，则给定 Y_W 的条件下，Y_v, Y_O 相互独立：

$$P(Y_v, Y_O | Y_W) = P(Y_v | Y_W) P(Y_O | Y_W)$$

(c) 全局马尔可夫性 (global Markov property)：结点集合 A, B, C，其中，结点集合 A, B 被结点集合 C 分开，则给定 Y_C 的条件下，Y_A, Y_B 相互独立：

$$P(Y_A, Y_B | Y_C) = P(Y_A | Y_C) P(Y_B | Y_C)$$

团与最大团：无向图中任何两个结点均有边连接的结点子集称为团 (clique)；若 C 是无向图 G 的一个团，并且不能再加进任何一个 G 的结点使其成为一个更大的团，则称 C 为最大团 (maximal clique)。

根据 Hammersley-Clifford 定理，概率无向图模型的联合概率分布可以表示成**因子分解 (factorization) 的形式**：

$$P(Y) = \frac{\prod_{C \in \mathcal{C}} \Psi_C(Y_C)}{Z}$$

其中，$\mathcal{C}$ 表示无向图 G 中所有的最大团集合； $\Psi_C(Y_C)$ 为势函数 (potential function)，严格为正，通常取：

$$\Psi_C(Y_C) = \exp\left(-\mathscr{E}(Y_C)\right)$$

其中，$\mathscr{E}(\cdot)$ 为能量函数，可以看到能量越大，概率越低。能量函数的常见形式为：

$$\mathscr{E}(Y_C) = \sum_{i,j\in C, i\neq j} \alpha_{ij} Y_i Y_j + \sum_{k\in C} \beta_k Y_k$$

其中，α_{ij}, β_k 是参数，上式中的第一项考虑每一对结点的关系，第二项则仅考虑单结点。 Z 为规范化因子：

$$Z = \int \prod_{C\in\mathcal{C}} \Psi_C(Y_C) \mathrm{d}Y$$

概率有向图模型和概率无向图模型均通过因子分解对概率分布进行表示，只不过前者的因子为条件概率，后者的因子为势函数。

4 推断问题

通过学习得到联合概率模型 $P(Y)$ 后，接下来的问题就是利用概率模型进行推断。记 $Y_Q \cup Y_E = Y$, $Y_Q \cap Y_E = \emptyset$ ，推断问题举例：

1. 计算条件分布 $P(Y_Q|Y_E)$ ；
2. 计算边际分布 $P(Y_Q) = \int P(Y_Q, Y_E)\mathrm{d}Y_E$ ；
3. 计算期望 $\mathbb{E}_Y(f(Y)) = \int f(\boldsymbol{y})P(\boldsymbol{y})\mathrm{d}\boldsymbol{y}$ ；

等等。注意到：

$$P(Y_Q|Y_E) = \frac{P(Y_Q, Y_E)}{P(Y_E)}$$

而 $P(Y_Q, Y_E)$ 已知， $P(Y_E) = \int P(Y_Q, Y_E)\mathrm{d}Y_Q$ ，因此和问题 2 一样，问题 1 的关键也在于边际分布的计算。可以看到，**推断问题的关键在于求积分。**

概率图模型的推断方法可分为两类，即**精确推断**和**近似推断。**当网络结点较多，连接稠密时，难以进行精确推断，此时就需要借助近似推断，通过降低精度要求，在有限时间内求得近似解。

4.1 精确推断

动态规划的思想在精确推断中很重要，能大幅提高计算效率，比如 HMM 中计算概率时的前向算法、后向算法，针对解码问题的维特比算法；信念传播算法等。

4.1.1 变量消去法

变量消去法：无论是概率有向图模型还是概率无向图模型，它们的因子分解都是将概率分布展开为因子的累积形式，而对概率的积分实际上就是做加法。变量消去法就是根据乘法对加法的分配率，把多变量积的求和 (也就是积分) 问题，转化为对部分变量交替进行求积与求和 (积分) 的问题，这种转化使得每次的求和与求积运算限制在局部，仅与部分变量有关，从而简化了计算。

说得那么玄乎，实际上就是求解多维积分时对积分变量依次进行积分，每积一次就消除一个变量，而每一个变量只出现在有限几个因子中，因此，积分得以在局部进行。

4.1.2 信念传播算法 (Belief Propagation，也称 Sum-Product 算法)

在需要计算多个边际分布时，重复使用变量消去法会造成大量的冗余计算，即有些部分会被计算多次，为此，信念传播算法将变量消去算法中的求和 (积分) 操作看作一个消息传递过程，较好地解决了这一问题。

从结点 i 传播消息到结点 j 的过程的递推公式如下：

$$m_{i\to j}(y_j)=\int \psi(y_i,y_j)\prod_{k\in n(i)\backslash j} m_{k\to i}(y_i)\mathrm{d}y_i$$

其中，$n(i)$ 表示结点 i 的邻接结点，$n(i)\backslash j$ 则表示除了 j 的邻接点。可以看到：

1. 上述递推式或者说消息传递的概念实际上是从求解多维积分的过程中抽象而来，是对积分步骤的一种分解。
2. $\psi(y_i,y_j)$ 是概率有向图或概率无向图的因子分解中与 y_i,y_j 相关的因子。不要认为条件概率中的条件部分是常量无法积分，条件部分也是变量，也可以被积分。因此，无论是概率有向图模型还是概率无向图模型，我们可以统一用 $\psi(y_i,y_j)$ 表示其因子。和无向图一样，对于有向图而言，消息传递方向与边的方向无关，并不是只能朝有向边的方向传递，反向也是可以的进行的，因此，对于消息传递，我们可以将有向边看作是无向边。
3. 消息传递 $m_{i\to j}(y_j)$ 就是对 y_i 进行积分以消除变量 y_i，所得结果只剩下变量 y_j，而由于因子分解的关系，其他部分压根就不存在 y_i，因此，局部的积分 (也就是消息传递) 将使得 y_i 在整个概率分布中消失。
4. $m_{i\to j}(y_j)$ 与 $m_{k\to i}(y_i)$ 的递推关系表明，结点 i 仅在接收到其他所有结点 k 的消息后才能向结点 j 发送消息。

边缘分布的积分计算可基于消息表示如下：

$$P(y_i)\propto\prod_{j\in n(i)} m_{j\to i}(y_i)$$

再规范化即可。可以看到，问题的关键在于消息的计算，而基于前面给出的消息的递推公式，我们使用**动态规划算法**，依次计算并保存相应的结果即可，注意， 对于每条边 (i,j)，无论是有向的还是无向的，均存在两个方向消息 $m_{j\to i}(y_i)$ 和 $m_{i\to j}(y_j)$。

若图结构中没有环，则信念传播算法对图经过两次扫描即可完成所有消息传递，进而能计算所有变量上的边际分布，**相应的动态规划算法如下**：

1. 指定一个根结点，从根结点开始向叶结点传递消息，直到所有叶结点均收到消息；
2. 从所有叶结点开始向根节点传递消息，直到根结点收到所有邻结点的消息。

4.2 近似推断

近似推断的方法大致可分为两大类：(1) 采样 (sampling)，通过使用随机化方法完成近似，如 MCMC 方法；(2) 确定性近似，典型代表为变分推断 (variational inference, VI)。

4.2.1 MCMC 采样

不论是计算概率还是计算期望，推断问题的关键在于求积分，以求期望为例，根据大数定律：

$$\mathbb{E}_Y(f(Y))=\int f(\boldsymbol{y})P(\boldsymbol{y})\mathrm{d}\boldsymbol{y}\approx\frac{1}{N}\sum_{i=1}^{N}f(y^i)$$

其中，上标 i 表示第 i 个样本点，而不是指 i 次幂。这种求解问题的方法即为蒙特卡洛法，**问题的关键即为如何对复杂的概率分布 $P(y)$ 进行采样，得到样本点 $\{y^i\}_1^N$ ，一种常用的采样方法就是马尔可夫蒙特卡洛 (Markov Chain Monte Carlo, MCMC) 方法。**

记要采样的分布为 p ，MCMC 通过构造平稳分布为 p 的马尔科夫链来产生样本：当时刻 t 足够大时，马尔科夫链会收敛于一个平稳分布，我们通过恰当的构造，使得该平稳分布恰为我们要采样的分布 p ，则达到平稳分布后产生的样本即为采样结果。马尔科夫链的平稳条件为：

$$p(\boldsymbol{y}^t)T(\boldsymbol{y}^{t-1}|\boldsymbol{y}^t) = p(\boldsymbol{y}^{t-1})T(\boldsymbol{y}^t|\boldsymbol{y}^{t-1})$$

其中，$T(\boldsymbol{y}'|\boldsymbol{y})$ 表示马尔可夫链从状态 $\boldsymbol{y}$ 到状态 $\boldsymbol{y}'$ 的转移概率。若时刻 t_1 平稳条件成立，则马尔可夫链下面产生的样本 $\boldsymbol{y}^{t_1}, \boldsymbol{y}^{t_1+1}, \boldsymbol{y}^{t_1+2}, \cdots$ 就是满足要求的采样，近似服从于分布 p 。

马尔科夫链不同的构造方法就产生了不同的 MCMC 算法，其中 Metropolis-Hastings (MH) 算法是 MCMC 的重要代表。MH 算法基于拒绝采样 (reject sampling) 来逼近分布 p ，达到平稳分布后，马氏链每个时刻产生的样本将一定的接受率被接受 (或者说，以一定的拒绝率被拒绝)。吉布斯采样 (Gibbs sampling) 有时被视为 MH 算法的特例，只不过其接受率为 1 。马尔可夫链通常需要很长时间才能趋于平稳分布，因此吉布斯采样算法的收敛速度较慢。此外，若存在极端概率 0 或 1 ，则不能保证马尔可夫链存在平稳分布，此时吉布斯采样会给出错误的估计结果。

4.2.2 确定性近似之变分推断

已知概率模型为 $p(\boldsymbol{y}, \boldsymbol{z}|\theta)$ ，其中 $\boldsymbol{y} = (y_1, \cdots, y_n)$ 为观测变量， $\boldsymbol{z} = (z_1, \cdots, z_m)$ 为隐变量。学习任务为估计参数 θ ，在得到参数估计值 θ^* 后，最终目的是为了推断，即得到概率 $p(\boldsymbol{z}|\boldsymbol{y}, \theta^*)$ 。

对于含有隐变量的参数估计问题，我们采用 EM 算法进行求解。最大化 F 函数 $F(q, \theta) = \int_{\mathbb{Z}} q(\boldsymbol{z}) \ln \left(p(\boldsymbol{y}, \boldsymbol{z}|\theta)/q(\boldsymbol{z})\right) \mathrm{d}\boldsymbol{z}$ 等价于最大化对数似然函数 $\ln p(\boldsymbol{y}|\theta)$ ，EM 算法实际上就是 F 函数的坐标上升法，其中：

1. E 步，以 $q(\boldsymbol{z})$ 为变量，最大化 F 函数，可得： $q(\boldsymbol{z}) = p(\boldsymbol{z}|\boldsymbol{y}, \theta^t)$;
2. M 步，取 $q(\boldsymbol{z}) = p(\boldsymbol{z}|\boldsymbol{y}, \theta^t)$ ，最大化 F 函数，得 θ^{t+1} :

$$\begin{aligned}\theta^{t+1} &= \arg\max_{\theta} \int_{\mathbb{Z}} p(\boldsymbol{z}|\boldsymbol{y}, \theta^t) \ln \left(p(\boldsymbol{y}, \boldsymbol{z}|\theta)/p(\boldsymbol{z}|\boldsymbol{y}, \theta^t)\right) \mathrm{d}\boldsymbol{z} \\ &= \arg\max_{\theta} \int_{\mathbb{Z}} p(\boldsymbol{z}|\boldsymbol{y}, \theta^t) \ln p(\boldsymbol{y}, \boldsymbol{z}|\theta) \mathrm{d}\boldsymbol{z}\end{aligned}$$

考虑到 E 步所得结果中分布 $p(\boldsymbol{z}|\boldsymbol{y}, \theta^t)$ 的形式可能过于复杂，变分推断假设 $q(\boldsymbol{z})$ 可拆解为一系列相互独立的多变量 $\boldsymbol{z}_i$， $i = 1, \cdots, M$ 的联合分布，即：

$$q(\boldsymbol{z}) = \prod_{i=1}^{M} q_i(\boldsymbol{z}_i)$$

$$\boldsymbol{z} = \cup_{i=1}^{M} \boldsymbol{z}_i, \quad \boldsymbol{z}_i \cap \boldsymbol{z}_j = \emptyset \ \ (i \neq j)$$

其中， q_i 表示变量集合 $\boldsymbol{z}_i$ 的概率分布函数。此外，我们还可以取 q_i 为相对简单或有很好的结构的分布，比如指数族分布。

接下来，在 $q(\boldsymbol{z})$ 的上述假设空间下 (即 $q(\boldsymbol{z}) = \prod_{i=1}^{M} q_i(\boldsymbol{z}_i)$)，对应 EM 算法的 E 步，我们以 $q(\boldsymbol{z})$ 为变量，最大化 F 函数，去推导 $q(\boldsymbol{z})$ 的解。可以看到，此时 $q(\boldsymbol{z})$ 的解就是原始解 $p(\boldsymbol{z}|\boldsymbol{y}, \theta^t)$ 更便于处理的近似。我们有：

$$\max_q F(q,\theta^t)=\max_q\int_{\mathbb{Z}}q(\boldsymbol{z})\ln\left(p(\boldsymbol{y},\boldsymbol{z}|\theta^t)/q(\boldsymbol{z})\right)\mathrm{d}\boldsymbol{z}$$

$$=\max_q\int_{\mathbb{Z}}\prod_{i=1}^M q_i(\boldsymbol{z}_i)\ln p(\boldsymbol{y},\boldsymbol{z}|\theta^t)\mathrm{d}\boldsymbol{z}-\int_{\mathbb{Z}}\prod_{i=1}^M q_i(\boldsymbol{z}_i)\ln\prod_{i=1}^M q_i(\boldsymbol{z}_i)\mathrm{d}\boldsymbol{z}$$

$$=\max_q\int_{\mathbb{Z}}\prod_{i=1}^M q_i(\boldsymbol{z}_i)\ln p(\boldsymbol{y},\boldsymbol{z}|\theta^t)\mathrm{d}\boldsymbol{z}-\sum_i^M\int_{\mathbb{Z}_i}q_i(\boldsymbol{z}_i)\ln q_i(\boldsymbol{z}_i)\mathrm{d}\boldsymbol{z}_i$$

在优化过程中，我们依然使用坐标上升法，依次对 q_i 进行优化，此时 $q_j,\ i\neq j$ 被视为常量：

$$\max_{q_i}\int_{\mathbb{Z}}\left(\prod_{i=1}^M q_i\right)\ln p(\boldsymbol{y},\boldsymbol{z}|\theta^t)\mathrm{d}\boldsymbol{z}-\int_{\mathbb{Z}_i}q_i\ln q_i\mathrm{d}\boldsymbol{z}_i$$

$$=\max_{q_i}\int_{\mathbb{Z}_i}q_i\left(\int_{\mathbb{Z}_{\sim i}}\ln p(\boldsymbol{y},\boldsymbol{z}|\theta^t)\prod_{j\neq i}q_j\mathrm{d}\boldsymbol{z}_{\sim i}\right)\mathrm{d}\boldsymbol{z}_i-\int_{\mathbb{Z}_i}q_i\ln q_i\mathrm{d}\boldsymbol{z}_i$$

$$=\max_{q_i}\int_{\mathbb{Z}_i}q_i\mathbb{E}_{\prod_{j\neq i}q_j}\{\ln p(\boldsymbol{y},\boldsymbol{z}|\theta^t)\}\mathrm{d}\boldsymbol{z}_i-\int_{\mathbb{Z}_i}q_i\ln q_i\mathrm{d}\boldsymbol{z}_i$$

其中，$\int_{\mathbb{Z}_{\sim i}}\ln p(\boldsymbol{y},\boldsymbol{z}|\theta^t)\prod_{j\neq i}q_j\mathrm{d}\boldsymbol{z}_{\sim i}=\mathbb{E}_{\prod_{j\neq i}q_j}\{\ln p(\boldsymbol{y},\boldsymbol{z}|\theta^t)\}\triangleq g(\boldsymbol{y},\boldsymbol{z}_i;\theta^t)$，是 $\boldsymbol{y},\boldsymbol{z}_i$ 的函数。**为了求解上述优化问题，我们试图凑出 KL 散度的形式，以利用其相关结论。**对 $\exp\left(g(\boldsymbol{y},\boldsymbol{z}_i;\theta^t)\right)$ $(\geqslant 0)$，以 $\boldsymbol{z}_i$ 为变量，$\boldsymbol{y}$ 为给定，将其视为一个非规范化的关于 $\boldsymbol{z}_i$ 的概率分布，并记为 $\hat{p}(\boldsymbol{z}_i|\boldsymbol{y},\theta^t)$，则：

$$\max_{q_i}\int_{\mathbb{Z}_i}q_i\mathbb{E}_{\prod_{j\neq i}q_j}\{\ln p(\boldsymbol{y},\boldsymbol{z}|\theta^t)\}\mathrm{d}\boldsymbol{z}_i-\int_{\mathbb{Z}_i}q_i\ln q_i\mathrm{d}\boldsymbol{z}_i$$

$$=\max_{q_i}\int_{\mathbb{Z}_i}q_i\ln\hat{p}(\boldsymbol{z}_i|\boldsymbol{y},\theta^t)\mathrm{d}\boldsymbol{z}_i-\int_{\mathbb{Z}_i}q_i\ln q_i\mathrm{d}\boldsymbol{z}_i$$

$$=\max_{q_i}-\int_{\mathbb{Z}_i}q_i\ln\frac{q_i}{\hat{p}(\boldsymbol{z}_i|\boldsymbol{y},\theta^t)}\mathrm{d}\boldsymbol{z}_i$$

$$=\min_{q_i}KL(q_i||\hat{p}(\boldsymbol{z}_i|\boldsymbol{y},\theta^t))$$

事实上，我们还有约束条件 $\int_{\mathbb{Z}_i}q_i\mathrm{d}\boldsymbol{z}_i=1$，也就是上式中的 q_i 实际上也是一个非规范化的概率，由 KL 散度的性质，当两个概率分布相等时，KL 散度取最小值 0，即 $q_i^t=\hat{p}(\boldsymbol{z}_i|\boldsymbol{y},\theta^t)$，规范化后即为：

$$q_i^t=\frac{\exp\left(\mathbb{E}_{\prod_{j\neq i}q_j}\{\ln p(\boldsymbol{y},\boldsymbol{z}|\theta^t\right)\})}{\int_{\mathbb{Z}_i}\exp\left(\mathbb{E}_{\prod_{j\neq i}q_j}\{\ln p(\boldsymbol{y},\boldsymbol{z}|\theta^t\right)\})\mathrm{d}\boldsymbol{z}_i}$$

而基于恰当分割的独立变量子集 $\{\boldsymbol{z}_i\}_1^M$ 和 q_i 合适的分布类型，$\mathbb{E}_{\prod_{j\neq i}q_j}\{\ln p(\boldsymbol{y},\boldsymbol{z}|\theta^t)\}$ 往往有闭式解。

上述内容就是变分推断的 E 步，而因为 E 步中我们需要求解 $\mathbb{E}_{\prod_{j\neq i}q_j}\{\ln p(\boldsymbol{y},\boldsymbol{z}|\theta^t)\}$ 也就是将变量 $\boldsymbol{z}_{\sim i}$ 积分掉，因此也称其为**基于平均场 (mean field) 的变分推断。**接下来，我们取 $q(\boldsymbol{z})=\prod_{i=1}^M q_i^t$，代入到 F 函数中以优化 θ，也就是变分推断的 M 步。交替进行 EM 步，完成参数的学习和对 $\boldsymbol{z}$ 后验分布 $q(\boldsymbol{z})$ 的推断。

可以看到，**变分推断实际上就是对 $q(\boldsymbol{z})$ 进行简化假设的 EM 算法，其学习过程和推断过程结合在一起，交替进行。由于对 $q(\boldsymbol{z})$ 的简化假设，最终我们会得到形式简单的 $\boldsymbol{z}$ 的后验分布 $q(\boldsymbol{z})$ 以便于推断，而影响变分推断效果的因素有对隐变量 $\boldsymbol{z}$ 的拆解和对变量子集 q_i 分布类型的假设。**
