变分推断

- 1.变分推断思路
- 2.分解概率分布推导过程
- 3.一元高斯分布实例
- 4.指数族分布
- 5.指数族分布+VI 的推导过程

一、变分推断思路

1.1 变分推断（VI， Variational Inference）背景：

贝叶斯公式：$p(\theta|X)=\frac{p(X|\theta)p(\theta)}{p(X)}$，我们通过似然函数，参数的先验分布和观测的数据求出后验概率。其分母部分通常是在数据中求积分得到，需要得到这部分数据很难。我们的问题集中在如何求后验概率上，关于如何得到后验概率一般分为两类方法：精确推断和近似推断。近似推断又分为确定性近似和随机性近似，下面介绍的变分推断则是确定性近似的一种，而随机性近似的方法有 Gibbs Sampling、Gibbs Sampling 等。
现在围绕如何根据可观测数据变量X来计算潜在变量Z的后验概率分布$p(Z|X)$展开讨论。假设有一个纯粹的贝叶斯模型，其中每个参数都有一个先验概率分布，这个模型也可以有隐含（潜在）变量及其参数，我们将所有的潜在变量和参数组成一个集合，记为Z，将观测变量的集合记为X，例如：$X=\{x_1,x_2,...x_N\}$，$Z=\{z_1,z_2,...z_N\}$，则可以确定联合概率分布$P(X,Z)$，我们的目标是要找到后验概率$P(Z|X)$的近似以及$P(X)$的近似。

1.2 数学框架

假设Z符合q分布，即$q(Z)$。在贝叶斯公式$p(X)=\frac{p(X,Z)}{p(Z|X)}$分子分母上下同除以$q(Z)$：

$$\log P(x)=\log P(x,z)-\log P(z|x) \quad =\log\frac{P(x,z)}{Q(z;\lambda)}-\log\frac{P(z|x)}{Q(z;\lambda)}$$

对Z取期望可得：

$$\ln p(X)=E_q[\ln\frac{p(X,Z)}{q(Z)}-\ln\frac{p(Z|X)}{q(Z)}]=E_q[\ln\frac{p(X,Z)}{q(Z)}]-E_q[\ln\frac{p(Z|X)}{q(Z)}]$$
$$=\int q(Z)\ln\frac{p(X,Z)}{q(Z)}dZ-\int q(Z)\ln\frac{p(Z|X)}{q(Z)}$$
$$=ELBO(q(Z))+KL(q(Z)||p(Z|X))\quad(1.1)$$

这个推导过程与EM算法里面是一样的，$ELBO$是**Evidence Lower BOund**，

> 为什么称之为 ELBO 呢？ $p(x)$一般被称之为 evidence ，又因为 $KL(q||p)>=0$, 所以 $p(x)>=E_{q(z;\lambda)}[\log p(x,z)-\log q(z;\lambda)]$, 这就是为什么被称为 ELBO

不同的是：
（1）EM算法里面有参数θ，而VI里面整合到随机变量Z里面去了。比如$\ln p(X)$在EM算法中是$\ln p(X|\theta)=ELBO(q(Z，\ \theta))+KL(q(Z)||p(Z|X,\theta))$。
（2）EM算法的目的是为了求解一个特定的θ优化值，使观测数据X的对数似然函数$\ln p(X|\theta)$最大,思路是不断提升ELBO来提升$\ln p(X|\theta)$：在E-step中首先给定当前θ^t，即设定概率分布$q(Z)^t=p(Z|X,\theta^t)$使上

式右边第二项KL散度为0，则θ^t使得下界ELOB达到了当前限定的值，为$\ln p(X|\theta)$; 在M-step中，将E-step获得的分布$q(Z)^t$代入到ELBO中，然后通过调节模型参数θ来最大化ELBO，得到下一轮的θ^{t+1}，若算法没收敛，此时新的KL(q||p)>0，因此$\ln p(X)$的提升值大于ELBO提升值。通过不断迭代E-step和M-step来不断提升ELBO从而最大化$\ln p(X|\theta)$。
而VI的思想是希望用概率分布$q(Z)$来近似模拟$p(Z|X)$，很简单的想法就是希望$KL(q||p)=0$，通过上式可知，当给定数据后，$\ln p(X)$是一个固定的上界，则期望通过最大化ELOB来最小化KL。VI的变分体现在调整输入到ELOB中的概率分布函数$q(Z)$来使得ELOB最大化。

> 什么是泛函？什么是变分法? [参见PRML附录D.变分法]
> 注意泛函与复合函数不同

1.3 思考

为了计算方便，却又能更充分、更灵活地提供对$P(Z|X)$的近似概率分布$q(Z)$，我们通常会选择一种限定类型的概率分布。核心思想包含两步：假设分布$q(Z;\lambda)$， 通过改变分布的参数 λ , 使 $q(z;\lambda)$ 靠近 $p(z|x)$ 总结成一句话就是，**为真实的后验分布引入了一个参数化的模型。** 即：用一个简单的分布 $q(Z_i;\lambda_i)$ 拟合复杂的分布 $p(Z|X)$ 。这种策略将计算 $p(z|x)$ 的问题转化成优化问题了

$$\lambda^* = \arg\min_{\lambda} \ divergence(p(z|x), q(z;\lambda))$$

收敛后，就可以用 $q(z;\lambda)$ 来代替 $p(z|x)$了。
也可以用基于平均随机场的VI方法，对$q(Z)$做了进一步的限定和简化，设定$q(Z)=\prod_{i=1}^{M} q_i(Z_i;\lambda_i)$，然后对所有单个概率分布进行自由形式的变分最优化。

二、分解概率分布推导过程

> 平均场理论：将数量巨大的互相作用的多体问题转化成每一个粒子处在一种弱周期场中的单体问题，即：对某个独立的小个体，所有其他个体对它产生的作用可以用一个平均的量给出，如此，简化后的模型成为一个单体问题。
> 总而言之，平均场理论是用于简化复杂模型的理论，譬如对于一个概率模型:$P(x_1,x_2,x_3,...,x_n)$,利用平均场理论，找到另一个模型$Q(x_1,x_2,x_3,...,x_n)=Q(x_1)Q(x_2)Q(x_3)...Q(x_n)$,使得$Q$尽量和$P$一致，并可以来近似代替$p(x_1,x_2,x_3,...,x_n)$

基于平均场理论，将Z的元素划分为若干个互不相交的组，记为Z_i或Z_j，其中 $i,j=1,2,3...M$。假设q的分布关于这些分组可以分解，则$q(Z)=\prod_{i=1}^{M} q_i(Z_i)$， 将$q(Z)=\prod_{i=1}^{M} q_i(Z_i)$代入ELBO式子:

$$L(q)=\int q(Z)\ln p(X,Z)dZ-\int q(Z)\ln q(Z)dZ$$

$$=\int\left(\prod_{i=1}^{M}q_i(Z_i)\right)\ln p(X,Z)dZ-\int\left(\prod_{i=1}^{M}q_i(Z_i)\right)\ln\prod_{i=1}^{n}q_i(Z_i)dZ$$

$$=\int\left(\prod_{i=1}^{M}q_i(Z_i)\right)\ln p(X,Z)dZ-\int\left(\prod_{i=1}^{M}q_i(Z_i)\right)\sum_{i=1}^{M}\ln q_i(Z_i)dZ$$

$$=L_1(q)-L_2(q)\ （2.1）$$

先处理第一项 $L_1(q)$:

$$L_1(q)=\int\Big(\prod_{i=1}^{M}q_i(Z_i)\Big)\ln p(X,Z)dZ$$

$$= \int_{Z_j} q_j(Z_j) \int_{\overline{Z}_j} \Big(\prod_{i=1, i\neq j}^{M} q_i(Z_i) \Big) \ln p(X, Z) d\overline{Z}_j dZ_j$$

$$= \int_{Z_j} q_j(Z_j) \Big[E_{q(\overline{Z}_j)} \ln p(X, Z) \Big] dZ_j$$

其中：$E_{q(\overline{Z}_j)} \ln p(X, Z) = \int_{\overline{Z}_j} \Big(\prod_{i=1, i\neq j}^{M} q_i(Z_i) \Big) \ln p(X, Z) d\overline{Z}_j$
为了方便计算，我们定义函数$\ln \check{p}(X, Z_j) = E_{i\neq j}[\ln p(X, Z)]$，这个式子将要刻画关于定义在所有$\overline{z}_j$ 上的q的概率分布。而函数$\ln \check{p}(X, Z_j)$是没有归一化的，其概率之和可能不等于1， 因此我们需要归一化$\ln \check{p}(X, Z_j)$，得到概率分布$q_j^*(Z_j)$:

$$q_j^*(Z_j) = \frac{\check{p}(X, Z_j)}{\int_{Z_j} \check{p}(X, Z_j) dZ_j} = \frac{e^{E_{i\neq j}[\ln p(X,Z)]}}{\int_{Z_j} e^{E_{i\neq j}[\ln p(X,Z)]} dZ_j} = \frac{\check{p}(X, Z_j)}{C_1}$$

其中$C_1 = \int_{Z_j} \check{p}(X, Z_j) dZ_j = \int_{Z_j} e^{E_{i\neq j}[\ln p(X,Z)]} dZ_j$，则:

$$\ln \check{p}(X, Z_j) = \ln q_j^*(Z_j) + \ln C_1$$

因此，第一项为:

$$L_1(q) = \int_{Z_j} q_j(Z_j) * \Big(\ln q_j^*(Z_j) + \ln C_1 \Big) dZ_j$$

$$= \int_{Z_j} q_j(Z_j) * \ln q_j^*(Z_j) dZ_j + \int_{Z_j} q_j(Z_j) * \ln C_1 dZ_j$$

$$= \int_{Z_j} q_j(Z_j) * \ln q_j^*(Z_j) dZ_j + \ln C_1 \quad (2.2)$$

再处理第二项, $L_2(q) = \int \Big(\prod_{i=1}^{M} q_i(Z_i) \Big) \sum_{i=1}^{M} \ln q_i(Z_i) dZ$,将其展开（注意每一个函数$q_i$只涉及到一个变量$Z_i$,函数$q_i$是泛函变量:

$$L_2(q) = \int_{Z_1} \int_{Z_2} \dots \int_{Z_M} \Big[q_1(Z_1) q_2(Z_2) \dots q_M(Z_M) \Big] * \Big[\ln q_1(Z_1) + \ln q_2(Z_2) \dots + \ln q_M(Z_M) \Big] dZ_1 \dots dZ_{M-1}$$

$$= \sum_{i=1}^{M} \Big(\int_{Z_1} \int_{Z_2} \dots \int_{Z_M} \ln q_i(Z_i) * \Big[q_1(Z_1) q_2(Z_2) \dots q_M(Z_M) \Big] dZ_1 \dots dZ_{M-1} dZ_M \Big)$$

对求和内的第i项来说:

$$\int_{Z_1} \int_{Z_2} \dots \int_{Z_M} \ln q_i(Z_i) * \Big[q_1(Z_1) q_2(Z_2) \dots q_M(Z_M) \Big] dZ_1 \dots dZ_{M-i} dZ_M$$

$$= \int_{Z_i} \Big[\int_{Z_{j\neq i}} \dots \int \ln q_i(Z_i) * \Big[q_1(Z_1) q_2(Z_2) \dots q_M(Z_M) \Big] \Big(\prod_{j\neq i}^{M} dZ_j \Big) \Big] dZ_i$$

$$= \int_{Z_i} \ln q_i(Z_i) * \Big[\int_{Z_{j\neq i}} \dots \int \Big[q_1(Z_1) q_2(Z_2) \dots q_M(Z_M) \Big] \Big(\prod_{j\neq i}^{M} dZ_j \Big) \Big] dZ_i$$

上式中，中括号内的部分，对不包括i内的$q(Z)$的所有其他维度积分，相当于计算$q(Z) = \prod_{i=1}^{M} q_i(Z_i)$的边缘概率，即：中括号内的值直接等于$q_i(Z_i)$。则上式转化为:

$$\int_{Z_i} \ln q_i(Z_i) * q_i(Z_i) dZ_i$$

因此我们将每部分的结果**代入**$L_2(q)$**中得到第二部分**:

$$L_2(q) = \sum_{i=1}^{M} \Big(\int_{Z_i} \ln q_i(Z_i) * q_i(Z_i) dZ_i \Big) = \int_{Z_j} \ln q_j(Z_j) * q_j(Z_j) dZ_j + \sum_{i \neq j} \Big(\int_{Z_i} \ln q_i(Z_i) * q_i(Z_i) dZ_i \Big)$$

$$= \int_{Z_j} \ln q_j(Z_j) * q_j(Z_j) dZ_j + \ln C_2 \quad (2.3)$$

因为我们将$q(Z) = \prod_{i=1}^{M} q_i(Z_i)$的每一项分开处理，现在只考虑第$j$项，则将上式中不涉及$Z_j$的项视为常数项$\ln C_2$。
将两部分结果（2.2）（2.3）代入（2.1）中，得到$L(q)$:

$$L(q) = L(q_j) = \int_{Z_j} q_j(Z_j) * \ln q_j^*(Z_j) dZ_j - \int_{Z_j} \ln q_j(Z_j) * q_j(Z_j) dZ_j + \ln C_1 - C_2$$

$$= \int_{Z_j} q_j(Z_j) * \ln \frac{\ln q_j^*(Z_j)}{q_j(Z_j)} dZ_j + C = -KL(q_j || q_j^*) + C$$

其中，$q_j^*(Z_j)$是归一化的$\breve{p}(X, Z_j)$，表示关于定义在所有$i \neq j$的Z_i上的q概率分布的期望。
我们假设让所有的$\{q_{i \neq j}\}$固定，分析上式，若要最大化$L(q)$，相当于最大化$KL(q_j || q_j^*)$的相反数，即最小化$KL(q_j || q_j^*)$,而当 $q_j(Z_j) = q_j^*(Z_j)$时KL散度最小。因此，最优解的一般表达式为

$$\ln q_j^*(Z_j) = E_{q(\overline{Z}_j)} \ln p(X, Z) - \ln C_1 \quad (2.4)$$

这个解表示为了得到因子q_j最优解的对数，我们只需要考虑所有隐含变量Z和可见变量x上的联合概率分布的对数，然后关于所有其他因子$\{q_{i \neq j}\}$取期望即可。

平均场方法通过恰当的分割独立变量子集$\{Z_i\}_{i=1}^{M}$，使得相应的$q_i(Z_i)$用闭合解。由于q_i的分布依赖于其他变量子集的分布，因此平均场是一个不断迭代的方法以保证各子集分布收敛到局部最优解(下界相对各子集分布是凸函数)。
至此我们已经从理论上找到了变分贝叶斯推断的通用公式求法：对于每一个$q_j(Z_j)$: 令$q_j(Z_j) = q_j^*(Z_j)$循环，直到收敛 。

虽然从理论上推导了变分推断的框架算法，但是在实践中，对于不同模型，最重要的是考虑如何对隐变量进行拆解，以及假设各变量子集服从何种分布。然后手动推导$q_j^*(Z_j)$，简要来说，**推导变分贝叶斯模型一般分为四个步骤**:
1.确定好研究模型各个参数的的共轭先验分布(如果想做full bayes model)
2.写出研究模型的联合分布$P(Z, X)$
3.根据联合分布确定变分分布的形式$q(Z)$
4.对于每个变分因子$q_j(Z_j)$求出$P(Z, X)$关于不包含变量Z_j的数学期望，再规整化为概率分布
当然这个过程并不简单，对于实际模型，其推导一般比较繁冗复杂，很容易出错，所以后来便有学者研究出更加一般更加自动化的基于概率图模型的算法框架——**VMP**(Variational Message Passing) 。如果模型是指数族的模型，都可以套用VMP自动得到算法求解。

三、实例

假设给定一个数据集$D = \{x_1, \cdots, x_N\}$，其中每一个观测的单变量数据x_n均独立的抽样于**高斯分布**。那么似然函数为:

$$p(D|\mu,\tau) = \left(\frac{\tau}{2\pi}\right)^{N/2} exp\{-\frac{\tau}{2}\sum_{i=1}^{N}(x_n-\mu)^2\}$$

其中$\tau = \frac{1}{\delta^2}$，表示**精度。** 我们现在的目标是估计参数$\mu$和$\tau$。我们引入两个参数的共轭先验分布

$$p(\mu|\tau) = N(\mu|\mu 0, (\lambda_0\tau)^{-1})$$

$$p(\tau) = Gam(\tau|a_0, b_0)$$

实质上，对于这个简单问题的后验概率是能精确计算出来的，由于共轭的性质，参数的后验分布也是其先验分布。这里，我们采用变分推断的思想进行后验概率的计算，首先我们假设：

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

则对于$q_\mu(\mu)$

$$\ln q_\mu^*(\mu) = E_\tau[\ln p(D|\mu,\tau) + \ln p(\mu|\tau)] + const = -\frac{E[r]}{2}\{\lambda_0(\mu-\mu_0)^2 + \sum_n(x_n-\mu)^2\} + const$$

注意我们只关注μ，其它无关项全融入常数项。由上式可以清楚推断$q_\mu(\mu)$服从高斯分布$N(\mu|\mu_N, \lambda_N^{-1})$，注意我们并没有事先预设其$q_\mu(\mu)$的分布形式，完全推断与似然函数和预设的先验分布。其中

$$\mu_N = \frac{\lambda_0\mu_0 + N\overline{x}}{\lambda_0 + N}$$

$$\lambda_N = (\lambda_0 + N)E[\tau]$$

综上，由μ的后验分布$q_\mu^*(\mu)$，可得知其期望为μ_N。注意当$N \to \infty$，$\mu_N = \overline{x}$。同理，τ的估计采用类似的方法。

四、指数族分布

4.1 指数族分布形式

如果一类分布可以写成如下形式，则它可以叫做 *指数族分布* (exponential family distributions) 。

$$P(X|\eta) = h(X)exp\{\eta^T\Phi(X) - A(\eta)\}\ （4.1）$$

η是自然参数（也叫canonical parameter），其中X是观测数据，$\Phi(X)$是观测数据的充分统计量(sufficient statistic), $A(\eta)$ 是对数归一化项（也叫对数分配函数，log partition function,log normalization ）。
其中：$h(X)$只含有X，而$A(\eta)$ 只含有自然参数η，$\eta^T\Phi(X)$相当于数据和参数的一个内积，这样把参数和数据分开使得很多计算变得简单很多，而且也产生了一定的规律和结论，后面会详细介绍。

> 大多数概率分布都是指数族分布，都可以表示成上面公式给出的形式：
> 1）伯努利分布：对0、1问题进行建模
> 2）多项式分布：对K个离散结果的事件建模
> 3）泊松分布：对计数过程进行建模
> 4）伽马分布与指数分布：对间隔的正数进行建模
> 5）Beta分布：对小数进行建模
> 6）Dirichlet分布：对小数进行建模

> 7）Wishart分布：对协方差进行建模
> 8）高斯分布

4.2 一元高斯分布函数化成指数族分布的形式

一元高斯分布函数形式如下：

$$P(X|\theta)=P(X|\mu,\delta)=\frac{1}{\sqrt{2\pi\delta^2}}e^{-\frac{(x-\mu)^2}{2\delta^2}}$$

目的是将此式子化为（4.1）式的样式。
观察（4.1）式，其实η也是参数的一个函数，这个地方目的是要将参数θ映射到对应的η中，并且找到对应的对数归一化项$A(\eta)$和前面的系数$h(X)$。
推导过程如下：

$$P(X|\theta)=\frac{1}{\sqrt{2\pi\delta^2}}e^{-\frac{(x-\mu)^2}{2\delta^2}}=e^{\log\frac{1}{\sqrt{2\pi\delta^2}}}*e^{-\frac{(x-\mu)^2}{2\delta^2}}$$

$$=exp\Big\{\log\frac{1}{\sqrt{2\pi\delta^2}}-\frac{(x^2-\mu x+\mu^2)}{2\delta^2}\Big\}$$

$$=exp\Big\{\begin{bmatrix}\frac{\mu}{\delta^2} & \frac{-1}{2\delta^2}\end{bmatrix}\begin{bmatrix}x\\x^2\end{bmatrix}+(\log 2\pi\delta^2)^{-1/2}-\frac{\mu^2}{\delta^2}\Big\}\quad（4.2）$$

可以看出式（4.1）中$h(X)=1$且：将η命为$\begin{bmatrix}\eta_1 & \eta_2\end{bmatrix}^T=\begin{bmatrix}\frac{\mu}{\delta^2} & \frac{-1}{2\delta^2}\end{bmatrix}^T$,将$\Phi(X)$命为$\begin{bmatrix}x\\x^2\end{bmatrix}$，反过来解得：

$$\mu=-\frac{\eta_1}{2\eta_2};\delta^2=-\frac{1}{2\eta_2}\quad（4.3）$$

将上式代入到（4.2）的后半部分$(\log 2\pi\delta^2)^{-1/2}-\frac{\mu^2}{\delta^2}$中，可以得到这部分关于$\eta$的表达式，如下：

$$A(\eta)=-\frac{\eta_1^2}{4\eta_2}+\frac{1}{2}\log\frac{-\pi}{\eta}\quad（4.4）$$

总结：根据上面的推导过程，将一元高斯分布分概率密度函数转化为了标准的指数族分布函数，其中$\eta=\begin{bmatrix}\eta_1 & \eta_2\end{bmatrix}^T$，$\Phi(X)=\begin{bmatrix}x\\x^2\end{bmatrix}$,$A(\eta)$为（4.4），$h(X)=1$。

4.3 关于$A(\eta)$

根据（4.1），将指数部分的$A(\eta)$分开，式子转化为：

$$P(X|\eta)=exp\{A(\eta)\}^{-1}h(X)exp\{\eta^T\Phi(X)\}\quad（4.5）$$

上式左右两边对x积分，可以得到：

$$1=exp\{A(\eta)\}^{-1}\int h(X)exp\{\eta^T\Phi(X)\}\mathrm{d}x$$

因此：

$$exp\{A(\eta)\} = \int h(X)exp\{\eta^T\Phi(X)\}\mathrm{d}x \quad (4.6)$$

这也是为什么**将$A(\eta)$称为对数归一化项的原因。**式子(4.6)中，左右两边都对η 求导可得：

$$A'(\eta)exp\{A(\eta)\} = \frac{\partial}{\partial\eta}\int h(X)exp\{\eta^T\Phi(X)\}\mathrm{d}x$$

$$A'(\eta)exp\{A(\eta)\} = \int h(X)exp\{\eta^T\Phi(X)\}\Phi(X)\mathrm{d}x$$

$$A'(\eta) = \frac{\int h(X)exp\{\eta^T\Phi(X)\}\Phi(X)\mathrm{d}x}{exp\{A(\eta)\}} = \int h(X)e^{\eta^T\Phi(X)-A(\eta)}\Phi(X)\mathrm{d}x$$

$$= \int P(X|\eta)\Phi(X)\mathrm{d}x = E_{P(X|\eta)}[\Phi(X)]$$

因此可以得到结论：

$$A'(\eta) = E_{P(X|\eta)}[\Phi(X)] \quad (4.7)$$

即：**对数归一化项$A(\eta)$的一阶导数等于充分统计量的期望。**
这里也可以理解为什么$\Phi(X)$称为充分统计量，因为我们只需要知道$\Phi(X)$就可以刻画这个分布了，比如一元高斯分布中的$\Phi(X) = \begin{bmatrix} x \\ x^2 \end{bmatrix}$，只需要知道$E_{x\ P(X|\eta)}[\Phi(x)]$和$E_{x\ P(X|\eta)}[\Phi(x^2)]$就可明确这个分布了。
继续推导还可以得出：

$$A''(\eta) = Var[\Phi(X)] \quad (4.8)$$

其中：$A'(\eta)$是凸函数。

> **对于式子4.7的理解：**
>
> > （1）在极大似然估计中，我们的目标是求参数η的值使得对数似然函数$P(X|\eta)$最大，此时$P(X|\eta)$对参数η求偏导然后令导数等于0，则可以求出η_{MLE}的值。对于指数族分布，我们可以证明：
> >
> > $$\frac{\partial}{\partial\eta}\sum_{i=1}^{n}[logP(X|\eta)] = \sum_{i=1}^{n}[\Phi(X_i) - A'(\eta)]$$
> >
> > 令上式为0，可以得到：
> >
> > $$A'(\eta) = E_{P(X|\eta)}[\Phi(X)]$$
> >
> > 与（4.7）的结论相同，因此我们可以得到求解极大似然估计的简单方法：
> > **对于指数族分布，先将分布的对数归一化项对参数求导得到$A'(\eta)$（等式左边），然后将充分统计量对分布$P(X|\eta)$求均值（等式右边），让两者相等，得到参数值η_{MLE}。**
>
> > （2）在贝叶斯推断中，对于不同的参数值η，η可以取η_1、η_2...η_N,不同的η_i决定不同的概率密度函数$P(X|\eta_i)$，所求出来的期望（等号右边）$E_{P(X|\eta)}[\Phi(X)]$也不同，即：等号右边的值是由参数η决定的。因此得到一个求期望的新方法：将$A'(\eta)$对η求导，然后将η_i的值代入，则可以的到等式右边的期望。这将应用在指数族分布的变分推断中。

4.4 指数族分布的共轭先验

根据贝叶斯公式可知：

$$P(\theta|X) \propto P(X|\theta) * P(\theta)$$

共轭表示：似然函数$P(X|\theta)$与先验分布$P(\theta)$共轭，即后验分布$P(\theta|X)$与先验分布$P(\theta)$的形式相同，只是参数不同。若$P(X|\theta)$服从指数族分布，则

$$P(X|\theta) = h(X)exp\{\theta^T\Phi(X) - A(\theta)\}$$

结论：理论上来说，任何一个指数族似然函数，都必定有一个指数族的先验分布与其共轭，使得后验分布与先验分布有相同的形式。而此时，只需要另先验分布的充分统计量$\Phi(X)$的第二部分 与 似然函数的对数归一化项$A_l(\theta)$相等即可。 下证此结论。
假设先验分布$P(\theta)$和似然函数都服从指数族分布，并为先验分布引入参数α，则先验分布为：

$$P(\theta|\alpha) = h(\theta)exp\{\alpha^T\Phi(\theta) - A_1(\alpha)\}（4.9）$$

似然函数为：

$$P(X|\theta) = h(X)exp\{\theta^T\Phi(X) - A_2(\theta)\}（4.10）$$

假设

$$\Phi(\theta) = \begin{bmatrix} \theta \\ -g(\theta) \end{bmatrix}，\ \alpha = \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix}$$

其中θ和α_1是标量，而$g(\theta)$和α_2是矢量且长度相同。我们的目的是让（4.9）和（4.10）相乘的结果有着标准的指数族分布的形式。（4.9）和（4.10）相乘(观测数据X视为常量,$A_1(\alpha)$视为常量：

$$\begin{aligned} P(\theta|\alpha)P(X|\theta) &= h(\theta)exp\{\alpha^T\Phi(\theta) - A_1(\alpha)\} * h(X)exp\{\theta^T\Phi(X) - A_2(\theta)\} \\ &\propto h(\theta)exp\{\alpha^T\Phi(\theta) - A_2(\theta) + \theta^T\Phi(X)\} \\ &\propto h(\theta)exp\{\alpha_1\theta - \alpha_2 g(\theta) - A_2(\theta) + \theta\Phi(X)\} \\ &= h(\theta)exp\{(\alpha_1 + \Phi(X))\theta - \alpha_2 g(\theta) - A_2(\theta)\} \end{aligned}$$

若 **令**$g(\theta) = A_2(\theta)$， 则上式可以化简为：

$$P(X|\theta)P(\theta|\alpha) \propto h(\theta)exp\{(\alpha_1 + \Phi(X))\theta - (\alpha_2 + 1)A_2(\theta)\}（4.11）$$

令$\hat{\alpha}_1 = (\alpha_1 + \Phi(X))$,$\hat{\alpha}_2 = (\alpha_2 + 1)$则：

$$P(X|\theta)P(\theta|\alpha) \propto h(\theta)exp\{\hat{\alpha}_1\theta - \hat{\alpha}_2 A_2(\theta)\} = h(\theta)exp\{\hat{\alpha}\Phi(\theta)\}（4.11）$$

式子（4.11）符合指数族分布的标准形式，即：若令$g(\theta) = A_1(\alpha)$两者相乘后得到的也是一个指数族分布，反过来也成立。证毕。
结论是：让参数的先验分布的的第二部分$g(\theta)$，取似然函数的充分统计量取似然函数的对数归一化项$A_2(\theta)$,则先验分布与似然函数共轭，得到的后验概率分布的参数为$\hat{\alpha}_1 = (\alpha_1 + \Phi(X))$,$\hat{\alpha}_2 = (\alpha_2 + 1)$。

五、指数族分布的变分推断

变分推断的目的是找到一个分布$q(Z)$去近似于后验概率$p(Z|X)$。即：将式子（1.1）中的KL散度最小化，而数据是固定的，即等价于将$ELBO$最大化：

$$ELBO: L(q) = \int q(Z)logp(X,Z)dZ - \int q(Z)logq(Z)dZ$$

$$= E_{q(Z)}logp(X,Z) - E_{q(Z)}logq(Z)（5.1）$$

在第二部分推导过程中，将模型中的变量分为了观测变量和隐含变量，现在将**隐含变量**拆开，比如拆解成两部分，例如：Z=：$\{Z,\beta\}$，其中我们选择的两个分布Z, β是独立的，类似于第二部分所述的$q(Z) = \prod_{i=1}^{M} q_i(Z_i)$。

对于$\{Z,\beta\}$，其后验概率有如下推导：

$$P(\beta,Z|X) = P(\beta|Z,X)P(Z|X) = P(Z|\beta,X)P(\beta|X)(5.2)$$

对于（5.2）中的$P(\beta|Z,X)$和$P(Z|\beta,X)$，现假设他们为指数族分布，将其写成指数族分布的标准形式：

$$P(\beta|\eta(Z,X)) = h(\beta)exp\{\eta(Z,X)^T\Phi(\beta) - A_1(\eta(Z,X))\}（5.3）$$

$$P(Z|\eta(\beta,X)) = h(Z)exp\{\eta(\beta,X)^T\Phi(Z) - A_2(\eta(\beta,X))\}（5.4）$$

我们的目标变换为：想用一个变分函数$q(\beta)$近似于$P(\beta|Z,X)$,其中$q(\beta)$也是指数族分布，**引入参数**λ，$q(\beta|\lambda)$**的形式如下**：

$$q(\beta|\lambda) = h(\beta)exp\{\lambda^T\Phi(\beta) - A(\lambda)\}（5.5）$$

不断优化λ的值，使式（5.5）趋近于式（5.3）
同理，想用一个变分函数想$q(Z)$近似于$P(Z|\beta,X)$,其中$q(Z)$也是指数族分布，**引入参数**γ，$q(Z|\gamma)$**的形式如下**：

$$q(Z|\gamma) = h(Z)exp\{\gamma^T\Phi(Z) - A(\gamma)\}（5.6）$$

不断优化γ的值，使式（5.6）趋近于式（5.4）

在式（5.1）中，$L(q)$中的q为隐含变量$\{Z,\beta\}$的函数，而Z依赖于参数λ,β依赖于参数γ,即：可以将ELBO记为：$L(q(\lambda,\gamma))$。对式子$L(q(\lambda,\gamma))$，若想优化参数，可以使用常用的迭代优化，即：先固定γ，优化λ，然后固定λ，优化γ,不停迭代。（这里是两个参数，同理，若有多个参数则需要固定其他的所有参数来优化其中一个参数）
式（5.1）将Z拆分，且是独立的，即$q(Z,\beta) = q(Z)q(\beta)$，并引入参数$\lambda,\gamma$，变换为:

$$L(q(\lambda,\gamma)) = E_{q(Z,\beta)}logp(X,Z,\beta) - E_{q(Z,\beta)}logq(Z,\beta)$$

$$= E_{q(Z,\beta)}log[P(\beta|Z,X)P(Z|X)P(X)] - E_{q(Z,\beta)}log[q(\beta)q(Z)]$$

$$= E_{q(Z,\beta)}logP(\beta|Z,X) + E_{q(Z,\beta)}logP(Z|X) + E_{q(Z,\beta)}logP(X) - E_{q(Z,\beta)}logq(\beta) - E_{q(Z,\beta)}logq(Z)（5$$

我们**首先固定**γ，**优化**$q(\beta|\lambda)$（此时由于固定了γ，$q(Z|\gamma)$也固定，将γ、$q(Z;\gamma)$视为常数），式子（5.7）中的第二项与λ（$q(\beta)$）无关，第三项也无关，第五项也无关，剩下的只有第一项和第四项，其他的都视为常数，则（5.7）化简为：

$$E_{q(Z,\beta)}logP(\beta|Z,X) - E_{q(Z,\beta)}logq(\beta) + constant（5.8）$$

想要使ELBO max,则优化式（5.8）。对式（5.8）的指数族分布展开，即代入式（5.3）(5.5):

$$(5.8) = E_{q(Z,\beta)}log\Big[h(\beta)exp\{\eta(Z,X)^T\Phi(\beta) - A_1(\eta(Z,X))\}\Big] - E_{q(Z,\beta)}log\Big[h(\beta)exp\{\lambda^T\Phi(\beta) - A(\lambda)\}\Big]$$

$$= \Big\{E_{q(Z,\beta)}log\Big[h(\beta)\Big] + E_{q(Z,\beta)}\Big[\eta(Z,X)^T\Phi(\beta)\Big] - E_{q(Z,\beta)}\Big[A_1(\eta(Z,X))\Big]\Big\} -$$

$$\left\{E_{q(Z,\beta)}log\left[h(\beta)\right]+E_{q(Z,\beta)}\left[\lambda^T\Phi(\beta)\right]-E_{q(Z,\beta)}\left[A(\lambda)\right]\right\}\ （5.9）$$

上式中，第一项和第四项约掉，第三项只有Z没有β,视为常数，上式化简为：

$$E_{q(Z,\beta)}\left[\eta(Z,X)^T\Phi(\beta)\right]-E_{q(Z,\beta)}\left[\lambda^T\Phi(\beta)\right]+E_{q(Z,\beta)}\left[A(\lambda)\right]+constant$$

$$\overset{将Z,\beta分开}{=}E_{q(Z)}\left[\eta(Z,X)^T\right]*E_{q(\beta)}log\left[\Phi(\beta)\right]-E_{q(\beta)}\left[\lambda^T\Phi(\beta)\right]+A(\lambda)\ （5.10）$$

上式中的相乘项的因子$E_{q(\beta)}log\left[\Phi(\beta)\right]$。由上一节的公式（4.7）可知，$E_{q(\beta)}log\left[\Phi(\beta)\right]=A'(\lambda)$，$E_{q(\beta)}\left[\lambda^T\Phi(\beta)\right]=\lambda^T*E_{q(\beta)}\left[\Phi(\beta)\right]=\lambda^TA'(\lambda)$。则，上式（5.10）化简为：

$$A'(\lambda)E_{q(Z)}\left[\eta(Z,X)^T\right]-\lambda^TA'(\lambda)+A(\lambda)+constant\ （5.11）$$

若要最大化（5.11），只需要对η求导，即：

$$\begin{aligned}\frac{\partial L(q(\lambda,\gamma))}{\partial\lambda}&=A''(\lambda)E_{q(Z)}\left[\eta(Z,X)^T\right]-\lambda^TA''(\lambda)\\&=A''(\lambda)\left(E_{q(Z)}\left[\eta(Z,X)^T\right]-\lambda^T\right)\end{aligned}$$

令上式为0，解得：

$$\lambda=E_{q(Z)}\left[\eta(Z,X)^T\right]\ （5.12）$$

其中，A'(λ)是凸函数。
这样**固定**γ**得到了**λ，同理，在（5.7）中通过**固定**λ**得到了**γ 的结果为：

$$\gamma=E_{q(\beta)}\left[\eta(\beta,X)^T\right]\ （5.13）$$

即：若要优化$q(\beta|\lambda)$，则将P的参数分解为独立的参数，将剩下的所有参数的分布求期望则得到想要优化的$q(\beta|\lambda)$的λ值。在第二部分中，我们将$q(Z)=\prod_{i=1}^{M}q_i(Z_i)$，这里我们若也将$Z$分解为独立的$Z_i$,不同的$q_i(Z_i)$由不同的参数决定，且与$q_i(\eta)$耦合，其中$\eta$是除去$Z_i$后的所有参数。我们可以使用迭代优化，固定其他所有的分布来优化一个分布的参数，其结果等于计算$E_{q_i(Z_i)}\left[\eta^T\right]$；然后再固定其他所有的分布优化另一个分布的参数，直至收敛，达到ELBO的最大值。

---

references:
1.PRML 10.1
2.PRML 2.4
3.PRML附录
4.https://blog.csdn.net/step_forward_ML/article/details/78077383
5.https://www.scribd.com/document/388036022/Notes-on-Expectation-Maximization-and-Variational-Inference-pdf
6.https://www.scribd.com/document/388036017/Notes-on-Varational-Inference-pdf
7.https://www.scribd.com/document/388036021/VI简明推导-pdf
8.https://www.youtube.com/watch?v=arMoli91OZE&list=PLFze15KrfxbF0n1zTNoFIaDpxnSyfgNgc