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

一、变分推断思路

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

在概率模型的应用中，我们经常要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$p(Z|X)$。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：

$$\ln p(X) = E_q[\ln \frac{p(X, Z)}{q(Z)} - \ln \frac{p(Z|X)}{q(Z)}] = E_q[\ln \frac{p(X, Z)}{q(Z)}] - E_q[\ln \frac{p(Z|X)}{q(Z)}]$$

$$= \int q(Z) \ln \frac{p(X, Z)}{q(Z)} dZ - \int q(Z) \ln \frac{p(Z|X)}{q(Z)}$$

$$= ELOB(q(Z)) + KL(q(Z)||p(Z|X))$$

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$\theta^t$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$\theta$来最大化ELBO，得到下一轮的$\theta^{t+1}$，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ELOB式子：

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

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

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

$$=L_1(q)-L_2(q)$$

先处理第一项 $L_1(q)$:

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

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

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

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:

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

$$=\int_{Z_j}q_j(Z_j)*\ln q_j^*(Z_j)dZ_j+\int_{Z_j}q_j(Z_j)*\ln C_1dZ_j$$

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

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:

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

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

对求和内的第i项来说:

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

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

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

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**:

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

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

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$\ln C_2$。
将两部分结果代入ELBO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$\check{p}(X,Z_j)$，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

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

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:

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

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

$$p(\mu|\tau)=N(\mu|\mu0,(\lambda_0\tau)^{-1})$$

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

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。

references:

1.PRML 10.1

2.PRML附录

3.https://blog.csdn.net/step_forward_ML/article/details/78077383

4.https://www.scribd.com/document/388036023/Examples-of-Variational-Inference-With-Gaussian-Gamma-Distribution