变分贝叶斯

Junnan Zhu

2016 年 10 月 10 日

变分贝叶斯是个神奇的理论，刚开始接触会对其中的大批数学公式有抵触感，但是当你真正领悟过来的时候却发现原来是个这么有趣的事情。这里先按部就班的给出维基百科上变分贝叶斯的定义，变分贝叶斯是一类用于近似贝叶斯估计以及机器学习中复杂积分的技术。大家都知道，涉及贝叶斯的统计模型往往会包含三部分：数据（或称观测变量）、未知参数、隐变量；在贝叶斯理论中，后两者可以统称为不可观测变量。变分贝叶斯主要有两种用途：

1°逼近（近似）不可观测变量的后验概率

2°对于一个特定的模型，给出观测变量的边缘似然函数的下界（evidence lower bound）。这主要用于模型选择的过程，因为通常认为边缘似然函数的值越高就表示模型对数据拟合程度越好，由该模型产生这批数据的概率越高。

看完上面大家可能还是云里雾里的，因为说的实在有点抽象。那么我们从简单的聊起，$P(X) = P(X,Z)/P(Z|X)$这个简单的式子相信大家看了很多遍了，就是一个简单的条件概率的表达。但是其中的P就是模型自身的概率分布，往往是预先假定的一个较为复杂的分布。而其中X为模型中的观测变量，Z为隐变量。那么易知$P(X|Z)$为似然函数，$P(Z|X)$为后验概率，那么$P(X)$叫什么呢？大家也许已经猜到了，就是上面提到过的边缘似然函数，因为其中已经将隐变量抹去。那么我们继续对上述那个等式进行扩展：

$$\begin{aligned}\ln P(X) &= \ln\left(P(X,Z)\right) - \ln P(Z|X) \\ &= \ln\left(\frac{P(X,Z)}{q(Z)}\right) - \ln\left(\frac{P(Z|X)}{q(Z)}\right) \\ &= \ln P(X,Z) - \ln q(Z) - \ln\left(\frac{P(Z|X)}{q(Z)}\right)\end{aligned}$$

上面的式子十分简单，只是引入了一个$q(Z)$，当然$q(Z)$可以是任意分布，这都无关紧要。现在我们利用类似EM算法的思想在$q(Z)$这个分布上对等式两边求期望即：

$$\begin{aligned}\int_Z q(Z)\ln P(X)\,dZ &= \int_Z q(Z)\ln P(X,Z)\,dZ - \int_Z q(Z)\ln q(Z)\,dZ - \int_Z q(Z)\ln\left(\frac{P(Z|X)}{q(Z)}\right)dZ \\ \ln P(X) &= \underbrace{\int_Z q(Z)\ln P(X,Z)\,dZ - \int_Z q(Z)\ln q(Z)\,dZ}_{\text{L(q):ELOB}} \underbrace{- \int_Z q(Z)\ln\left(\frac{P(Z|X)}{q(Z)}\right)dZ}_{\text{KL(q(Z)||P(X,Z))}}\end{aligned}$$

如在公式中标识的，右边那一项为$q(Z)$和$P(Z|X)$的KL散度，KL散度可以用来度量两个概率分布之间的距离，而且KL散度满足非负性但不满足对称和三角不等式。乍一看可能并不能

看出KL散度具有非负性，下面我们证明一下，由于$-\ln x$为凸函数，所以可以根据Jensen's inequality:

$$\begin{aligned} D_{KL}(Q||P) &= -\int Q(x)\ln\frac{P(x)}{Q(x)}\,dx \\ &= E_{Q(x)}(-\ln\frac{P(x)}{Q(x)}) \\ &\geq -\ln E_{Q(x)}(\frac{P(x)}{Q(x)}) \\ &= -ln\int Q(x)\frac{P(x)}{Q(x)}\,dx = 0 \end{aligned}$$

所以我们现在回过头看看$\ln P(X)$的表达式，$lnP(X) = L(q)+KL$由于$KL(q(Z)||P(X,Z)) \geq 0$,所以可得$\ln P(X) \geq L(q)$,这也恰好解释了为何$L(q)$被称为下界的原因，那么等号取得的条件就是$q(Z)$和$P(X,Z)$的KL散度为0，也就是这两个概率分布一模一样的时候。这就恰好变成了之前我们所说过的变分贝叶斯的两个目的：（1）提供边缘似然函数的下界；（2）近似不可观测变量的后验概率。注意到，我们在上面的推导中完全没有引入其他假设，完全是一个自然而然的结果，只是引入了一个任意的概率分布$q(Z)$，在一个特定的模型中，P的概率分布往往是预先假定的一个较为复杂的分布，而观测数据也是给定的，所以$\ln P(X)$可以视作一个常量。我们的目的现在很明确了，要去近似后验概率分布那么就要尽量使得下界尽可能接近真实值。之所以要近似这个后验分布，是因为由于P的复杂性过高，真实的后验概率$P(Z|X)$往往也是十分复杂的，我们搞不定，所以我们迫切的想要找到一种能够近似这种复杂分布的简洁的分布$q(Z)$，而这个$q(Z)$是我们有把握搞定的分布。$P(Z)$中隐变量往往不是独立的即：

$$P(Z) \neq P_1(Z_1)P_2(Z_2)\cdots P_M(Z_M) \tag{1}$$

但是我们希望我们自己选择用来近似$P(Z)$的$q(Z)$可以有$q_1(Z_1)q_2(Z_2)\cdots q_M(Z_M)$的形式，即每个维度独立分布的分布。所以我们先假设$q(Z)$具有上述形式,我们的主要任务就是让$L(q)$尽可能大。这看上去可能比较棘手，因为$L(q)$是个泛函啊，也就是函数的函数，当然这有别于复合函数，因为复合函数是值到值的映射，而泛函是函数到值的映射。但别怕，我们先做下去。

$$\begin{aligned} L(q) &= \int_Z q(Z)\ln P(X,Z)\,dZ - \int_Z q(Z)\ln q(Z)\,dZ \\ &= \underbrace{\int_Z \prod_{i=1}^{M} q_i(Z_i)\ln P(X,Z)\,dZ}_{L_1} - \underbrace{\int_Z \prod_{i=1}^{M} q_i(Z_i)\ln\prod_{i=1}^{M} q_i(Z_i)\,dZ}_{L_2} \\ L_1 &= \int_Z \prod_{i=1}^{M} q_i(Z_i)\ln P(X,Z)\,dZ \\ &= \int_{Z_1,Z_2\cdots Z_M}\cdots\int \prod_{i=1}^{M} q_i(Z_i)\ln P(X,Z)\,dZ_1dZ_2\ldots dZ_M \end{aligned}$$

假设我们只关心其中的第j个部分，L_1可以写成：

$$\begin{aligned} L_1 &= \int_{Z_j} q_j(Z_j)(\int_{Z_{i\neq j}} \cdots \int q_i(Z_i) \prod_{i\neq j}^{M} q_i(Z_i) \ln P(X,Z) \prod_{i\neq j}^{M} dZ_i) dZ_j \\ &= \int_{Z_j} q_j(Z_j)(E_{i\neq j}(\ln P(X,Z)))\, dZ_j \end{aligned}$$

注意，上式利用了积分顺序的交换，但也并非所有的积分都能交换顺序，需要满足富比尼定理（https://zh.wikipedia.org/zh-hans/富比尼定理）。为了化简L_2，我们注意到对于任意一个概率分布$p(x_1, x_2)$都会有：

$$\begin{aligned} \iint_{x_1\, x_2} (f(x_1) + f(x_2)) p(x_1,x_2)\, dx_1 dx_2 &= \iint_{x_1\, x_2} f(x_1) p(x_1,x_2)\, dx_1 dx_2 + \iint_{x_1\, x_2} f(x_2) p(x_1,x_2)\, dx_1 dx_2 \\ &= \int_{x_1} f(x_1) \int_{x_2} p(x_1,x_2)\, dx_1 dx_2 + \int_{x_2} f(x_2) \int_{x_1} p(x_1,x_2)\, dx_1 dx_2 \\ &= \int_{x_1} f(x_1) p(x_1)\, dx_1 + \int_{x_2} f(x_2) p(x_2)\, dx_2 \end{aligned}$$

所以根据这个结论：

$$\begin{aligned} L_2 &= \int \prod_{i=1}^{M} q_i(Z_i) \sum_{i=1}^{M} \ln q_i(Z_i)\, dZ \\ &= \sum_{i=1}^{M} (\int_{Z_i} q_i(Z_i) \ln q_i(Z_i)\, dZ_i) \end{aligned}$$

同理，假设我们只关心其中的第j个部分，其他部分可视为常数：

$$L_2 = \int_{Z_j} q_j(Z_j) \ln q_j(Z_j)\, dZ_j + const \tag{2}$$

所以最终$L(q)$可以简化成：

$$\begin{aligned} L(q) &= L_1 - L_2 \\ &= \int_{Z_j} q_j(Z_j)(E_{i\neq j}(\ln P(X,Z)))\, dZ_j - \int_{Z_j} q_j(Z_j) \ln q_j(Z_j)\, dZ_j \\ &= \int_{Z_j} q_j(Z_j) \frac{E_{i\neq j}(\ln P(X,Z))}{\ln q_j(Z_j)}\, dZ_j \end{aligned}$$

这个结果，大家想必不会陌生，因为这个式子出现的太频繁了，这不就很像KL散度的定义么，虽然应该还有个负号。但是KL散度应该是度量两个分布的距离，对吧？那么我们这里就定义一个伪分布$\tilde{P}(X, Z_j)$满足：

$$\ln \tilde{P}(X, Z_j) = E_{i\neq j}(\ln P(X,Z)) \tag{3}$$

所以:

$$L(q) = \int_{Z_j} q_j(Z_j) \ln \frac{\tilde{P}(X,Z_j)}{q_j(Z_j)}\, dZ_j + const = -D_{KL}(q_j(Z_j) || \tilde{P}(X,Z_j)) + const \tag{4}$$

欲使$L(q)$最大，需满足$\ln q_j^*(Z_j) = E_{i\neq j}(\ln P(X,Z)) = E_{q_1,\ldots,q_M(Z_M)/q_j(Z_j)}(\ln P(X,Z))$,所以根据这个结论可以迭代的更新$q(Z)$中的第$j$个元素直至收敛就能得到近似$P(Z)$的分布,那么变分贝叶斯的思想可以总结为下述迭代的算法结构。

Algorithm 1 Variational Bayes Inference

1: *initialize* $q(Z) = q(Z_1)q(Z_2)\cdots q(Z_M)$
2: **while** *not Converge* **do**
3: **for** $i = 1$ to M **do**
4: $\ln q_j^*(Z_j) = \int\cdots\int_{q_1\ldots q_M/q_j} \ln P(X,Z)\prod_{i\neq j} q_i(Z_i)\prod_{i\neq j} dZ_i$
5: *update the* $q_j^*(Z_j)$ *in the* $q(Z)$
6: **end for**
7: **end while**

最后我们再来探讨一个有趣的事，变分贝叶斯可以使$KL(q(Z)||P(X,Z))$尽可能的小，那如果我们需要最小化reverse Kullback-Leibler divergence即$KL(P(X,Z)||q(Z))$的话会变成怎样呢？当然这并不是变分贝叶斯的目的，我们可以通过与这种方式来对比一下变分贝叶斯中的结论。那么我们还是按之前的思路去做。

$$
\begin{aligned}
KL(P(X,Z)||q(Z)) &= -\int P(X,Z)\ln\frac{q(Z)}{P(X,Z)}dZ \\
&= -\int P(X,Z)\ln q(Z)dZ + \int P(X,Z)\ln P(X,Z)dZ \\
&= -\int P(X,Z)\ln q(Z)dZ + const \\
&= -\int P(X,Z)\sum_{j=1}^{M}\ln q_j(Z_j)dZ + const \\
&= -\sum_{j=1}^{M}\int_{Z_j}\ln q_j(Z_j)P(X,Z_j)dZ_j + const \\
&= -\int_{Z_j}\ln q_j(Z_j)P(X,Z_j)dZ_j + const \\
&= -(\int_{Z_j}\ln q_j(Z_j)P(X,Z_j)dZ_j - \int_{Z_j}\ln P(X,Z_j)P(X,Z_j)dZ_j + \\
&\quad \int_{Z_j}\ln P(X,Z_j)P(X,Z_j)dZ_j) + const \\
&= KL(P(X,Z_j)||q_j(Z_j)) + \int_{Z_j}\ln P(X,Z_j)P(X,Z_j)dZ_j + const \\
&= KL(P(X,Z_j)||q_j(Z_j)) + const
\end{aligned}
$$

所以欲最小化上式，只需$q_j^*(Z_j) = P(X,Z_j)$,可以看到此时$q_j^*(Z_j)$为数据真实分布的边缘分布时，KL距离最小。