PRML（Pattern Recognition And Machine Learning）读书会

第十章　Approximate Inference

主讲人 戴玮

（新浪微博：@戴玮_CASIA）

QQ 群 177217565

读书会微信公众平台请扫描下面的二维码

Wilbur_中博(1954123) 20:02:04

我们在前面看到，概率推断的核心任务就是计算某分布下的某个函数的期望、或者计算边缘概率分布、条件概率分布等等。 比如前面在第九章尼采兄讲 EM 时，我们就计算了对数似然函数在隐变量后验分布下的期望。这些任务往往需要积分或求和操作。 但在很多情况下，计算这些东西往往不那么容易。因为首先，我们积分中涉及的分布可能有很复杂的形式，这样就无法直接得到解析解，而我们当然希望分布是类似指数族分布这样具有共轭分布、容易得到解析解的分布形式；其次，我们要积分的变量空间可能有很高的维度，这样就把我们做数值积分的路都给堵死了。因为这两个原因，我们进行精确计算往往是不可行的。为了解决这一问题，我们需要引入一些近似计算方法。

近似计算有随机和确定两条路子。随机方法也就是 MCMC 之类的采样法，我们会在讲第十一章的时候专门讲到，而确定近似法就是我们这一章讲的变分。变分法的优点主要是：有解析解、计算开销较小、易于在大规模问题中应用。但它的缺点是推导出想要的形式比较困难。也就是说，人琢磨的部分比较复杂，而机器算的部分比较简单。这和第十一章的采样法的优缺点恰好有互补性。所以我们可以在不同的场合应用变分法或采样法。这里我的一个问题是：是否可以结合二者的优点，使得人也不用考虑太多、机器算起来也比较简单？

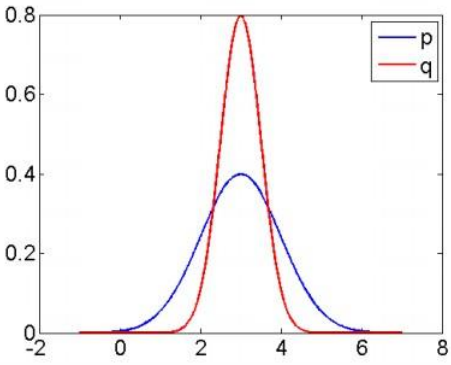
变分法相当于把微积分从变量推广到函数上。我们都知道，微积分是用来分析变量变化、也就是函数性质的，这里函数定义为 f: x -> f(x)，而导数则是 df/dx；与之相对，变分用到了泛函的概念：F: f -> F(f)，也就是把函数映射为某个值，而相应地，也有导数 dF/df，衡量函数是如何变化的。比如我们熟悉的信息论中的熵，就是把概率分布这个函数映射到熵这个值上。和微积分一样，我们也可以通过导数为 0 的条件求解无约束极值问题，以及引入拉格朗日乘子来求解有约束极值问题。比如说，我们可以通过概率分布积分为 1 的约束，求解最大熵的变分问题。PRML 的附录 D 和 E 有比较详细的解释，我们后面也还会看到，这里就不多说了。

变分法这名字听起来比较可怕，但它的核心思想，就是从某个函数空间中找到满足某些条件或约束的函数。我们在统计推断当中用到的变分法，实际上就是用形式简单的分布，去近似形式复杂、不易计算的分布，这样再做积分运算就会容易很多。 比如，我们可以在所有高斯分布当中，选一个和目标分布最相似的分布，这样后面做进一步计算时就容易获得解析解。此外，我们还可以假设多元分布的各变量之间独立，这样积分的时候就可以把它们变成多个一元积分，从而解决高维问题。这也是最简单的两种近似。

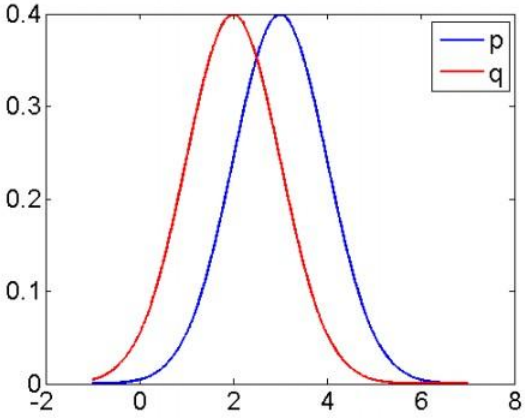
概率推断中的变分近似方法，最根本的思想，就是想用形式简单的分布去近似形式复杂、不易计算的分布。比如，我们可以在指数族函数空间当中，选一个和目标分布最相像的分布，这样计算起来就方便多了。显然，我们这里需要一个衡量分布之间相似性或差异性的度量，然后我们才能针对这个度量进行最优化，求相似性最大或差异性最小的分布。一般情况下，我们会选用 KL 散度：

$$KL(q||p) = E_{q(x)}[\log(q(x)) - log(p(x))]$$

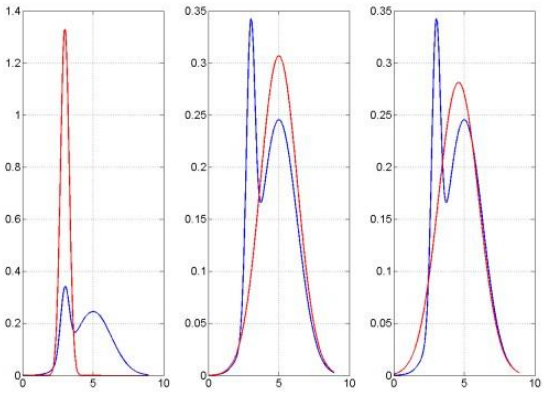
或者 $$KL(q||p) = \int q(x)\log\frac{q(x)}{p(x)}dx$$ ，当然离散分布就不是积分而是在离散状态上求和。这个值是非负的，而且只在两分布完全相同的情况下取 0，所以可以看成两分布之间的距离。但这种度量是不对称的，也就是 $KL(q||p) \neq KL(p||q)$ ，而我们在优化的时候，这两种度量实际上都可以使用。这样一来，我们后面也会看到，会造成一些有趣且奇怪的现象。有了这个度量，我们就可以对某个给定的概率分布，求一个在某些条件下和它最相似或距离最小的分布。这里我们看几个例子，直观地认识一下 KL 散度的不对称性、以及产生这种不对称性的原因。这是两个方差不同的一元高斯分布，其中方差较小的是 q（红色曲线），方差较大的是 p（蓝色曲线）：



根据KL散度的公式 $$KL(q||p) = \int q(x)\log\frac{q(x)}{p(x)}dx$$,我们能否估计一下 ,是KL(q||p)较大 ,还是KL(p||q)较大？我们可以看到，在曲线的中间部分，q(x) > p(x)，因此，如果光考虑这部分，显然 KL(q||p)会比较大。但是，考虑两边 q(x) < p(x) 的部分，我们可以看到，q(x) 很快趋近于 0，此时 p(x)/q(x) 会变得很大，比中间部分要大得多（打个比方，0.8/0.4 和 0.01/0.001）。尽管还要考虑 log 前面的 q(x)，但当 q(x) 不等于 0 时，分母趋近于 0 造成的影响还是压倒性的。所以综合考虑，KL(q||p)要小于 KL(p||q)。它们的精确值分别为 0.32 和 0.81。另一个例子是，如果两个高斯分布方差相等，则 KL 散度也会相等：

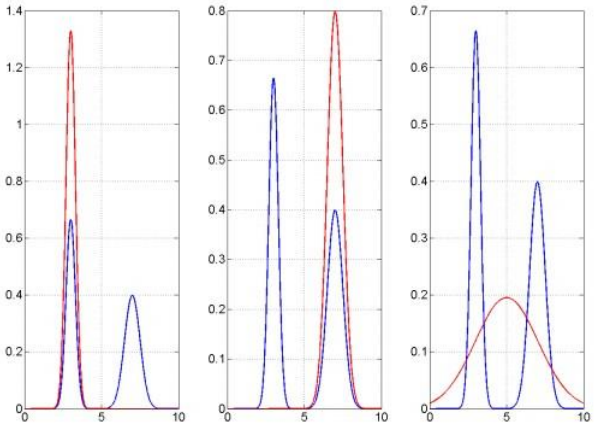


这一点很容易理解。再来看一个复杂一点的例子。在这个例子中，q 是单峰高斯分布，p 是双峰高斯分布：



这三种情况中，p 的两个峰没有分开，有一定粘连，而 q 则分别拟合了 p 的左峰、右峰（见 PRML 4.4 节的拉普拉斯近似，上次读书会也简单介绍过，可参看上次读书会的总结），以及拟合 p 的均值和方差（即单峰高斯分布的两个参数）。三种拟合情况对应左、中、右三图。对于这三种情况，KL(q||p)分别为 1.17、0.09、0.07，KL(p||q)分别为 23.2、0.12、0.07。可以看到，无论是哪一种 KL 散度，在 p 的双峰没有完全分开的情况下，用单峰高斯 q 去近似双峰高斯 p 得到的最优解，都相当于拟合 p 的均值和方差。如果 p

的两个峰分开的话，情况会如何呢？

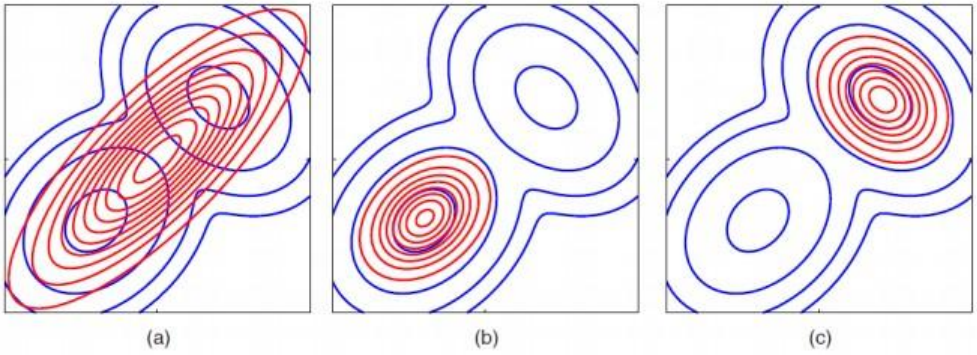


和前一个例子一样，我们分别拟合 p 的左峰、右峰，以及均值和方差。显然，这里由于 p 中间有一段概率密度为 0 的区域，所以可以想见，KL(q||p)可能会比较大。实际情况也是如此：KL(q||p)分别为 0.69、0.69、3.45，KL(p||q)分别为 43.9、15.4、0.97。可以看到，如果用 KL(p||q)做最优化，结果和双峰粘连时一样，仍然是拟合 p 的均值和方差，也就是所谓的 moment-matching；而用 KL(q||p)做最优化，结果则会有所变化：会拟合双峰的其中一峰，也就是所谓的 mode-seeking。

我们从前面这几个例子中，可以总结一个规律：用 KL(q||p)做最优化，是希望 p(x)为 0 的地方 q(x)也要为 0，否则 q(x)/p(x)就会很大，刚才例子的右图在中间部分（5 附近）就违背了这一点；反之，如果用 KL(p||q)做最优化，就要尽量避免 p(x)不为 0 而 q(x)用 0 去拟合的情况，或者说 p(x)不为 0 的地方 q(x)也不要为 0，刚才例子的左、中两图也违反了这一点。

所以，KL(q||p)得到的近似分布 q(x)会比较窄，因为它希望 q(x)为 0 的地方可能比较多；而 KL(p||q)得到的近似分布 q(x)会比较宽，因为它希望 q(x)不为 0 的地方比较多。

最后看一个多元高斯分布的例子，书上的图 10.3：



即有了前面的讲解，我们可以猜一下，哪些图是 KL(q||p)得到的最优解，哪些图是 KL(p||q)得到的最优解。由于 KL(q||p)至少可以拟合到其中的一个峰上，而 KL(p||q)拟合的结果，其概率密度最大的地方可能没什么意义，所以很多情况下，KL(q||p)得到的结果更符合我们的需要。到这里有什么问题吗。。理解理解。。KL 散度这东西。

==========讨论==========

飞羽(346723494) 20:24:23

KL(q||p) 就是相当于用 q 去拟合 p？

Yuli(764794071) 20:25:31

KL 就是 KL Divergence（相对熵）吧 用信息论来解释的话 是用来衡量两个正函数是否相似

飞羽(346723494) 20:25:57

对， 就是相对熵

Wilbur_中博(1954123) 20:27:06
嗯，我们现在有一个分布 p，很多时候是后验分布，但它形式复杂，所以想用形式比较简单的 q 去近似 p。其实也可以直接用后验分布的统计量，比如 mode 或 mean 去代替整个分布，进行进一步计算，比如最大后验什么的。但现在如果用近似分布去做预测的话，性能会好得多。
linbo-phd-bayesian(99878724) 20:27:15
请问为何 KL(q||p)》=0，为何没有《0 啊，有知道的吗？
飞羽(346723494) 20:28:06

相对熵的值为非负数：

$$D_{\mathrm{KL}}(P\|Q) \geq 0,$$

由吉布斯不等式（en:Gibbs' inequality）可知，当且仅当$P = Q$时$D_{\mathrm{KL}}(P|Q)$为零。

尽管从直觉上KL散度是个度量或距离函数，但是它实际上并不是一个真正的度量或距离。因为KL散度不具有对称性：从分布P到Q的距离（或度量）通常并不等于从Q到P的距离（或度量）。

$$D_{\mathrm{KL}}(P\|Q) \neq D_{\mathrm{KL}}(Q\|P)$$

Wilbur_中博(1954123) 20:29:21
那个不太难证，利用 ln 凹函数性质可以证出来。。不过细节我忘记了，呵呵。查一查吧。。应该很多地方都有的。
逸风(421723497) 20:30:44
PRML P56
Wilbur_中博(1954123) 20:31:50
总之就是利用 KL 作为目标函数，去做最优化。。找到和已知复杂分布最相近的一个近似分布。这一章的基本思路就是这样。具体动机最开始的时候已经提到过了。
逸风(421723497) 20:35:31
为什么要用 KL 散度这样一个不具备对称性的“距离”，而不采用对称性的测度呢？有什么好处?
Wilbur_中博(1954123) 20:37:15
似乎没有特别好的对称的度量？PRML 的公式(10.20)提过一种叫 Hellinger distance 的度量，是对称的，但后来也没有用这个。不知道为什么。不容易优化？有没有了解原因的朋友。。比如说，为啥不用 (p(x) - q(x) ^ 2 做积分作为度量？ 或者其他什么的。
WayneZhang(824976094) 20:41:52
我感觉是优化求解过程中近似时自然而然导出了 KL 这个度量。
karnon(447457116) 20:42:24
KL 算的是熵的增益，所以一定是那种形式 ，这取决于你怎么定义“近似”，认为信息增益最少就是“近似”也是一种合理的定义
Wilbur_中博(1954123) 20:43:07
这里目的是为了找近似分布
==========讨论结束==========
我们在 PRML 这本书的 4.4 节，其实看到过一种简单的近似方法，或者可以说是最简单的近似方法之一：拉普拉斯近似。它是用高斯分布去近似目标分布的极值点也就是 mode。这里并没有涉及到变分的概念。它只是要求高斯分布的 mode 和目标分布的 mode 位置相同，方法就是把目标分布在 mode 处做泰勒级数展开到第二阶，然后用对应的高斯分布去代替，就是把未知系数给凑出来：

$$\mathcal{L}(\theta^*) + \frac{1}{2}\,\mathcal{L}''(\theta^*)\,(\theta - \theta^*)^2$$

这是目标分布在\theta^*（mode）的二阶泰勒展开：

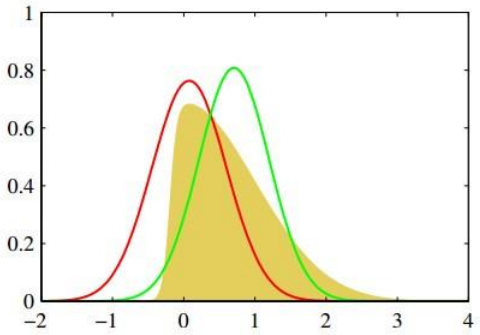
$$\begin{aligned}\ln \mathcal{N}(\theta|\mu,\eta^{-1}) &= \frac{1}{2}\ln\eta - \frac{1}{2}\ln 2\pi - \frac{\eta}{2}(\theta-\mu)^2 \\ &= \frac{1}{2}\ln\frac{\eta}{2\pi} + \frac{1}{2}(-\eta)(\theta-\mu)^2\end{aligned}$$

一比较就知道高斯分布的两个参数应该取：

$$\mu = \theta^*$$

$$\eta = -\mathcal{L}''(\theta^*)$$

也就是 PRML 图 10.1 的红线：



棕色部分是目标分布，绿线是我们用变分近似，在高斯分布中选一个和目标函数的 KL 散度最小的分布。反正就均值和方差两个未知参数，优化起来应该不难。

下面开始讲 10.1.1 可分解分布，这一节非常重要，可以说是本章的基础和最重点的部分。基本思想就是，我们把近似分布限制在可分解分布的范围内，也就是满足(10.5)式：

$$q(\mathbf{Z}) = \prod_{i=1}^{M} q_i(\mathbf{Z}_i)$$

可以说，这个分布的各组变量 Z_i 互相之间是独立的。这样一来，我们计算这个分布的积分时，就可以变成多个较低维度的积分，就算用数值积分什么的也会简单很多。 在统计物理当中，这种可分解形式的变分近似方法称作平均场（mean field）方法，这个名字实际上是很直观的，和它最后得到的解的形式有关，我们马上会看到。不过现在不仅在统计物理领域，机器学习很多时候也就管它叫 mean field 了。现在很火的 RBM 什么的，求参数时经常能看到这个术语。

上一章曾经讲过，最小化 KL 距离，和最大化下界 L(q)是一回事，也就是(10.2)到(10.4)这三个式子：

$$\ln p(\mathbf{X}) = \mathcal{L}(q) + \mathrm{KL}(q\|p)$$

$$\mathcal{L}(q) = \int q(\mathbf{Z}) \ln\left\{\frac{p(\mathbf{X},\mathbf{Z})}{q(\mathbf{Z})}\right\} \mathrm{d}\mathbf{Z}$$

$$\mathrm{KL}(q\|p) = -\int q(\mathbf{Z}) \ln\left\{\frac{p(\mathbf{Z}|\mathbf{X})}{q(\mathbf{Z})}\right\} \mathrm{d}\mathbf{Z}.$$

这和 9.4 节当中(9.70)到(9.72)实际上是一样的，区别在于 Z 不仅是隐变量还把参数吸收了进来。等式左边那项和我们想求的 Z 无关，所以可以看成常数，而右边的 p(Z|X)是我们想去近似的，不知道具体形式，所以可以间接通过最大化右边第一项来达到最小化右边第二项也就是 KL 散度的目的。

根据上面的(10.5)式会得到公式(10.6)：

$$\begin{aligned}\mathcal{L}(q) &= \int \prod_i q_i \left\{ \ln p(\mathbf{X}, \mathbf{Z}) - \sum_i \ln q_i \right\} \mathrm{d}\mathbf{Z} \\ &= \int q_j \left\{ \int \ln p(\mathbf{X}, \mathbf{Z}) \prod_{i \neq j} q_i \, \mathrm{d}\mathbf{Z}_i \right\} \mathrm{d}\mathbf{Z}_j - \int q_j \ln q_j \, \mathrm{d}\mathbf{Z}_j + \text{const} \\ &= \int q_j \ln \widetilde{p}(\mathbf{X}, \mathbf{Z}_j) \, \mathrm{d}\mathbf{Z}_j - \int q_j \ln q_j \, \mathrm{d}\mathbf{Z}_j + \text{const} \qquad (10.6)\end{aligned}$$

我们这里也可以看 MLAPP 的(21.28)到(21.31)：

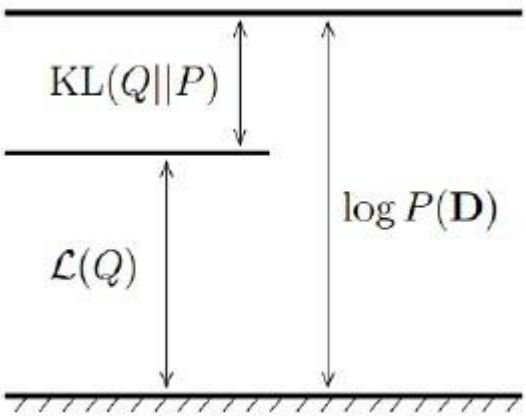
$$\begin{aligned}L(q_j) &= \sum_{\mathbf{x}} \prod_i q_i(\mathbf{x}_i) \left[\log \tilde{p}(\mathbf{x}) - \sum_k \log q_k(\mathbf{x}_k) \right] \\ &= \sum_{\mathbf{x}_j} \sum_{\mathbf{x}_{-j}} q_j(\mathbf{x}_j) \prod_{i \neq j} q_i(\mathbf{x}_i) \left[\log \tilde{p}(\mathbf{x}) - \sum_k \log q_k(\mathbf{x}_k) \right] \\ &= \sum_{\mathbf{x}_j} q_j(\mathbf{x}_j) \sum_{\mathbf{x}_{-j}} \prod_{i \neq j} q_i(\mathbf{x}_i) \log \tilde{p}(\mathbf{x}) \\ &\quad - \sum_{\mathbf{x}_j} q_j(\mathbf{x}_j) \sum_{\mathbf{x}_{-j}} \prod_{i \neq j} q_i(\mathbf{x}_i) \left[\sum_{k \neq j} \log q_k(\mathbf{x}_k) + q_j(\mathbf{x}_j) \right] \\ &= \sum_{\mathbf{x}_j} q_j(\mathbf{x}_j) \log f_j(\mathbf{x}_j) - \sum_{\mathbf{x}_j} q_j(\mathbf{x}_j) \log q_j(\mathbf{x}_j) + \text{const}\end{aligned}$$

推导得要详细得多。。所以多备几本参考书是必要的。

MLAPP 是 Machine Learning - A Probabilistic Perspective 的缩写。。群共享里应该有吧。很不错的机器学习书。

huajh7(284696304) 21:02:10

插一句。这里优化的目标其实是最大化 low bound L(q) （log P(D)是对数证据，常数，KL(Q||P)=0 时，L(Q)最大）。也就是找到一个最合适的 q 分布，而不是优化参数。 优化过程中，求导，拉格朗日什么，是针对 q 分布的，也就是泛函。 这是为什么叫变分法：



Wilbur_中博(1954123) 21:03:04

好，谢谢。我看了你的博客 http://www.blog.huajh7.com/variational-bayes/，文章写得很好。你好像毕业论文就是专门做这个的吧？ 也许你下次可以再专门讲一讲你对变分近似的心得体会，呵呵。

简单说，这里的推导就是每一步只看 q_j 相关的那些项，和 q_j 无关的项全都归到常数里去。比如(21.30)的这部分：

$$\sum_{\mathbf{x}_j} q_j(\mathbf{x}_j) \sum_{\mathbf{x}_{-j}} \prod_{i \neq j} q_i(\mathbf{x}_i) \left| \sum_{k \neq j} \log q_k(\mathbf{x}_k) \right.$$

实际上就全扔到常数里去了。哦。。还少了个(21.32)：

where

$$\log f_j(\mathbf{x}_j) \triangleq \sum_{\mathbf{x}_{-j}} \prod_{i \neq j} q_i(\mathbf{x}_i) \log \tilde{p}(\mathbf{x}) = \mathbb{E}_{-q_j} [\log \tilde{p}(\mathbf{x})]$$

这里我们是在除了 x_j 之外的其他 x_i 上求期望，也就是这个东西：

$\mathbb{E}_{-q_j} [\log \tilde{p}(\mathbf{x})]$ ，它是关于 x_j 的函数。

下面讲一下 10.1.1 的可分解分布，也就是刚才说过的，假设多元分布可分解为多个一元分布的乘积，即

$q(\mathbf{x}) = \prod_i q_i(\mathbf{x}_i)$

用 去近似 p(x)。由于各个变量之间是解耦的，所以我们可以每次只关注单个变量的最优化，也就是用所谓坐标下降（coordinate descent）的方式来做最优化。具体做法，就是把最小化 KL 散度转化为最大化 L(q)（参见公式(10.2)到(10.4)），然后把公式(10.5)代入(10.3)，每次把 L(q)其中一个 q_j 当做变量，而把其他 q_i 当做常数，对 L(q)进行最优化：

$$\begin{aligned}
\mathcal{L}(q) &= \int \prod_i q_i \left\{ \ln p(\mathbf{X}, \mathbf{Z}) - \sum_i \ln q_i \right\} \mathrm{d}\mathbf{Z} \\
&= \int q_j \left\{ \int \ln p(\mathbf{X}, \mathbf{Z}) \prod_{i \neq j} q_i \, \mathrm{d}\mathbf{Z}_i \right\} \mathrm{d}\mathbf{Z}_j - \int q_j \ln q_j \, \mathrm{d}\mathbf{Z}_j + \text{const} \\
&= \int q_j \ln \widetilde{p}(\mathbf{X}, \mathbf{Z}_j) \, \mathrm{d}\mathbf{Z}_j - \int q_j \ln q_j \, \mathrm{d}\mathbf{Z}_j + \text{const} \qquad (10.6)
\end{aligned}$$

这里： $\ln \widetilde{p}(\mathbf{X}, \mathbf{Z}_j) = \mathbb{E}_{i \neq j}[\ln p(\mathbf{X}, \mathbf{Z})] + \text{const.} \qquad (10.7)$

前面讲过，KL 散度也可以写成： $KL(q||p) = E_{q(x)}[\log(q(x)) - log(p(x))]$ ，可以看到，(10.6)最后得到的这个 $\int q_j \ln \widetilde{p}(\mathbf{X}, \mathbf{Z}_j) \, \mathrm{d}\mathbf{Z}_j - \int q_j \ln q_j \, \mathrm{d}\mathbf{Z}_j$ ，恰好是负 KL 散度的形式。我们知道，KL 散度为 0 也就是最小的时候，两分布恰好相同，因此每一步的最优化结果可得到：

$$\ln q_j^\star(\mathbf{Z}_j) = \mathbb{E}_{i \neq j}[\ln p(\mathbf{X}, \mathbf{Z})] + \text{const.}$$

也就是每一步更新的结果，可得到分解出来的变量的分布为：

$$q_j^\star(\mathbf{Z}_j) = \frac{\exp\left(\mathbb{E}_{i \neq j}[\ln p(\mathbf{X}, \mathbf{Z})]\right)}{\int \exp\left(\mathbb{E}_{i \neq j}[\ln p(\mathbf{X}, \mathbf{Z})]\right) \, \mathrm{d}\mathbf{Z}_j}.$$

就是两边都取 exp 然后归一化。由于是以其他变量的均值作为当前变量分布的形式，所以这种方法也称作 mean-field。这部分内容也可以参见 MLAPP 的 21.3.1，那一节讲得感觉比 PRML 清楚一些。 那个公式是比较头疼。。不过只要记住只有一个 q_j 是变量，其他都当成常数，推一推应该也 ok。

重新回顾下前面的内容：

变分推断的核心思想是：用形式比较简单、做积分运算比较容易的分布，去替代原先形式复杂、不易求积分的分布。因此，这里的主要问题就是：如何找到和原分布近似程度较高的简单分布。前面我们讲了一些变分推断的背景知识和 KL divergence（KLD）的相关知识，还稍微讲了讲假设分布可分解时是如何推导出 mean field 形式的。KLD 是衡量两个分布差异大小的方式之一，KLD 越大则差异越大，反之则两分布越相似。因此，我们可以将 KL(q||p)作为目标函数，并限定 q 为较简单的分布形式，找到这类分布中最接近原分布 p 的那个分布。我们这里主要关注的近似对象是后验分布。因为我们前面一直在讲如何求后验分布，但后验分布求出来的形式往往不那么好用，所以需要用简单分布去近似。然而，计算 p(Z|X)需要计算归一化因子 p(X)。p(X)是边缘分布，需要对 p(Z,X)做积分，而 p(Z,X)又不那么容易积分。因此，我们可以直接用未归一化的 p(Z,X)作为近似计算的目标，也就是下面这个关系：

$$\ln p(\mathbf{X}) = \mathcal{L}(q) + \mathrm{KL}(q\|p) \tag{10.2}$$

其中：

$$\mathcal{L}(q) = \int q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{X},\mathbf{Z})}{q(\mathbf{Z})} \right\} \mathrm{d}\mathbf{Z} \tag{10.3}$$

$$\mathrm{KL}(q\|p) = -\int q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{Z}|\mathbf{X})}{q(\mathbf{Z})} \right\} \mathrm{d}\mathbf{Z}. \tag{10.4}$$

这里 ln(p(X))只是个常数，所以极小化 KLD 和极大化 L 得到的结果是一样的，但对 L 做最优化可直接用联合概率分布去做、而不用归一化。:我们想要得到的简单分布具有什么样的形式？我们喜欢的一种简单分布是可分解分布，就是说，我们可以假设各个隐变量 Z_i 之间是独立的，因此可拆成各隐变量分布的乘积：

$$q(\mathbf{Z}) = \prod_{i=1}^{M} q_i(\mathbf{Z}_i). \tag{10.5}$$

那么，各个隐变量的 L 可写为：

$$\begin{aligned}
\mathcal{L}(q) &= \int \prod_i q_i \left\{ \ln p(\mathbf{X},\mathbf{Z}) - \sum_i \ln q_i \right\} \mathrm{d}\mathbf{Z} \\
&= \int q_j \left\{ \int \ln p(\mathbf{X},\mathbf{Z}) \prod_{i\neq j} q_i \, \mathrm{d}\mathbf{Z}_i \right\} \mathrm{d}\mathbf{Z}_j - \int q_j \ln q_j \, \mathrm{d}\mathbf{Z}_j + \mathrm{const} \\
&= \int q_j \ln \widetilde{p}(\mathbf{X},\mathbf{Z}_j) \, \mathrm{d}\mathbf{Z}_j - \int q_j \ln q_j \, \mathrm{d}\mathbf{Z}_j + \mathrm{const}
\end{aligned} \tag{10.6}$$

其中

$$\ln \widetilde{p}(\mathbf{X},\mathbf{Z}_j) = \mathbb{E}_{i\neq j}[\ln p(\mathbf{X},\mathbf{Z})] + \mathrm{const}. \tag{10.7}$$

这里 $\mathbb{E}_{i\neq j}$ 表示：

$$\mathbb{E}_{i\neq j}[\ln p(\mathbf{X},\mathbf{Z})] = \int \ln p(\mathbf{X},\mathbf{Z}) \prod_{i\neq j} q_i \, \mathrm{d}\mathbf{Z}_i. \tag{10.8}$$

这是对 Z_j 之外的其他所有随机变量求期望，也叫做 mean field。极大化 L 相当于极小化 $KL(q_j||\tilde{p})$，显然 q_j 取和 $\widetilde{p}$ 完全相同的形式时，KLD 极小，同时 L 极大。所以我们有最优解：

$$\ln q_j^\star(\mathbf{Z}_j) = \mathbb{E}_{i\neq j}[\ln p(\mathbf{X},\mathbf{Z})] + \text{const.} \qquad (10.9)$$

这里 p 是已知的，所以可对它做积分。对除 Z_j 以外的随机变量求期望得到的分布，就是分解出来的 q_j 的分布。我们每一步迭代都对每一个分解分布 q_j 进行求解。这种方法也称做 coordinate descent。

==========讨论==========

一夏 吕(992463596) 21:15:06

10.6 后面有，10.7 的 const 没有必要吧？我当时好像看懂了 做的笔记 现在一下看不懂了。。 后面那个很简单 是因为 其他的 Zi 积分为 1。我记起来了 ，把后面的 lnqi 的和拆开，只把 j 的那一项留着，其他的都可以积分积掉，划到 const 里，这里主要是吧 j 的那一项拿出来表示，其他的 不相干的都不管。

huajh7(284696304) 21:25:23

有必要吧。否则不相等了，这里 const 表示归一化常量。实际上需要特别注意 const，尤其自己推导的时候，const 更多是表示与 z_j 无关的量，而不是指一个常量。在概率图中就是不在 z_j 的马尔科夫毯上的量。

阿邦(1549614810) 21:26:08

mean filed 看 koller 的最清楚

一夏 吕(992463596) 21:26:57

注意 Z 是大写，所以 j 的那个积分里 其他的 i 都积分为 1 了。

huajh7(284696304) 21:27:48

const 有必要。exp(E_{i~=j}[..]) 是没有归一化的。

一夏 吕(992463596) 21:28:23

哪里有 exp

huajh7(284696304) 21:28:32

ln .

软件所-张巍<zh3f@qq.com> 21:26:42

问个问题：用分解的分布去近似原始分布，精度怎么保证，有没有直观点的解释。

Wilbur_中博(1954123) 21:28:32

@软件所-张巍 的问题是好问题啊。。一般来说，似乎是把变分近似看作在 MAP 和贝叶斯推断（用整个分布）之间的一种 trade-off？

huajh7(284696304) 21:29:54

给个图 ：

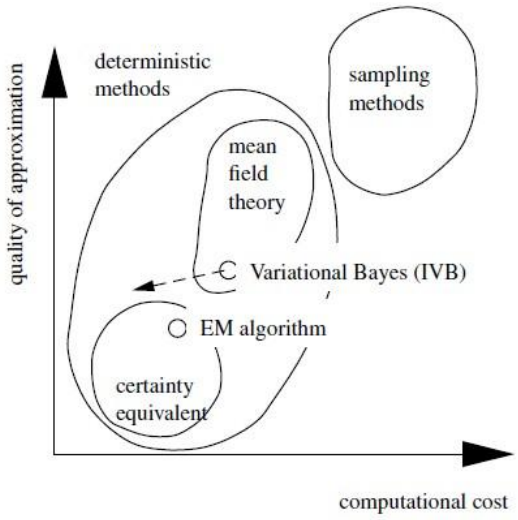


Fig. 1.3. The accuracy-vs-complexity trade-off in the VB method.

Wilbur_中博(1954123) 21:30:01
因为一个是用后验分布的点估计，一个是用整个分布，不错，这是哪里的图？
huajh7(284696304) 21:32:23
variational bayeian 可以说是分布式 distributional approximation，也就是 wilbur 说的，用的是整个分布。 The Variational Bayes Method in Signal Processing 这本书的 第 9 页。
李笑一(94755934) 21:32:46
@张巍，我记得变分法能保证收敛到 local minimum。一般情况下，最大似然是 non convex 的，但是变分下界却是 convex，下界的 minimum 就是下一步要前进的方向。
一夏 吕(992463596) 21:33:23
但是变分法的前提是把 dependence 去掉了 ,这样才能把总概率拆开成各自概率的积。即使是 convex 的 ，也只是逼近原先 intractable 的形式。10.7 的那个我还是觉得 const 没必要。
Wilbur_中博(1954123) 21:36:04
其实就是没归一化的，所以要加个 const，(10.9)那个也是这样
一夏 吕(992463596) 21:36:17
后面那个是求期望，就是上面那个花括号里的
李笑一(94755934) 21:36:30
@huajh7，图上看来，EM 更好使？？？
一夏 吕(992463596) 21:37:16
EM 是可解的时候用的，只是有隐变量
秦淮/sun 人家(76961223) 21:37:31
EM 是可以求得精确地后验
Happy(467594602) 21:38:30
直观解释 请参照 jordan 写的 introduction
huajh7(284696304) 21:38:33
10.6-10.9 就是利用 KL（q||p）=0

$$q_j^\star(\mathbf{Z}_j) = \frac{\exp\left(\mathbb{E}_{i\neq j}[\ln p(\mathbf{X},\mathbf{Z})]\right)}{\int \exp\left(\mathbb{E}_{i\neq j}[\ln p(\mathbf{X},\mathbf{Z})]\right)\,\mathrm{d}\mathbf{Z}_j}.$$

分母就是 const，VB 也可以看成是 EM。
一夏 吕(992463596) 21:41:00
@Happy jordan 的 introduction 是他的那本书吗？
Happy(467594602) 21:41:20
introduction to variational methods in graphical model
用简单分布的族 把复杂分布包裹起来 ，然后复杂分布的每一点都有一个简单分布的参数来近似
一夏 吕(992463596) 21:42:47
thanks 他还有一本书 是 Graphical Models, Exponential Families, and Variational Inference
huajh7(284696304) 21:43:25
Neal,HintonA view of the EM algorithm that justifies incremental, sparse,and other variants.pdf
这篇文章 说 EM，其实就是变分贝叶斯。
Happy(467594602) 21:43:25
后一本太难了。。
李笑一(94755934) 21:43:26

@huajh，弱弱的问一下，分母将 Z marginalize 掉这步只是在推导中出现是吧，编程的时候不会出现实际的过程？
huajh7(284696304) 21:44:06
写程序的时候，还是还归一化的。比如，GMM 中的隐变量，全部算出来之后，然后再归一化。
一夏 吕(992463596) 21:45:37
如果隐变量很多不是 exponential 个组合了
huajh7(284696304) 21:46:01
就转化为 exponential
Wilbur_中博(1954123) 21:46:01
mean field 的过程中呢？每个 Z_j 的分布也都要归一化么？@huajh7
Happy(467594602) 21:46:16
我咋记得不用归一化。。mean-field
一夏 吕(992463596) 21:46:18
那就很费时间
huajh7(284696304) 21:46:45
后来会知道。算的是充分统计量。
一夏 吕(992463596) 21:46:46
如果有 64 个，每个 01 分布就是 2^64 次方
李笑一(94755934) 21:47:33
恩，partition function 永远是问题
Happy(467594602) 21:47:35
程序里面没有归一化步骤吧，推导中体现了
李笑一(94755934) 21:48:12
啥叫充分统计量？
huajh7(284696304) 21:48:16
概率才归一化啊。
Happy(467594602) 21:48:30
指数族里面有
huajh7(284696304) 21:48:35
充分统计量能完全表示一个分布。对
一夏 吕(992463596) 21:49:29
不用归一化吧，看 10.9 10.10 中间那个公式下面那段话
huajh7(284696304) 21:49:42
为什么是指数族。。。一个最主要的原因就是其充分统计量是可计算的
Happy(467594602) 21:50:03
这个 jordan 后面那个书有深入介绍。。
一夏 吕(992463596) 21:51:09
通常不需要求出分布，而是得到分布的类型和参数
huajh7(284696304) 21:51:33
@一夏 吕 可能理解不一样。归一化不是指计算那个积分(partition function)..
一夏 吕(992463596) 21:51:41
通常就是指数族，自然服从积分为 1
Happy(467594602) 21:52:25
没有自然哈 也有归一化系数

一夏 吕(992463596) 21:53:00
恩 但是那个是和分布本身有关的，知道了参数就可以推，比如高斯的方差
李笑一(94755934) 21:56:38
这部分有没有类似的书写的不错的。直接讲替代教材得了
Happy(467594602) 21:57:29
jordan 那个不错，不过主要是针对 graphical model 的
Wilbur_中博(1954123) 21:58:14
我除了 PRML 和 MLAPP，还看了一下 Bayesian Reasoning and Machine Learning 的最后一章
一夏 吕(992463596) 21:58:16
有看多 lda 的吗 那个里面的 variational inference 和这个方法完全不同
Happy(467594602) 21:58:28
肿么不同。。
秦淮/sun 人家(76961223) 21:58:33
@一夏 吕 其实是一样的
huajh7(284696304) 21:58:33
bishop 不喜欢详细推导的。讲清楚就行。这里有篇:
A Tutorial on Variational Bayesian Inference（ http://staffwww.dcs.shef.ac.uk/people/c.fox/fox_vbtut.pdf） 还是很清楚的 。LDA 。其实是一样的。。建立的图模型上，比较直观。
秦淮/sun 人家(76961223) 21:59:37
LDA 那篇文章就是使用的 mean field
一夏 吕(992463596) 21:59:38
blei 用拉格朗日乘子法做的。。
Happy(467594602) 21:59:46
一样的啊。。
秦淮/sun 人家(76961223) 22:00:02
不同的优化方法而已……
huajh7(284696304) 22:00:03
嗯。其实是一样的。
秦淮/sun 人家(76961223) 22:00:12
本质是一样的
Happy(467594602) 22:00:21
直觉一致
秦淮/sun 人家(76961223) 22:00:26
不一样的是 Expectation propagation 那篇
huajh7(284696304) 22:00:37
对。那个感觉有些难。
一夏 吕(992463596) 22:00:40
恩 也是搞 kl 距离 方法各不相同
Happy(467594602) 22:01:52
对这些有兴趣就看 jordan 的大作吧，这些全部都归到架构里去了
一夏 吕(992463596) 22:05:45
http://www.cs.princeton.edu/courses/archive/fall11/cos597C/lectures/variational-inference-i.pdf
推荐这个 blei 的讲义

一夏 吕(992463596) 22:09:18
variational inference 是不是只是对指数族的才有用?
Happy(467594602) 22:09:26
一样的 统计模型下
一夏 吕(992463596) 22:10:04
我一般只在贝叶斯学派的文章里见到，一般都用 Gibbs sampling
Happy(467594602) 22:10:16
也不一定。。
一夏 吕(992463596) 22:11:00
比如 rbm 就不能用 variational inference
Happy(467594602) 22:11:21
可以啊，mean-field 必须可以用
天际一线(1002621044) 22:19:40
lda 那个话题模型 谁有完整的算法啊
Happy(467594602) 22:23:48
lda 老模型了吧。。程序应该多如牛毛
秦淮/sun 人家(76961223) 22:24:22
对啊，mean field ，expectation propagation ，gibbs sampling，distributed ，online 的都有一堆
Matrix(690119781) 22:28:18
https://github.com/sudar/Yahoo_LDA 这个可能满足你的要求
陪你听风(407365525) 22:31:18
在效果上，variational inference，gibbs sampling 两个谁更好呢
秦淮/sun 人家(76961223) 22:38:35
sampling 近似效果好，慢，不好分布式计算
陪你听风(407365525) 22:39:08
vb 比较容易分布式吗
huajh7(284696304) 22:40:24
噗。VB 是可以很自然地分布式的。。
李笑一(94755934) 22:42:16
弱问。。VB 为啥自然可以用分布式
huajh7(284696304) 22:45:22
利用 variational message passing 框架下即可。。。节点之间传递充分统计量。充分统计量(一阶矩，二阶矩) 的 consensus 或 diffusion 是有较多 paper 的。图模型中的 BP 或 loopy BP 算一种分布式嘛？
李笑一(94755934) 22:48:28
有个问题，不同问题的 vb 是否需要自己推导出来？不能随意套用别人的推导呢？
huajh7(284696304) 22:49:05
推导框架。如出一辙。。但自己推并不容易的。
李笑一(94755934) 23:04:01
karnon，一篇 jmlr 的文章，在一个问题上证了 vb 的全局解
Global Analytic Solution of Fully-observed Variational Bayesian Matrix Factorization
看明白了给讲讲。。。
huajh7(284696304) 23:10:47
2,3 年前就出来了。。这篇估计是 combined and extended。

light.(513617306) 23:15:13
这个是证明了在矩阵分解这个问题上的全局最有，不证明在其他模型上也是这样。》？
karnon(447457116) 23:15:34
这就已经很牛了
李笑一(94755934) 23:17:43
vb 对于不同问题有不同的解，我觉得除非熟到一定程度了，否则不可能拿来一个问题就能用 vb 的
karnon(447457116) 23:21:29
我看看，我知道最近有些文章研究全局收敛的矩阵分解问题，粗翻了一下，好像说的是把 vb 转成一个等价的 svd 问题？
==========讨论结束==========

接着主要讲几个变分推断的例子，试图阐述清楚变分推断到底是如何应用的。首先是二元高斯分布的近似。我们假设二元高斯分布是可分解的，也就是两变量之间独立。

二元高斯分布 $p(\mathbf{z}) = \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1})$

其中

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \qquad \boldsymbol{\Lambda} = \begin{pmatrix} \Lambda_{11} & \Lambda_{12} \\ \Lambda_{21} & \Lambda_{22} \end{pmatrix} \tag{10.10}$$

可分解形式为：$q(\mathbf{z}) = q_1(z_1)q_2(z_2)$

我们想用 q(z)去近似 p(z)，用前面推导出来的(10.9)：

$$\begin{aligned} \ln q_1^\star(z_1) &= \mathbb{E}_{z_2}[\ln p(\mathbf{z})] + \text{const} \\ &= \mathbb{E}_{z_2}\left[-\frac{1}{2}(z_1-\mu_1)^2\Lambda_{11} - (z_1-\mu_1)\Lambda_{12}(z_2-\mu_2)\right] + \text{const} \\ &= -\frac{1}{2}z_1^2\Lambda_{11} + z_1\mu_1\Lambda_{11} - z_1\Lambda_{12}\left(\mathbb{E}[z_2]-\mu_2\right) + \text{const.} \end{aligned} \tag{10.11}$$

因为是求 z1 的分布，所以按(10.9)，我们在 z2 上求期望，得到(10.11)。然后，我们就可以祭出第二章修炼的法宝——配方法，从(10.11)得到高斯分布：

$$q^\star(z_1) = \mathcal{N}(z_1|m_1, \Lambda_{11}^{-1}) \tag{10.12}$$

其中

$$m_1 = \mu_1 - \Lambda_{11}^{-1}\Lambda_{12}\left(\mathbb{E}[z_2]-\mu_2\right). \tag{10.13}$$

同样，z2 的分布也可如法炮制：

$$q_2^\star(z_2) = \mathcal{N}(z_2|m_2, \Lambda_{22}^{-1}) \tag{10.14}$$

其中

$$m_2 = \mu_2 - \Lambda_{22}^{-1}\Lambda_{21}\left(\mathbb{E}[z_1]-\mu_1\right). \tag{10.15}$$

它们是完全对称的。因为 m1 里有 z2 的期望，而 m2 里又有 z1 的期望，所以我们可以设一个初始值，然后迭代求解。但实际上这两个式子恰好有解析解：$\mathbb{E}[z_1] = \mu_1$和$\mathbb{E}[z_2] = \mu_2$，我们可把它们代入(10.13)和(10.15)验证一下。

下面我们重点看一下参数推断问题，但其核心思想实际上和前面讲的例子区别不大。同样还是先看一下高斯分布：

我们想推断后验高斯分布的均值μ和精度τ

假如我们观察到 N 个数据$\mathcal{D} = \{x_1, \ldots, x_N\}$，那么似然函数就是：

$$p(\mathcal{D}|\mu,\tau) = \left(\frac{\tau}{2\pi}\right)^{N/2} \exp\left\{-\frac{\tau}{2}\sum_{n=1}^{N}(x_n-\mu)^2\right\}. \qquad (10.21)$$

另外引入先验分布，均值服从高斯分布、精度服从 Gamma 分布：

$$p(\mu|\tau) = \mathcal{N}\left(\mu|\mu_0,(\lambda_0\tau)^{-1}\right) \qquad (10.22)$$
$$p(\tau) = \mathrm{Gam}(\tau|a_0,b_0) \qquad (10.23)$$

其实这个问题我们前面第二章就讲过，不用变分推断也能直接求出来，但这里用变分推断实际上增加了更多的灵活性，因为如果先验和似然的形式不是高斯-Gamma 的形式，而是更加复杂，那么我们也可以利用变分推断来算参数，这是非常方便的。我们这里只是用我们熟悉的高斯分布来举例子，把这个弄明白，以后再推广到其他例子上就容易多了。

利用 mean field 形式(10.9)，我们可计算出μ的分布：

$$\begin{aligned}\ln q_\mu^\star(\mu) &= \mathbb{E}_\tau\left[\ln p(\mathcal{D}|\mu,\tau)+\ln p(\mu|\tau)\right]+\text{const}\\ &= -\frac{\mathbb{E}[\tau]}{2}\left\{\lambda_0(\mu-\mu_0)^2+\sum_{n=1}^{N}(x_n-\mu)^2\right\}+\text{const.} \qquad (10.25)\end{aligned}$$

可以看到，μ服从高斯分布形式$\mathcal{N}\left(\mu|\mu_N,\lambda_N^{-1}\right)$，且通过配方，可得到该分布参数为：

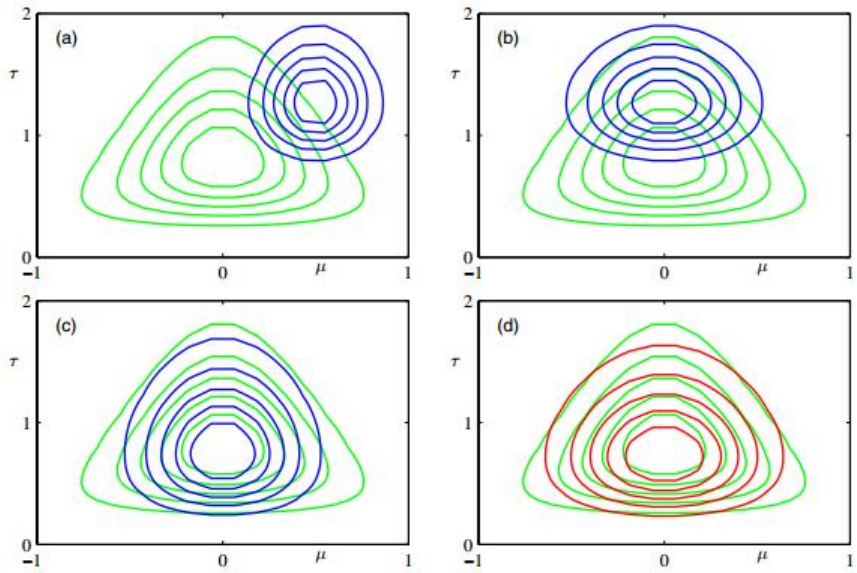
$$\mu_N = \frac{\lambda_0\mu_0+N\overline{x}}{\lambda_0+N} \qquad (10.26)$$
$$\lambda_N = (\lambda_0+N)\mathbb{E}[\tau]. \qquad (10.27)$$

注意到，样本越多也就是 N 越大时，均值会趋向于样本均值$\mu_N = \overline{x}$，同时精度趋向于无穷大。同样

可用(10.9)计算τ的分布，得到：

$$\begin{aligned}\ln q_\tau^\star(\tau) &= \mathbb{E}_\mu\left[\ln p(\mathcal{D}|\mu,\tau)+\ln p(\mu|\tau)\right]+\ln p(\tau)+\text{const}\\ &= (a_0-1)\ln\tau-b_0\tau+\frac{N}{2}\ln\tau\\ &\quad -\frac{\tau}{2}\mathbb{E}_\mu\left[\sum_{n=1}^{N}(x_n-\mu)^2+\lambda_0(\mu-\mu_0)^2\right]+\text{const} \qquad (10.28)\end{aligned}$$

它服从 Gamma 分布形式$\mathrm{Gam}(\tau|a_N, b_N)$，可以看到，(10.27)和(10.30)里，仍然有和另一分布相关的期望需要计算，所以我们可以设定初始值，然后迭代计算。迭代过程和收敛后的结果图书上 10.4 所示：



再看一个例子，是用变分推断计算线性回归的参数。线性回归的参数 w，有似然和先验如下：

$$p(\mathbf{t}|\mathbf{w}) = \prod_{n=1}^{N} \mathcal{N}(t_n|\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}_n, \beta^{-1}) \qquad (10.87)$$

$$p(\mathbf{w}|\alpha) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I}) \qquad (10.88)$$

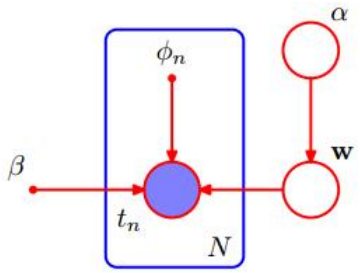
2.3.6 讲过，α的共轭先验是 Gamma 分布：

$$p(\alpha) = \mathrm{Gam}(\alpha|a_0, b_0) \qquad (10.89)$$

这样联合分布就是：

$$p(\mathbf{t}, \mathbf{w}, \alpha) = p(\mathbf{t}|\mathbf{w})p(\mathbf{w}|\alpha)p(\alpha). \qquad (10.90)$$

其概率图模型为图 10.8：



利用变分推断来计算 w 和α，同样是假设它们有可分解形式：

$$q(\mathbf{w}, \alpha) = q(\mathbf{w})q(\alpha). \qquad (10.91)$$

再用(10.9)（这个绝对是看家法宝）来搞，得到：

$$\begin{aligned}\ln q^\star(\alpha) &= \ln p(\alpha) + \mathbb{E}_{\mathbf{w}}\left[\ln p(\mathbf{w}|\alpha)\right] + \text{const} \\ &= (a_0 - 1)\ln\alpha - b_0\alpha + \frac{M}{2}\ln\alpha - \frac{\alpha}{2}\mathbb{E}[\mathbf{w}^{\mathrm{T}}\mathbf{w}] + \text{const.} \qquad (10.92)\end{aligned}$$

可看到它服从 Gamma 分布：

$$q^\star(\alpha) = \mathrm{Gam}(\alpha|a_N, b_N) \qquad (10.93)$$

其中

$$a_N = a_0 + \frac{M}{2} \qquad (10.94)$$

$$b_N = b_0 + \frac{1}{2}\mathbb{E}[\mathbf{w}^{\mathrm{T}}\mathbf{w}]. \qquad (10.95)$$

以及：

$$\ln q^\star(\mathbf{w}) = \ln p(\mathbf{t}|\mathbf{w}) + \mathbb{E}_\alpha\left[\ln p(\mathbf{w}|\alpha)\right] + \text{const} \qquad (10.96)$$

$$= -\frac{\beta}{2}\sum_{n=1}^{N}\{\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}_n - t_n\}^2 - \frac{1}{2}\mathbb{E}[\alpha]\mathbf{w}^{\mathrm{T}}\mathbf{w} + \text{const} \qquad (10.97)$$

$$= -\frac{1}{2}\mathbf{w}^{\mathrm{T}}\left(\mathbb{E}[\alpha]\mathbf{I} + \beta\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}\right)\mathbf{w} + \beta\mathbf{w}^{\mathrm{T}}\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{t} + \text{const.} \qquad (10.98)$$

可看到它服从高斯分布：

$$q^\star(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N) \qquad (10.99)$$

其中

$$\mathbf{m}_N = \beta\mathbf{S}_N\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{t} \qquad (10.100)$$

$$\mathbf{S}_N = \left(\mathbb{E}[\alpha]\mathbf{I} + \beta\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}\right)^{-1}. \qquad (10.101)$$

(10.95)和(10.97)里还有奇怪的东西$\mathbb{E}[\alpha]$和$\mathbb{E}[\mathbf{w}\mathbf{w}^{\mathrm{T}}]$，从附录 B 可知，它们分别是：

$$\mathbb{E}[\alpha] = a_N/b_N \qquad (10.102)$$

$$\mathbb{E}[\mathbf{w}\mathbf{w}^{\mathrm{T}}] = \mathbf{m}_N\mathbf{m}_N^{\mathrm{T}} + \mathbf{S}_N. \qquad (10.103)$$

所以我们仍然可以迭代计算：给初始值，每一步都算出 a_N、b_N 和 m_N、S_N，代入求解。

掌握了上面的三个例子，我想推广到其他情况也都没有太大难度了。其实书中还有一个例子也非常重要，就是 10.2 所讲的用变分推断计算高斯混合模型的参数。不过我想尼采兄讲第九章时已经打下了很好的基础，再加上刚才讲的这一章的例子，看懂这部分应该不难。

后面还有一些有趣的内容，比如 Expectation Propagation，是说对$\mathrm{KL}(p\|q)$做极小化，而不是$\mathrm{KL}(q\|p)$。因为积分里前面那项变成了 p(Z)而不是 q(Z)，而 p(Z)又是复杂分布，所以这里处理方式有所不同。感兴趣的朋友可以看看 10.7 节是如何做的。

我讲的内容就到这里。我个人的一点心得体会就是：高斯分布以及其他常用分布的形式、还有第二章讲到的配方法一定要掌握好，这是识别分布和直接计算分布参数的最大利器。然后就是这一章的(10.9)，也就是用可分解分布去做近似得到的 mean field，这也是比较常用的。其实群里有不少对变分推断很了解的高手，比如@huajh7 ，大家对这一块有什么问题也可以找他们交流讨论。

==========讨论==========

数据挖掘(983272906) 21:44:16

$$p(\mathbf{t}, \mathbf{w}, \alpha) = p(\mathbf{t}|\mathbf{w})p(\mathbf{w}|\alpha)p(\alpha). \quad (10.90)$$这种分解有没有什么限制条件

Wilbur_中博(1954123) 21:45:20

这不是分解，是从先验和似然算联合分布。可分解的简单分布形式是(10.91)。

Y(414474527) 21:47:49

变分推断怎么应用到实际问题中呢

tzk<tangzk2011@163.com> 21:48:29

LDA 的原始论文用的也是变分呢。。

<(523723864) 21:48:43

10.9 式一定是 tractable 的吗？

zeno(117127143) 21:52:58

平均场假设可以有效减少参数。

Wilbur_中博(1954123) 21:53:54

@Y 实际问题吗？我觉得就是作为一种工具，求解模型参数的时候会比较简单吧。之前在稀疏编码里看到过一些，我觉得这篇文章不错：

http://ipg.epfl.ch/~seeger/lapmalmainweb/papers/seeger_wipf_spm10.pdf 。另外 RBM 似乎也有用这个的。

zeno(117127143) 21:54:19

变分把推断变为求极值问题，怎么求是另外一门课

Wilbur_中博(1954123) 21:54:43

@< 我觉得不一定。。还得看 p(X,Z)是什么样的。

@zeno 嗯

<(523723864) 21:55:01

按照 10.9 主要是推式子咯，事先不知道 qj 的分布

Wilbur_中博(1954123) 21:55:57

嗯，应该是。。但是一般来说都可以想办法搞出来吧，(10.9)的积分。

karnon(447457116) 21:59:27

为什么一开始又要用复杂分布呢，建模时用那些复杂的模型，最后到求解时都退化成 naive 模型，所以事实上，和 naive 模型一样

Wilbur_中博(1954123) 22:02:31

可能一开始就用简单分布的话，推出后验分布有连锁效应，就会越来越差吧。现在搞出后验分布再用简单分布去近似，我觉得道理上还是能说得通。

zeno(117127143) 22:03:27

那为啥有泰勒展开，展开把高次舍弃，不都不是原来函数了吗

<(523723864) 22:04:21

关键是每次迭代的时候 lower bound 会不会上去

karnon(447457116) 22:04:22

如果你要用 taylor 展开来近似，那就得证明近似后你的解的性质不变 ，所以不是任何问题都能随便近似

Wilbur_中博(1954123) 22:04:43

@< 是，我觉得这个蛮关键的

karnon(447457116) 22:06:49
就是你的解为什么好，它好在哪，近似之后，这些好处是不是还保留着，这在变分法中，完全没有讨论
zeno(117127143) 22:10:58
要是有 kl 跟概率差异定量关系就没问题了，平均场本来就是假设，变分推断是合理的，kl 嘛，不好说，反正不像熟悉的欧式度量，pgm 不只变分一种推断方法，所以也不能建成简单模型。说实在如果能解决一类小问题效果不错就已经很好了，mrf，hmm，crf，都能算到 pgm 中。pgm 解决不少问题。
阿邦(1549614810) 23:41:20
推断方法不坑，主要还是模型的问题
karnon(447457116) 0:02:10
我总感觉，一定有基于非概率模型的方法
弹指一瞬间(337595903) 6:31:34
昨晚大家讨论的好热闹啊。@karnon：我觉得近似推理对原模型的好处还是保留着的。虽然求解的时候是在简单模型上做，但是简单模型的求解目标是去近似原模型的最优而不是简单模型的最优。这个和一上来就做简单模型假设是不大一样的。近似推理可以理解为在最优解附近找一个次优解，但总体目标还是原模型最优解的方向。而简单模型求解可能目标就不一样了。相比之下，还是用近似推理来解原问题比较好。

（个人理解不一定对，欢迎跟帖  ）

zeno(117127143) 6:52:44
我喜欢概率模型，概率既能对不确定性建模更能对未知建模。做单选题 25%表达的是学生对答案的未知，同样的题对老师就是已知的。同样问题用非概率解你需要知道的更多。同样四道单选题三道不会，其他三道分别选 a，b，c。第四道用概率方法根据一定先验会尽量选 d。不用概率方法根本做不了这种问题。
同样如果知道了答案，肯定不会用概率方法，概率比通常非概率方法麻烦。
karnon(447457116) 7:33:16
这只是理想的情况，概率模型的缺点，在于它需要精确地刻划细节。