PRML（Pattern Recognition And Machine Learning）读书会

第十一章　Sampling Methods

主讲人 网络上的尼采

（新浪微博：@Nietzsche_复杂网络机器学习）

QQ 群 177217565

读书会微信公众平台请扫描下面的二维码

网络上的尼采(813394698) 9:05:00

今天的主要内容：Markov Chain Monte Carlo，Metropolis-Hastings，Gibbs Sampling，Slice Sampling，Hybrid Monte Carlo。

上一章讲到的平均场是统计物理学中常用的一种思想，将无法处理的复杂多体问题分解成可以处理的单体问题来近似，变分推断便是在平均场的假设约束下求泛函 L(Q)极值的最优化问题，好处在于求解过程中可以推出精致的解析解。变分是从最优化的角度通过坐标上升法收敛到局部最优，这一章我们将通过计算从动力学角度见证 Markov Chain Monte Carlo 收敛到平稳分布。

先说 sampling 的原因 ,因为统计学中经常会遇到对复杂的分布做加和与积分 ,这往往是 intractable 的。MCMC 方法出现后贝叶斯方法才得以发展，因为在那之前对不可观测变量（包括隐变量和参数）后验分布积分非常困难，对于这个问题上一章变分用的解决办法是通过最优化方法寻找一个和不可观测变量后验分布 p(Z|X)近似的分布，这一章我们看下 sampling 的解决方法，举个简单的例子：比如我们遇到这种形式

$$\mathbb{E}[f] = \int f(\mathbf{z}) p(\mathbf{z})\, \mathrm{d}\mathbf{z}$$

，z 是个连续随机变量，p(z)是它的分布，我们求 f(z)的期望。如果我们从 p(z)中 sampling 一个数据集 $z^{(l)}$，然后再求个平均 $\widehat{f} = \frac{1}{L}\sum_{l=1}^{L} f(\mathbf{z}^{(l)})$ 来近似 f(z)的期望， so,问题就解决了，关键是如何从 p(z)中做无偏的 sampling。

为了说明 sampling 的作用，我们先举个 EM 的例子，最大似然计算中求分布的积分问题，我们在第九章提到了，完整数据的 log 似然函数是对隐变量 Z 的积分：

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{\mathrm{old}}) = \int p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\mathrm{old}}) \ln p(\mathbf{Z}, \mathbf{X}|\boldsymbol{\theta})\, \mathrm{d}\mathbf{Z}$$

如果 Z 是比较复杂的分布，我们就需要对 Z 进行采样，从而得到：

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{\mathrm{old}}) \simeq \frac{1}{L}\sum_{l=1}^{L} \ln p(\mathbf{Z}^{(l)}, \mathbf{X}|\boldsymbol{\theta}).$$

具体就是从 Z 的后验分布 $\text{posterior distribution } p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\mathrm{old}})$ 中进行采样。

如果我们从贝叶斯的观点，把 EM 参数 theta 也当成一个分布的话，有下面一个 IP 算法:

IP Algorithm

I-step. We wish to sample from $p(\mathbf{Z}|\mathbf{X})$ but we cannot do this directly. We therefore note the relation

$$p(\mathbf{Z}|\mathbf{X}) = \int p(\mathbf{Z}|\boldsymbol{\theta}, \mathbf{X}) p(\boldsymbol{\theta}|\mathbf{X})\, \mathrm{d}\boldsymbol{\theta} \tag{11.30}$$

and hence for $l = 1, \ldots, L$ we first draw a sample $\boldsymbol{\theta}^{(l)}$ from the current estimate for $p(\boldsymbol{\theta}|\mathbf{X})$, and then use this to draw a sample $\mathbf{Z}^{(l)}$ from $p(\mathbf{Z}|\boldsymbol{\theta}^{(l)}, \mathbf{X})$.

P-step. Given the relation

$$p(\boldsymbol{\theta}|\mathbf{X}) = \int p(\boldsymbol{\theta}|\mathbf{Z}, \mathbf{X}) p(\mathbf{Z}|\mathbf{X})\, \mathrm{d}\mathbf{Z} \tag{11.31}$$

we use the samples $\{\mathbf{Z}^{(l)}\}$ obtained from the I-step to compute a revised estimate of the posterior distribution over $\boldsymbol{\theta}$ given by

$$p(\boldsymbol{\theta}|\mathbf{X}) \simeq \frac{1}{L}\sum_{l=1}^{L} p(\boldsymbol{\theta}|\mathbf{Z}^{(l)}, \mathbf{X}). \tag{11.32}$$

By assumption, it will be feasible to sample from this approximation in the I-step.

I 步，我们无法直接对 P(Z|X)取样，我们可以先对 P(theta|X)取样$\boldsymbol{\theta}^{(l)}$，然后再对 Z 的后验分布进行取样：

$$\text{draw a sample } \mathbf{Z}^{(l)} \text{ from } p(\mathbf{Z}|\boldsymbol{\theta}^{(l)}, \mathbf{X}).$$

P 步，利用上一步对 P(Z|X)的取样，来确定新的参数：

$$p(\boldsymbol{\theta}|\mathbf{X}) \simeq \frac{1}{L}\sum_{l=1}^{L} p(\boldsymbol{\theta}|\mathbf{Z}^{(l)}, \mathbf{X}).$$

然后按这个 I 步和 P 步的方式迭代。

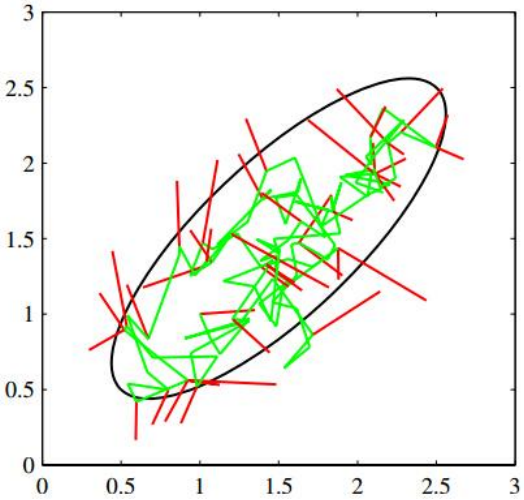
接下来我们讲 sampling methods，为了节省时间，两个基本的 rejection sampling 和 Importance sampling 不讲了，这两种方法在高维下都会失效，我们直奔主题 MCMC（Markov Chain Monte Carlo ）。蒙特卡洛，是个地中海海滨城市，气候宜人，欧洲富人们的聚集地，更重要的是它是世界三大赌城之一，用这个命名就知道这种方法是基于随机的，过会我们会讲到。马尔科夫大神就不用多说了，他当时用自己的名字命名马尔科夫链是预见到这个模型巨大作用。他还有一个师弟叫李雅普洛夫，控制论里面的李雅普洛夫函数说的就是这位，他们的老师叫契比雪夫，都是圣彼得堡学派的。俄国数学家对人类的贡献是无价的 orz

最早的 MCMC 方法是美国科学家在研制原子弹时算积分发明的。我们先介绍一个最基本的 Metropolis 方法，这种方法的接受率是：

$$A(\mathbf{z}^{\star}, \mathbf{z}^{(\tau)}) = \min\left(1, \frac{\widetilde{p}(\mathbf{z}^{\star})}{\widetilde{p}(\mathbf{z}^{(\tau)})}\right)$$

但有个要求，就是 proposal distribution 满足$q(\mathbf{z}_A|\mathbf{z}_B) = q(\mathbf{z}_B|\mathbf{z}_A)$。过程很简单，我们先找个比较容易采样的分布即 proposal 分布，然后从这个分布中取一个样本 Z*，如果$\frac{\widetilde{p}(\mathbf{z}^{\star})}{\widetilde{p}(\mathbf{z}^{(\tau)})}$大于 1 直接接受，如果小于 1 就接着算出接受率，并且从（0，1）之间取一个随机数和这个接受率做比较来决定是否接受这个样本。过会会在 Metropolis-Hastings algorithm 方法中具体说。下图是一个简单的例子，对高斯分布做采样，绿线是表示接受的步骤，红线表示拒绝的：

A simple illustration using Metropolis algorithm to sample from a Gaussian distribution whose one standard-deviation contour is shown by the ellipse. The proposal distribution is an isotropic Gaussian distribution whose standard deviation is 0.2. Steps that are accepted are shown as green lines, and rejected steps are shown in red. A total of 150 candidate samples are generated, of which 43 are rejected.



讲 Metropolis-Hastings 方法前，我们先来回顾下马尔科夫链的性质，这个很重要。markov chains 最基本的性质就是无后效性，就是这条链的下一个节点的状态由当前节点状态完全决定：

$$p(\mathbf{z}^{(m+1)}|\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(m)}) = p(\mathbf{z}^{(m+1)}|\mathbf{z}^{(m)})$$

特定的齐次马尔科夫链可以收敛到平稳分布，也就是经过相当长的一段时间转移，收敛到的分布和初始值无关，转移核起着决定的作用。关于马尔科夫链的收敛，**我们将介绍一个充分条件：detailed balance 细致平稳条件。**

先介绍两个公式，马尔科夫链节点状态的 marginal distribution 计算公式，由于无后效性我们可以得到

$$p(\mathbf{z}^{(m+1)}) = \sum_{\mathbf{z}^{(m)}} p(\mathbf{z}^{(m+1)}|\mathbf{z}^{(m)})p(\mathbf{z}^{(m)}). \qquad (11.38)$$

上面公式的加和结合马儿科夫链的状态转移矩阵是比较容易理解。

平稳分布的定义就是下面的形式：

$$p^\star(\mathbf{z}) = \sum_{\mathbf{z}'} T(\mathbf{z}', \mathbf{z})p^\star(\mathbf{z}'). \qquad (11.39)$$

其中$T(\mathbf{z}', \mathbf{z})$是 z'到 z 的转移概率，上面的公式不难理解，就是转移后分布不再发生变化。

下面我们给出细致平稳条件的公式：

$$p^\star(\mathbf{z})T(\mathbf{z}, \mathbf{z}') = p^\star(\mathbf{z}')T(\mathbf{z}', \mathbf{z}) \qquad (11.40)$$

满足细致平稳条件就能收敛到平稳分布，下面是推导：

$$\sum_{\mathbf{z}'} p^\star(\mathbf{z}')T(\mathbf{z}', \mathbf{z}) = \sum_{\mathbf{z}'} p^\star(\mathbf{z})T(\mathbf{z}, \mathbf{z}') = p^\star(\mathbf{z}) \sum_{\mathbf{z}'} p(\mathbf{z}'|\mathbf{z}) = p^\star(\mathbf{z}). \qquad (11.41)$$

我们把上面细致平稳条件的公式 11.40 代入公式 11.41 最左边 ,从左朝右推导就是平稳分布的公式 11.39。

细致平稳条件的好处，就是我们能控制马尔科夫链收敛到我们指定的分布$p^\star(\mathbf{z})$。以后的 Metropolis-Hastings 方法及改进都是基于这个基础的。

前面我们提到，Metropolis 方法需要先选一个比较容易取样的 proposal distribution，从这个分布里取样，然后通过接受率决定是否采用这个样本。一个简单的例子就是对于 proposal distribution 我们可以采用 Gaussian centred on the current state，其实很好理解，就是上一步节点的值可以做下一步节点需要采样的 proposal distribution 即高斯分布的均值，这样下一步节点的状态由上一步完全决定，这就是一个马尔科夫链。马尔科夫链有了，我们怎么保证能收敛到目标分布呢？就是前面说的细致平稳条件，我们可以通过设置接受率的形式来满足这个条件。Metropolis-Hastings 接受率的形式：

$$A_k(\mathbf{z}^\star, \mathbf{z}^{(\tau)}) = \min\left(1, \frac{\widetilde{p}(\mathbf{z}^\star)q_k(\mathbf{z}^{(\tau)}|\mathbf{z}^\star)}{\widetilde{p}(\mathbf{z}^{(\tau)})q_k(\mathbf{z}^\star|\mathbf{z}^{(\tau)})}\right). \qquad (11.44)$$

$\widetilde{p}(\mathbf{z})$来自于$p(\mathbf{z}) = \widetilde{p}(\mathbf{z})/Z_p$.，分布 q 便是 proposal distribution。

范涛@推荐系统(289765648) 10:49:15

Zp 是什么？

网络上的尼采(813394698) 10:52:04

Zp 是分布中和 z 无关的部分。

为了使各位有个形象的理解，我描述一下过程，我们把$\mathbf{z}^{(\tau)}$当做高斯分布的均值，方差是固定的。然后从这个分布取一个样本就是$\mathbf{z}^\star$，如果$\frac{\widetilde{p}(\mathbf{z}^\star)q_k(\mathbf{z}^{(\tau)}|\mathbf{z}^\star)}{\widetilde{p}(\mathbf{z}^{(\tau)})q_k(\mathbf{z}^\star|\mathbf{z}^{(\tau)})}$大于 1 肯定接受，如果小于 1，我们便从（0，1）

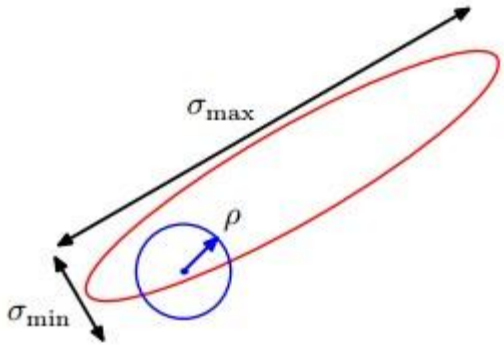
之间取一个随机数，和这个接受率做比较，如果接受率大于这个随机数便接受，反之便拒绝。接受率这么设就能满足细致平稳条件的原因，看这个（11.45）公式：

$$\begin{aligned} p(\mathbf{z})q_k(\mathbf{z}'|\mathbf{z})A_k(\mathbf{z}',\mathbf{z}) &= \min\left(p(\mathbf{z})q_k(\mathbf{z}'|\mathbf{z}), p(\mathbf{z}')q_k(\mathbf{z}|\mathbf{z}')\right) \\ &= \min\left(p(\mathbf{z}')q_k(\mathbf{z}|\mathbf{z}'), p(\mathbf{z})q_k(\mathbf{z}'|\mathbf{z})\right) \\ &= p(\mathbf{z}')q_k(\mathbf{z}|\mathbf{z}')A_k(\mathbf{z},\mathbf{z}') \end{aligned}$$

$$\begin{aligned} p(\mathbf{z})q_k(\mathbf{z}'|\mathbf{z})A_k(\mathbf{z}',\mathbf{z}) &= \min\left(p(\mathbf{z})q_k(\mathbf{z}'|\mathbf{z}), p(\mathbf{z}')q_k(\mathbf{z}|\mathbf{z}')\right) \\ &= \min\left(p(\mathbf{z}')q_k(\mathbf{z}|\mathbf{z}'), p(\mathbf{z})q_k(\mathbf{z}'|\mathbf{z})\right) \\ &= p(\mathbf{z}')q_k(\mathbf{z}|\mathbf{z}')A_k(\mathbf{z},\mathbf{z}') \end{aligned}$$

我们把接受率公式 11.44 代入上面的公式的左边，会推出左右两边就是细致平稳条件的形式，红框部分便是细致平稳条件公式 11.40 的转移核，书上的公式明显错了，上面的这个是勘误过的。

刚才说了 proposal distribution 一般采用 Gaussian centred on the current state，高斯分布的方差是固定的，其实方差就是步长，如何选择步长这是一个 state of the art 问题，步子太小扩散太慢，步子太大，拒绝率会很高，原地踏步。书中的一个例子，当用 Gaussian centred on the current state 作 proposal distribution 时，步长设为目标高斯分布的最小标准差最合适：



下面讲 Gibbs Sampling，Gibbs Sampling 其实是每次只对一个维度的变量进行采样，固定住其他维度的变量，然后迭代，可以看做是 Metropolis-Hastings 的特例，它的接受率一直是 1.

Gibbs Sampling

1. Initialize $\{z_i : i = 1, \ldots, M\}$
2. For $\tau = 1, \ldots, T$:
   - Sample $z_1^{(\tau+1)} \sim p(z_1|z_2^{(\tau)}, z_3^{(\tau)}, \ldots, z_M^{(\tau)})$.
   - Sample $z_2^{(\tau+1)} \sim p(z_2|z_1^{(\tau+1)}, z_3^{(\tau)}, \ldots, z_M^{(\tau)})$.
   - $\vdots$
   - Sample $z_j^{(\tau+1)} \sim p(z_j|z_1^{(\tau+1)}, \ldots, z_{j-1}^{(\tau+1)}, z_{j+1}^{(\tau)}, \ldots, z_M^{(\tau)})$.
   - $\vdots$
   - Sample $z_M^{(\tau+1)} \sim p(z_M|z_1^{(\tau+1)}, z_2^{(\tau+1)}, \ldots, z_{M-1}^{(\tau+1)})$.

步骤是比较容易理解的，跟上一章的变分法的有相似之处。假设有三个变量的分布$p(z_1, z_2, z_3)$，

先固定住 z2 z3 对 z1 进行采样，$p(z_1|z_2^{(\tau)}, z_3^{(\tau)})$；

然后固定住 z1 z3 对 z2 进行采样，$p(z_2|z_1^{(\tau+1)}, z_3^{(\tau)})$；

然后是 Z3，$p(z_3|z_1^{(\tau+1)}, z_2^{(\tau+1)})$

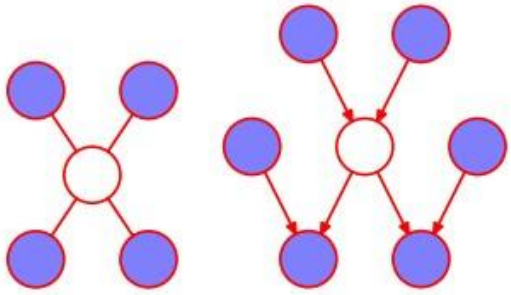
如此迭代。

根据 Metropolis-Hastings，它的接受率恒为 1。看下面的推导：

$$A(\mathbf{z}^\star, \mathbf{z}) = \frac{p(\mathbf{z}^\star)q_k(\mathbf{z}|\mathbf{z}^\star)}{p(\mathbf{z})q_k(\mathbf{z}^\star|\mathbf{z})} = \frac{p(z_k^\star|\mathbf{z}_{\setminus k}^\star)p(\mathbf{z}_{\setminus k}^\star)p(z_k|\mathbf{z}_{\setminus k}^\star)}{p(z_k|\mathbf{z}_{\setminus k})p(\mathbf{z}_{\setminus k})p(z_k^\star|\mathbf{z}_{\setminus k})} = 1$$

因为其他维度是固定不变的，所以 $\mathbf{z}_{\setminus k}^\star = \mathbf{z}_{\setminus k}$，代入上式就都约去了，等于 1.

最后对于图模型采用 gibbs sampling，条件概率 $p(z_k|\mathbf{z}_{\setminus k})$可以根据马尔科夫毯获得，下面一个是无向图，一个是有向图，蓝色的节点是和要采样的变量有关的其他变量：



关于更多的 gibbs sampling 的内容可以看 MLAPP，里面有 blocked gibbs 和 collapsed gibbs。

刚才提到 Metropolis-Hastings 对步长敏感，针对这个问题，下面介绍两个增加辅助变量的方法，这些方法也是满足细致平稳条件的。先介绍 slice sampling，这种方法增加了一个变量 U，可以根据分布的特征自动调整步长：

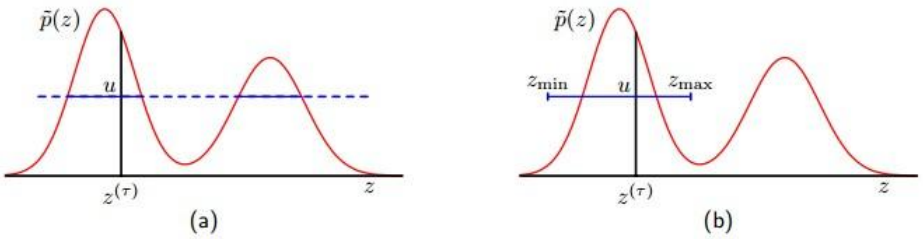


Figure 11.13 Illustration of slice sampling. (a) For a given value $z^{(\tau)}$, a value of u is chosen uniformly in the region $0 \leqslant u \leqslant \tilde{p}(z^{(\tau)})$, which then defines a 'slice' through the distribution, shown by the solid horizontal lines. (b) Because it is infeasible to sample directly from a slice, a new sample of z is drawn from a region $z_{\min} \leqslant z \leqslant z_{\max}$, which contains the previous value $z^{(\tau)}$.

步骤很简单：在$z^{(\tau)}$与$\tilde{p}(z)$之间的这段距离随机取个值 U，然后通过 U 画个横线，然后在包含$z^{(\tau)}$并且$\{z : \widetilde{p}(z) > u\}$这段横线对 z 进行随机采样，然后按这种方式迭代。图(b)为了实际中便于操作，有时还需要多出那么一段，因为我们事先不知道目标分布的具体形式，所以包含$z^{(\tau)}$并且$\{z : \widetilde{p}(z) > u\}$这段横线没法确定，只能朝外延伸加单位长度进行试，最后会多出来一段，这一点书上并没有介绍详细。

下面介绍 The Hybrid Monte Carlo Algorithm（Hamiltonian MCMC）：哈密顿，神童，经典力学三巨头之一，这个算法引入了哈密顿动力系统的概念，计算接受率时考虑的是系统的总能量。
Hybrid Monte Carlo 定义了势能和动能两种能量，它们的和便是系统总能量哈密顿量。先看势能，分布可以写成这种形式：

$$p(\mathbf{z}) = \frac{1}{Z_p} \exp\left(-E(\mathbf{z})\right)$$

E(z)便是系统的势能。

另外增加一个变量，状态变量变化的速率：$r_i = \frac{\mathrm{d}z_i}{\mathrm{d}\tau}$

系统的动能便是：

$$K(\mathbf{r}) = \frac{1}{2}\|\mathbf{r}\|^2 = \frac{1}{2}\sum_i r_i^2$$

总的能量便是：$H(\mathbf{z}, \mathbf{r}) = E(\mathbf{z}) + K(\mathbf{r})$
比如高斯分布的哈密顿量就可表示为：

$$H(\mathbf{z}, \mathbf{r}) = \frac{1}{2}\sum_i \frac{1}{\sigma_i^2} z_i^2 + \frac{1}{2}\sum_i r_i^2.$$

下面这个公式便是 Hybrid Monte Carlo 的接受率：

$$\min\left(1, \exp\{H(\mathbf{z}, \mathbf{r}) - H(\mathbf{z}^\star, \mathbf{r}^\star)\}\right)$$

可以证明 这种接受率是满足 detailed balance 条件的。
ORC(267270520) 12:14:03
推荐一本相关的书 Introducing Monte Carlo Methods with R (use R) ，PS：R 做 MCMC 很方便。
赞尼采讲的很精彩，学习了，嘿嘿
网络上的尼采(813394698) 12:36:37
Markov Chain Monte Carlo In Practice(Gilks)这本书也挺不错。
红烧鱼(403774317) 12:38:06
这本读研的时候生读过，非常实用，随书附带 code
网络上的尼采(813394698)
最后需要补充的是：判断 MCMC 的 burn-in 何时收敛是个问题，koller 介绍了两种方法，即同一条链上设置不同的时间窗做比较，另一种同时跑多条链然后作比较。当然也有一条链跑到黑的。