从 MC 到 MCMC

jamescao(曹孝卿)*

2017年1月20日
v1.0

MC 指的是 Monte Carlo，MCMC 这个词中，第一个 MC 指的是 Markov Chain，第二个 MC 指的是 Monte Carlo。本文档旨在用通俗的语言描述 MC 和 MCMC 的基本思想和方法，并且受限于笔者的知识水平，文档中可能有很多不准确的措辞和描述。因此，读者若是追求数学层次上完全准确的算法原理，应该参考相关的专业书籍。

*jamescao@tencent.com

目录

1 什么是 MC

在贝叶斯统计分析中，一个统计模型一般包含三个部分，首先是先验分布 $\pi(\theta)$，其次是样本分布（或者称似然函数）$p(\mathbf{x}|\theta)$，其中的 $\mathbf{x}$ 是样本，θ 是待估计的参数，最后是后验概率分布 $\pi(\theta|\mathbf{x})$。这三部分通过贝叶斯定理联系在一起：

$$\begin{aligned}\pi(\theta|\mathbf{x}) &= \frac{\pi(\theta)p(\mathbf{x}|\theta)}{\int_\Theta \pi(\theta)p(\mathbf{x}|\theta)\mathbf{d}\theta} \\ &\propto \pi(\theta)p(\mathbf{x}|\theta)\end{aligned}$$

那么我们就能得到 θ 在二次损失函数下的贝叶斯估计值

$$\begin{aligned}\hat{\theta} &= \int_\Theta \theta\pi(\theta|\mathbf{x})\mathbf{d}\theta \\ &= \frac{\int_\Theta \theta\pi(\theta)p(\mathbf{x}|\theta)\mathbf{d}\theta}{\int_\Theta \pi(\theta)p(\mathbf{x}|\theta)\mathbf{d}\theta}.\end{aligned}$$

或者对于 θ 的函数 $g(\theta)$ 的贝叶斯估计值

$$\begin{aligned}\hat{g}(\theta) &= \int_\Theta g(\theta)\pi(\theta|\mathbf{x})\mathbf{d}\theta \\ &= \frac{\int_\Theta g(\theta)\pi(\theta)p(\mathbf{x}|\theta)\mathbf{d}\theta}{\int_\Theta \pi(\theta)p(\mathbf{x}|\theta)\mathbf{d}\theta}.\end{aligned}$$

但是要得到上面的这些估计值，就需要计算上面式子中的积分，但这些积分通常是很难得到具体表达式的。在得不到显示表达式的情况下，如果参数的维数不高，还可以通过数值积分方法做数值计算，然而在参数纬度很大时，数值积分方法也很难实现，这个时候就需要 MC 方法乃至 MCMC 方法出马了。

举个栗子，比如我们要计算 $g(\theta)$ 的估计，理论上我们是要通过下式来计算的：

$$\hat{g}(\theta) = \int_\Theta g(\theta)\pi(\theta|\mathbf{x})\mathbf{d}\theta \tag{1}$$

这个估计值就是 $g(\theta)$ 的后验均值 $E(g(\theta)|\mathbf{x})$，如果上面的积分求不出来，我们也可以用下面的方法来求近似值：

$$\bar{g}(\theta) = \frac{1}{m}\sum_{i=1}^{m} g(\theta^{(i)}) \tag{2}$$

其中，$\theta^{(i)}$ 是从 θ 的后验分布 $\pi(\theta|\mathbf{x})$ 里抽取的样本，如果这些样本都是独立的，那么由大数定律，样本的均值 $\bar{g}(\theta)$ 依概率收敛到 $g(\theta)$ 的后验均值 $E(g(\theta)|\mathbf{x})$。因此，只要样本抽的足够多，我们可以得到任意所需的精度。这种方法就称为 Monte Carlo 估计。目前，MC 方法已经成为了贝叶斯统计分析中最为常用的近似方法。

2 再加一个 MC

我们看到，在 MC 方法中有一个要求，那就是抽取的样本都要求是独立的，但是在一些问题中，从 $\pi(\theta|\mathbf{x})$ 中抽取独立样本是非常困难的。这时就要靠 MCMC 方法出马了，MCMC 方法的基本思路就是虽然不能从 $\pi(\theta|\mathbf{x})$ 抽取独立样本，但是可以设计一种策略，抽取一系列的非独立“样本”[1]，这些样本能具有一些非常好的性质，它们与从 $\pi(\theta|\mathbf{x})$ 抽取的独立样本的性质是一样的，那么就可以继续用上面的 MC 方法近似计算了。以上设计策略的过程就是 MCMC 方法中第一个 MC 所做的事情。

所以总结起来再说一遍，MCMC 方法就是，为了在无法抽取独立样本的情况下继续愉快地使用 Monte Carlo 方法，需要设计一个策略抽取一些非独立的“样本”，这些样本其实就是 Markov Chain，它们具有好的性质且与独立样本有同样的作用，然后就可以用这些样本代替独立样本继续用 Monte Carlo 方法。

3 为什么 MCMC 可以

MCMC 方法设计策略从 $\pi(\theta|\mathbf{x})$ 抽取一系列样本并不是随意抽取的，抽取的样本链 $\{\theta^{(0)}, \theta^{(1)}, \theta^{(2)}...\}$ 要能够使用，是需要这个链满足很多条件的。这些条件包括马氏性、平稳性、正常返性、不可约性、非周期性、遍历性。这些条件的定义和性质请参考教科书吧，这里就不赘述了。当抽取的样本链满足上面这些条件后，下面的一些定理从理论上保证了 MCMC 方法的正确性，让大家可以放心大胆地用。

定理 3.1 设 $\theta^{(t)}, t \geq 0$ 是一不可约的非周期的马氏链，π 是其平稳分布，π_0 是其初始分布，则有：

$$\pi_t \to \pi, t \to \infty. \tag{3}$$

其中，π_t 是马氏链在时刻 t 的边际分布。

这个定理表明，当马氏链运行充分长的时间后，$\theta^{(t)}$ 的分布近似为 π。当然还有很多其它的定理一起保证了 MCMC 方法的理论正确性，这里就不多说了，感兴趣的读者去看教科书。

[1]严格来讲这一系列的值是某种随机过程在一些状态下的值

以上就是说在 MCMC 方法中，我们需要设计策略构造一个以后验分布 $\pi(\theta|\mathbf{x})$ 为平稳分布的马氏链，虽然样本间不是独立的，但让它运行足够长的时间后再取这个链的值，就可以近似当作独立的用了。

4 MCMC 采样法之 Gibbs 采样

Gibbs 采样最早由 Geman 提出，并用于 Gibbs 格子点分布，由此得名。Gibbs 采样通常应用于目标分布是多维的场合。下面以超级简单的方式说明下这个方法的原理，假设参数就是 2 维的好了，$\theta = (\theta_{\mathbf{1}}, \theta_{\mathbf{2}})$，Gibbs 采样利用各个参数的满条件分布迭代采样的步骤如下：

- 给定 θ_1 的初始值；
- 从 $\pi(\theta_2|\theta_1^{(t)}, \mathbf{x})$ 中采样 $\theta_2^{(t+1)}$；
- 从 $\pi(\theta_1|\theta_2^{(t+1)}, \mathbf{x})$ 中采样 $\theta_1^{(t+1)}$.

重复后面的两步就可以得到一个 $\theta^{(t)} = (\theta_1^{(t)}, \theta_2^{(t)})$ 的马氏链，当这个马氏链达到平稳状态后其值就可以视为从联合分布 $\pi((\theta_1, \theta_2)|\mathbf{x})$ 中得到的独立样本。

好了，上面基本上就描述清楚了 Gibbs 采样的基本思想。至于什么是参数的满条件后验分布，我觉得就不妨看成边缘后验分布吧。不过你应该很快就提问，如果某个参数的满条件分布还是不太容易采样又该怎么办呢？一般来说主要有以下几种解决方案：

- 如果满条件后验分布满足对数上凸的，则可以用 Gilks 和 Wild 提出的自适应拒绝抽样方法进行抽样，效率奇高。
- Metropolis-Hastings 算法。
- 通过引入辅助变量，拆分后验分布中复杂的项，使得辅助变量与模型参数的满条件后验分布变得容易抽样，这实际上是一种参数空间扩充的方法。
- 切片抽样(Slice Sampling)，这是另一种参数空间扩充的方法。

上面这些方案中有些是很复杂的，比如切片采样。这里我们只捡最简单的了解下，因为这些简单的方法已经可以解决绝大部分问题了。

5 MCMC 采样法之 Metropolis-Hastings 算法

Metropolis-Hastings 算法（简称 MH 算法或者 MH 采样法），最先由 Metropolis 等人提出，后由 Hastings 进行推广。上一节讲的 Gibbs 采样法其实是一种特殊的 MH 采样法，下面再介绍三种特殊又常用的 MH 采样法：Metropolis 采样法、独立性采样和随机游动 Metropolis 采样法。

MCMC 方法的精髓就是构造合适的马氏链，使得其平稳分布就是待抽样的目标分布。在贝叶斯分析中此目标分布就是后验分布 $\pi(\theta|\mathbf{x})$。因此 MH 算法的主要任务也就是产生满足上述要求的马氏链 $\theta^{(t)}, t = 0, 1, 2, ...$，即在给定状态 $\theta^{(t)}$ 下产生下一个状态 $\theta^{(t+1)}$。所有 MH 算法的构造框架如下：

- 构造合适的建议分布（proposal distribution，又称跳跃分布，jumping distribution）$q(\cdot|\theta^{(t)})$;
- 从 $q(\cdot|\theta^{(t)})$ 产生候选点 θ';
- 按一定的接受概率形成的准则判断是否应该接受 θ'。若接受，则 $\theta^{(t+1)} = \theta'$，否则令 $\theta^{(t+1)} = \theta^{(t)}$.

建议分布的选取是 MH 算法实现的关键，理论上来说，任何在参数空间上可产生不可约非周期马氏链的建议分布都是可行的，它可使得产生的马氏链的平稳分布为目标抽样分布，在本文档所讨论的范围内即后验分布 $\pi(\theta|\mathbf{x})$。但是在实际中，建议分布的好坏会影响 MCMC 采样的效率，它可以直接通过接受概率的大小来反映。说了这么多，下面就给出 MH 算法产生马氏链的过程：

1 构造合适的建议分布 $q(\cdot|\theta^{(t)})$;

2 从个某个分布中产生 $\theta^{(0)}$（通常是直接给定的）;

3 从 $q(\cdot|\theta^{(t)})$ 产生候选点 θ';

4 从均匀分布 $U(0, 1)$ 产生 U;

5 判断：若

$$U \leq r(\theta^{(t)}, \theta') \doteq \frac{\pi(\theta'|\mathbf{x})\mathbf{q}(\theta^{(\mathbf{t})}|\theta')}{\pi(\theta^{(t)}|\mathbf{x})\mathbf{q}(\theta'|\theta^{(\mathbf{t})})}, \tag{4}$$

则接受 θ'，且令 $\theta^{(t+1)} = \theta'$，否则令 $\theta^{(t+1)} = \theta^{(t)}$.

6 $t \to t + 1$，回到第 3 步重复.

上述算法的接受概率为

$$a(\theta^{(t)},\theta')=\min\left(1,\frac{\pi(\theta'|\mathbf{x})\mathbf{q}(\theta^{(\mathbf{t})}|\theta')}{\pi(\theta^{(t)}|\mathbf{x})\mathbf{q}(\theta'|\theta^{(\mathbf{t})})}\right). \tag{5}$$

Gibbs 采样是一种特殊的 MH 采样法，其中的每一个候选点都会被接受，下面介绍几种另外的特殊又常见的 MH 算法。

5.1 Metropolis 采样

Metropolis 采样的建议分布是对称的，即满足

$$q(X|Y)=q(Y|X). \tag{6}$$

相应的接受概率为

$$a(\theta^{(t)},\theta')=\min\left(1,\frac{\pi(\theta'|\mathbf{x})}{\pi(\theta^{(t)}|\mathbf{x})}\right). \tag{7}$$

5.2 随机游动 Metropolis 采样

随机游动 Metropolis 采样的建议分布为

$$q(X|Y)=q(|X-Y|). \tag{8}$$

实际使用时可先从 $q(\cdot)$ 中产生一个增量 Z，然后取候选点为 $\theta'=\theta^{(t)}+Z$。比如，从分布 $N(\theta^{(t)},\sigma^2)$ 中产生的候选点 θ' 可表示为 $\theta'=\theta^{(t)}+\sigma Z$，其中 Z 从标准正态分布中产生，$\sigma^2>0$ 已知[2]。

5.3 独立性采样

独立性采样法[3]的建议分布中并不依赖前面的历史值，即 $q(\cdot|\theta^{(t)})=q(\cdot)$，这时的接受概率为

$$a(\theta^{(t)},\theta')=\min\left(1,\frac{\pi(\theta'|\mathbf{x})\mathbf{q}(\theta^{(\mathbf{t})})}{\pi(\theta^{(t)}|\mathbf{x})\mathbf{q}(\theta')}\right). \tag{9}$$

[2](1)由大样本性质，后验分布通常都具有较好的正态性，因此常常选择正态分布为 MH 算法的建议分布，其均值为上一个状态的值，而方差的大小决定了所得马氏链在参数空间支撑上的混合程度。因此建议分布的好坏常受此参数的影响。

(2)在随机游动 Metropolis 采样中，当增量的方差太大时，大部分的候选点会被拒绝，导致算法的效率很低；而当增量的方差太小时，几乎所有的候选点都会被接受，这时候得到的链就几乎是随机游动。因此建议分布过大或过小的方差都会导致算法效率较低，通常的方法是在实施采样时监控接受概率，有专家建议接受概率在[0.15,0.5]区间内时，链具有较好的性质。

[3](1)要实现几何收敛，建议分布的尾部必须比目标分布的尾部更厚；(2)独立性采样法容易实施，但仅在建议分布与目标分布很接近时表现较好，因此在实际中很少单独使用。

6 MCMC 收敛性诊断

不管是一般的 MH 采样方法还是特殊的 Gibbs 采样方法，都需要确定所得到的马氏链已经收敛了才能使用采样得到的值，即需要确定马氏链达到收敛时迭代的次数（之前的一段链称为 burn-in 样本）。这通常没有一个全能统一的方法，在有些问题中马氏链的收敛速度会很慢，特别是多参数的情形；有些时候因为初始点选择不当会产生虚假收敛性，因为马氏链可能陷入目标分布的一个局部支撑上。大体上收敛性诊断主要有两类方法，一是从图形角度判断，二是从数量上判断。

6.1 收敛性诊断图

理论上，MCMC 方法采样得到的马氏链的平稳分布与初始值的选取无关，但马氏链的收敛速度会对初始点较敏感，因此可以用下面两种图形方法来判断收敛性：

（1）样本路径图

将生产的马氏链按迭代次数作图就是这个马氏链的样本路径图（trace plot）啦。从不同初始点产生多条马氏链，都画出来它们的样本路径图，如果这些样本路径图在一段时间后都稳定下来并混在一起无法区别，就可以认为采样已经收敛了。

（2）遍历均值图

MCMC 方法的理论基础是遍历均值定理，因此我们可以监控遍历均值是否达到收敛。将所生成的马氏链的累积均值对迭代次数作图就是此链的遍历均值图（ergodic mean plot）。达到平稳状态后的遍历均值会趋于一条水平的直线。同样，可以换不同的初始点多画几个。

6.2 收敛性指标

主要方法就是：**（1）MC-误差 （2）Gelman-Rubin 法**

看到这里，如果还是简单的写写思想，估计已经很难满足读者的需求了，所以为了不误导读者，请大家还是去参考相关专业书籍吧，尤其是推荐茆(mǎo)诗松老先生的《贝叶斯统计》第二版的第七章“贝叶斯计算”，讲的那是通俗易懂，深入浅出，又有大量例子帮助理解，强烈推荐。