第 19 章马尔可夫蒙特卡洛

统计学习方法

2024 年 10 月 20 日

前言

- 蒙特卡罗法，是通过从概率模型的随机抽样进行近似数值计算的方法，又称为统计模拟方法
- 马尔可夫链蒙特卡罗法（Markov Chain Monte Carlo, MCMC），则是以马尔可夫链为概率模型的蒙特卡罗法
  - M-H 算法：Metropolis 等人 1953 年提出原始算法；Hastings 在 1970 年改进，形成了现在的形式
  - Gibbs 抽样：1984年由 S. Geman 和 D. Geman 提出，是更简单、使用更广泛的 MCMC 方法
  - MCMC 应用于**概率分布的估计、定积分的近似计算、最优化问题的近似求解**等问题，特别是被应用于**统计学习中概率模型的学习与推理**

目录

解决的问题

- **统计学：** 通过收集和分析数据来推断总体的特征
- **机器学习：** 进一步，它通过训练模型来学习数据的潜在模式和结构，来逼近这个分布
- **蒙特卡罗：** 假设概率分布的定义已知，通过抽样获得概率分布的随机样本，并通过得到的随机样本对概率分布的特征进行分析

因此，蒙特卡罗法的核心是随机抽样 (random sampling)，一般方法有

- 直接抽样
- 接受-拒绝抽样
- 重要性抽样

接受-拒绝抽样、重要性抽样适合于概率密度函数复杂（如密度函数含有多个变量，各变量相互不独立，密度函数形式复杂），不能直接抽样的情况。这里只介绍接受-拒绝抽样（accept-reject sampling method）。

接受-拒绝抽样

假设有随机变量 x, 取值 $x \in \mathcal{X}$, 其概率密度函数为 $p(x)$ （不可以直接抽样），目标是得到该概率分布的随机样本,。接受-拒绝法的基本想法：

- 找一个可以直接抽样的分布 $q(x)$ ，称为建议分布（proposal distribution），要求 $cq(x) \geq p(x), c > 0$
  - $q(x)$ 的选择：与 $p(x)$ 外形相近，容易采样
  - c 的取值：$c \geq \max\{\frac{p(x)}{q(x)}\}$
- 用 $q(x)$ 进行抽样，假设得到结果是 x^* ，再以接受率 $\alpha(x^*) = \frac{p(x^*)}{cq(x^*)}$ 决定是否接受 x^* 。

直观上，落到 $p(x^*)$ 范围内的就接受，落到 $p(x^*)$ 范围外的就拒绝。

Algorithm 1: 接受-拒绝法

输入: 目标分布的概率密度函数 $p(x)$
输出: 随机样本 $x_1, x_2, \cdots, x_n$
参数: 样本数 n

1 选择建议分布 $q(x)$ 满足 $cq(x) \geqslant p(x)$，其中 $c > 0$
2 **while** 样本数 $< n$ **do**
3 　按照建议分布 $q(x)$ 随机抽样得到样本 x^*，再按照均匀分布在 $(0, 1)$ 范围内抽样得到 u
4 　**if** $u \leqslant \alpha(x^*)$ **then**
5 　　将 x^* 作为抽样结果;
6 　　样本数 $+1$;
7 　**else**
8 　　回到步骤 2;
9 　**end**
10 **end**

优点：容易实现。缺点：如果 $p(x)$ 的涵盖体积占 $cq(x)$ 的涵盖体积的比例很低，就会导致拒绝的比例很高，抽样效率很低。注意，一般是在高维空间进行抽样，即使 $p(x)$ 与 $cq(x)$ 很接近，两者涵盖体积的差异也可能很大。

数学期望估计

一般的蒙特卡罗法可以用于**数学期望估计**。假设 $f(x)$ 为定义在样本空间 $\mathcal{X}$ 上的函数，目标是求函数 $f(x)$ 关于密度函数 $p(x)$ 的数学期望 $E_{p(x)}[f(x)]$ 。

针对这个问题，蒙特卡罗法用以上方法按照概率分布 $p(x)$ 独立地抽取 n 个样本 $x_1, x_2, \cdots, x_n$ ，计算函数 $f(x)$ 的样本均值 $\hat{f}_n$

$$\hat{f}_n = \frac{1}{n} \sum_{i=1}^{n} f(x_i) \tag{1}$$

作为数学期望 $E_{p(x)}[f(x)]$ 的近似值。根据大数定律可知，当 $n \to \infty$ 时，样本均值以概率 1 收敛于数学期望：

$$P(\lim_{n\to\infty} \hat{f}_n(x) = E_{p(x)}[f(x)]) = 1 \quad 或 \quad \hat{f}_n \to a.s. E_{p(x)}[f(x)] \tag{2}$$

这样就得到了数学期望的近似计算方法：

$$E_{p(x)}[f(x)] \approx \frac{1}{n} \sum_{i=1}^{n} f(x_i) \tag{3}$$

积分计算

一般的蒙特卡罗法也可以用于**定积分的近似计算**，称为**蒙特卡罗积分**。假设有一个函数 $h(x)$，目标是计算该函数的（*Lebesgue*）积分 $\int_{\mathcal{X}} h(x)\mathrm{d}x$。

将函数 $h(x)$ 分解成一个函数 $f(x)$ 和一个概率密度函数 $p(x)$ 的乘积的形式，那么就有

$$\int_{\mathcal{X}} h(x)\mathrm{d}x = \int_{\mathcal{X}} f(x)p(x)\mathrm{d}x = E_{p(x)}[f(x)] \tag{4}$$

就是说，任何一个函数的积分都可以表示为某一个函数的数学期望，而函数的数学期望又可以通过函数的样本均值估计。实际上，给定一个概率密度函数 $p(x)$ ，只要取 $f(x) = \frac{h(x)}{p(x)}$ ，就可得式（4）

$$\int_{\mathcal{X}} h(x)\mathrm{d}x = \int_{\mathcal{X}} \frac{h(x)}{p(x)}p(x)\mathrm{d}x = E_{p(x)}[\frac{h(x)}{p(x)}] \approx \frac{1}{n}\sum_{i=1}^{n}\frac{h(x_i)}{p(x_i)} \tag{5}$$

例1 用蒙特卡罗积分法求 $\int_0^1 e^{-x^2/2}\,\mathrm{d}x$

令 $f(x)=\mathrm{e}^{-x^2/2}$ ，$p(x)=1 \quad (0<x<1)$

也就是说，假设随机变量 x 在 $(0,1)$ 区间遵循均匀分布。使用蒙特卡罗积分，在 $(0,1)$ 区间按照均匀分布抽取 10 个随机样本 $x_1,x_2,\cdots,x_{10}$ ，计算样本的函数均值 $\hat{f}_{10}=\frac{1}{10}\sum_{i=1}^{10}\mathrm{e}^{-x_i^2/2}=0.832$

例2 用蒙特卡罗积分法求 $\int_{-\infty}^{\infty} x\frac{1}{\sqrt{2\pi}}\exp\left(\frac{-x^2}{2}\right)\mathrm{d}x$

令 $f(x)=x$ ，$p(x)=\frac{1}{\sqrt{2\pi}}\exp\left(\frac{-x^2}{2}\right)$

$$\int_{-\infty}^{\infty} xp(x)\mathrm{d}x = E_{p(x)}(x)$$

使用蒙特卡罗积分，按照标准正态分布在区间 $(-\infty,\infty)$ 抽样 $x_1,x_2,\cdots,x_n$ ，取其平均值, 就得到要求的积分值。我们知道，当样本增大时, 积分值趋于 0 （标准正态分布均值）。

存在问题：一般的蒙特卡罗法，如接受-拒绝抽样法、重要性抽样法，能解决不能直接抽样的情形，**但抽样效率不高**。下面要提出的MCMC算法能够提高抽样效率，同样也适合于概率密度函数复杂，不能直接抽样的情况。

定义 (随机过程)

考虑一个随机变量的序列 $\{X_t, t \geq 0\}$ ，X_t 表示时刻 t 的随机变量。每个随机变量的取值集合相同，称为状态空间 $\mathcal{S}$ 。时间 T 、状态空间 $\mathcal{S}$、随机变量X 可以是离散的，也可以是连续的。以上随机变量的序列构成随机过程（stochastic process）。

- 时间：离散 $0, 1, 2, \ldots,$，即时间取离散值，通常等间隔 ;连续 $[0, \infty)$，连续区间取任意值
- （状态）空间：离散 $\mathcal{Z}$ ，即取有限、可数个值；连续 $\mathcal{R}$ ，实数轴上任意值
- 随机变量：固定时刻 t ，随机过程就是一个随机变量，又分连续、离散随机变量

定义 (马尔科夫链)

假设在时刻 0 的随机变量 X_0 遵循概率分布 $P(X_0) = \pi_0$ ，称为初始状态分布。在某个时刻 $t \geqslant 0$ 的随机变量 X_{t+1} 与前一个时刻的随机变量 X_t 之间有条件分布 $P(X_{t+1} \mid X_t)$ ，如果 X_{t+1} 只依赖于 X_t ，而不依赖于过去的随机变量 $\{X_0, X_1, \cdots, X_{t-1}\}$, 这一性质称为**马尔可夫性**, 即

$$P(X_{t+1} \mid X_t, \cdots, X_1, X_0) = P(X_{t+1} \mid X_t), \quad t = 0, 1, 2, \cdots \tag{6}$$

具有马尔可夫性的随机序列 $\{X_t, t \geq 0\}$ 称为马尔可夫过程（Markov process）。条件概率分布 $P(X_t \mid X_{t-1})$ 称为马尔可夫链的转移概率分布，刻画了马尔可夫链的特性。

定义 (时间齐次马尔可夫链)

若转移概率分布 $P(X_{t+1} \mid X_t)$ 与 t 无关, 即

$$P(X_{t+1+s} \mid X_{t+s}) = P(X_{t+1} \mid X_t), \quad t = 1, 2, \cdots; \quad s = 1, 2, \cdots \tag{7}$$

则称该马尔可夫链为时间齐次的马尔可夫链（time homogenous Markov chain）。

时间齐次描述了：**转移概率不随时间变化，只与当前状态、下一个状态有关**，这就简化了模型。本书中提到的马尔可夫链都是时间齐次的。

定义 (n 阶马尔可夫链)

以上定义的是一阶马尔可夫链，可以扩展到 n 阶马尔可夫链，满足 n 阶马尔可夫性

$$P(X_{t+1} \mid X_t X_{t-1} \cdots X_1 X_0) = P\left(X_{t+1} \mid X_t X_{t-1} \cdots X_{t-(n-1)}\right) \tag{8}$$

n 阶马氏性是说：**未来依赖于现在，以及以前 $n-1$ 个状态**。本书中提到的马尔可夫链都是一阶的。

转移概率矩阵

离散状态马尔可夫链 $\{Xt, t \geq 0\}$，随机变量 $X_t(t = 0, 1, 2, \cdots)$定义在离散状态空间 $\mathcal{S}$，转移概率分布可以由转移概率矩阵 P 表示。

若马尔可夫链在时刻 t 处于状态 i, 在时刻 $t+1$ 移动到状态 j, 将转移概率记作

$$p_{ij} = (X_{t+1} = j \mid X_t = i), \quad i = 0, 1, \cdots, n; \quad j = 0, 1, \cdots, n \tag{9}$$

满足$p_{ij} \geqslant 0, \quad \sum_j p_{ij} = 1$,马尔可夫链的转移概率 p_{ij} 可以由矩阵表示，即

$$P = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1n} \\ p_{21} & p_{22} & \cdots & p_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ p_{n1} & p_{n2} & \cdots & p_{nn} \end{bmatrix}$$

称为马尔可夫链的转移概率矩阵 P，满足条件 $p_{ij} \geqslant 0, \sum_j p_{ij} = 1$ （行和为 1 ）。满足这两个条件的矩阵称为随机矩阵（stochastic matrix）。

状态分布

考虑马尔可夫链 $\{X_t, t \geq 0\}$ 在时刻 t 的概率分布，称为状态分布，记作

$$\pi(t) = \begin{bmatrix} \pi_1(t) & \pi_2(t) & \cdots & \pi_n(t) \end{bmatrix}$$

其中 $\pi_i(t) = P(X_t = i)$ 表示时刻 t 状态为 i 的概率。通常初始分布 $\pi(0)$ 的向量只有一个分量是 1，其余分量都是 0，表示马尔可夫链从一个具体状态开始。**有限离散状态的马尔可夫链可以由有向图表示。结点表示状态，边表示状态之间的转移，边上的数值表示转移概率。**

例 3 假设某地的天气具有一定的规律，天气的变化具有马氏性，即明天的天气只依赖于今天的天气，而与昨天及以前的天气无关。具体地，如果今天是晴天，那么明天是晴天的概率是 0.9，是雨天的概率是 0.1；如果今天是雨天，那么明天是晴天的概率是 0.5，是雨天的概率也是 0.5。我们可以画出这个有向图。

马尔可夫链在时刻 t 的状态分布, 可以由在时刻 $t-1$ 的状态分布以及转移概率分布决定

$$\pi(t)=\pi(t-1)P \tag{10}$$

这是因为

$$\begin{aligned}\pi_j(t)=P\left(X_t=j\right)&=\sum_i P\left(X_{t-1}=i, X_t=j\right)\\&=\sum_i P\left(X_{t-1}=i\right) P\left(X_t=j \mid X_{t-1}=i\right)\\&=\sum_i \pi_i(t-1) p_{ij}\end{aligned}$$

马尔可夫链在时刻 t 的状态分布, 由Chapman-Kolmogorov 方程（用于计算 n 步转移概率), 于是递推得到

$$\pi(t)=\pi(0)P^t \tag{11}$$

这里的 P^t 称为 t 步转移概率矩阵，矩阵元素 $P_{ij}^t=P\left(X_t=j \mid X_0=i\right)$ 表示 0 时刻从状态 i 出发, t 时刻达到状态 j 的 t 步转移概率， P^t 自然也是随机矩阵。式（11）说明，马尔可夫链的状态分布可以直接由初始分布和转移概率分布决定。

例4 对于上个例子（天气）的马尔可夫链，转移矩阵为

$$P = \begin{bmatrix} 0.9 & 0.1 \\ 0.5 & 0.5 \end{bmatrix}$$

如果第一天是晴天的话，其天气概率分布（初始状态分布）如下：

$$\pi(0) = \begin{bmatrix} 1 & 0 \end{bmatrix}$$

根据这个马尔可夫链模型，可以计算第二天、第三天及之后的天气概率分布（状态分布）

$$\pi(1) = \pi(0)P = \begin{bmatrix} 1 & 0 \end{bmatrix} \begin{bmatrix} 0.9 & 0.1 \\ 0.5 & 0.5 \end{bmatrix} = \begin{bmatrix} 0.9 & 0.1 \end{bmatrix}$$

$$\pi(2) = \pi(0)P^2 = \begin{bmatrix} 1 & 0 \end{bmatrix} \begin{bmatrix} 0.9 & 0.1 \\ 0.5 & 0.5 \end{bmatrix}^2 = \begin{bmatrix} 0.86 & 0.14 \end{bmatrix}$$

平稳分布

设有马尔可夫链 $\{X_t, t \geq 0\}$，状态空间为 $\mathcal{S}$，转移概率矩阵为 $P = (p_{ij})$

定理 (平稳分布)

如果存在状态空间 $\mathcal{S}$ 上的一个分布 π

$$\pi = \begin{bmatrix} \pi_1 & \pi_2 & \cdots & \pi_n \end{bmatrix}$$

使得 $\pi P = \pi$，则称 π 为马尔可夫链的平稳分布。

引理

分布 $\pi = (\pi_1, \pi_2, \cdots, \pi_n)$ 为 X 的平稳分布的充分必要条件是 π 是下列方程组的解:

$$\pi_j = \sum_i \pi_i p_{ij}, \quad j = 1, 2, \cdots, n \tag{12}$$

$$\pi_j \geqslant 0, \sum_j \pi_j = 1 \quad j = 1, 2, \cdots, n \tag{13}$$

证明 必要性。假设 $\pi=(\pi_1,\pi_2,\cdots,\pi_n)$ 是平稳分布，显然满足式（13）。又

$$\pi_j=\sum_i \pi_i p_{ij}, \quad j=1,2,\cdots,n$$

即 $\pi=(\pi_1,\pi_2,\cdots,\pi_n)$ 满足式（12）。

充分性。由式 (13) 知 $\pi=(\pi_1,\pi_2,\cdots,\pi_n)$ 是一概率分布。假设 π 为 X_t 的分布, 则

$$P\left(X_t=j\right)=\pi_j=\sum_i \pi_i p_{ij}=\sum_i P\left(X_{t-1}=i\right) p_{ij}, \quad j=1,2,\cdots,n$$

说明 $\pi=(\pi_1,\pi_2,\cdots,\pi_n)$ 也为 X_{t-1} 的分布。这对任意 t 成立，所以 $\pi=(\pi_1,\pi_2,\cdots,\pi_n)$是马尔可夫链的平稳分布。

例5 设有如下所示马尔可夫链，其转移概率矩阵如下，求其平稳分布

$$P = \begin{bmatrix} 1/2 & 1/4 & 1/4 \\ 1/2 & 0 & 1/2 \\ 1/4 & 1/4 & 1/2 \end{bmatrix}$$

解 设平稳分布为 $\pi = (\pi_1, \pi_2, \pi_3)$，则由式 (12) 、(13) 有

$$\pi_1 = \frac{1}{2}\pi + \frac{1}{2}\pi_2 + \frac{1}{4}\pi_3$$

$$\pi_2 = \frac{1}{4}\pi_1 + \frac{1}{4}\pi_3$$

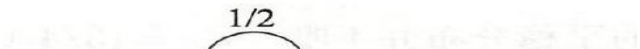
$$\pi_3 = \frac{1}{4}\pi_1 + \frac{1}{2}\pi_2 + \frac{1}{2}\pi_3$$

$$\pi_1 + \pi_2 + \pi_3 = 1$$

$$\pi_i \geqslant 0, \quad i = 1, 2, 3$$

解方程组，得到唯一的平稳分布

$$\pi = (2/5 \quad 1/5 \quad 2/5)$$



例 6 设有图所示马尔可夫链，其转移概率分布如下，求其平稳分布。

$$\begin{bmatrix} 1 & 1/3 & 0 \\ 0 & 1/3 & 0 \\ 0 & 1/3 & 1 \end{bmatrix}$$

解 这个马尔可夫链的平稳分布并不唯一， $\pi = \begin{pmatrix} \frac{3}{4} & 0 & \frac{1}{4} \end{pmatrix}, \pi = \begin{pmatrix} \frac{2}{3} & 0 & \frac{1}{3} \end{pmatrix}$等皆为其平稳分布。

马尔可夫链可能存在唯一平稳分布，无穷多个平稳分布，或不存在平稳分布。

连续状态马尔可夫链 $\{X_t, t \geq 0\}$, 随机变量 $X_t(t=0,1,2,\cdots)$ 定义在连续状态空间 $\mathcal{S}$ ，转移概率分布由概率转移核或转移核（transition kernel）表示。

对任意的状态（集合） $x \in \mathcal{S}, A \subset \mathcal{S}$ ，转移核 $P(x, A)$ 定义为

$$P(x, A)=\int_A p(x, y) \mathrm{d} y \tag{14}$$

其中 $p(x, \bullet)$ 是概率密度函数, 满足 $p(x, \cdot) \geqslant 0, P(x, \mathcal{S})=\int_{\mathcal{S}} p(x, y) \mathrm{d} y=1$ 。转移核 $P(x, A)$ 表示从 $x \sim A$ 的转移概率

$$P(x, A)=P\left(X_{t+1}=A \mid X_t=x\right) \tag{15}$$

若马尔可夫链的状态空间 $\mathcal{S}$ 上的概率分布 $\pi(x)$ 满足条件

$$\pi(y)=\int \pi(x) p(x, y) \mathrm{d} x, \quad \forall y \in \mathcal{S} \tag{16}$$

则称分布 $\pi(x)$ 为该马尔可夫链的平稳分布。等价地,

$$\pi(A)=\int \pi(x) P(x, A) \mathrm{d} x, \quad \forall A \subset \mathcal{S} \tag{17}$$

不可约性

马氏链有这么多状态，难道每个状态我们要逐个分析吗？显然不是，所以我们需要对状态进行分类。

定义 (不可约)

如果存在一个时刻 $t(t>0)$ 满足

$$P\left(X_t=j \mid X_0=i\right)=p_{ij}^t>0 \tag{18}$$

那么对于状态 i 来说，状态 j 是可达（accessible）。进一步地，如果对于状态 j ，状态 i 是可达的，我们就说两个状态互通（communication），在同一个类里面。如果一个马氏链只存在一个类，即里面的所有状态都是互通的（自然也是相互可达的），那么就称这个马氏链是不可约的。

例7 如图所示马尔可夫链是可约的

解

$$P = \begin{bmatrix} 0 & 1 & 0 \\ 1/2 & 0 & 1/2 \\ 0 & 0 & 1 \end{bmatrix}$$

平稳分布 $\pi = \begin{pmatrix} 0 & 0 & 1 \end{pmatrix}$ 。该马尔可夫链, 转移到状态 3 后, 就在该状态上循环跳转,不能到达状态 1 和状态 2 ，最终停留在状态 3 （吸收态）。

非周期

定义 (非周期)

如果时刻 0 从状态 i 出发， t 时刻返回状态的所有时间 $\left\{t : P\left(X_t = i \mid X_0 = i\right) = p_{ii}^t > 0\right\}$ 的最大公约数是 1, 则称此马尔可夫链 X 是非周期的 (aperiodic), 否则称马尔可夫链是周期的 (periodic)。

例 8 如图所示的马尔可夫链是周期的

解 转移概率矩阵

$$\begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}$$

其平稳分布是 $\pi = \begin{pmatrix} 1/3 & 1/3 & 1/3 \end{pmatrix}$ 。此马尔可夫链从每个状态出发，返回该状态的时刻都是 3 的倍数，$\{3, 6, 9\}$，具有周期性，最终停留在每个状态的概率都为 $1/3$ 。

正常返

定义

首达概率：对于任意状态 $i, j \in \mathcal{S}$，定义概率 p_{ij}^t 为时刻 0 从状态 i 出发，时刻 t 首次转移到状态 j 的概率，即

$$f_{ij}^t = P\left(X_t = j, X_s \neq j, s = 1, 2, \cdots, t-1 \mid X_0 = i\right), t = 1, 2, \cdots$$

常返性：令 $f_{ij} = \sum_{n=1}^{\infty} f_{ij}^t$ ，f_{ij}表示过程开始处于 i，迟早到达 j 的概率。如果 $f_{jj} = 1$ ，我们就称状态 j 是常返的，否则就是非常返的。

正常返与零常返：如果状态是常返的，令 $\mu_j = \sum_{t=1}^{\infty} t f_{jj}^t$ 表示从 j 返回到 j 的期望转移次数

- 如果这个级数收敛（期望存在），也就是 $\mu_j < \infty$，那么该状态是正常返的
- 如果级数不收敛，也就是$\mu_j = \infty$，那么该状态是零常返的

定理

若 ij 互通，若 j 非周期，则 $\lim\limits_{t \to \infty} P_{ij}^t = 1/\mu_{jj}$

若 ij 互通，若 j 有周期 d，则 $\lim\limits_{t \to \infty} P_{ij}^{td} = d/\mu_{jj}$

推论

- 令
$$\pi_j = \lim_{t\to\infty} P_{ij}^{td}$$
，若 $\pi_j > 0$，那么常返态 j 是正常返的（**书上的定义实际是这个推论**，而且是用 td 步转移概率 P_{ij}^{td} 定义，而非 t 步的首达概率 f_{ij}^{t} ）
- 有限状态、不可约的马氏链都是正常返的，于是方程组 $\pi P = \pi, \pi \mathbf{1}^T = 1$ 有解（平稳分布存在）
- 不可约、非周期的马氏链要么是所有状态都是非常返的或都是零常返的，要么都是正常返的（非常返、零常返不存在平稳分布）
- 有限状态、不可约、非周期的马氏链，方程组就有唯一解（存在唯一平稳分布），此时平稳分布又是极限分布，即
$$\pi_j = \lim_{t\to\infty} P_{ij}^{t}$$

例 9 如图所示无限状态马尔可夫链，当 $p > q$ 时是正常返的，当 $p \leqslant q$ 不是正常返的。

解 转移概率矩阵

$$\begin{bmatrix} p & q & 0 & 0 & \\ p & 0 & q & 0 & \cdots \\ 0 & p & 0 & q & \\ 0 & 0 & p & 0 & \\ & \vdots & & & \ddots \end{bmatrix}$$

当 $p > q$ 时，平稳分布是 $\pi_i = \left(\frac{q}{p}\right)^i \left(\frac{p-q}{p}\right), \quad i = 1, 2, \cdots$ 。当时间趋于无穷时，转移到任何一个状态的概率不为 0，马尔可夫链是正常返的。

当 $p \leqslant q$ 时，不存在平稳分布，马尔可夫链不是正常返的

马氏链的遍历定理

遍历理论是数学的一个重要分支，主要研究**动态系统的长时间行为与统计性质**，与物理、统计领域关系密切。

- 1930s，George D. Birkhoff 发表了遍历定理，证明了在某些条件下，时间平均和空间平均相等的结果。这是遍历理论最重要的成果之一，标志着遍历理论的建立。
- 1950s-1970s ，K. Itô 将遍历理论与马氏链结合，研究随机系统的长期行为。

定理 (遍历定理)

设 X_n 是不可约、正常返马氏链, 则 $\{X_n\}$有平稳分布 $\left\{\pi_i = \frac{1}{\mu_i}\right\}$, μ_j 是从 j 出发返回 j 的平均时间, 有

$$\lim_{N\to\infty} \frac{1}{N}\sum_{n=1}^{N} I_{\{X_n=i\}} = \pi_i,\ a.s.$$

说明平稳分布 π_i 是马氏链长期处于状态 i 的时间比例.

推论

设 $\{X_n\}$ 是不可约、正常返马氏链, 则有 $\{X_n\}$ 平稳分布 $\{\pi_i\}$, 设随机变量 Y 分布为 $\{\pi_i\}$, 函数 $f(x)$ 定义在状态空间 S 上, 满足 $E|f(Y)| < \infty$, 则

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} f(X_n) = Ef(Y), \quad a.s.$$

（书）设 $\{X_n\}$ 是不可约、非周期、正常返的马氏链, 则 $\{X_n\}$ 存在唯一的平稳分布 $\{\pi_i\}$ *(*也是极限分布*)*, 设定义在状态空间 S 上的函数 $f(\cdot)$ 满足 $E|f(Y)| < \infty$, 则

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} f(X_n) = Ef(Y), \quad a.s.$$

这个定理说明, 可以用模拟马氏链的方法来估计与平稳分布 $\{\pi\}$ 有关的数字特征。设随机向量 Y 服从分布 $\{\pi\}, E[f(Y)]$ 很难计算, 直接生成 Y 的简单随机样本也很困难, 就可以设计马氏链 $\{X_n\}$, 使得 $\{X_n\}$ 遍历且以 $\{\pi\}$ 为平稳分布（也是极限分布）, 则可以用 $\frac{1}{N}\sum_{n=1}^{N} f(X_n)$ 估计 $E[f(Y)]$ 。

时间平均等于空间平均：当时间趋于无穷时，样本均值可认为是时间平均（因为对时间取平均），数学期望可以认为是空间平均（系统在各状态出现的加权平均）。反映了在长期的时间演化过程中，系统的统计行为趋向于反映其状态空间的概率分布。

举个例子：记录摇骰子的平均点数

- 时间平均：通过多次摇骰子，记录每次结果，最终计算这些结果的平均值。随着摇骰子次数的增加，这个平均值会趋近于数学期望 3.5
- 空间平均：直接通过计算所有可能结果（1到6）的加权平均

$$\langle X\rangle_{\text{space}} = 1\cdot\frac{1}{6}+2\cdot\frac{1}{6}+3\cdot\frac{1}{6}+4\cdot\frac{1}{6}+5\cdot\frac{1}{6}+6\cdot\frac{1}{6}=\frac{21}{6}=3.5$$

理论上并不知道经过多少次迭代，马尔可夫链的状态分布才能接近于平稳分布，在实际应用遍历定理时，取一个足够大的整数 m ，经过 m 次迭代之后认为状态分布就是平稳分布，这时计算从第 $m+1$ 次迭代到第 n 次迭代的均值，称为遍历均值，即

$$\hat{E}f=\frac{1}{n-m}\sum_{i=m+1}^{n} f\left(x_i\right) \tag{19}$$

可逆马氏链

设有马尔可夫链 $\{X_t, t \geq 0\}$，对于任意状态 $i, j \in \mathcal{S}$, 对任意一个时刻 t 满足

$$\pi_i P\left(X_t = j \mid X_{t-1} = i\right) = P\left(X_{t-1} = i \mid X_t = j\right) \pi_j, \quad i, j = 1, 2, \cdots \tag{20}$$

或简写为

$$\pi_i p_{ij} = \pi_j p_{ji}, \quad i, j = 1, 2, \cdots \tag{21}$$

则称为可逆马尔可夫链（Reversible Markov chain），式（21）称为细致平衡方程（detailed balance equation）。直观上，如果有可逆的马尔可夫链，那么以该马尔可夫链的平稳分布作为初始分布，进行随机状态转移，无论是面向未来还是面向过去，任何一个时刻的状态分布都是该平稳分布。

定理

满足细致平衡方程的状态分布 π 就是该马尔可夫链的平稳分布，即 $\pi P = \pi$

该定理说明，可逆马尔可夫链一定有平稳分布（而且还是唯一的平稳分布，可逆马氏链一定是不可约）给出了一个马尔可夫链有平稳分布的充分条件 (不是必要条件)

例 10 如图所示马尔可夫链是不可逆的

解 转移概率矩阵

$$\begin{bmatrix} 1/4 & 1/4 & 1/2 \\ 1/2 & 0 & 1/2 \\ 1/4 & 1/2 & 1/4 \end{bmatrix}$$

平稳分布 $\pi = \begin{pmatrix} 8/25 & 7/25 & 2/5 \end{pmatrix}$，不满足细致平稳方程。

细致平稳方程理解：从方程来看，从状态 i 转移出去的概率等于外界流入i 的概率，说明系统达到了平衡态。

- 物理场景：一个具有大量交互作用粒子的系统达到平衡后，同一时刻，同一区域流出的粒子数量等于外界流入的粒子。加盐加糖！
- 化学反应：对于一个可逆化学反应，分析其正负反应速率，达到化学平衡后，正反应速率等于负反应速率。酯化反应！

基本步骤

对比一般的蒙特卡洛算法，MCMC 方法更适合于**随机变量是多元的、密度函数是非标准形式的、随机变量各分量不独立**等情况。

假设多元随机变量 x，满足 $x \in \mathcal{X}$，其概率密度函数为 $p(x), f(x)$ 为定义在 $x \in \mathcal{X}$ 上的函数，目标是获得概率分布 $p(x)$ 的样本集合，以及求函数 $f(x)$ 的数学期望 $E_{p(x)}[f(x)]$ 。

MCMC 方法概括为以下三步：

- 首先，在随机变量 x 的状态空间 $\mathcal{S}$ 上构造一个满足遍历定理的马尔可夫链，使其平稳分布为目标分布 $p(x)$
- 从状态空间的某一点 x_0 出发，用构造的马尔可夫链进行随机游走，产生样本序列 $x_0, x_1, \cdots, x_t, \cdots$ 。
- 应用马尔可夫链的遍历定理，确定正整数 m 和 $n, (m < n)$，得到样本集合 $\{x_{m+1}, x_{m+2}, \cdots, x_n\}$，求得函数 $f(x)$ 的均值 (遍历均值)

$$\hat{E} f = \frac{1}{n-m} \sum_{i=m+1}^{n} f\left(x_i\right) \tag{22}$$

重要问题

（1）如何定义马尔可夫链，保证马尔可夫链蒙特卡罗法的条件成立。

构建满足细致平稳方程的可逆马氏链，保证遍历定理成立。

（2）如何确定收敛步数 m，保证样本抽样的无偏性。

- 每隔一段时间取一次样本，计算样本均值，均值稳定后就认为马氏链收敛
- 并行运行多个马氏链，比较各个马氏链的样本均值，如果接近一致就认为马氏链收敛

（3）如何确定迭代步数 n，保证遍历均值计算的精度。

迭代次数一般是几千到几万次不等，

- 基于有效样本量（Effective Sample Size, ESS），是估计样本独立性后的样本大小。因为 MCMC 得到的样本是相关的，若需要独立样本，则需要在样本序列中再次进行随机抽样
- 计算标准误差：根据所需的置信区间或标准误差来决定所需的步数

马尔可夫链蒙特卡罗法在统计学习，特别是贝叶斯学习中，起着重要的作用。假设观测数据由随机变量 $y \in \mathcal{Y}$ 表示，模型由随机变量 $x \in \mathcal{X}$ 表示，贝叶斯学习通过贝叶斯定理计算给定数据条件下模型的后验概率，并选择后验概率最大的模型。后验概率

$$p(x \mid y)=\frac{p(x)p(y \mid x)}{\int_{\mathcal{X}} p\left(y \mid x^{\prime}\right) p\left(x^{\prime}\right) \mathrm{d} x^{\prime}} \tag{23}$$

贝叶斯学习中经常需要进行三种积分运算：

- 归范化计算

$$\int_{\mathcal{X}} p\left(y \mid x^{\prime}\right) p\left(x^{\prime}\right) \mathrm{d} x^{\prime}$$

- 如果有隐变量 $z \in \mathcal{Z}$，求边缘分布

$$p(x \mid y)=\int_{\mathcal{Z}} p(x, z \mid y) \mathrm{d} z$$

- 函数数学期望

$$E_{P(x \mid y)}[f(x)]=\int_{\mathcal{X}} f(x) p(x \mid y) \mathrm{d} x$$

当观测数据和模型都很复杂的时候，以上的积分计算变得困难，MCMC 方法提供了一个通用的有效计算方案。

假设要抽样的概率分布为 $p(x)$ 。Metropolis-Hastings 算法采用转移核为 $p(x,x')$的马尔可夫链:

$$p\left(x,x'\right)=q\left(x,x'\right)\alpha\left(x,x'\right) \tag{24}$$

其中 $q(x,x')$ 和 $\alpha(x,x')$ 分别称为建议分布和接受分布。

- 建议分布 $q(x,x')$ 是另一个马尔可夫链的转移核，是不可约的，即其概率值恒不为 0 ，同时是一个容易抽样的分布
- 接受分布 $\alpha(x,x')=\min\left\{1,\frac{p(x')q(x',x)}{p(x)q(x,x')}\right\}$

那么，转移核 $p(x,x')$ 可以写成

$$p\left(x,x'\right)=\begin{cases}q\left(x,x'\right), & p\left(x'\right)q\left(x',x\right)\geqslant p(x)q\left(x,x'\right)\\ q\left(x',x\right)\frac{p(x')}{p(x)}, & p\left(x'\right)q\left(x',x\right)<p(x)q\left(x,x'\right)\end{cases} \tag{25}$$

然后马尔可夫链按转移概率进行随机游走。如果在时刻 $(t-1)$ 处于状态 x，即 $x_{t-1}=x$，则先按建议分布 $q(x,x')$ 抽样产生一个候选状态 x',然后按照接受分布 $\alpha(x,x')$ 抽样决定是否接受状态 x' 。以概率 $\alpha(x,x')$ 接受 x' ，决定时刻 t 转移到状态 x' ，而以概率 $1-\alpha(x,x')$ 拒绝 x' ，决定时刻 t 仍停留在状态 x 。

一般来说，$p(x)q(x,x') \neq p(x')q(x',x)$，我们希望两边各乘以一个系数（称为接受率 α）后相等：

$$p(x)q(x,x')\alpha(x,x') = p(x')q(x',x)\alpha(x',x) \tag{26}$$

其实只要令 $\alpha(x,x') = p(x')q(x',x)$，$\alpha(x',x) = p(x)q(x,x')$ 即可划上等号。所以令 $p(x,x') = q(x,x')\alpha(x,x')$，则满足细致平稳方程：

$$p(x)p(x,x') = p(x')p(x',x) \tag{27}$$

存在问题：若接受率 $\alpha(x,x')$，$\alpha(x',x)$ 太小，算法效率就很低，因为采集的样本总是被拒绝。

改进接受率！两边等比例扩大。若$\alpha(x,x') \geq \alpha(x',x)$ 则令 $\alpha(x,x')$ 扩大到 1，那么由式（26）

$$\alpha(x',x) = \frac{p(x)q(x,x')}{p(x')q(x',x)} \tag{28}$$

反之，$\alpha(x',x) = 1, \alpha(x,x') = \frac{p(x')q(x',x)}{p(x)q(x',x)}$。于是接受分布就有了该形式：

$$\alpha\left(x,x'\right) = \min\left\{1, \frac{p\left(x'\right)q\left(x',x\right)}{p(x)q\left(x,x'\right)}\right\}$$

定理

由转移核式（*25*）构成的马尔可夫链是可逆的，即

$$p(x)p\left(x, x'\right) = p\left(x'\right) p\left(x', x\right) \tag{29}$$

并且 $p(x)$ 是该马尔可夫链的平稳分布。

建议分布

建议分布 $q(x, x')$ 有多种可能的形式，介绍两种常用形式：

- 对称形式，即对任意的 x 和 x' 有 $q(x, x') = q(x', x)$。接受分布简化为 $\alpha(x, x') = \min\left\{1, \frac{p(x')}{p(x)}\right\}$
- 独立抽样，假设 $q(x, x')$ 与当前状态 x 无关，即 $q(x, x') = q(x')$。接受分布简化为 $\alpha(x, x') = \min\left\{1, \frac{w(x')}{w(x)}\right\}$

满条件分布

MCMC 的目标分布通常是多元联合概率分布 $p(x)=p\left(x_1, x_2, \cdots, x_k\right)$, 其中 $x=\left(x_1, x_2, \cdots, x_k\right)^{\mathrm{T}}$ 为 k 维随机变量。如果条件概率分布 $p\left(x_I \mid x_{-I}\right)$ 中所有 k 个变量全部出现，其中 $x_I=\left\{x_i, i \in I\right\}, x_{-I}=\left\{x_i, i \notin I\right\}, I \subset K=\{1,2, \cdots, k\}$ ，那么称这种条件概率分布为满条件分布。

满条件分布有以下性质：对任意的 $x, x^{\prime} \in \mathcal{X}$ 和任意的 $I \subset K$ ，有

$$p\left(x_I \mid x_{-I}\right)=\frac{p(x)}{\int p(x) \mathrm{d} x_I} \propto p(x) \tag{30}$$

而且, 对任意的 $x, x^{\prime} \in \mathcal{X}$ 和任意的 $I \subset K$ ，有

$$\frac{p\left(x_I^{\prime} \mid x_{-I}^{\prime}\right)}{p\left(x_I \mid x_{-I}\right)}=\frac{p\left(x^{\prime}\right)}{p(x)} \tag{31}$$

Metropolis-Hastings 算法中, 可以利用性质（31）, 简化计算, 提高计算效率。具体地，通过满条件分布概率的比 $\frac{p\left(x_I^{\prime} \mid x_{-I}^{\prime}\right)}{p\left(x_I \mid x_{-I}\right)}$ 计算联合概率的比 $\frac{p\left(x^{\prime}\right)}{p(x)}$ ，而前者更容易计算。

Metropolis-Hastings 算法

输入：目标分布的密度函数 $p(x)$，函数 $f(x)$；

输出：$p(x)$ 的随机样本 $x_{m+1}, x_{m+2}, \cdots, x_n$，函数样本均值 f_{mn}；

参数：收敛步数 m，迭代步数 n。

- (1) 任意选择一个初始值 x_0
- (2) 对 $i = 1, 2, \cdots, n$ 循环执行
  - 设状态 $x_{i-1} = x$，按照建议分布 $q(x, x')$ 随机抽取一个候选状态 x'
  - 计算接受概率

$$\alpha(x, x') = \min\left\{1, \frac{p(x') q(x', x)}{p(x) q(x, x')}\right\}$$

  - 从区间 $(0,1)$ 中按均匀分布随机抽取一个数 u。若 $u \leqslant \alpha(x, x')$，则状态 $x_i = x'$；否则，状态 $x_i = x$
- (3) 得到样本集合 $\{x_{m+1}, x_{m+2}, \cdots, x_n\}$ 计算

$$f_{mn} = \frac{1}{n-m} \sum_{i=m+1}^{n} f(x_i)$$

在 Metropolis-Hastings 算法中，通常需要对多元变量分布 $p(x)$ 进行抽样，有时抽样是困难的。可以对**多元变量的每一变量的条件分布依次分别进行抽样，从而实现对整个多元变量的一次抽样**，这就是单分量 Metropolis-Hastings 算法。

假设马尔可夫链的状态由 k 维随机变量表示

$$x = (x_1, x_2, \cdots, x_k)^{\mathrm{T}}$$

其中 x_j 表示随机变量 x 的第 j 个分量，$j = 1, 2, \cdots, k$，而 $x^{(i)}$ 表示马尔可夫链在时刻 i 的状态

$$x^{(i)} = \left(x_1^{(i)}, x_2^{(i)}, \cdots, x_k^{(i)}\right)^{\mathrm{T}}, \quad i = 1, 2, \cdots, n$$

其中 $x_j^{(i)}$ 是随机变量 $x^{(i)}$ 的第 j 个分量，$j = 1, 2, \cdots, k$.

为了生成容量为 n 的样本集合 $\left\{x^{(1)}, x^{(2)}, \cdots, x^{(n)}\right\}$，单分量 Metropolis-Hastings 算法由下面的 k 步迭代实现 Metropolis-Hastings 算法的一次迭代。

设在第 $(i-1)$ 次迭代结束时分量 x_j 的取值为 $x_j^{(i-1)}$ ，在第 i 次迭代的第 j 步，对分量 x_j 根据 Metropolis-Hastings 算法更新，得到其新的取值 $x_j^{(i)}$ 。

- 首先，由建议分布 $q\left(x_j^{(i-1)}, x_j \mid x_{-j}^{(i)}\right)$ 抽样产生分量 x_j 的候选值 $x_j^{\prime(i)}$ ，这里 $x_{-j}^{(i)}$ 表示在第 i 次迭代的第 $(j-1)$ 步后的 $x^{(i)}$ 除去 $x_j^{(i-1)}$ 的所有值，即

$$x_{-j}^{(i)} = \left(x_1^{(i)}, \cdots, x_{j-1}^{(i)}, x_{j+1}^{(i-1)}, \cdots, x_k^{(i-1)}\right)^{\mathrm{T}}$$

其中分量 $1, 2, \cdots, j-1$ 已经更新。

- 然后，按照接受概率

$$\alpha\left(x_j^{(i-1)}, x_j^{\prime(i)} \mid x_{-j}^{(i)}\right) = \min\left\{1, \frac{p\left(x_j^{\prime(i)} \mid x_{-j}^{(i)}\right) q\left(x_j^{\prime(i)}, x_j^{(i-1)} \mid x_{-j}^{(i)}\right)}{p\left(x_j^{(i-1)} \mid x_{-j}^{(i)}\right) q\left(x_j^{(i-1)}, x_j^{(i)} \mid x_{-j}^{(i)}\right)}\right\} \tag{32}$$

决定是否接受候选值 $x^{\prime(i)}$ 。如果 $x_j^{\prime(i)}$ 被接受，则令 $x_j^{(i)} = x_j^{\prime(i)}$；否则令 $x_j^{(i)} = x_j^{(i-1)}$ ，其余分量在第 j 步不改变。于是，马尔可夫链的转移概率为

$$p\left(x_j^{(i-1)}, x_j^{\prime(i)} \mid x_{-j}^{(i)}\right) = \alpha\left(x_j^{(i-1)}, x_j^{\prime(i)} \mid x_{-j}^{(i)}\right) q\left(x_j^{(i-1)}, x_j^{\prime(i)} \mid x_{-j}^{(i)}\right) \tag{33}$$

如图所示，单分量 Metropolis-Hastings 算法的迭代过程。目标是对含有两个变量的随机变量 x 进行抽样。如果变量 x_1 或 x_2 更新，那么在水平或垂直方向产生一个移动，连续水平和垂直移动产生一个新的样本点。注意由于建议分布可能不被接受，Metropolis-Hastings 算法可能在一些相邻的时刻不产生移动。

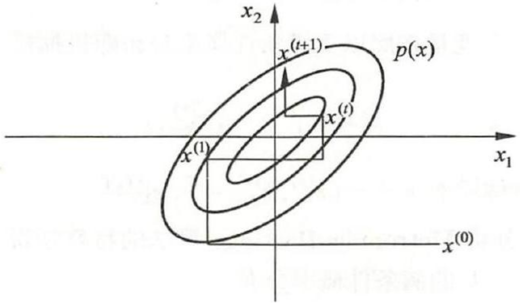


图 2:

基本原理

吉布斯抽样, 用于多元变量联合分布的抽样和估计，是 MetropolisHastings 算法的特殊情况，但是更容易实现，因而被广泛使用。

基本做法：从**联合概率分布定义满条件概率分布，依次对满条件概率分布进行抽样，**得到样本的序列。

假设多元变量的联合概率分布为 $p(x)=p\left(x_1, x_2, \cdots, x_k\right)$ 。吉布斯抽样从一个初始样本 $x^{(0)}=\left(x_1^{(0)}, x_2^{(0)}, \cdots, x_k^{(0)}\right)^{\mathrm{T}}$ 出发，不断进行迭代，每一次迭代得到联合分布的一个样本 $x^{(i)}=\left(x_1^{(i)}, x_2^{(i)}, \cdots, x_k^{(i)}\right)^{\mathrm{T}}$ 。最终得到样本序列 $\left\{x^{(0)}, x^{(1)}, \cdots, x^{(n)}\right\}$ 。

在每次迭代中，依次对 k 个随机变量中的一个变量进行随机抽样。如果在第 i 次迭代中，对第 j 个变量进行随机抽样，那么抽样的分布是满条件概率分布 $p\left(x_j \mid x_{-j}^{(i)}\right)$ ，这里 $x_{-j}^{(i)}$ 表示第 i 次迭代中，变量 j 以外的其他变量。

具体怎么做的呢？

设在第 $(i-1)$ 步得到样本 $\left(x_1^{(i-1)}, x_2^{(i-1)}, \cdots, x_k^{(i-1)}\right)^{\mathrm{T}}$，在第 i 步，首先对第一个变量按照以下满条件概率分布随机抽样

$$p\left(x_1 \mid x_2^{(t-1)}, \cdots, x_k^{(t-1)}\right)$$

得到 $x_1^{(i)}$, 之后依次对第 j 个变量按照以下满条件概率分布随机抽样

$$p\left(x_j \mid x_1^{(i)}, \cdots, x_{j-1}^{(i)}, x_{j+1}^{(i-1)}, \cdots, x_k^{(i-1)}\right), \quad j = 2, \cdots, k-1$$

得到 $x_j^{(i)}$，最后对第 k 个变量按照以下满条件概率分布随机抽样

$$p\left(x_k \mid x_1^{(i)}, \cdots, x_{k-1}^{(i)}\right)$$

得到 $x_k^{(i)}$, 于是得到整体样本 $x^{(i)} = \left(x_1^{(i)}, x_2^{(i)}, \cdots, x_k^{(i)}\right)^{\mathrm{T}}$

吉布斯抽样是单分量 Metropolis-Hastings 算法的特殊情况：建议分布是当前变量 $x_j, j=1,2,\cdots,k$ 的满条件概率分布

$$q\left(x, x^{\prime}\right)=p\left(x_j^{\prime} \mid x_{-j}\right) \tag{19.49}$$

这时，接受概率 $\alpha=1$ ，

$$\begin{aligned}
\alpha\left(x, x^{\prime}\right) & =\min \left\{1, \frac{p\left(x^{\prime}\right) q\left(x^{\prime}, x\right)}{p(x) q\left(x, x^{\prime}\right)}\right\} & (34)\\
& =\min \left\{1, \frac{p\left(x^{\prime}{ }_{-j}\right) p\left(x^{\prime}{ }_j \mid x^{\prime}{ }_{-j}\right) p\left(x_j \mid x^{\prime}{ }_{-j}\right)}{p\left(x_{-j}\right) p\left(x_j \mid x_{-j}\right) p\left(x^{\prime}{ }_j \mid x_{-j}\right)}\right\}=1 & (35)
\end{aligned}$$

转移核就是满条件概率分布（接受率为 1）

$$p\left(x, x^{\prime}\right)=p\left(x^{\prime}{ }_j \mid x_{-j}\right) \tag{36}$$

也就是说依次按照单变量的满条件概率分布 $p\left(x_j^{\prime} \mid x_{-j}\right)$ 进行随机抽样, 就能实现单分量 Metropolis-Hastings 算法。

Gibbs 抽样算法

输入：目标概率分布的密度函数 $p(x)$，函数 $f(x)$；

输出：$p(x)$ 的随机样本 $x_{m+1}, x_{m+2}, \cdots, x_n$，函数样本均值 f_{mn}；

参数：收敛步数 m，迭代步数 n。

- （1）初始化。给出初始样本 $x^{(0)}=\left(x_1^{(0)}, x_2^{(0)}, \cdots, x_k^{(0)}\right)^{\mathrm{T}}$
- （2）对 i 循环执行

设第 $(i-1)$ 次迭代结束时的样本为 $x^{(i-1)}=\left(x_1^{(i-1)}, x_2^{(i-1)}, \cdots, x_k^{(i-1)}\right)^{\mathrm{T}}$，则第 i 次迭代进行如下几步操作：

（1）由满条件分布 $p\left(x_1 \mid x_2^{(i-1)}, \cdots, x_k^{(i-1)}\right)$ 抽取 $x_1^{(i)}$

（j）由满条件分布 $p\left(x_j \mid x_1^{(i)}, \cdots, x_{j-1}^{(i)}, x_{j+1}^{(i-1)}, \cdots, x_k^{(i-1)}\right)$ 抽取 $x_j^{(i)}$ ⋮

（k）由满条件分布 $p\left(x_k \mid x_1^{(i)}, \cdots, x_{k-1}^{(i)}\right)$ 抽取 $x_k^{(i)}$ 得到第 i 次迭代值 $x^{(i)}=\left(x_1^{(i)}, x_2^{(i)}, \cdots, x_k^{(i)}\right)^{\mathrm{T}}$。

- (3) 得到样本集合 $\left\{x^{(m+1)}, x^{(m+2)}, \cdots, x^{(n)}\right\}$
- (4) 计算 $f_{mn}=\frac{1}{n-m} \sum_{i=m+1}^{n} f\left(x^{(i)}\right)$

例 11 用吉布斯抽样从以下二元正态分布中抽取随机样本。

$$x=(x_1,x_2)^{\mathrm{T}}\sim N(0,\Sigma)$$

$$\Sigma=\left[\begin{array}{cc}1 & \rho\\ \rho & 1\end{array}\right]$$

解 条件概率分布为一元正态分布

$$p\left(x_1\mid x_2\right)=N\left(\rho x_2,\left(1-\rho^2\right)\right)$$

$$p\left(x_2\mid x_1\right)=N\left(\rho x_1,\left(1-\rho^2\right)\right)$$

假设初始样本为 $x^{(0)}=\left(x_1^{(0)},x_2^{(0)}\right)$, 通过吉布斯抽样, 可以得到以下样本序列:

迭代次数	对 x_1 抽样	对 x_2 抽样	产生样本
1	$x_1 \sim N\left(\rho x_2^{(0)}, (1-\rho^2)\right)$, 得到 $x_1^{(1)}$	$x_2 \sim N\left(\rho x_1^{(1)}, (1-\rho^2)\right)$, 得到 $x_2^{(1)}$	$x^{(1)} = \left(x_1^{(1)}, x_2^{(1)}\right)$
	$\vdots$	$\vdots$	$\vdots$
t	$x_1 \sim N\left(\rho x_2^{(t-1)}, (1-\rho^2)\right)$, 得到 $x_1^{(t)}$	$x_2 \sim N\left(\rho x_1^{(t)}, (1-\rho^2)\right)$, 得到 $x_2^{(t)}$	$x^{(t)} = \left(x_1^{(t)}, x_2^{(t)}\right)$
	$\vdots$	$\vdots$	$\vdots$

得到的样本集合 $\left\{x^{(m+1)}, x^{(m+2)}, \cdots, x^{(n)}\right\}, m < n$ 就是二元正态分布的随机抽样。

抽样计算

吉布斯抽样中需要对满条件概率分布进行重复多次抽样，可以利用概率分布的性质提高抽样的效率。设 y 表示观测数据， α, θ, z 分别表示超参数、模型参数、未观测数据， $x = (\alpha, \theta, z)$ ，如图所示。贝叶斯学习的目的是估计后验概率分布 $p(x \mid y)$, 求后验概率最大的模型。

$$p(x \mid y) = p(\alpha, \theta, z \mid y) \propto p(z, y \mid \theta) p(\theta \mid \alpha) p(\alpha) \tag{37}$$

式中 $p(\alpha)$ 是超参数分布, $p(\theta \mid \alpha)$ 是先验分布, $p(z, y \mid \theta)$ 是完全数据的分布。

现在用吉布斯抽样估计 $p(x \mid y)$ ，其中 y 已知， $x=(\alpha, \theta, z)$ 未知。吉布斯抽样中各个变量 α, θ, z 的满条件分布有以下关系：

$$
\begin{aligned}
& p\left(\alpha_i \mid \alpha_{-i}, \theta, z, y\right) \propto p(\theta \mid \alpha) p(\alpha) && (38) \\
& p\left(\theta_j \mid \theta_{-j}, \alpha, z, y\right) \propto p(z, y \mid \theta) p(\theta \mid \alpha) && (39) \\
& p\left(z_k \mid z_{-k}, \alpha, \theta, y\right) \propto p(z, y \mid \theta) && (40)
\end{aligned}
$$

满条件概率分布与若干条件概率分布的乘积成正比，各个条件概率分布只由少量的相关变量组成（图模型中相邻结点表示的变量）。所以，依满条件概率分布的抽样可以通过依这些条件概率分布的乘积的抽样进行。这样可以大幅减少抽样的计算复杂度，因为计算只涉及部分变量。

参考文献

[1] 李航. 统计学习方法（第2 版）. 清华大学出版社,2019.

[2] 张波. 应用随机过程（第 6 版）.中国人民大学出版社,2023.

[3] Hastie, Trevor, et al. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd ed., Springer, 2009.

[4] Sheldon M. Ross. Introduction to Probability Models . 12th ed., Academic press, 2019. `https://doi.org/https://doi.org/10.1016/C2017-0-01324-1.`

[5] Sheldon M. Ross. Stochastic Processes. 3rd ed., Wiley, 2014.

Thanks!