清華大學
Tsinghua University

第十九章
马尔可夫链蒙特卡罗法

马尔可夫链蒙特卡罗法

- 蒙特卡罗法（Monte Carlo method），也称为统计模拟方法（statistical simulation method），是通过从概率模型的随机抽样进行近似数值计算的方法。

- 马尔可夫链蒙特卡罗法（Markov Chain Monte Carlo, MCMC），则是以马尔可夫链（Markov chain)为概率模型的蒙特卡罗法。

- 马尔可夫链蒙特卡罗法构建一个马尔可夫链，使其平稳分布就是要进行抽样的分布，首先基于该马尔可夫链进行随机游走，产生样本的序列，之后使用该平稳分布的样本进行近似数值计算

马尔可夫链蒙特卡罗法

- Metropolis-Hastings算法是最基本的马尔可夫链蒙特卡罗法
- 吉布斯抽样（Gibbs sampling）是更简单、使用更广泛的马尔可夫链蒙特卡罗法，
- 马尔可夫链蒙特卡罗法被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，是重要的统计学习计算方法。

蒙特卡罗法

随机抽样

- 蒙特卡罗法要解决的问题是，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。

- 比如，从样本得到经验分布，从而估计总体分布
- 或者从样本计算出样本均值，从而估计总体期望

- 所以蒙特卡罗法的核心是随机抽样(random sampling)。

随机抽样

- 蒙特卡罗法
  - 直接抽样法
  - 接受-拒绝抽样法
  - 重要性抽样法

- 接受-拒绝抽样法、重要性抽样法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。

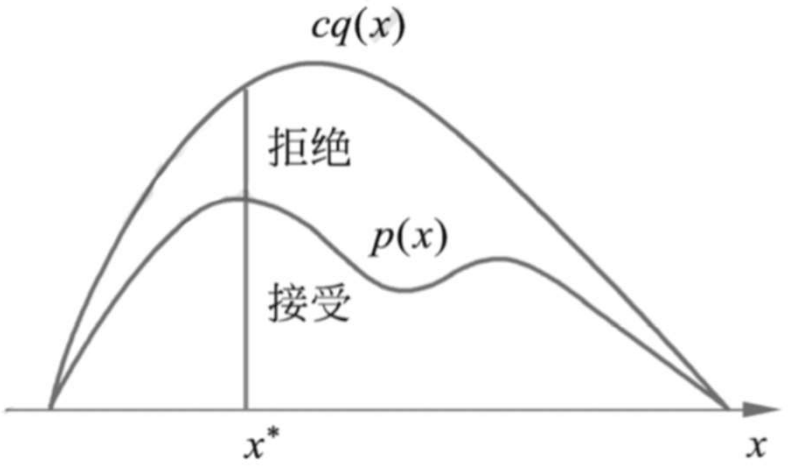
随机抽样

- 接受-拒绝抽样法(accept-reject sampling method)
- 假设有随机变量 x，取值 $x \in X$，其概率密度函数为 p(x)
- 目标是得到该概率分布的随机样本，以对这个概率分布进行分析

随机抽样

- 假设 p(x) 不可以直接抽样。找一个可以直接抽样的分布，称为建议分布(proposal distribution)

- 假设 q(x) 是建议分布的概率密度函 数，并且有 q(x) 的 c 倍一定大于等于 p(x)，其中 c > 0，如图中所示



接受-拒绝法

输入：抽样的目标概率分布的概率密度函数 $p(x)$；

输出：概率分布的随机样本 $x_1, x_2, \cdots, x_n$。

参数：样本数 n

（1）选择概率密度函数为 $q(x)$ 的概率分布，作为建议分布，使其对任一 x 满足 $cq(x) \geqslant p(x)$，其中 $c > 0$。

（2）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。

（3）如果 $u \leqslant \dfrac{p(x^*)}{cq(x^*)}$，则将 x^* 作为抽样结果；否则，回到步骤 (2)。

（4）直至得到 n 个随机样本，结束。

接受-拒绝法

- 接受-拒绝法的优点是容易实现，缺点是效率可能不高
- 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

数学期望佔计

- 一般的蒙特卡罗法也可以用于数学期望估计（estimation of mathematical expectation)。

- 假设有随机变量x，取值$x \in \mathcal{X}$，其概率密度函数为p(x), f(x)为定义在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 $E_{p(x)}[f(x)]$

数学期望佔计

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

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

- 作为数学期望 $E_{p(x)}[f(x)]$ 的近似值
- 根据大数定律可知，当样本容量增大时，样本均值以概率1收敛于数学期望：

$$\hat{f}_n \to E_{p(x)}[f(x)], \quad n \to \infty$$

- 这样就得到了数学期望的近似计算方法：

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

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（Monte Carlo integration）

- 假设有一个函数h(x)，目标是计算该函数的积分 $\int_{\mathcal{X}} h(x)\,\mathrm{d}x$
- 如果能够将函数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

$$\int_{\mathcal{X}} h(x)\,\mathrm{d}x = \int_{\mathcal{X}} f(x)\,p(x)\,\mathrm{d}x = E_{p(x)}\left[f(x)\right]$$

- 于是函数h(x)的积分可以表示为函数f(x)关于概率密度函数p(x)的数学期望。

积分计算

- 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

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

- 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的形式，而函数的数学期望 又可以通过函数的样本均值估计

- 于是，就可以利用样本均值来近似计算积分

例

- 用蒙特卡罗积分法求

$$\int_0^1 \mathrm{e}^{-x^2/2}\mathrm{d}x$$

- 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

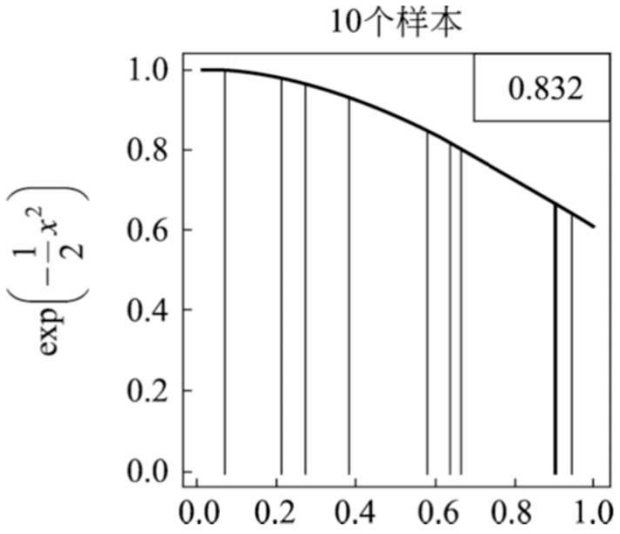
例

- 使用蒙特卡罗积分法，如图所示，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 $\hat{f}_{10}$

$$\hat{f}_{10} = \frac{1}{10}\sum_{i=1}^{10} \mathrm{e}^{-x_i^2/2} = 0.832$$

- 也就是积分的近似

- 随机样本数越大，计算就越精确



例

- 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 $f(x) = x$

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

- p(x) 是标准正态分布的密度函数

- 使用蒙特卡罗积分法，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。当样本增大时，积分值趋于0

马尔可夫链

基本定义

定义 19.1（马尔可夫链） 考虑一个随机变量的序列 $X = \{X_0, X_1, \cdots, X_t, \cdots\}$，这里 X_t 表示时刻 t 的随机变量，$t = 0, 1, 2, \cdots$。每个随机变量 X_t（$t = 0, 1, 2, \cdots$）的取值集合相同，称为状态空间，表示为 $\mathcal{S}$。随机变量可以是离散的，也可以是连续的。以上随机变量的序列构成随机过程（stochastic process）。

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 $t \geqslant 1$ 的随机变量 X_t 与前一个时刻的随机变量 X_{t-1} 之间有条件分布 $P(X_t|X_{t-1})$，如果 X_t 只依赖于 X_{t-1}，而不依赖于过去的随机变量 $\{X_0, X_1, \cdots, X_{t-2}\}$，这一性质称为马尔可夫性，即

$$P(X_t|X_0, X_1, \cdots, X_{t-1}) = P(X_t|X_{t-1}), \quad t = 1, 2, \cdots \tag{19.6}$$

具有马尔可夫性的随机序列 $X = \{X_0, X_1, \cdots, X_t, \cdots\}$ 称为马尔可夫链（Markov chain），或马尔可夫过程（Markov process）。条件概率分布 $P(X_t|X_{t-1})$ 称为马尔可夫链的转移概率分布。转移概率分布决定了马尔可夫链的特性。

基本定义

- 马尔可夫性的直观解释是“未来只依赖于现在（假设现在已知），而与过去无关”

- 若转移概率分布 $P(X_t|X_{t-1})$与t无关，即

$$P(X_{t+s}|X_{t-1+s}) = P(X_t|X_{t-1}), \quad t = 1, 2, \cdots; \quad s = 1, 2, \cdots$$

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

$$P(X_t|X_0X_1 \cdots X_{t-2}X_{t-1}) = P(X_t|X_{t-n} \cdots X_{t-2}X_{t-1})$$

离散状态马尔可夫链

- 转移概率矩阵和状态分布

- 离散状态马尔可夫链 $X = \{X_0, X_1, \cdots, X_t, \cdots\}$，随机变量 X_t $(t = 0, 1, 2, \cdots)$ 定义在离散空间S，转移概率分布可以由矩阵表示

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

$$p_{ij} = (X_t = i | X_{t-1} = j), \quad i = 1, 2, \cdots; \quad j = 1, 2, \cdots$$

- 满足

$$p_{ij} \geqslant 0, \quad \sum_i p_{ij} = 1$$

转移概率矩阵和状态分布

- 马尔可夫链的转移概率 p_{ij} 可以由矩阵表示，即

$$P = \begin{bmatrix} p_{11} & p_{12} & p_{13} & \cdots \\ p_{21} & p_{22} & p_{23} & \cdots \\ p_{31} & p_{32} & p_{33} & \cdots \\ \cdots & \cdots & \cdots & \cdots \end{bmatrix}$$

- 称为马尔可夫链的转移概率矩阵，转移概率矩阵P满足条件

$$p_{ij} \geqslant 0, \sum_i p_{ij} = 1$$

- 这两个条件的矩阵称为随机矩阵（stochastic matrix)
- 矩阵列元素之和为1

转移概率矩阵和状态分布

- 考虑马尔可夫链 $X=\{X_0,X_1,\cdots,X_t,\cdots\}$，在时刻 $t(t=0,1,2,\cdots)$ 的概率分布，称为时刻t的状态分布，记作

$$\pi(t)=\begin{bmatrix}\pi_1(t)\\ \pi_2(t)\\ \vdots\end{bmatrix}$$

- 其中 $\pi_i(t)$ 其中时刻t状态为i的概率 $P(X_t=i)$

$$\pi_i(t)=P(X_t=i),\quad i=1,2,\cdots$$

转移概率矩阵和状态分布

- 特别地，马尔可夫链的初始状态分布可以表示为

$$\pi(0) = \begin{bmatrix} \pi_1(0) \\ \pi_2(0) \\ \vdots \end{bmatrix}$$

- 其中 $\pi_i(0)$ 表示时刻0状态为i的概率 $P(X_0 = i)$ ，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。

转移概率矩阵和状态分布

- 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
- 从一个初始状态出发，根据有向边上定义的概率在状态之间随机跳转（或随机转移），就可以产生状态的序列
- 马尔可夫链实际上是刻 画随时间在状态之间转移的模型，假设未来的转移状态只依赖于现在的状态，而与过去的状态无关。

例

- 自然语言处理、语音处理中经常用到语言模型(language model)，是建立在词表上的n阶马尔可夫链
- 比如，在英语语音识别 中，语音模型产生出两个候选：“How to recognize speech”与“How to wreck a nice beach” 要判断哪个可能性更大
- 显然从语义的角度前者的可能性更大，语言模型可以帮助做出这个判断

例

- 假设每个单词只依赖于其前面出现的单词，也就是说单词序列具有马尔可夫性， 那么可以定义一阶马尔可夫链，即语言模型，如下计算语句的概率

$$
\begin{aligned}
& P(w_1 w_2 \cdots w_s) \\
&= P(w_1)P(w_2|w_1)P(w_3|w_1 w_2)\cdots P(w_i|w_1 w_2 \cdots w_{i-1})\cdots P(w_s|w_1 w_2 \cdots w_{s-1}) \\
&= P(w_1)P(w_2|w_1)P(w_3|w_2)\cdots P(w_i|w_{i-1})\cdots P(w_s|w_{s-1})
\end{aligned}
$$

- 这里第三个等式基于马尔可夫链假设。这个马尔可夫链中，状态空间为词表，一个位置上单词的产生只依赖于前一个位置的单词，而不依赖于更前面的单词。
- 以上是一阶 马尔可夫链，一般可以扩展到 n 阶马尔可夫链。

例

- 语言模型的学习等价于确定马尔可夫链中的转移概率值，如果有充分的语料，转移概率可以直接从语料中估计

- 直观上，“wreck a nice”出现之后，下面出现 “beach” 的概率极低，所以第二个语句的概率应该更小，从语言模型的角度看第一个语句的可能性更大

转移概率矩阵和状态分布

- 马尔可夫链 X 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

$$\pi(t) = P\pi(t-1)$$

- 这是因为

$$\begin{aligned}\pi_i(t) &= P(X_t = i) \\ &= \sum_m P(X_t = i | X_{t-1} = m) P(X_{t-1} = m) \\ &= \sum_m p_{im} \pi_m(t-1)\end{aligned}$$

转移概率矩阵和状态分布

- 马尔可夫链在时刻t的状态分布，可以通过递推得到。由

$$\pi(t) = P\pi(t-1) = P(P\pi(t-2)) = P^2\pi(t-2)$$

- 递推得到 $$\pi(t) = P^t\pi(0)$$
- 这里的P^t称为t步转移概率矩阵，

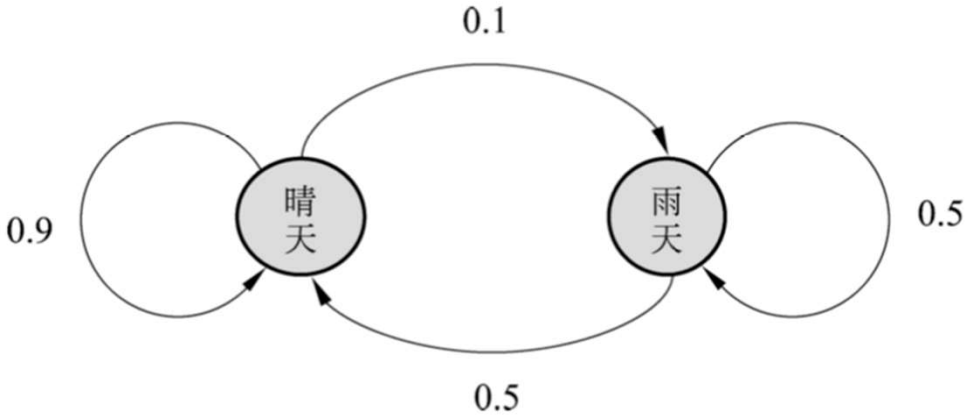
$$P_{ij}^t = P(X_t = i | X_0 = j)$$

- 表示时刻0从状态j出发，时刻t达到状态i的t步转移概率
- P^t也是随机矩阵。马尔可夫链的状态分布由初始分布和转移概率分布决定。

例

- 假设观察某地的天气，按日依次是“晴，雨，晴，晴，晴，雨，晴……”，具有一定的规律。

- 假设天气的变化具有马尔可夫性，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。



例

- 转移矩阵为

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

- 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：

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

例

- 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）

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

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

平稳分布

定义 19.2（平稳分布） 设有马尔可夫链 $X = \{X_0, X_1, \cdots, X_t, \cdots\}$，其状态空间为 $\mathcal{S}$，转移概率矩阵为 $P = (p_{ij})$，如果存在状态空间 $\mathcal{S}$ 上的一个分布

$$\pi = \begin{bmatrix} \pi_1 \\ \pi_2 \\ \vdots \end{bmatrix}$$

使得

$$\pi = P\pi \tag{19.15}$$

则称 π 为马尔可夫链 $X = \{X_0, X_1, \cdots, X_t, \cdots\}$ 的平稳分布。

- 直观上，如果马尔可夫链的平稳分布存在，那么以该平稳分布作为初始分布，面向未来进行随机状态转移，之后任何一个时刻的状态分布都是该平稳分布

平稳分布

引理 19.1 给定一个马尔可夫链 $X = \{X_0, X_1, \cdots, X_t, \cdots\}$，状态空间为 $\mathcal{S}$，转移概率矩阵为 $P = (p_{ij})$，则分布 $\pi = (\pi_1, \pi_2, \cdots)^{\mathrm{T}}$ 为 X 的平稳分布的充分必要条件是 $\pi = (\pi_1, \pi_2, \cdots)^{\mathrm{T}}$ 是下列方程组的解:

$$x_i = \sum_j p_{ij} x_j, \quad i = 1, 2, \cdots \tag{19.16}$$

$$x_i \geqslant 0, \quad i = 1, 2, \cdots \tag{19.17}$$

$$\sum_i x_i = 1 \tag{19.18}$$

平稳分布

- 证明 - 必要性
- 假设 $\pi = (\pi_1, \pi_2, \cdots)^{\mathrm{T}}$ 是平稳分布，显然满足式(19.17)和式(19.18)。又

$$\pi_i = \sum_j p_{ij}\pi_j, \quad i = 1, 2, \cdots$$

- 即

$$\pi = (\pi_1, \pi_2, \cdots)^{\mathrm{T}}$$

- 满足式(19.16)

平稳分布

- 证明 – 充分性

- 由式（19.17）和式（19.18）知 $\pi = (\pi_1, \pi_2, \cdots)^{\mathrm{T}}$ 是一概率分布
- 假设 $\pi = (\pi_1, \pi_2, \cdots)^{\mathrm{T}}$ 为X_t的分布，则

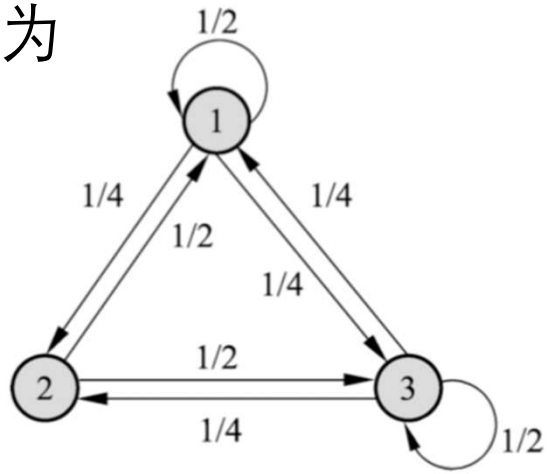
$$P(X_t = i) = \pi_i = \sum_j p_{ij}\pi_j = \sum_j p_{ij} P(X_{t-1} = j), \quad i = 1, 2, \cdots$$

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

例

- 设有图上所示马尔可夫链，其转移概率矩阵为

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



- 求其平稳分布

例

- 设平稳分布为 $\pi=(x_1,x_2,x_3)^{\mathrm{T}}$，则由式 (19.16)~式 (19.18) 有

$$x_1=\frac{1}{2}x_1+\frac{1}{2}x_2+\frac{1}{4}x_3$$

$$x_2=\frac{1}{4}x_1+\frac{1}{4}x_3$$

$$x_3=\frac{1}{4}x_1+\frac{1}{2}x_2+\frac{1}{2}x_3$$

$$x_1+x_2+x_3=1$$

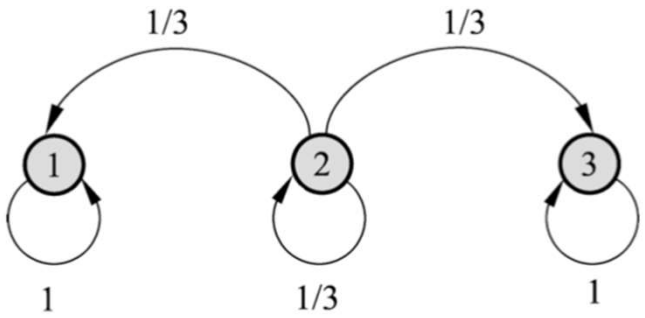
$$x_i\geqslant 0,\quad i=1,2,3$$

- 解方程组，得到唯一的平稳分布 $\pi=(2/5\quad 1/5\quad 2/5)^{\mathrm{T}}$

例

- 设有图上所示马尔可夫链，其转移概率分布如下，求其平稳分布。

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



例

- 这个马尔可夫链的平稳分布并不唯一

$$\pi=(3/4\ \ 0\ \ 1/4)^{\mathrm{T}},\ \pi=(2/3\ \ 0\ \ 1/3)^{\mathrm{T}}$$

- 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

- 连续状态马尔可夫链 $X=\{X_0, X_1, \cdots, X_t, \cdots\}$，随机变量 $X_t(t=0,1,2,\cdots)$ 定义在连续状态空间S
- 转移概率分布由概率转移核或转移核(transition kernel）表示。

- 设S是连续状态空间，对任意的 $x \in \mathcal{S}, A \subset \mathcal{S}$ 定义为

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

- 其中 $p(x, \cdot)$ 是概率密度函数，满足 $p(x, \cdot) \geqslant 0$，$P(x, \mathcal{S})=\int_{\mathcal{S}} p(x, y)\,\mathrm{d}y=1$.

连续状态马尔可夫链

- 转移核 $P(x, A)$ 表示从 x ~ A的转移概率

$$P\left(X_t = A | X_{t-1} = x\right) = P(x, A)$$

- 有时也将概率密度函数 $p(x, \bullet)$ 称为转移核

- 若马尔可夫链的状态空间S上的概率分布 $\pi(x)$ 满足条件

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

- 则称分部 $\pi(x)$ 为该马尔可夫链的平稳分布。等价地，$\pi = P\pi$ 或写为

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

马尔可夫链的性质

定义 19.3（不可约） 设有马尔可夫链 $X = \{X_0, X_1, \cdots, X_t, \cdots\}$，状态空间为 $\mathcal{S}$，对于任意状态 $i, j \in \mathcal{S}$，如果存在一个时刻 $t(t > 0)$ 满足

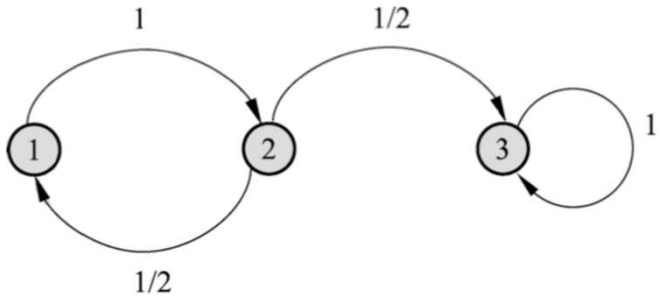
$$P(X_t = i | X_0 = j) > 0 \tag{19.24}$$

也就是说，时刻 0 从状态 j 出发，时刻 t 到达状态 i 的概率大于 0，则称此马尔可夫链 X 是不可约的（irreducible），否则称马尔可夫链是可约的（reducible）。

- 直观上，一个不可约的马尔可夫链，从任意状态出发，当经过充分长时间后，可以到达任意状态

例

- 图上所示马尔可夫链是可约的



- 转移概率矩阵 $\begin{bmatrix} 0 & 1/2 & 0 \\ 1 & 0 & 0 \\ 0 & 1/2 & 1 \end{bmatrix}$ 平稳分布 $\pi = (0 \quad 0 \quad 1)^{\mathrm{T}}$

- 此马尔可夫链，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

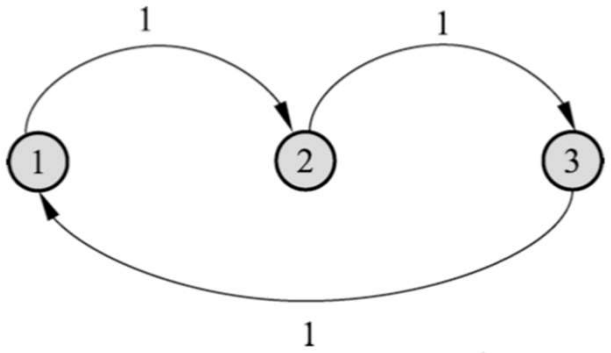
马尔可夫链的性质

定义 19.4（非周期） 设有马尔可夫链 $X = \{X_0, X_1, \cdots, X_t, \cdots\}$，状态空间为 $\mathcal{S}$，对于任意状态 $i \in \mathcal{S}$，如果时刻 0 从状态 i 出发，t 时刻返回状态的所有时间长 $\{t : P(X_t = i \,|\, X_0 = i) > 0\}$ 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）。

- 直观上，一个非周期性的马尔可夫链，不存在一个状态，从这一个状态出发，再返回到这个状态时所经历的时间长呈一定的周期性

例

- 图上所示的马尔可夫链是周期的



例

- 转移概率矩阵 $\begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$

- 其平稳分布是 $\pi = (1/3 \quad 1/3 \quad 1/3)^{\mathrm{T}}$。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.

定理 19.2 不可约且非周期的有限状态马尔可夫链，有唯一平稳分布存在。

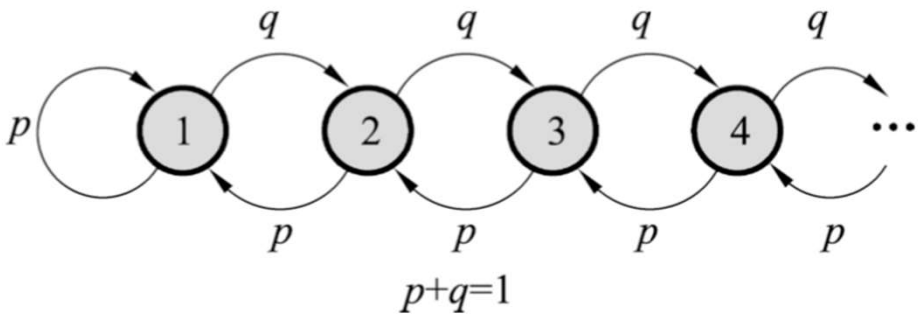
马尔可夫链的性质

定义 19.5（正常返） 设有马尔可夫链 $X=\{X_0,X_1,\cdots,X_t,\cdots\}$，状态空间为 $\mathcal{S}$，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 j 出发，时刻 t 首次转移到状态 i 的概率，即 $p_{ij}^t=P\left(X_t=i,X_s\neq i,s=1,2,\cdots,t-1|X_0=j\right),t=1,2,\cdots$。若对所有状态 i,j 都满足 $\lim\limits_{t\to\infty}p_{ij}^t>0$，则称马尔可夫链 X 是正常返的（positive recurrent）。

- 直观上，一个正常返的马尔可夫链，其中任意一个状态，从其他任意一个状态出发，当时间趋于无穷时，首次转移到这个状态的概率不为0。

例

- 图上所示无限状态马尔可夫链，当p>q时是正常返的，当p≤q不是正常返的。



例

- 转移概率矩阵

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

- 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≤q 时，不存在平稳分布，马尔可夫链不是正常返的。

定理 19.3 不可约、非周期且正常返的马尔可夫链，有唯一平稳分布存在。

马尔可夫链的性质

定理 19.4（遍历定理） 设有马尔可夫链 $X = \{X_0, X_1, \cdots, X_t, \cdots\}$，状态空间为 $\mathcal{S}$，若马尔可夫链 X 是不可约、非周期且正常返的，则该马尔可夫链有唯一平稳分布 $\pi = (\pi_1, \pi_2, \cdots)^{\mathrm{T}}$，并且转移概率的极限分布是马尔可夫链的平稳分布

$$\lim_{t\to\infty} P(X_t = i | X_0 = j) = \pi_i, \quad i = 1, 2, \cdots; \quad j = 1, 2, \cdots \tag{19.25}$$

若 $f(X)$ 是定义在状态空间上的函数，$E_\pi[|f(X)|] < \infty$，则

$$P\{\hat{f}_t \to E_\pi[f(X)]\} = 1 \tag{19.26}$$

这里

$$\hat{f}_t = \frac{1}{t}\sum_{s=1}^{t} f(x_s)$$

$E_\pi[f(X)] = \sum_i f(i)\pi_i$ 是 $f(X)$ 关于平稳分布 $\pi = (\pi_1, \pi_2, \cdots)^{\mathrm{T}}$ 的数学期望，式 (19.26) 表示

$$\hat{f}_t \to E_\pi[f(X)], \quad t \to \infty \tag{19.27}$$

几乎处处成立或以概率 1 成立。

马尔可夫链的性质

- 遍历定理的直观解释:
- 满足相应条件的马尔可夫链，当时间趋于无穷时，马尔可夫链的状态分布趋近于平稳分布，随机变量的函数的样本均值以概率1收敛于该函数的数学期望。
- 样本均值可以认为是时间均值，而数学期望是空间均值。遍历定理实际表述了遍历性的含义：当时间趋于无穷时，时间均值等于空间均值。
- 遍历定理的三个 条件：不可约、非周期、正常返，保证了当时间趋于无穷时达到任意一个状态的概率不为0

马尔可夫链的性质

- 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，即

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

马尔可夫链的性质

定义 19.6（可逆马尔可夫链） 设有马尔可夫链 $X = \{X_0, X_1, \cdots, X_t, \cdots\}$，状态空间为 $\mathcal{S}$，转移概率矩阵为 P，如果有状态分布 $\pi = (\pi_1, \pi_2, \cdots)^{\mathrm{T}}$，对于任意状态 $i, j \in \mathcal{S}$，对任意一个时刻 t 满足

$$P(X_t = i | X_{t-1} = j)\pi_j = P(X_{t-1} = j | X_t = i)\pi_i, \quad i, j = 1, 2, \cdots \tag{19.29}$$

或简写为

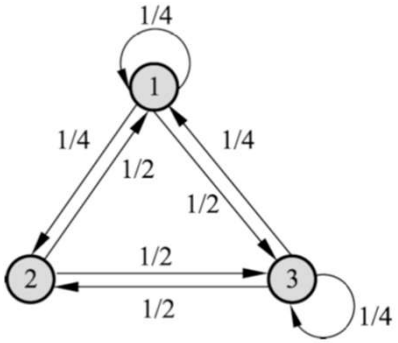
$$p_{ji}\pi_j = p_{ij}\pi_i, \quad i, j = 1, 2, \cdots \tag{19.30}$$

则称此马尔可夫链 X 为可逆马尔可夫链（reversible Markov chain），式 (19.30) 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。

例

- 图上所示马尔可夫链是不可逆的



转移概率矩阵

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

- 平稳分布 $\pi = (8/25 \quad 7/25 \quad 2/5)^{\mathrm{T}}$
- 不满足细致平稳方程

马尔可夫链的性质

定理 19.5（细致平衡方程） 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

$$P\pi = \pi$$

证明 事实上

$$(P\pi)_i = \sum_j p_{ij}\pi_j = \sum_j p_{ji}\pi_i = \pi_i \sum_j p_{ji} = \pi_i, \quad i = 1, 2, \cdots \tag{19.31}$$

- 可逆马尔可夫链一定有唯一平稳分布，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)。
- 也就是说，可逆马尔可夫链满足遍历定理 19.4 的条件。

马尔可夫链蒙特卡罗法

基本想法

- 马尔可夫链蒙特卡罗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

$$E_{p(x)}[f(x)]$$

-

基本想法

- 马尔可夫链蒙特卡罗法的基本想法:

- 在随机变量x的状态空间S上定义一个满足遍历定理的马尔可夫链 $X = \{X_0, X_1, \cdots, X_t, \cdots\}$ ， 使其平稳分布就是抽样的目标分布p(x)
- 然后在这个马尔可夫链上进行随机游走，每个时刻得到一个样本

- 根据遍历定理，当时间趋于无穷时，样本的分布趋近平稳分布，样本的函数均值趋近函数的数学期望

基本想法

- 所以，当时间足够长时（时刻大于某个正整数m），在之后的时间（时刻小于等于某个正整数n，n > m）里随机游走得到的样本集合 $\{x_{m+1}, x_{m+2}, \cdots, x_n\}$ 就是目标概率分布的抽样结果

- 得到的函数均值（遍历均值） 就是要计算的数学期望值:

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

- 到时刻m为止的时间段称为燃烧期

基本想法

- 构建具体的马尔可夫链:
- 连续变量的时候，需要定义转移核函数
- 离散变量的时候，需要定义转移矩阵

- 一个方法是定义特殊的转移核函数或者转移矩阵，构建可逆马尔可夫链，这样可以保证遍历定理成立。

- 常用的马尔可夫链蒙特卡罗法有Metropolis-Hastings算法、吉布斯抽样。
- 由于这个马尔可夫链满足遍历定理，随机游走的起始点并不影响得到的结果，即从不同的起始点出发，都会收敛到同一平稳分布。

基本想法

- 马尔可夫链蒙特卡罗法的收敛性的判断通常是经验性的
- 比如，在马尔可夫链上进行随机游走，检验遍历均值是否收敛

- 具体地，每隔一段时间取一次样本，得到多个样本以后，计算遍历均值
- 当计算的均值稳定后，认为马尔可夫链已经收敛

- 再比如，在马尔可夫链上并行进行多个随机游走，比较各个随机游走的遍历均值是否接近 一致。

基本想法

- 马尔可夫链蒙特卡罗法中得到的样本序列，相邻的样本点是相关的，而不是独立的
- 因此，在需要独立样本时，可以在该样本序列中再次进行随机抽样
- 比如每隔一段时间取一次样本，将这样得到的子样本集合作为独立样本集合。

- 马尔可夫链蒙特卡罗法比接受-拒绝法更容易实现，因为只需要定义马尔可夫链， 而不需要定义建议分布

- 一般来说马尔可夫链蒙特卡罗法比接受-拒绝法效率更高，没有大量被拒绝的样本，虽然燃烧期的样本也要抛弃

基本步骤

- 可以将马尔可夫链蒙特卡罗法概括为以下三步:
- （1）首先，在随机变量x的状态空间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)
- （2）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 $x_0, x_1, \cdots, x_t, \cdots$。
- （3）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）

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

马尔可夫链蒙特卡罗法与统计学习

- 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 $$p(x|y) = \frac{p(x)p(y|x)}{\int_{\mathcal{X}} p(y|x')p(x')\mathrm{d}x'}$$

- 贝叶斯学习中经常需要进行三种积分运算:
  - 归范化（normalization)
  - 边缘 化（marginalization)
  - 数学期望（expectation)

马尔可夫链蒙特卡罗法与统计学习

- 后验概率计算中需要归范化计算：$\int_{\mathcal{X}} p(y|x')p(x')\mathrm{d}x'$
- 如果有隐变量 $z \in \mathcal{Z}$ ，后验概率的计算需要边缘化计算：

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

- 如果有一个函数f(x)，可以计算该函数的关于后验概率分布的数学期望：

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

- 马尔可夫链蒙特卡罗法为这些计算提供了一个通用的有效解决方案

Metropolis-Hastings算法

基本原理

- 1. 马尔科夫链

- 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

$$p(x, x') = q(x, x')\alpha(x, x')$$

- 其中q(x, x’)和α(x, x’)分别称为建议分布（proposal distribution）和接受分布 (acceptance distribution)

基本原理

- 建议分布q(x, x′)是另一个马尔可夫链的转移核，并且q(x, x′)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α(x, x′) 是

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

- 这时，转移核p(x, x′)可以写成

$$p(x, x') = \begin{cases} q(x, x'), & p(x')q(x', x) \geqslant p(x)q(x, x') \\ q(x', x)\dfrac{p(x')}{p(x)}, & p(x')q(x', x) < p(x)q(x, x') \end{cases}$$

基本原理

- 转移核为p(x, x')的马尔可夫链上的随机游走以以下方式进行

- 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α(x, x')抽样决定是否接受状态x'
- 以概率α(x, x')接受了x'，决定时刻t转移到状态x'，而以概率1- α(x, x')拒绝x'，决定时刻t仍停留在状态x

基本原理

- 具体地，从区间(0,1)上的均匀分布中抽取一个随机数u，决定时刻t的状态。

$$x_t = \begin{cases} x', & u \leqslant \alpha(x, x') \\ x, & u > \alpha(x, x') \end{cases}$$

- 可以证明，转移核为p(x, x')的马尔可夫链是可逆马尔可夫链（满足遍历定理），其平稳分布就是p(x)，即要抽样的目标分布。

- 也就是说这是马尔可夫链蒙特卡罗法的一个具体实现。

基本原理

定理 19.6 由转移核 (19.38)~(19.40) 构成的马尔可夫链是可逆的，即

$$p(x)p(x,x') = p(x')p(x',x)$$

并且 $p(x)$ 是该马尔可夫链的平稳分布。

基本原理

- 证明:

- 若x = x′，则式（19.41)显然成立。
- 若x ≠ x′，则

$$
\begin{aligned}
p(x)p(x,x') &= p(x)q(x,x')\min\left\{1, \frac{p(x')q(x',x)}{p(x)q(x,x')}\right\} \\
&= \min\{p(x)q(x,x'), p(x')q(x',x)\} \\
&= p(x')q(x',x)\min\left\{\frac{p(x)q(x,x')}{p(x')q(x',x)}, 1\right\} \\
&= p(x')p(x',x)
\end{aligned}
$$

- 式（19.41)

基本原理

- 由式（19.41)知，

$$\begin{aligned}\int p(x)p(x,x')\mathrm{d}x &= \int p(x')p(x',x)\mathrm{d}x \\ &= p(x')\int p(x',x)\mathrm{d}x \\ &= p(x')\end{aligned}$$

- 根据平稳分布的定义，p(x)是马尔可夫链的平稳分布。

基本原理

- 2. 建议分部

- 建议分部q(x, x′) 有多种可能的形式，这里介绍两种常用形式

- 第一种形式，假设建议分布是对称的，即对任意的x和x′有

$$q(x, x') = q(x', x)$$

- 这样的建议分布称为Metropolis选择。这时，接受分部α(x, x′)简化为

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

基本原理

- Metropolis选择的一个特例是q(x, x′) 取条件概率分布p(x′|x)，定义为多元正态 分布，其均值是x，其协方差矩阵是常数矩阵。

- Metropolis选择的另一个特例是令q(x, x′) = q(|x - x′|)，这时算法称为随机游走 Metropolis算法。例如，

$$q(x, x') \propto \exp(-\frac{(x' - x)^2}{2})$$

- Metropolis选择的特点是当x′与x接近时， q(x, x′) 的概率值高，否则q(x, x′) 的概率值低。状态转移在附近点的可能性更大。

基本原理

- 第二种形式称为独立抽样。假设q(x, x′) 与当前状态x 无关， 即q(x, x′) = q(x′)

- 建议分布的计算按照q(x′) 独立抽样进行。此时， 接受分布α(x, x′)可以写成

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

- 其中

$$w(x') = p(x')/q(x'),\ w(x) = p(x)/q(x)$$

- 独立抽样实现简单，但可能收敛速度慢，通常选择接近目标分布p(x)的分布作为建议分布q(x)

基本原理

- 3. 满条件分部

- 马尔可夫链蒙特卡罗法的目标分布通常是多元联合概率分布 $p(x) = p(x_1, x_2, \cdots, x_k)$，其中 $x = (x_1, x_2, \cdots, x_k)^{\mathrm{T}}$ 为k维随机变量

- 如果条件概率分布 $p(x_I|x_{-I})$ 中所有 k个变量全部出现，其中 $x_I = \{x_i, i \in I\}$, $x_{-I} = \{x_i, i \notin I\}$, $I \subset K = \{1, 2, \cdots, k\}$ 那么称这种条件概率分布为满条件分布（full conditional distribution)。

基本原理

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

$$p\left(x_I | x_{-I}\right) = \frac{p(x)}{\frac{p(x'_I | x'_{-I})}{p(x_I | x_{-I})}\, x_I} \propto p(x)$$

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

$$\frac{p(x'_I | x'_{-I})}{p(x_I | x_{-I})} = \frac{p(x')}{p(x)}$$

基本原理

- Metropolis-Hastings算法中，可以利用上述性质，简化计算，提高计算效率。具体地，通过满条件分布概率的比

$$\frac{p(x'_I|x'_{-I})}{p(x_I|x_{-I})}$$

- 计算联合概率的比

$$\frac{p(x')}{p(x)}$$

- 而前者更容易计算

例

- 设x_1和x_2的联合概率分布的密度函数为

$$p(x_1, x_2) \propto \exp\left\{-\frac{1}{2}(x_1 - 1)^2(x_2 - 1)^2\right\}$$

- 求其满条件分部

例

- 由满条件分布的定义有

$$p(x_1|x_2) \propto p(x_1, x_2)$$

$$\propto \exp\left\{-\frac{1}{2}(x_1-1)^2(x_2-1)^2\right\}$$

$$\propto N(1, (x_2-1)^{-2})$$

- 这里 $N(1, (x_2-1)^{-2})$ 是均值为1，方差为 $(x_2-1)^{-2}$ 的正态分布，这时x_1是变量，x_2是参数。同样可得 $p(x_2|x_1) \propto p(x_1, x_2)$

$$\propto \exp\left\{-\frac{1}{2}(x_2-1)^2(x_1-1)^2\right\}$$

$$\propto N(1, (x_1-1)^{-2})$$

Metropolis-Hastings算法

算法 19.2（Metropolis-Hastings 算法）

输入：抽样的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

Metropolis-Hastings算法

（2）对 $i = 1, 2, \cdots, n$ 循环执行

（a）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'。

（b）计算接受概率

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

（3）得到样本集合 $\{x_{m+1}, x_{m+2}, \cdots, x_n\}$

计算

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

单分量Metropolis- Hastings算法

- 在Metropolis-Hastings算法中，通常需要对多元变量分布进行抽样，有时对多元变量分布的抽样是困难的

- 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 (single-component Metropolis-Hastings）算法。

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

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

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

单分量Metropolis- Hastings算法

- 为了生成容量为的样本集合 $\{x^{(1)}, x^{(2)}, \cdots, x^{(n)}\}$，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 $x_j^{(i)}$

单分量Metropolis- Hastings算法

- 首先，由建议分布 $q(x_j^{(i-1)}, x_j|x_{-j}^{(i)})$ 抽样产生分量 x_j 的候选值 $x_j'^{(i)}$，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

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

- 其中分量1,2,…,j—1已经更新。然后，按照接受概率

$$\alpha(x_j^{(i-1)}, x'^{(i)}_j|x_{-j}^{(i)}) = \min\left\{1, \frac{p(x'^{(i)}_j|x_{-j}^{(i)})q(x'^{(i)}_j, x_j^{(i-1)}|x_{-j}^{(i)})}{p(x_j^{(i-1)}|x_{-j}^{(i)})q(x_j^{(i-1)}, x'^{(i)}_j|x_{-j}^{(i)})}\right\}$$

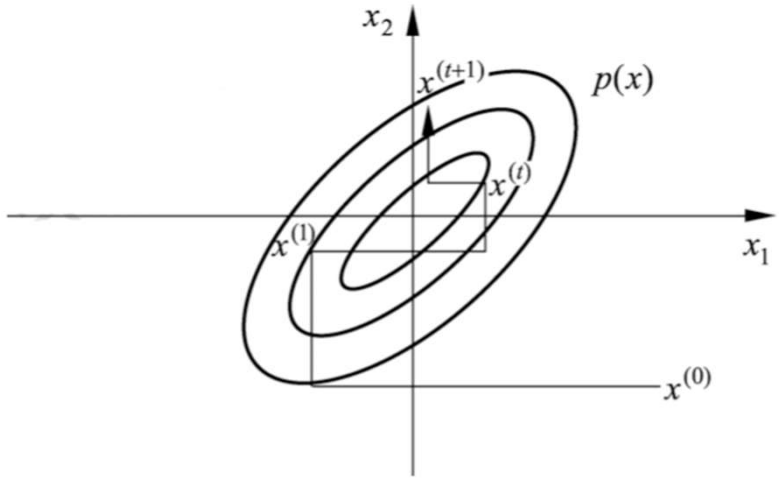
- 抽样决定是否接受候选值 $x_j'^{(i)}$。如果$x_j'^{(i)}$ 被接受，则令 $x_j^{(i)} = x'^{(i)}_j$
- 否则令 $x_j^{(i)} = x_j^{(i-1)}$。其余分量在第j步不改变

单分量Metropolis- Hastings算法

- 马尔可夫链的转移概率为

$$p\left(x_j^{(i-1)}, x'^{(i)}_j | x_{-j}^{(i)}\right) = \alpha(x_j^{(i-1)}, x'^{(i)}_j | x_{-j}^{(i)}) q(x_j^{(i-1)}, x'^{(i)}_j | x_{-j}^{(i)})$$

- 右图示意单分量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。



吉布斯抽样

吉布斯抽样

- 吉布斯抽样是马尔可夫链蒙特卡罗法的常用算法吉布斯抽样

- 可以认为是Metropolis-Hastings算法的特殊情况，但是更容易实现，因而被广泛使用。

基本原理

- 吉布斯抽样（Gibbs sampling）用于多元变量联合分布的抽样和估计
- 其基本做法是，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
- 可以证明这样的抽样过程是在一个马尔可夫链上的随机游走，每一个样本对应着马尔可夫链的状态，平稳分布就是目标的联合分布。
- 整体成为一个马尔可夫 链蒙特卡罗法，燃烧期之后的样本就是联合分布的随机样本

基本原理

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

基本原理

- 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 $p(x_j|x_{-j}^{(i)})$，这里 $x_{-j}^{(i)}$ 表示第i次迭代中，变量j以外的其他变量

基本原理

- 设在第（i-1）步得到样本 $(x_1^{(i-1)}, x_2^{(i-1)}, \cdots, x_k^{(i-1)})^{\mathrm{T}}$ ， 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

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

- 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

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

- 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

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

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

基本原理

- 吉布斯抽样是单分量Metropolis-Hastings算法的特殊情况

- 定义建议分布是当前变量x_j，j = 1,2,…，k的满条件概率分布

$$q(x, x') = p(x'_j | x_{-j})$$

- 这时，接受概率 $\alpha = 1$

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

$$= \min\left\{1, \frac{p(x'_{-j})p(x'_j|x'_{-j})p(x_j|x'_{-j})}{p(x_{-j})p(x_j|x_{-j})p(x'_j|x_{-j})}\right\} = 1$$

- 这里用到 $p(x_{-j}) = p(x'_{-j})$ 和 $p(\bullet|x_{-j}) = p(\bullet|x'_{-j})$

基本原理

- 转移核就是满条件概率分布

$$p(x, x') = p(x'_j | x_{-j})$$

- 也就是说依次按照单变量的满条件概率分布 $p(x'_j | x_{-j})$ 进行随机抽样，就能实现单分量Metropolis-Hastings算法
- 吉布斯抽样对每次抽样的结果都接受，没有拒绝，这一点和一般的Metropolis-Hastings算法不同

- 这里，假设满条件概率分布 $p(x'_j | x_{-j})$ 不为0，即马尔可夫链是不可约的

吉布斯抽样算法

算法 19.3 (吉布斯抽样)

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 $x^{(0)} = (x_1^{(0)}, x_2^{(0)}, \cdots, x_k^{(0)})^{\mathrm{T}}$。

（2）对 i 循环执行

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

$$
\begin{cases}
\text{（1）由满条件分布 } p(x_1 | x_2^{(i-1)}, \cdots, x_k^{(i-1)}) \text{ 抽取 } x_1^{(i)} \\
\quad \vdots \\
\text{（j）由满条件分布 } p(x_j | x_1^{(i)}, \cdots, x_{j-1}^{(i)}, x_{j+1}^{(i-1)}, \cdots, x_k^{(i-1)}) \text{ 抽取 } x_j^{(i)} \\
\quad \vdots \\
\text{（k）由满条件分布 } p(x_k | x_1^{(i)}, \cdots, x_{k-1}^{(i)}) \text{ 抽取 } x_k^{(i)}
\end{cases}
$$

得到第 i 次迭代值 $x^{(i)} = (x_1^{(i)}, x_2^{(i)}, \cdots, x_k^{(i)})^{\mathrm{T}}$。

吉布斯抽样算法

（3）得到样本集合

$$\{x^{(m+1)}, x^{(m+2)}, \cdots, x^{(n)}\}$$

（4）计算

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

例

- 用吉布斯抽样从以下二元正态分布中抽取随机样本

$$x = (x_1, x_2)^{\mathrm{T}} \sim p(x_1, x_2)$$

$$p(x_1, x_2) = N(0, \Sigma), \quad \Sigma = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$$

例

- 条件概率分布为一元正态分布

$$p(x_1|x_2) = N(\rho x_2, (1-\rho^2))$$

$$p(x_2|x_1) = N(\rho x_1, (1-\rho^2))$$

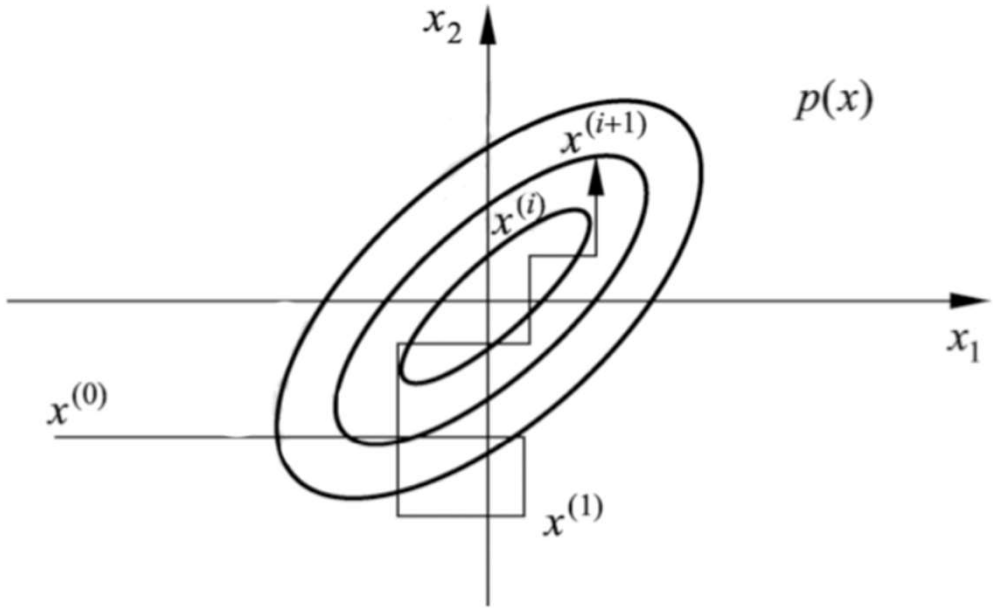
- 假设初始样本为 $x^{(0)} = (x_1^{(0)}, x_2^{(0)})$，通过吉布斯抽样，可以得到以下样本序列：

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

例

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

- 右图示意吉布斯抽样的过程



吉布斯抽样算法

- 单分量Metropolis-Hastings 算法
  - 抽样会在样本点之间移动，但其间可能在某一些样本点上停留（由于抽样被拒绝）
  - 适合于满条件概率分布不容易抽样的情况，使用容易抽样的条件分 作建议分布

- 吉布斯抽样算法
  - 抽样会在样本点之间持续移动
  - 适合于满条件概率分布容易抽样的情况

抽样计算

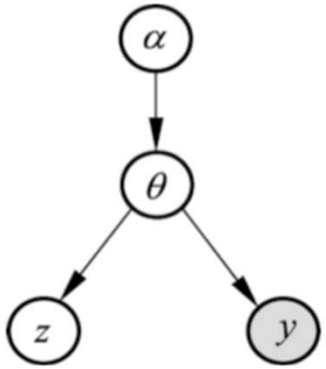
- 吉布斯抽样中需要对满条件概率分布进行重复多次抽样
- 可以利用概率分布的性质提高抽样的效率

- 下面以贝叶斯学习为例介绍这个技巧

抽样计算

- 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，

$x = (\alpha, \theta, z)$



- 贝叶斯学习的目的是估计后验概率分布 $p(x|y)$，求后验概率最大的模型

$$p(x|y) = p(\alpha, \theta, z|y) \propto p(z, y|\theta)p(\theta|\alpha)p(\alpha)$$

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

抽样计算

- 现在用吉布斯抽样估计 $p(x|y)$，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

$$p(\alpha_i|\alpha_{-i},\theta,z,y) \propto p(\theta|\alpha)p(\alpha)$$

$$p(\theta_j|\theta_{-j},\alpha,z,y) \propto p(z,y|\theta)p(\theta|\alpha)$$

$$p(z_k|z_{-k},\alpha,\theta,y) \propto p(z,y|\theta)$$

- 其中 α_{-i} 表示变量 α_i 以外的所有变量，θ_{-j} 和 z_{-k} 类似
- 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。