[MCMC(一)蒙特卡罗方法]

MCMC(一)蒙特卡罗方法

[MCMC(二)马尔科夫链]

[MCMC(三)MCMC采样和M-H采样]

[MCMC(四)Gibbs采样]

作为一种随机采样方法，马尔科夫链蒙特卡罗（Markov Chain Monte Carlo，以下简称MCMC）在机器学习,深度学习以及自然语言处理等领域都有广泛的应用，是很多复杂算法求解的基础。比如我们前面讲到的[分解机(Factorization Machines)推荐算法]，还有前面讲到的[受限玻尔兹曼机（RBM）原理总结]，都用到了MCMC来做一些复杂运算的近似求解。下面我们就对MCMC的原理做一个总结。

1. MCMC概述

从名字我们可以看出，MCMC由两个MC组成，即蒙特卡罗方法（Monte Carlo Simulation，简称MC）和马尔科夫链（Markov Chain ，也简称MC）。要弄懂MCMC的原理我们首先得搞清楚蒙特卡罗方法和马尔科夫链的原理。我们将用三篇来完整学习MCMC。在本篇，我们关注于蒙特卡罗方法。

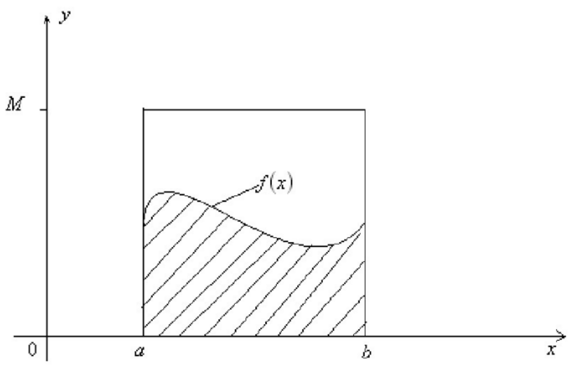
2. 蒙特卡罗方法引入

蒙特卡罗原来是一个赌场的名称，用它作为名字大概是因为蒙特卡罗方法是一种随机模拟的方法，这很像赌博场里面的扔骰子的过程。最早的蒙特卡罗方法都是为了求解一些不太好求解的求和或者积分问题。比如积分：

$$\theta = \int_a^b f(x)dx$$

θ=∫abf(x)dx

如果我们很难求解出$f(x)$ f(x)的原函数，那么这个积分比较难求解。当然我们可以通过蒙特卡罗方法来模拟求解近似值。如何模拟呢？假设我们函数图像如下图:



则一个简单的近似求解方法是在[a,b]之间随机的采样一个点。比如x_0x0,然后用$f(x_0)$ f(x0)代表在[a,b]区间上所有的$f(x)$ f(x)的值。那么上面的定积分的近似求解为:

$$(b-a)f(x_0)$$

(b−a)f(x0)

当然，用一个值代表[a,b]区间上所有的$f(x)$ f(x)的值，这个假设太粗糙。那么我们可以采样[a,b]区间的n个值：$x_0, x_1, \ldots x_{n-1}$ x0,x1,...xn−1,用它们的均值来代表[a,b]区间上所有的$f(x)$ f(x)的值。这样我们上面的定积分的近似求解为:

$$\frac{b-a}{n}\sum_{i=0}^{n-1} f(x_i)$$

b−an∑i=0n−1f(xi)

虽然上面的方法可以一定程度上求解出近似的解，但是它隐含了一个假定，即x在[a,b]之间是均匀分布的，而绝大部分情况，x在[a,b]之间不是均匀分布的。如果我们用上面的方法，则模拟求出的结果很可能和真实值相差甚远。

怎么解决这个问题呢？ 如果我们可以得到x在[a,b]的概率分布函数$p(x)$，那么我们的定积分求和可以这样进行：

$$\theta = \int_a^b f(x)dx = \int_a^b \frac{f(x)}{p(x)} p(x)dx \approx \frac{1}{n}\sum_{i=0}^{n-1}\frac{f(x_i)}{p(x_i)}$$

θ=∫abf(x)dx=∫abf(x)p(x)p(x)dx≈1n∑i=0n−1f(xi)p(xi)

上式最右边的这个形式就是蒙特卡罗方法的一般形式。当然这里是连续函数形式的蒙特卡罗方法，但是在离散时一样成立。

可以看出，最上面我们假设x在[a,b]之间是均匀分布的时候，$p(x_i) = 1/(b-a)$，带入我们有概率分布的蒙特卡罗积分的上式，可以得到：

$$\frac{1}{n}\sum_{i=0}^{n-1}\frac{f(x_i)}{1/(b-a)} = \frac{b-a}{n}\sum_{i=0}^{n-1} f(x_i)$$

1n∑i=0n−1f(xi)1/(b−a)=b−an∑i=0n−1f(xi)

也就是说，我们最上面的均匀分布也可以作为一般概率分布函数$p(x)$在均匀分布时候的特例。那么我们现在的问题转到了如何求出x的分布$p(x)$对应的若干个样本上来。

3. 概率分布采样

上一节我们讲到蒙特卡罗方法的关键是得到x的概率分布。如果求出了x的概率分布，我们可以基于概率分布去采样基于这个概率分布的n个x的样本集，带入蒙特卡罗求和的式子即可求解。但是还有一个关键的问题需要解决，即如何基于概率分布去采样基于这个概率分布的n个x的样本集。

对于常见的均匀分布$uniform(0,1)$是非常容易采样样本的，一般通过线性同余发生器可以很方便的生成(0,1)之间的伪随机数样本。而其他常见的概率分布，无论是离散的分布还是连续的分布，它们的样本都可以通过$uniform(0,1)$的样本转换而得。比如二维正态分布的样本(Z_1, Z_2)可以通过通过独立采样得到的$uniform(0,1)$样本对(X_1, X_2)通过如下的式子转换而得：

$$Z_1 = \sqrt{-2\ln X_1}\cos(2\pi X_2)$$

Z1=−2lnX1cos(2πX2)

$$Z_2 = \sqrt{-2\ln X_1}\sin(2\pi X_2)$$

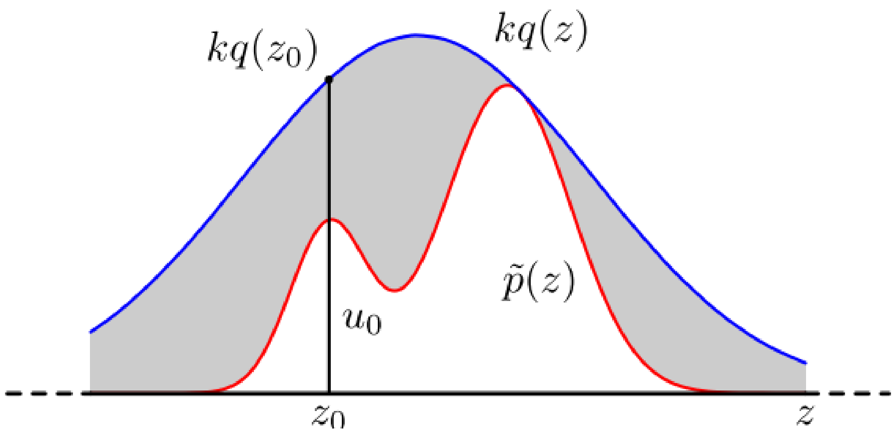
Z2=−2lnX1sin(2πX2)

其他一些常见的连续分布，比如t分布，F分布，Beta分布，Gamma分布等，都可以通过类似的方式从$uniform(0,1)$得到的采样样本转化得到。在python的numpy，scikit-learn等类库中，都有生成这些常用分布样本的函数可以使用。

不过很多时候，我们的x的概率分布不是常见的分布，这意味着我们没法方便的得到这些非常见的概率分布的样本集。那这个问题怎么解决呢？

4. 接受-拒绝采样

对于概率分布不是常见的分布，一个可行的办法是采用接受-拒绝采样来得到该分布的样本。既然 $p(x)$ 太复杂在程序中没法直接采样，那么我设定一个程序可采样的分布 $q(x)$ 比如高斯分布，然后按照一定的方法拒绝某些样本，以达到接近 $p(x)$ 分布的目的，其中$q(x)$叫做 proposal distribution。



具体采用过程如下，设定一个方便采样的常用概率分布函数 $q(x)$q(x)，以及一个常量 kk，使得 $p(x)$p(x) 总在 $kq(x)$kq(x) 的下方。如上图。

首先，采样得到$q(x)$q(x)的一个样本z_0z0，采样方法如第三节。然后，从均匀分布（$0, kq(z_0)$) (0,kq(z0))中采样得到一个值uu。如果uu落在了上图中的灰色区域，则拒绝这次抽样，否则接受这个样本z_0z0。重复以上过程得到n个接受的样本$z_0, z_1, \ldots z_{n-1}$ z0,z1,...zn−1,则最后的蒙特卡罗方法求解结果为：

$$\frac{1}{n}\sum_{i=0}^{n-1}\frac{f(z_i)}{p(z_i)}$$

1n∑i=0n−1f(zi)p(zi)

整个过程中，我们通过一系列的接受拒绝决策来达到用$q(x)$q(x)模拟$p(x)$p(x)概率分布的目的。

5. 蒙特卡罗方法小结

使用接受-拒绝采样，我们可以解决一些概率分布不是常见的分布的时候，得到其采样集并用蒙特卡罗方法求和的目的。但是接受-拒绝采样也只能部分满足我们的需求，在很多时候我们还是很难得到我们的概率分布的样本集。比如：

1）对于一些二维分布$p(x, y)$p(x,y)，有时候我们只能得到条件分布$p(x|y)$p(x|y)和$p(y|x)$p(y|x)和,却很难得到二维分布$p(x, y)$p(x,y)一般形式，这时我们无法用接受-拒绝采样得到其样本集。

2）对于一些高维的复杂非常见分布$p(x_1, x_2, \ldots, x_n)$p(x1,x2,...,xn)，我们要找到一个合适的$q(x)$q(x)和kk非常困难。

从上面可以看出，要想将蒙特卡罗方法作为一个通用的采样模拟求和的方法，必须解决如何方便得到各种复杂概率分布的对应的采样样本集的问题。而我们下一篇要讲到的马尔科夫链就是帮助找到这些复杂概率分布的对应的采样样本集的白衣骑士。下一篇我们来总结马尔科夫链的原理。

« 上一篇：受限玻尔兹曼机（RBM）原理总结
» 下一篇：MCMC(二)马尔科夫链

评论列表

#1楼 2017-03-27 17:13 codesnippet.info
加油!
支持(0)反对(0)

#2楼 2017-03-27 20:05 xulu1352
大神又开更了，，大神后期能不能讲讲特征工程的东东，想看看老司机怎么做的
支持(0)反对(0)

#3楼[楼主] 2017-03-28 10:11 刘建平Pinard
@ xulu1352
写作提纲里有特征工程部分，应该在写完自然语言处理部分就会开始。
支持(0)反对(0)

#4楼 2017-03-28 19:04 桂。
感谢分享，学到了许多干货!
支持(0)反对(0)

#5楼 2017-10-31 09:54 开拓者亮仔
@刘建平Pinard 怎么解决这个问题呢？ 如果我们可以得到x在[a,b]的概率分布函数p(x)，那么我们的定积分求和可以这样进行. 那面的式子的得出不是很明白，可否详细说下。
支持(0)反对(0)

#6楼[楼主] 2017-10-31 10:37 刘建平Pinard
@ 开拓者亮仔
你说的是这个式子吧。

$$\theta = \int_a^b f(x)dx = \int_a^b \frac{f(x)}{p(x)}p(x)dx \approx \frac{1}{n}\sum_{i=0}^{n-1}\frac{f(x_i)}{p(x_i)}$$

θ=∫abf(x)dx=∫abf(x)p(x)p(x)dx≈1n∑i=0n−1f(xi)p(xi)

猜你的问题在最后一步转换。由于$\int_a^b \frac{f(x)}{p(x)}p(x)dx$∫abf(x)p(x)p(x)dx可以看做是$\frac{f(x)}{p(x)}$f(x)p(x)基于概率分布$p(x)$p(x)的期望，那么我们可以用期望的方法来求这个式子的值。而计算期望的一个近似方法是取$\frac{f(x)}{p(x)}$f(x)p(x)的若干个基于分布$p(x)$p(x)的采样点，然后求平均值得到。
支持(10)反对(0)

#7楼 2018-02-08 11:29 stillriver
@ 刘建平Pinard

> 引用
>
> @开拓者亮仔
> 你说的是这个式子吧。
> θ=∫abf(x)dx=∫abf(x)p(x)p(x)dx≈1n∑i=0n−1f(xi)p(xi)" role="presentation" style="text-align: center; position: relative;">θ=∫baf(x)dx=∫baf(x)p(x)p(x)dx≈1n∑i=0n−1f(xi)p(xi)θ=∫abf(x)dx=∫abf(x)p(x)p(x)dx≈1n∑i=0n−1f(xi)p(xi...

大数定理
支持(0)反对(0)

#8楼 2018-02-10 15:39 扼杀
是的，基于分布p(x)的实际采样时，样本的密度不一样，f(x)/p(x)刚好是对应的值。当采样的数量足够多，就大数定理，基本符合p(x)的分布了。 采样，是一个从分布函数 到 n个采样点的映射。 刚一看，真是不理解，说明以前看的也没有仔细问题
支持(0)反对(0)

#9楼 2018-04-14 15:51 ppen2018
你好博主，有个问题我有点晕。
如果x不是均匀分布，x符合p分布。那么 ∫fx * px dx不就是代表了x不均匀分布的时候在a-b区间上的积分么？为啥文章里是∫ fx/px * px dx呢 ?
支持(0)反对(0)

[#10楼]()[楼主] 2018-04-14 23:37 [刘建平Pinard]()
[@]() ppen2018
你好，fx不均匀分布的时候在a-b区间上的积分应该是：

$$\int_a^b f(x)dx$$

∫abf(x)dx

,而不是

$$\int_a^b f(x)p(x)dx$$

∫abf(x)p(x)dx

后面这个式子是$f(x)$ f(x)的期望。
[支持(0)]()[反对(1)]()

[#11楼]() 2018-04-20 09:47 [hapjin]()
第二节末尾这句话：“那么我们现在的问题转到了如何求出x的分布p(x)的若干和样本上来。”有点不通顺哎？
[支持(0)]()[反对(0)]()

[#12楼]()[楼主] 2018-04-20 12:30 [刘建平Pinard]()
[@]() hapjin
的确不通顺，改为：“那么我们现在的问题转到了如何求出x的分布p(x)对应的若干个样本上来”。
感谢指出。
[支持(0)]()[反对(0)]()

[#13楼]() 2018-04-20 17:08 [hapjin]()
请教大神一个问题：我们的目标是求解积分

$$\int_a^b f(x)\,dx$$

∫abf(x)dx

由于直接积分积不出来，于是转而求解 **和式**：

$$\frac{1}{n}\sum_{i=0}^{n-1}\frac{f(x_i)}{p(x_i)}$$

1n∑i=0n−1f(xi)p(xi)

假设随机变量X X是一个连续型随机变量，它的概率密度是$p(x_i)$p(xi)。但是由于$p(x_i)$p(xi)不好采样，于是使用 容易采样的概率密度函数$q(x_i)$q(xi)（比如高斯分布/均匀分布），当采得样本之后代入那个**和式**里面去。
由于**和式**里面有：$p(x_i)$p(xi)，那这是不是意味着 $p(x)$p(x)的表达式是已知的？因为这样才能代入到 **和式** 里面 把累加和求解出来。
那在现实中，为什么 $p(x)$p(x) 是已经的呢？也许是我没有太多实践经验，这个问题问得比较傻。
[支持(0)]()[反对(0)]()

[#14楼]() 2018-04-20 17:15 [hapjin]()
补充一下：因为您在文章中第4节说了：“。既然 $p(x)$p(x) 太复杂在程序中没法直接采样......”，$p(x)$p(x) 太复杂，指的是哪方面的“特征”太复杂？
上一条评论中，我说的$p(x)$p(x)的表达式是已知的，比如说：

$$p(x) = x^2 + 1$$

p(x)=x2+1

当然，我这只是举个例子，方便大神理解。
[支持(0)]()[反对(0)]()

[#15楼]()[楼主] 2018-04-20 22:46 [刘建平Pinard]()
[@]() hapjin
你好。
首先要理解容易采样的分布基本都是从均匀分布采样转化而得。均匀分布式最好采样的。

有些分布$p(x)$p(x)虽然表达式已知，但是还是很难采样。比如有些含有无法直接求原函数的积分对应的概率分布，就很难直接从均匀分布转化而得，那么也就很难采样了。

支持(0)反对(0)

#16楼 2018-05-09 11:37 钱艺铭

你好，大神，看了你的博客很多次了，感觉写的很好，想请教一下，你这些知识的总结是从哪些书籍学习到的，现在属于入门阶段，想系统学习下，谢谢~

支持(0)反对(0)

#17楼[楼主] 2018-05-10 10:20 刘建平Pinard

@ 钱艺铭

你好，书的话 李航的 统计学习方法 和 周志华的 机器学习 这两本我觉得还不错。建议关注一些比较新机器学习的博客和一些科普论文。

支持(0)反对(0)

#18楼 2018-05-10 11:01 钱艺铭

@ 刘建平Pinard

好的，非常感谢~会一直关注你的博客进行学习的~感觉受益很大~

支持(0)反对(0)

#19楼 2018-05-10 16:04 linshizuowei

> 引用
>
> 如果我们可以得到xx在[a,b]的概率分布函数p(x)p(x)，那么我们的定积分求和可以这样进行

请问博主，我们一般如何得到x在[a, b]上的概率分布函数p(x)?

支持(0)反对(0)

#20楼[楼主] 2018-05-10 21:58 刘建平Pinard

@ linshizuowei

你好，在数据分析中概率分布我们一般是假定的，比如正态分布。我们一般会把数据分布近似的看做某一种常见的分布。

当然，如果概率分布是给定已知的，那就更好了。

支持(0)反对(0)

#21楼 2018-05-11 09:51 linshizuowei

@ 刘建平Pinard

OK明白了，读你的博客受益匪浅，非常感谢~

支持(0)反对(0)

#22楼 2018-05-18 16:03 蜉蝣2015

你这个MCMC讲的太好了，逻辑性强，目标清晰，大神可否提供下参考资料来源?

支持(0)反对(0)

#23楼[楼主] 2018-05-19 12:07 刘建平Pinard

@ 蜉蝣2015

你可以参考下<LDA数学八卦>这篇文章，网上找得到

支持(0)反对(0)

#24楼 2018-09-09 10:00 Asber

博主大大 请教两个问题

第一个是

求定积分的时候 您提到x在[a,b]之间是均匀分布的

请问是指f(x)在[a,b]是均匀分布的吗 因为感觉x这个自变量就是线性增长的 f(x）才是随之变化的 一个有分布的随机变量

之后说的 x在[a,b]的概率分布函数p(x)

是否也是说f(x)在x~a,b间的概率分布函数呢

第二个问题是

看后面拒绝接受的算法说明 因为要比较，那么明显是已知Z0点的f（x）的值了，请问如果这个值已知，为什么不能用微积分的无穷小逼近思想求积分呢

问题有点白痴 不好意思哈

支持(0)反对(0)

#25楼[楼主] 2018-09-09 13:55 刘建平Pinard

@ Asber

你好!

1. 这里的分布是指的x在[a,b]之间的分布。也就是x在[a,b]之间某一个点出现的概率相同。如果x在[a,b]之间某些位置出现的概率高，其他位置出现的概率低，那么就不是均匀分布了。

2. x在[a,b]的概率分布函数p(x)，仍然说的是x的概率分布，而f(x) 只是x依概率出现在[a,b]之间一个确定的位置后，对应的函数值。

3. 微积分的无穷小逼近思想,假设了x在数轴所有的位置出现的概率均等，如果x在数轴各个位置出现的概率不一样，那么就不能用你说的微积分逼近了。
支持(0)反对(0)

#26楼 2018-09-19 09:57 aaronwang123
@ 刘建平Pinard
θ=∫baf(x)dx=∫baf(x)p(x)p(x)dx≈1n∑i=0n−1f(xi)p(xi)
你好，博主：

这个期望感觉不能用 求和取均值吧。
取均值（除以n）的前提是概率密度p（x）是均匀分布。
显然事实不是均匀分布。
支持(0)反对(0)

#27楼[楼主] 2018-09-19 11:13 刘建平Pinard
@ aaronwang123
你好，你说的应该是这个式子：

$$\theta = \int_a^b f(x)dx = \int_a^b \frac{f(x)}{p(x)}p(x)dx \approx \frac{1}{n}\sum_{i=0}^{n-1}\frac{f(x_i)}{p(x_i)}$$

θ=∫abf(x)dx=∫abf(x)p(x)p(x)dx≈1n∑i=0n−1f(xi)p(xi)

这里 $p(x)$ p(x)的确不是均匀分布，但是 $\frac{f(x_i)}{p(x_i)}$ f(xi)p(xi)里的这个分母已经考虑了你说的情况。如果没有这个分母，那就是我们假定了是均匀分布。
支持(0)反对(0)

#28楼 2018-09-19 11:26 aaronwang123
@ 刘建平Pinard
支持(0)反对(0)

#29楼[楼主] 2018-09-19 13:56 刘建平Pinard
@ aaronwang123
你好，这个式子不能按你图中的方法去理解。你那个是等于的情况，而我这里是近似等于。

举个例子，$f(x)$ f(x)的取值只有2个，$x = 1，2$x=1，2.对应的yy值分别是$f(1) = 1, f(2) = 4$ f(1)=1,f(2)=4。其中x x的取值不是平均的，取1的概率$p(1) = 0.25$p(1)=0.25,取2的概率是0.75。

那么严格来说，对应的$f(x)$ f(x)的积分等于$4 + 1 = 5$4+1=5。

此时我们去采样三次，期望求近似结果。发现第一次采样到1，第二次和第三次采样到2，那么最后的近似结果是 $\frac{1}{3}(\frac{1}{0.25} + (\frac{4}{0.75} + (\frac{4}{0.75}) = 4.89$13(10.25+(40.75+(40.75)=4.89。

这个4.89就是我们5的近似。虽然有些距离，但是是是由于采样太少的原因。

假设我们采样100次，得到26次1, 74次2，那么最后的近似结果是 $\frac{1}{100}(\frac{1}{0.25} \times 26 + (\frac{4}{0.75} \times 74) = 4.99$1100(10.25×26+(40.75×74)=4.99

可见越来越接近的。接近的原因是随着采样数的增多，采样的样本的分布越来越接近于x本来的分布
支持(1)反对(0)

#30楼 2018-09-19 14:28 aaronwang123
@ 刘建平Pinard
谢谢博主详细解答！
支持(0)反对(0)

[MCMC(二)马尔科夫链]

在[MCMC(一)蒙特卡罗方法]中，我们讲到了如何用蒙特卡罗方法来随机模拟求解一些复杂的连续积分或者离散求和的方法，但是这个方法需要得到对应的概率分布的样本集，而想得到这样的样本集很困难。因此我们需要本篇讲到的马尔科夫链来帮忙。

1. 马尔科夫链概述

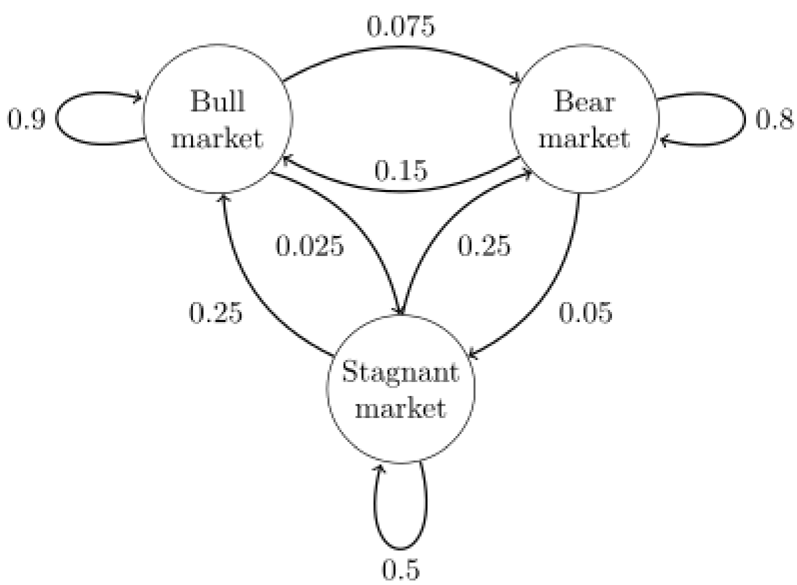
马尔科夫链定义本身比较简单，它假设某一时刻状态转移的概率只依赖于它的前一个状态。举个形象的比喻，假如每天的天气是一个状态的话，那个今天是不是晴天只依赖于昨天的天气，而和前天的天气没有任何关系。当然这么说可能有些武断，但是这样做可以大大简化模型的复杂度，因此马尔科夫链在很多时间序列模型中得到广泛的应用，比如循环神经网络RNN，隐式马尔科夫模型HMM等，当然MCMC也需要它。

如果用精确的数学定义来描述，则假设我们的序列状态是 $\ldots X_{t-2}, X_{t-1}, X_t, X_{t+1}, \ldots$，那么我们的在时刻 X_{t+1} 的状态的条件概率仅仅依赖于时刻 X_t，即：

$$P(X_{t+1}|\ldots X_{t-2}, X_{t-1}, X_t) = P(X_{t+1}|X_t)$$

P(Xt+1|...Xt−2,Xt−1,Xt)=P(Xt+1|Xt)

既然某一时刻状态转移的概率只依赖于它的前一个状态，那么我们只要能求出系统中任意两个状态之间的转换概率，这个马尔科夫链的模型就定了。我们来看看下图这个马尔科夫链模型的具体的例子(来源于维基百科)。



这个马尔科夫链是表示股市模型的，共有三种状态：牛市（Bull market）, 熊市（Bear market）和横盘（Stagnant market）。每一个状态都以一定的概率转化到下一个状态。比如，牛市以0.025的概率转化到横盘的状态。这个状态概率转化图可以以矩阵的形式表示。如果我们定义矩阵阵P P某一位置$P(i, j)$ P(i,j)的值为$P(j|i)$ P(j|i),即从状态i转化到状态j的概率，并定义牛市为状态0， 熊市为状态1, 横盘为状态2. 这样我们得到了马尔科夫链模型的状态转移矩阵为：

$$P = \begin{pmatrix} 0.9 & 0.075 & 0.025 \\ 0.15 & 0.8 & 0.05 \\ 0.25 & 0.25 & 0.5 \end{pmatrix}$$

P=(0.90.0750.0250.150.80.050.250.250.5)

讲了这么多，那么马尔科夫链模型的状态转移矩阵和我们蒙特卡罗方法需要的概率分布样本集有什么关系呢？这需要从马尔科夫链模型的状态转移矩阵的性质讲起。

2. 马尔科夫链模型状态转移矩阵的性质

得到了马尔科夫链模型的状态转移矩阵，我们来看看马尔科夫链模型的状态转移矩阵的性质。

完整代码参见我的github:https://github.com/ljpzzz/machinelearning/blob/master/mathematics/mcmc_2.ipynb

仍然以上面的这个状态转移矩阵为例。假设我们当前股市的概率分布为：[0.3,0.4,0.3],即30%概率的牛市，40%概率的熊盘与30%的横盘。然后这个状态作为序列概率分布的初始状态t_0t0，将其带入这个状态转移矩阵计算$t_1, t_2, t_3...$t1,t2,t3...的状态。代码如下：

```
import numpy as np
matrix = np.matrix([[0.9,0.075,0.025],[0.15,0.8,0.05],[0.25,0.25,0.5]], dtype=float)
vector1 = np.matrix([[0.3,0.4,0.3]], dtype=float)
for i in range(100):
    vector1 = vector1*matrix
    print "Current round:" , i+1
    print vector1
```

部分输出结果如下：

```
Current round: 1
[[ 0.405    0.4175   0.1775]]
Current round: 2
[[ 0.4715    0.40875   0.11975]]
Current round: 3
[[ 0.5156   0.3923   0.0921]]
Current round: 4
[[ 0.54591    0.375535   0.078555]]
。。。。。。
Current round: 58
[[ 0.62499999   0.31250001   0.0625     ]]
Current round: 59
[[ 0.62499999   0.3125       0.0625     ]]
Current round: 60
[[ 0.625    0.3125   0.0625]]
。。。。。。
Current round: 99
[[ 0.625    0.3125   0.0625]]
```

```
Current round: 100
[[ 0.625    0.3125   0.0625]]
```

可以发现，从第60轮开始，我们的状态概率分布就不变了，一直保持在[0.625　0.3125　0.0625]，即62.5%的牛市，31.25%的熊市与6.25%的横盘。那么这个是巧合吗？

我们现在换一个初始概率分布试一试，现在我们用[0.7,0.1,0.2]作为初始概率分布，然后这个状态作为序列概率分布的初始状态t_0t0，将其带入这个状态转移矩阵计算$t_1, t_2, t_3...$t1,t2,t3...的状态。代码如下：

```
matrix = np.matrix([[0.9,0.075,0.025],[0.15,0.8,0.05],[0.25,0.25,0.5]], dtype=float)
vector1 = np.matrix([[0.7,0.1,0.2]], dtype=float)
for i in range(100):
    vector1 = vector1*matrix
    print "Current round:" , i+1
    print vector1
```

部分输出结果如下：

```
Current round: 1
[[ 0.695    0.1825   0.1225]]
Current round: 2
[[ 0.6835    0.22875   0.08775]]
Current round: 3
[[ 0.6714   0.2562   0.0724]]
Current round: 4
[[ 0.66079    0.273415   0.065795]]
。。。。。。。
Current round: 55
[[ 0.62500001   0.31249999   0.0625     ]]
Current round: 56
[[ 0.62500001   0.31249999   0.0625     ]]
Current round: 57
[[ 0.625    0.3125   0.0625]]
。。。。。。。
Current round: 99
[[ 0.625    0.3125   0.0625]]
Current round: 100
[[ 0.625    0.3125   0.0625]]
```

可以看出，尽管这次我们采用了不同初始概率分布，最终状态的概率分布趋于同一个稳定的概率分布[0.625　0.3125　0.0625]， 也就是说我们的马尔科夫链模型的状态转移矩阵收敛到的稳定概率分布与我们的初始状态概率分布无关。这是一个非常好的性质，也就是说，如果我们得到了这个稳定概率分布对应的马尔科夫链模型的状态转移矩阵，则我们可以用任意的概率分布样本开始，带入马尔科夫链模型的状态转移矩阵，这样经过一些序列的转换，最终就可以得到符合对应稳定概率分布的样本。

这个性质不光对我们上面的状态转移矩阵有效，对于绝大多数的其他的马尔科夫链模型的状态转移矩阵也有效。同时不光是离散状态，连续状态时也成立。

同时，对于一个确定的状态转移矩阵P P，它的n次幂P^nPn在当n大于一定的值的时候也可以发现是确定的，我们还是以上面的例子为例，计算代码如下：

```
matrix = np.matrix([[0.9,0.075,0.025],[0.15,0.8,0.05],[0.25,0.25,0.5]], dtype=float)
for i in range(10):
    matrix = matrix*matrix
    print "Current round:" , i+1
    print matrix
```

输出结果如下：

```
Current round: 1
[[ 0.8275   0.13375  0.03875]
 [ 0.2675   0.66375  0.06875]
 [ 0.3875   0.34375  0.26875]]
Current round: 2
[[ 0.73555   0.212775  0.051675]
 [ 0.42555   0.499975  0.074475]
 [ 0.51675   0.372375  0.110875]]
。。。。。。
Current round: 5
[[ 0.62502532  0.31247685  0.06249783]
 [ 0.6249537   0.31254233  0.06250397]
 [ 0.62497828  0.31251986  0.06250186]]
Current round: 6
[[ 0.625   0.3125  0.0625]
 [ 0.625   0.3125  0.0625]
 [ 0.625   0.3125  0.0625]]
Current round: 7
[[ 0.625   0.3125  0.0625]
 [ 0.625   0.3125  0.0625]
 [ 0.625   0.3125  0.0625]]
。。。。。。
Current round: 9
[[ 0.625   0.3125  0.0625]
 [ 0.625   0.3125  0.0625]
 [ 0.625   0.3125  0.0625]]
Current round: 10
[[ 0.625   0.3125  0.0625]
 [ 0.625   0.3125  0.0625]
 [ 0.625   0.3125  0.0625]]
```

我们可以发现，在$n \geq 6$n≥6以后，P^nPn的值稳定不再变化，而且每一行都为[0.625 0.3125 0.0625]，这和我们前面的稳定分布是一致的。这个性质同样不光是离散状态，连续状态时也成立。

好了，现在我们可以用数学语言总结下马尔科夫链的收敛性质了：

如果一个非周期的马尔科夫链有状态转移矩阵P P, 并且它的任何两个状态是连通的，那么$\lim_{n \to \infty} P_{ij}^n$limn→∞Pijn与i无关，我们有：

1）

$$\lim_{n \to \infty} P_{ij}^n = \pi(j)$$

limn→∞Pijn=π(j)

2)

$$\lim_{n \to \infty} P^n = \begin{pmatrix} \pi(1) & \pi(2) & \ldots & \pi(j) & \ldots \\ \pi(1) & \pi(2) & \ldots & \pi(j) & \ldots \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ \pi(1) & \pi(2) & \ldots & \pi(j) & \ldots \\ \ldots & \ldots & \ldots & \ldots & \ldots \end{pmatrix}$$

limn→∞Pn=(π(1)π(2)…π(j)…π(1)π(2)…π(j)……………π(1)π(2)…π(j)……………)

3）

$$\pi(j) = \sum_{i=0}^{\infty} \pi(i)P_{ij}$$

π(j)=∑i=0∞π(i)Pij

4) π是方程$\pi P = \pi$的唯一非负解，其中：

$$\pi = [\pi(1), \pi(2), \ldots, \pi(j), \ldots] \quad \sum_{i=0}^{\infty} \pi(i) = 1$$

π=[π(1),π(2),...,π(j),...]∑i=0∞π(i)=1

上面的性质中需要解释的有：

1）非周期的马尔科夫链：这个主要是指马尔科夫链的状态转化不是循环的，如果是循环的则永远不会收敛。幸运的是我们遇到的马尔科夫链一般都是非周期性的。用数学方式表述则是：对于任意某一状态i，d为集合$\{n \mid n \geq 1, P_{ii}^n > 0\}$ 的最大公约数，如果$d = 1$，则该状态为非周期的

2）任何两个状态是连通的：这个指的是从任意一个状态可以通过有限步到达其他的任意一个状态，不会出现条件概率一直为0导致不可达的情况。

3）马尔科夫链的状态数可以是有限的，也可以是无限的。因此可以用于连续概率分布和离散概率分布。、

4）π通常称为马尔科夫链的平稳分布。

3. 基于马尔科夫链采样

如果我们得到了某个平稳分布所对应的马尔科夫链状态转移矩阵，我们就很容易采用出这个平稳分布的样本集。

假设我们任意初始的概率分布是$\pi_0(x)$, 经过第一轮马尔科夫链状态转移后的概率分布是$\pi_1(x)$，。。。第i轮的概率分布是$\pi_i(x)$。假设经过n轮后马尔科夫链收敛到我们的平稳分布$\pi(x)$，即：

$$\pi_n(x) = \pi_{n+1}(x) = \pi_{n+2}(x) = \ldots = \pi(x)$$

πn(x)=πn+1(x)=πn+2(x)=...=π(x)

对于每个分布$\pi_i(x)$，我们有：

$$\pi_i(x) = \pi_{i-1}(x)P = \pi_{i-2}(x)P^2 = \pi_0(x)P^i$$

πi(x)=πi−1(x)P=πi−2(x)P2=π0(x)Pi

现在我们可以开始采样了，首先，基于初始任意简单概率分布比如高斯分布$\pi_0(x)$采样得到状态值x_0，基于条件概率分布$P(x|x_0)$采样状态值x_1，一直进行下去，当状态转移进行到一定的次数时，比如到n次时，我们认为此时的采样集$(x_n, x_{n+1}, x_{n+2}, \ldots)$即是符合我们的平稳分布的对应样本集，可以用来做蒙特卡罗模拟求和了。

总结下基于马尔科夫链的采样过程：

1）输入马尔科夫链状态转移矩阵P，设定状态转移次数阈值n_1，需要的样本个数n_2

2）从任意简单概率分布采样得到初始状态值x_0

3）for $t = 0$ to $n_1 + n_2 - 1$: 从条件概率分布$P(x|x_t)$中采样得到样本x_{t+1}

样本集$(x_{n_1}, x_{n_1+1}, ..., x_{n_1+n_2-1})$即为我们需要的平稳分布对应的样本集。

4. 马尔科夫链采样小结

如果假定我们可以得到我们需要采样样本的平稳分布所对应的马尔科夫链状态转移矩阵，那么我们就可以用马尔科夫链采样得到我们需要的样本集，进而进行蒙特卡罗模拟。但是一个重要的问题是，随意给定一个平稳分布π,如何得到它所对应的马尔科夫链状态转移矩阵P呢？这是个大问题。我们绕了一圈似乎还是没有解决任意概率分布采样样本集的问题。

幸运的是，MCMC采样通过迂回的方式解决了上面这个大问题，我们在下一篇来讨论MCMC的采样，以及它的使用改进版采样: M-H采样和Gibbs采样.

（欢迎转载，转载请注明出处。欢迎沟通交流： liujianping-ok@163.com）

« 上一篇：MCMC(一)蒙特卡罗方法
» 下一篇：MCMC(三)MCMC采样和M-H采样

评论列表

#1楼 2017-11-03 11:48 JoeLee2017
该说什么好呢，写得简直不要太通俗易懂了。哈哈哈哈哈哈，谢谢您的付出
支持(7)反对(0)

#2楼 2018-05-19 12:11 turing袁
您的每一篇文章我都在研读中！ 感谢！ 大神我想请问一下 我不太理解这个采样 均匀分布采样我还能想到是怎么做的 那最后这里“从条件概率分布P(x|xt)中采样得到样本xt+1” 这是怎么做到的？？
支持(2)反对(0)

#3楼[楼主] 2018-05-19 13:37 刘建平Pinard
@ turing袁
你好，这里只是按概率采样一个样本而已。由于我们有状态转移矩阵P，所以每一次，在现有$x_t$xt情况下，我们可以得到待采样样本在各个可能取值的分布。比如有3个取值的时候，取值概率分别是0.1,0.2,0.7，那么采样是很容易的。
比如基于均匀采样，那么在[0,1]之间采样到小于等于0.1的则是第一个取值，以此类推。
支持(3)反对(0)

#4楼 2018-05-19 13:43 turing袁
@ 刘建平Pinard
感谢！
支持(0)反对(0)

#5楼 2018-06-27 09:34 yalesaleng
博主你好，请问第三小节中，由于概率是稳定分布的，那么是不是可以说平稳采样集(xn,xn+1,xn+2....)中的值全都是一样的，那这样的采样意义何在呢？
支持(0)反对(0)

#6楼[楼主] 2018-06-27 22:12 刘建平Pinard
@ yalesaleng
你好，平稳采样集(xn,xn+1,xn+2....)中的值符合我们需要采样的概率分布律，并不是说这些值都是一样的。如果平稳采样集都是一样的，那这个概率分布就是
$P(x = a) = 1, P(x \neq a) = 0$ P(x=a)=1,P(x≠a)=0
支持(0)反对(0)

#7楼 2018-06-29 18:12 renminghuang
老师您好：
如果是从条件概率分布P(x|xt)中采样得到样本xt+1，那为什么一定要等到n1轮（概率分布收敛于平稳分布）后才采用抽样样本呢？我直接从第1轮抽样并采用也一样啊（毕竟对于有限个抽样取值，下一轮的抽样概率分布即是状态转移矩阵的某一行，且保持不变），完全没必要等到n1轮。
如果是从n1轮后才采用抽样样本，那每轮抽样应该是根据pi（x）才对，这样才满足n1轮后pi（x）的概率分布收敛于平稳分布，后续抽样值才可用。
支持(0)反对(0)

#8楼[楼主] 2018-07-01 22:09 刘建平Pinard
@ renminghuang
你好，第一轮抽样的$\pi_1(x)$π1(x)和最后稳定的$\pi(x)$π(x)是不一致的。
你可以看我第二节举的第一个例子。如果你使用开始的股市的概率分布为：[0.3,0.4,0.3]，那么采样得到的数据大概率是3:4:3的分布，不是我们真正趋于稳定时候的分布[0.625 0.3125 0.0625]， 而我们期望采样的分布是从[0.625 0.3125 0.0625]里面得到的，即采样得到的样本的比例大概符合0.625: 0.3125 :0.0625
支持(0)反对(0)

#9楼 2018-07-02 10:36 renminghuang
@ 刘建平Pinard
老师您好：
可能我没表述清楚意思。
以股市为例，如您所说，我们的目标是当概率分布区域稳定时抽样并采用，如[0.625 0.3125

0.0625]，即pi(x)
但在下述基于马尔科夫链采样过程中，
3）for t=0 to n1+n2−1: 从条件概率分布P(x|xt)中采样得到样本xt+1
抽样是基于条件概率分布P(x|xt)的，在股市的例子中，抽样根据xt的取值，按照[0.9 0.15 0.25]，或[0.075 0.8 0.25]，或[0.025 0.05 0.5]分布进行。
我的困惑是，此时的抽样方法，或者说抽样所依据的概率分布，和pi(x)不一致，为什么不直接根据pi(x)抽样，而是根据P(x|xt)抽样?
支持(0)反对(0)

#10楼[楼主] 2018-07-03 10:56 刘建平Pinard
@ renminghuang
你好，理解你的意思了。这里的关键是很多时候我们无法直接对pi(x)抽样，这时候需要MCMC这样的方法来得到采样样本。

对于上面那个股市的例子来说，的确，我们根本不需要那么麻烦去MCMC采样，直接采样就搞定了。这里这么做主要是为了用一个简单的例子理解MCMC采样的流程。
支持(0)反对(0)

#11楼 2018-07-04 21:00 爱学习的三日
老师，你好！在马尔科夫链的收敛性质（3）中，该式不是全概率公式吗?难道不是对于任何迭代步都满足而不是仅当收敛的时候才满足吗?
支持(0)反对(0)

#12楼 2018-07-04 22:04 爱学习的三日
@ 刘建平Pinard
老师，你好！在马尔科夫链的收敛性质（3）中，该式不是全概率公式吗?难道不是对于任何迭代步都满足而不是仅当收敛的时候才满足吗?
支持(0)反对(0)

#13楼[楼主] 2018-07-05 11:36 刘建平Pinard
@ 爱学习的三日
你好，这里不是全概率公式，注意这里做的是状态转移概率，而不是概率分布概率。
"难道不是对于任何迭代步都满足而不是仅当收敛的时候才满足吗？"： 只有收敛的时候乘以状态转移矩阵概率分布不变，其他时候乘了以后是会变的。
支持(0)反对(0)

#14楼 2018-07-06 19:18 爱学习的三日
@ 刘建平Pinard
谢谢老师！我所理解的马尔科夫链状态转移矩阵Pi,j是P(j|i)，也就是状态j在状态i下的条件慨率，请问这里是不能理解为条件概率吗?
支持(0)反对(0)

#15楼[楼主] 2018-07-08 22:20 刘建平Pinard
@ 爱学习的三日
你好，的确是状态本身转移的条件概率。和全概率公式里的还是稍有不同。全概率公式里的条件概率一般指的是两种不同事件之间条件概率。
支持(0)反对(0)

#16楼 2018-08-09 00:30 一直在用心

博主好，感谢您写的所有东西，我一路看过来的。
这里想说的是，其实这个状态矩阵P这样一直乘，最后收敛到一个向量，这是数值分析课程里的使用幂法求主特征值和对应特征向量，和这里是相通的。
支持(0)反对(0)

#17楼[楼主] 2018-08-09 11:04 刘建平Pinard
@ 一直在用心
你好，幂法求矩阵特征值的思路的确和这样非常类似的。感谢指出。
支持(0)反对(0)

#18楼 2018-08-29 13:55 飞扬扬
@刘建平Pinard
博主你好！有个问题想请教一下。复制一下马尔科夫链的采样过程：
1）输入马尔科夫链状态转移矩阵P，设定状态转移次数阈值n1，需要的样本个数n2
2）从任意简单概率分布采样得到初始状态值x0
3）for t=0 to n1+n2?1: 从条件概率分布P(x|xt)中采样得到样本xt+1，样本集(xn1,xn1+1,...,xn1+n2?1)即为我们需要的平稳分布对应的样本集。

第二步的简单概率分布只是用于生成初始状态值，并未作为输入直接参与到链的后续生成，而且下一状态的采样只取决于当前状态和对应的条件概率分布，而转移矩阵对初始概率分布的一步步改变直到收敛这个过程，并没有直接的体现。请问这个过程是如何实现的？一是在只输入x0的情况下链如何判断用的什么初始分布？二是整个收敛过程难道是一个和链的生成同步进行的隐藏过程吗？
支持(0)反对(0)

#19楼[楼主] 2018-08-30 10:24 刘建平Pinard
@ 飞扬扬
你好，这个是否收敛只与状态转移矩阵P有关，只要P是确定的，那么最后收敛的概率分布是确定的，那么采样的分布也是确定的。

所以在初始化的时候，x_0x0来自于什么分布我们不关心。因为这个初始分布不影响最终的收敛。

请问这个过程是如何实现的？----->请参考第二节的那个python程序实例过程。

一是在只输入x0的情况下链如何判断用的什么初始分布？---->理论上没有要求，均匀分布都行。初始分布影响的只是收敛的速度。

二是整个收敛过程难道是一个和链的生成同步进行的隐藏过程吗？---> 是的，随着链的生成，到某一个节点，采样分布就已经收敛了。
支持(0)反对(0)

#20楼 2018-09-16 21:02 ichengs
老师您好！
我想问一下，按文中说的，Xn1状态有可能回到X0状态，当第n1个状态又回到了第0个状态了，那岂不是前面0到n1次的跳转白做了，这样一来为何不直接从第0个状态开始采样呢
支持(0)反对(0)

#21楼 2018-09-16 21:05 ichengs
希望老师能用python写个采样的例子，将过程简单的模拟一遍，这样更能理解，谢谢！
支持(0)反对(0)

#22楼[楼主] 2018-09-17 11:09 刘建平Pinard
@ ichengs
你好，我们采样的是一个数据分布，即数据整体的分布情况，而不会特意关注于分布中某一个数据的取值。因此就算重新出现了x_0x0,也不能说前面0到n1次的跳转白做了。

另一个角度，我们容易给一个二维的数据（0.1,0.3），我们根本无法知道它是从一个什么分布中采样得到的。可能是均匀分布，正态分布，也可以是其他分布。

采样的例子其实依照我第二节的这个矩阵P，就可以很容易写出来，有机会我写一个，你也可以试一试，并不难，参考我第三节的算法流程即可。
支持(0)反对(0)

#23楼 2018-09-18 12:02 南鱼知刃
@ 刘建平Pinard
@JoeLee2017
您好，应该是根据P^n(x|xt)采样，而不是P(x|xt)吧
支持(0)反对(0)

#24楼[楼主] 2018-09-18 14:29 刘建平Pinard
@ 南鱼知刃
你好，采样还是基于$P(x|x_t)$ P(x|xt)的，而不是$P^n(x|x_t)$ Pn(x|xt)。使用的是性质3.
上面出现的$P^n(x|x_t)$ Pn(x|xt)主要是为了描述马尔科夫链的状态转移矩阵的性质1。实际采样并未使用。
支持(0)反对(0)

#25楼 2018-09-18 14:33 南鱼知刃
@ 刘建平Pinard
您好，这样的话，从您的算法描述中没看到P的更新啊，那该如何收敛到平稳分布呢
支持(0)反对(0)

#26楼 2018-09-18 14:45 南鱼知刃
@ 刘建平Pinard
还是说，算法中每对$P(x|x_t)$ P(x|xt)采样一次就意味着模拟了一次
$\pi_i(x) = \pi_{i-1}(x)P$ πi(x)=πi−1(x)P操作
支持(0)反对(0)

#27楼[楼主] 2018-09-19 10:09 刘建平Pinard
@ 南鱼知刃
你好，你说的对，每次采样就模型了一次状态分布ππ转化的过程。在此过程中，状态转化矩阵P没有变化。
支持(0)反对(0)
刷新评论刷新页面返回顶部
注册用户登录后才能发表评论，请 登录 或 注册，访问网站首页。

MCMC(三)MCMC采样和M-H采样

在MCMC(二)马尔科夫链中我们讲到给定一个概率平稳分布π, 很难直接找到对应的马尔科夫链状态转移矩阵P。而只要解决这个问题，我们就可以找到一种通用的概率分布采样方法，进而用于蒙特卡罗模拟。本篇我们就讨论解决这个问题的办法：MCMC采样和它的易用版M-H采样。

1. 马尔科夫链的细致平稳条件

在解决从平稳分布π, 找到对应的马尔科夫链状态转移矩阵P之前，我们还需要先看看马尔科夫链的细致平稳条件。定义如下：

如果非周期马尔科夫链的状态转移矩阵P和概率分布$\pi(x)$对于所有的i,j满足：

$$\pi(i)P(i,j) = \pi(j)P(j,i)$$

π(i)P(i,j)=π(j)P(j,i)

则称概率分布$\pi(x)$是状态转移矩阵P的平稳分布。

证明很简单,由细致平稳条件有：

$$\sum_{i=1}^{\infty}\pi(i)P(i,j) = \sum_{i=1}^{\infty}\pi(j)P(j,i) = \pi(j)\sum_{i=1}^{\infty}P(j,i) = \pi(j)$$

∑i=1∞π(i)P(i,j)=∑i=1∞π(j)P(j,i)=π(j)∑i=1∞P(j,i)=π(j)

将上式用矩阵表示即为：

$$\pi P = \pi$$

πP=π

即满足马尔可夫链的收敛性质。也就是说，只要我们找到了可以使概率分布$\pi(x)$满足细致平稳分布的矩阵P即可。这给了我们寻找从平稳分布π, 找到对应的马尔科夫链状态转移矩阵P的新思路。

不过不幸的是，仅仅从细致平稳条件还是很难找到合适的矩阵P。比如我们的目标平稳分布是$\pi(x)$,随机找一个马尔科夫链状态转移矩阵Q,它是很难满足细致平稳条件的，即：

$$\pi(i)Q(i,j) \neq \pi(j)Q(j,i)$$

π(i)Q(i,j)≠π(j)Q(j,i)

那么如何使这个等式满足呢？下面我们来看MCMC采样如何解决这个问题。

2. MCMC采样

由于一般情况下，目标平稳分布$\pi(x)$和某一个马尔科夫链状态转移矩阵Q不满足细致平稳条件，即

$$\pi(i)Q(i,j) \neq \pi(j)Q(j,i)$$
π(i)Q(i,j)≠π(j)Q(j,i)

我们可以对上式做一个改造，使细致平稳条件成立。方法是引入一个$\alpha(i,j)$,使上式可以取等号，即：

$$\pi(i)Q(i,j)\alpha(i,j) = \pi(j)Q(j,i)\alpha(j,i)$$
π(i)Q(i,j)α(i,j)=π(j)Q(j,i)α(j,i)

问题是什么样的$\alpha(i,j)$可以使等式成立呢？其实很简单，只要满足下两式即可：

$$\alpha(i,j) = \pi(j)Q(j,i)$$
α(i,j)=π(j)Q(j,i)
$$\alpha(j,i) = \pi(i)Q(i,j)$$
α(j,i)=π(i)Q(i,j)

这样，我们就得到了我们的分布$\pi(x)$对应的马尔科夫链状态转移矩阵P，满足：

$$P(i,j) = Q(i,j)\alpha(i,j)$$
P(i,j)=Q(i,j)α(i,j)

也就是说，我们的目标矩阵P可以通过任意一个马尔科夫链状态转移矩阵Q乘以$\alpha(i,j)$得到。$\alpha(i,j)$我们有一般称之为接受率。取值在[0,1]之间，可以理解为一个概率值。即目标矩阵P可以通过任意一个马尔科夫链状态转移矩阵Q以一定的接受率获得。这个很像我们在MCMC(一)蒙特卡罗方法第4节讲到的接受-拒绝采样，那里是以一个常用分布通过一定的接受-拒绝概率得到一个非常见分布，这里是以一个常见的马尔科夫链状态转移矩阵Q通过一定的接受-拒绝概率得到目标转移矩阵P,两者的解决问题思路是类似的。

好了，现在我们来总结下MCMC的采样过程。

1）输入我们任意选定的马尔科夫链状态转移矩阵Q，平稳分布$\pi(x)$，设定状态转移次数阈值n_1，需要的样本个数n_2

2）从任意简单概率分布采样得到初始状态值x_0

3）for $t = 0$ to $n_1 + n_2 - 1$:

a) 从条件概率分布$Q(x|x_t)$中采样得到样本x_*

b) 从均匀分布采样$u \sim uniform[0,1]$

c) 如果$u < \alpha(x_t, x_*) = \pi(x_*)Q(x_*, x_t)$, 则接受转移$x_t \rightarrow x_*$，即$x_{t+1} = x_*$

d) 否则不接受转移，$t = max(t-1, 0)$

样本集$(x_{n_1}, x_{n_1+1}, ..., x_{n_1+n_2-1})$即为我们需要的平稳分布对应的样本集。

上面这个过程基本上就是MCMC采样的完整采样理论了，但是这个采样算法还是比较难在实际中应用，为什么呢？问题在上面第三步的c步骤，接受率这儿。由于$\alpha(x_t, x_*)$可能非常的小，比如0.1，导致我们大部分的采样值都被拒绝转移，采样效率很低。有可能我们采样了上百万次马尔可夫链还没有收敛，也就是上面这个n_1要非常非常的大，这让人难以接受，怎么办呢？这时就轮到我们的M-H采样出场了。

3. M-H采样

M-H采样是Metropolis-Hastings采样的简称，这个算法首先由Metropolis提出，被Hastings改进，因此被称之为Metropolis-Hastings采样或M-H采样

M-H采样解决了我们上一节MCMC采样接受率过低的问题。

我们回到MCMC采样的细致平稳条件：

$$\pi(i)Q(i,j)\alpha(i,j) = \pi(j)Q(j,i)\alpha(j,i)$$

π(i)Q(i,j)α(i,j)=π(j)Q(j,i)α(j,i)

我们采样效率低的原因是$\alpha(i,j)$太小了，比如为0.1，而$\alpha(j,i)$为0.2。即：

$$\pi(i)Q(i,j) \times 0.1 = \pi(j)Q(j,i) \times 0.2$$

π(i)Q(i,j)×0.1=π(j)Q(j,i)×0.2

这时我们可以看到，如果两边同时扩大五倍，接受率提高到了0.5，但是细致平稳条件却仍然是满足的，即：

$$\pi(i)Q(i,j) \times 0.5 = \pi(j)Q(j,i) \times 1$$

π(i)Q(i,j)×0.5=π(j)Q(j,i)×1

这样我们的接受率可以做如下改进，即：

$$\alpha(i,j) = \min\{\frac{\pi(j)Q(j,i)}{\pi(i)Q(i,j)}, 1\}$$

α(i,j)=min{π(j)Q(j,i)π(i)Q(i,j),1}

通过这个微小的改造，我们就得到了可以在实际应用中使用的M-H采样算法过程如下：

1）输入我们任意选定的马尔科夫链状态转移矩阵Q，平稳分布$\pi(x)$，设定状态转移次数阈值n_1，需要的样本个数n_2

2）从任意简单概率分布采样得到初始状态值x_0

3）for $t = 0$ to $n_1 + n_2 - 1$:

a) 从条件概率分布$Q(x|x_t)$中采样得到样本x_*

b) 从均匀分布采样$u \sim uniform[0,1]$

c) 如果$u < \alpha(x_t, x_*) = \min\{\frac{\pi(j)Q(j,i)}{\pi(i)Q(i,j)}, 1\}$ u<α(xt,x∗)=min{π(j)Q(j,i)π(i)Q(i,j),1} , 则接受转移$x_t \rightarrow x_*$ xt→x∗ ，即$x_{t+1} = x_*$ xt+1=x∗

d) 否则不接受转移，$t = \max(t-1, 0)$ t=max(t−1,0)

样本集$(x_{n_1}, x_{n_1+1}, ..., x_{n_1+n_2-1})$ (xn1,xn1+1,...,xn1+n2−1)即为我们需要的平稳分布对应的样本集。

很多时候，我们选择的马尔科夫链状态转移矩阵Q Q如果是对称的，即满足$Q(i,j) = Q(j,i)$ Q(i,j)=Q(j,i),这时我们的接受率可以进一步简化为：

$$\alpha(i,j) = \min\{\frac{\pi(j)}{\pi(i)}, 1\}$$

α(i,j)=min{π(j)π(i),1}

4. M-H采样实例

为了更容易理解，这里给出一个M-H采样的实例。

完整代码参见我的github: https://github.com/ljpzzz/machinelearning/blob/master/mathematics/mcmc_3_4.ipynb

在例子里，我们的目标平稳分布是一个均值3，标准差2的正态分布，而选择的马尔可夫链状态转移矩阵$Q(i,j)$ Q(i,j)的条件转移概率是以i i为均值,方差1的正态分布在位置j j的值。这个例子仅仅用来让大家加深对M-H采样过程的理解。毕竟一个普通的一维正态分布用不着去用M-H采样来获得样本。

代码如下：

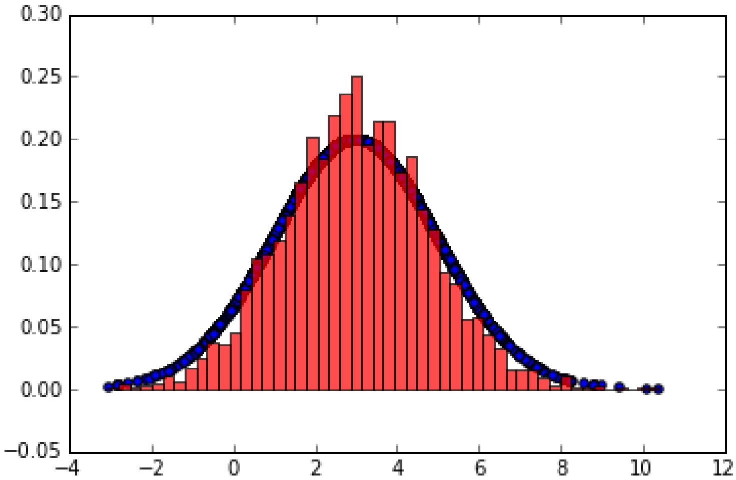
复制代码

```
import random
import math
from scipy.stats import norm
import matplotlib.pyplot as plt
%matplotlib inline

def norm_dist_prob(theta):
    y = norm.pdf(theta, loc=3, scale=2)
    return y

T = 5000
pi = [0 for i in range(T)]
sigma = 1
t = 0
while t < T-1:
    t = t + 1
    pi_star = norm.rvs(loc=pi[t - 1], scale=sigma, size=1, random_state=None)
    alpha = min(1, (norm_dist_prob(pi_star[0]) / norm_dist_prob(pi[t - 1])))

    u = random.uniform(0, 1)
    if u < alpha:
        pi[t] = pi_star[0]
    else:
        pi[t] = pi[t - 1]
```

```
plt.scatter(pi, norm.pdf(pi, loc=3, scale=2))
num_bins = 50
plt.hist(pi, num_bins, normed=1, facecolor='red', alpha=0.7)
plt.show()
```

[复制代码]

输出的图中可以看到采样值的分布与真实的分布之间的关系如下，采样集还是比较拟合对应分布的。



5. M-H采样总结

M-H采样完整解决了使用蒙特卡罗方法需要的任意概率分布样本集的问题，因此在实际生产环境得到了广泛的应用。

但是在大数据时代，M-H采样面临着两大难题：

1）我们的数据特征非常的多，M-H采样由于接受率计算式 $\frac{\pi(j)Q(j,i)}{\pi(i)Q(i,j)}$ π(j)Q(j,i)π(i)Q(i,j)的存在，在高维时需要的计算时间非常的可观，算法效率很低。同时 $\alpha(i,j)$ α(i,j)一般小于1，有时候辛苦计算出来却被拒绝了。能不能做到不拒绝转移呢？

2）由于特征维度大，很多时候我们甚至很难求出目标的各特征维度联合分布，但是可以方便求出各个特征之间的条件概率分布。这时候我们能不能只有各维度之间条件概率分布的情况下方便的采样呢？

Gibbs采样解决了上面两个问题，因此在大数据时代，MCMC采样基本是Gibbs采样的天下，下一篇我们就来讨论Gibbs采样。

（欢迎转载，转载请注明出处。欢迎沟通交流： liujianping-ok@163.com）

分类: [0040. 数学统计学]
标签: [机器学习中的数学]
[好文要顶] [关注我] [收藏该文]
[刘建平Pinard]
[关注 - 14]

« 上一篇：MCMC(二)马尔科夫链
» 下一篇：MCMC(四)Gibbs采样

评论列表

#1楼 2017-08-03 18:20 pmonkeychen
您好，请问第二第三章在叙述算法过程时，3)中的a)里应该都是从Q中生成样本而不是P吧？
支持(0)反对(0)

#2楼[楼主] 2017-08-03 18:24 刘建平Pinard
@ pmonkeychen
你好，这儿的确写错了，已改正。感谢指出错误。
支持(0)反对(0)

#3楼 2017-09-10 18:22 DTStor
你好，请问代码
pi_star = norm.rvs(loc=theta[t - 1], scale=sigma, size=1, random_state=None)
中的theta在哪儿定义的？谢谢
支持(0)反对(0)

#4楼 2017-09-10 19:20 DTStor
知道了，
pi_star = norm.rvs(loc=theta[t - 1], scale=sigma, size=1, random_state=None)
中的theta应该是pi
支持(0)反对(0)

#5楼[楼主] 2017-09-11 10:30 刘建平Pinard
@ DTStor
你好，这里的确如你所说，应该是pi,已修改，感谢指出错误。
支持(0)反对(0)

#6楼 2017-10-03 00:00 dethrive
请问，第2和第三部分的算法中“d) 否则不接受转移，t = t - 1”，如果 t = 0 时候，算法就直接拒绝了，那么 t = -1，后面的算法还能进行下去吗？或者说这里的 t 是不是要理解为第 t 次有效的样本采集啊？
支持(0)反对(0)

#7楼 2017-10-03 00:16 dethrive
@ 刘建平Pinard
还有应该是均值为3，标准差为2
支持(0)反对(0)

#8楼[楼主] 2017-10-09 11:06 刘建平Pinard
@ dethrive
你好，这里我写的的确有点问题。改成了这个样子，避免初始拒绝的问题。

$t = \max(t-1, 0)$t=max(t−1,0)
那个均值和标准差的笔误也改了。
感谢指出这么多的错误。
支持(0)反对(0)

#9楼 2017-10-31 10:42 王小熊
你好，我想请问程序中在求alpha的时候，为什么没有乘以pi，公式里是有的
支持(0)反对(0)

#10楼[楼主] 2017-10-31 11:23 刘建平Pinard
@ 王小熊
你好，这里采用的是这个公式计算的αα

$$\alpha(i,j) = \min\{\frac{\pi(j)}{\pi(i)}, 1\}$$

α(i,j)=min{π(j)π(i),1}

即假定了$Q(i,j) = Q(j,i)$ Q(i,j)=Q(j,i)

当然我上面的例子里面Q并不对称。这个例子实际使用的是最早的Metropolis采样的公式，这个公式并不管Q对称与否，直接用了上式。

你可以修改下我的程序，按照我上面的M-H采样公式更新αα，再画图看看结果。：)
支持(0)反对(0)

#11楼 2017-10-31 11:24 王小熊
@ 刘建平Pinard
谢谢!
支持(0)反对(0)

#12楼 2017-12-12 16:34 爱吃香菜的JQ
你好，算法 3）：d）否则不接受转移，t=max(t−1,0) 这个是拒绝本次采样，那程序中为什么是else: pi[t] = pi[t - 1]?
支持(1)反对(0)

#13楼[楼主] 2017-12-12 17:17 刘建平Pinard
@ 爱吃香菜的JQ
你好，这里不接受转移的话，t就是t-1，取max(t−1,0)是为了避免在t=0的时候的t变成负数。
不接受转移的话，t就是t-1，那么pi[t] = pi[t - 1]
支持(0)反对(0)

#14楼 2017-12-12 18:03 爱吃香菜的JQ
@ 刘建平Pinard
我认为t=t-1不等于pi[t] = pi[t - 1]。t=t-1意味者拒绝这次采样，pi[t] = pi[t - 1]意味着 连续两个pi是同一个值，就如下：
序号 类型 数值
23 float64 4.63203414485
24 float64 4.63203414485
支持(0)反对(0)

#15楼[楼主] 2017-12-13 10:48 刘建平Pinard

@ 爱吃香菜的JQ
你好，这里你说的的确有道理，这里算法和下面的例子上逻辑稍有不同。

在算法里，维度目标是得到n_1n1个真正的转移后，后面的n_2n2个样本就是我们的需要的数据，所以我使用了t=t-1, 这样被拒绝的样本就不进入采样样本序列。

而在下面的例子里面，就算不是真正的转移（被决绝），我们也把它放入了样本序列，所以有些mismatch。

不过这个不同对算法的理解影响并不大。
支持(0)反对(0)

#16楼 2018-02-10 23:31 扼杀
α(i,j)α(i,j)我们有一般称之为接受率。取值在[0,1]之间，可以理解为一个概率值。 π(j)*Q(j,i)这个可以得到一个数，而不是矩阵吗?
支持(0)反对(0)

#17楼[楼主] 2018-02-11 13:50 刘建平Pinard
@ 扼杀
你好，QQ是一个矩阵，而$Q(j,i)$ Q(j,i)是矩阵对应位置的一个标量，所以$\pi(j)*Q(j,i)$ π(j)∗Q(j,i)可以得到一个数，而$\pi(j)*Q$ π(j)∗Q可以得到一个一个矩阵。
支持(2)反对(0)

#18楼 2018-02-22 22:41 扼杀
@ 刘建平Pinard
谢谢博主大神
支持(0)反对(0)

#19楼 2018-03-12 21:26 泡泡糖nana
谢谢博主大神，对我很有帮助!
支持(0)反对(0)

#20楼 2018-04-30 09:14 kylin0228
博主您好，想请问一下细致平稳条件一般不成立是不是说：这是pai(x)的充分不必要条件? ？就是说满足这个条件一定是平稳分布了，但是不满足也不代表不是?
支持(0)反对(0)

#21楼[楼主] 2018-05-01 00:01 刘建平Pinard
@ kylin0228
我说的应该不是这个意思。
满足细致平稳条件，那么这作为一个充分条件，可以推导出$\pi(x)$ π(x)。

细致平稳条件一般不成立，这是因为这个矩阵很难直接找出来。也就是说，我随便给你一个概率分布，比如[0.3,0.5,0.2](只是举个例子)，你很难给出它所对应的满足细致平稳条件的转移矩阵。
支持(0)反对(0)

#22楼 2018-05-01 08:35 kylin0228
博主您好
后来我查阅了一些资料
细致平稳条件是一个充分不必要条件

即不满足细致平稳条件不一定不是平稳分布
参考链接：https://www.zhihu.com/question/37000071/answer/131542485
支持(0)反对(0)

#23楼[楼主] 2018-05-01 23:09 刘建平Pinard
@ kylin0228
你好，细致平稳条件是一个充分不必要条件，我21楼也说了"这作为一个充分条件"。

只是"细致平稳条件一般不成立",不是因为它是一个充分不必要条件。而是这样的矩阵难以直接找出。
支持(0)反对(0)

#24楼 2018-05-02 11:28 kylin0228
嗯嗯 get到博主的意思了
谢谢博主大神!
支持(0)反对(0)

#25楼 2018-05-19 09:34 蜉蝣2015
@ 刘建平Pinard
从您的代码上来看，我是这样理解的 1）存储状态序列里面的值均为接收值，只是有发生重复的时候 2）这里这个序列没有体现出来 超过 n1阈值的地方。
支持(0)反对(0)

#26楼[楼主] 2018-05-19 12:09 刘建平Pinard
@ 蜉蝣2015
1）是这样的。
2） 在实际工程应用中，我们一般不判断阈值，都是默认采样到某一个数量，后面认为就是我们需要的样本。如果计算量大的时候，这个“数量阈值”有时候都不管了。一般问题不大。
支持(1)反对(0)

#27楼 2018-05-19 19:37 蜉蝣2015
还是有些不明白。1）根据文中所述，MCMC和MH两种方法中的detail balance condition，对所有的用分布 Q生成的样本点都成立，这点您同意么?
2）为什么通过 比较 u和接收率a的值，就能确定Q 生成的样本点，也可以作为P的样本点？除了从 rejection sampling 方法中，感性的比较去理解，有啥具体的证明或者定量的分析不?
支持(0)反对(0)

#28楼[楼主] 2018-05-19 22:53 刘建平Pinard
@ 蜉蝣2015
你好，1） 是的，通过一定的拒绝率，我们用Q模拟得到了满足细致平稳条件的分布P。
2） 如果你对MCMC进一步的数学原理和演进感兴趣，推荐你读这篇文章：
http://stat.wharton.upenn.edu/~stjensen/stat542/lecture14.mcmchistory.pdf
支持(1)反对(0)

#29楼 2018-05-20 09:08 蜉蝣2015
@ 刘建平Pinard
好的 谢谢。
支持(0)反对(0)

#30楼 2018-06-22 11:40 忆梦涟

请问，之前的MCMC和M-H算法是基于联合概率吗，看起来中间用到的都是条件概率，有些不明白，谢谢！
支持(0)反对(0)

#31楼[楼主] 2018-06-22 23:27 刘建平Pinard
@ 忆梦涟
你好，这里是条件概率。
你可以理解为有一个概率分布函数P(A)，当前有一个点在分布的位置A1，下一个点采样得到分布位置A2的概率。也就是P(A2|A1)。这个P(A2|A1)是分布P(A)中对应的两点之间的条件概率。你可以理解为是一个状态转移概率。
支持(1)反对(0)

#32楼 2018-06-29 16:34 renminghuang
老师，关于细致平稳条件，有个地方不懂。
pi(i)是随机变量状态为i时的概率，P(i,j)是随机变量从状态i转移至状态j的概率，它是一个条件概率，那么两者相乘即随机变量在状态i和状态j时的联合分布概率，那么细致平稳条件等式两侧应该是恒相等的，为什么会大多数情况下不相等，而是需要再乘以个接受率呢？
谢谢。

编辑：
是不是因为状态存有时序的原因？即时刻t状态i时刻t+1状态j的联合概率分布不一定等于时刻t状态j时刻t+1状态i的联合概率分布。
支持(0)反对(0)

#33楼[楼主] 2018-07-01 22:04 刘建平Pinard
@ renminghuang
你好，这里的原因应该是这个P矩阵很难寻找，我们使用的是另外一个Q矩阵，这个Q矩阵和P矩阵是不一样的，所以需要一个接受率来用Q慢慢拟合P。

如果我们直接拿到了P矩阵，那么的确就是恒等了。
支持(0)反对(0)

#34楼 2018-07-07 16:02 潇洒走着
有两个疑问希望博主能解答下：
1. 为什么要设置u来跟α比较，判断是否接受。
2. 代码中，在上一个的采样值作为均值，1为方差的高斯分布上做采用， 为什么用上一个采样的样本做当前采样的高斯分布均值？
支持(0)反对(0)

#35楼[楼主] 2018-07-08 22:25 刘建平Pinard
@ 潇洒走着
你好，第一个问题其实我第二节已经讲过了。你可以看看，简言之就是Q不是我们想要的P，通过一定的接收率来让Q近似于P。

第二个问题，这里其实是我们假设的，上面我也写了，是： “选择的马尔可夫链状态转移矩阵Q(i,j)的条件转移概率是以i为均值,方差1的正态分布在位置j的值”
支持(1)反对(0)

#36楼 2018-07-26 10:49 华师大小公主

@ 刘建平Pinard
博主您好，MH算法中，平稳分布π没有更新，转移概率q(i,j)也没有更新，相应的α(i,j)也不会改变，而且从开始采样的时候，由于加入了α就满足平稳分布了，为什么还需要采样到一定阈值之后的样本，才算是分布的真实样本？不应该从t=1次采样开始就是分布的真实样本吗?
支持(0)反对(0)

#37楼[楼主] 2018-07-26 11:02 刘建平Pinard
@ 华师大小公主
你好！加入了接受率后，的确可以使Q近似于P了。但是如果你看了上一届马尔科夫链后，你会发现，就算拿到了P，你也得采样到一定的数目，让采样分布收敛后才能得到你想要的样本。
支持(0)反对(0)

#38楼 2018-07-26 11:18 华师大小公主
博主，我去看了上节的，第三小节您确实提到采样n次后的样本才是平稳分布的对应样本集，但是似乎没有解释原因（可能没理解到）。

所以我还是有点不明白，为什么需要采样到一定数目后才是想要的？因为并没有看到哪个分布在更新，那么所谓采样分布收敛，是如何收敛的?

您指的采样分布是指π呢还是转移概率q？如果是π，在上面MH的算法中，π作为输入，是已知的；如果是q，算法中没有看到q的更新？所以，何来的采样分布收敛呢?
支持(0)反对(0)

#39楼[楼主] 2018-07-27 10:40 刘建平Pinard
@ 华师大小公主
你好！ 原因你可以看我上一篇的python代码，也可以跑一遍。采样n次后，我们的分布ππ就稳定了，这个稳定分布的ππ就是我们想要采样的分布。

具体收敛我没有证明，不过通过一个实际的例子和代码展示了。

采样的分布是ππ，这个ππ在乘以转换矩阵P后分布会变，当ππ稳定后就是采样分布收敛
支持(0)反对(0)

#40楼 2018-07-27 10:59 华师大小公主
@ 刘建平Pinard
π是平稳分布，按照定义πp=π，为什么π乘以p后π还会变呢?
支持(0)反对(0)

#41楼 2018-07-27 11:19 华师大小公主
@ 刘建平Pinard
看完LDA，我在想，是不是因为每次采样完，条件概率分布Q(x|xt)发生了改变，所以需要等到Q(x|xt)收敛后，采样才算有效?
支持(0)反对(0)

#42楼[楼主] 2018-07-27 11:35 刘建平Pinard
@ 华师大小公主
你好，最开始你采样使用的初始ππ是任意选择的，不是我们的平稳分布ππ。所以乘了以后会变。
条件概率分布Q(x|xt)的分布函数是不变的，但是每次$X_t$xt会变，这样采样得到的Xx也是会变的。不过根本原因还是你最开始使用的ππ不是平稳分布，所以采样无效。
支持(0)反对(0)

#43楼 2018-09-18 14:13 南鱼知刃
@ 刘建平Pinard
@ 华师大小公主
" 你好！加入了接受率后，的确可以使Q近似于P了。但是如果你看了上一届马尔科夫链后，你会发现，就算拿到了P，你也得采样到一定的数目，让采样分布收敛后才能得到你想要的样本。"

您好，上面是您回复这位网友的一段话，个人认为上一节马尔科夫链中的P一开始是没有收敛的（即一开始不满足细致平稳条件），所以需要采样多次（不过在上一节最后MH的算法描述中确实没看到P更新，很奇怪）；但这一节中的P一开始就是满足细致平稳条件的（只是不能直接得到P，而是用Q通过拒绝-接受来模拟P而已），所以在MCMC采样中应该一开始的采样样本就能用啊；
以上两点疑问，希望您有空的时候帮忙解答下，感激不尽！
支持(0)反对(0)

#44楼[楼主] 2018-09-18 14:33 刘建平Pinard
@ 南鱼知刃
你好！
首先，马尔科夫状态转移矩阵没有收敛一说，只有采样的概率分布有收敛一说。对应上一篇中的符号，就是采样的概率分布开始是$\pi(i)$ π(i),随着采样的进行，收敛到我们需要的采样分布$\pi(j)$ π(j),即：

$$\pi(j) = \sum_{i=0}^{\infty} \pi(i)P_{ij}$$

π(j)=∑i=0∞π(i)Pij

采样多次的原因是希望$\pi(i)$ π(i)慢慢向$\pi(j)$ π(j)靠拢。
支持(0)反对(0)

#45楼 2018-09-27 11:22 南鱼知刃
@ 刘建平Pinard
您好，在MCMC采样中，既然已知了任意转移矩阵Q和和平稳分别$\pi(x)$ π(x)，就可以计算得到目标矩阵P，那在第三步中为什么不直接根据$P(x|x_t)$ P(x|xt)采样，而要根据$Q(x|x_t)$ Q(x|xt)采样？
支持(0)反对(0)

#46楼[楼主] 2018-09-28 10:26 刘建平Pinard
@ 南鱼知刃
你好，目标矩阵P不是计算得到的，而是采样$Q(x|x_t)$ Q(x|xt),并通过接受拒绝这样的方式间接得到的。没法直接采样$P(x|x_t)$ P(x|xt)
支持(0)反对(0)
刷新评论刷新页面返回顶部
注册用户登录后才能发表评论，请 登录 或 注册，访问网站首页。

[MCMC(四)Gibbs采样]

[MCMC(一)蒙特卡罗方法]

[MCMC(二)马尔科夫链]

[MCMC(三)MCMC采样和M-H采样]

MCMC(四)Gibbs采样

在[MCMC(三)MCMC采样和M-H采样]中，我们讲到了M-H采样已经可以很好的解决蒙特卡罗方法需要的任意概率分布的样本集的问题。但是M-H采样有两个缺点：一是需要计算接受率，在高维时计算量大。并且由于接受率的原因导致算法收敛时间变长。二是有些高维数据，特征的条件概率分布好求，但是特征的联合分布不好求。因此需要一个好的方法来改进M-H采样，这就是我们下面讲到的Gibbs采样。

1. 重新寻找合适的细致平稳条件

在上一篇中，我们讲到了细致平稳条件：如果非周期马尔科夫链的状态转移矩阵P P和概率分布$\pi(x)$π(x)对于所有的i,j i,j满足：

$$\pi(i)P(i,j) = \pi(j)P(j,i)$$
π(i)P(i,j)=π(j)P(j,i)

则称概率分布$\pi(x)$π(x)是状态转移矩阵P P的平稳分布。

在M-H采样中我们通过引入接受率使细致平稳条件满足。现在我们换一个思路。

从二维的数据分布开始，假设$\pi(x_1,x_2)$π(x1,x2)是一个二维联合数据分布，观察第一个特征维度相同的两个点$A(x_1^{(1)},x_2^{(1)})$ A(x1(1),x2(1))和$B(x_1^{(1)},x_2^{(2)})$ B(x1(1),x2(2))，容易发现下面两式成立：

$$\pi(x_1^{(1)},x_2^{(1)})\pi(x_2^{(2)}|x_1^{(1)}) = \pi(x_1^{(1)})\pi(x_2^{(1)}|x_1^{(1)})\pi(x_2^{(2)}|x_1^{(1)})$$
π(x1(1),x2(1))π(x2(2)|x1(1))=π(x1(1))π(x2(1)|x1(1))π(x2(2)|x1(1))

$$\pi(x_1^{(1)},x_2^{(2)})\pi(x_2^{(1)}|x_1^{(1)}) = \pi(x_1^{(1)})\pi(x_2^{(2)}|x_1^{(1)})\pi(x_2^{(1)}|x_1^{(1)})$$
π(x1(1),x2(2))π(x2(1)|x1(1))=π(x1(1))π(x2(2)|x1(1))π(x2(1)|x1(1))

由于两式的右边相等，因此我们有：

$$\pi(x_1^{(1)},x_2^{(1)})\pi(x_2^{(2)}|x_1^{(1)}) = \pi(x_1^{(1)},x_2^{(2)})\pi(x_2^{(1)}|x_1^{(1)})$$
π(x1(1),x2(1))π(x2(2)|x1(1))=π(x1(1),x2(2))π(x2(1)|x1(1))

也就是：

$$\pi(A)\pi(x_2^{(2)}|x_1^{(1)}) = \pi(B)\pi(x_2^{(1)}|x_1^{(1)})$$
π(A)π(x2(2)|x1(1))=π(B)π(x2(1)|x1(1))

观察上式再观察细致平稳条件的公式，我们发现在$x_1 = x_1^{(1)}$ x1=x1(1)这条直线上，如果用条件概率分布$\pi(x_2 \mid x_1^{(1)})$ π(x2|x1(1))作为马尔科夫链的状态转移概率，则任意两个点之间的转移满足细致平稳条件！这真是一个开心的发现，同样的道理，在在$x_2 = x_2^{(1)}$ x2=x2(1)这条直线上，如果用条件概率分布$\pi(x_1 \mid x_2^{(1)})$ π(x1|x2(1))作为马尔科夫链的状态转移概率，则任意两个点之间的转移也满足细致平稳条件。那是因为假如有一点$C(x_1^{(2)},x_2^{(1)})$ C(x1(2),x2(1)),我们可以得到：

$$\pi(A)\pi(x_1^{(2)}|x_2^{(1)}) = \pi(C)\pi(x_1^{(1)}|x_2^{(1)})$$
π(A)π(x1(2)|x2(1))=π(C)π(x1(1)|x2(1))

基于上面的发现，我们可以这样构造分布$\pi(x_1, x_2)$π(x1,x2)的马尔可夫链对应的状态转移矩阵P P：

$$P(A \to B) = \pi(x_2^{(B)}|x_1^{(1)}) \;\; if \; x_1^{(A)} = x_1^{(B)} = x_1^{(1)}$$
P(A→B)=π(x2(B)|x1(1))ifx1(A)=x1(B)=x1(1)

$$P(A \to C) = \pi(x_1^{(C)}|x_2^{(1)}) \;\; if \; x_2^{(A)} = x_2^{(C)} = x_2^{(1)}$$
P(A→C)=π(x1(C)|x2(1))ifx2(A)=x2(C)=x2(1)

$$P(A \to D) = 0 \;\; else$$
P(A→D)=0else

有了上面这个状态转移矩阵，我们很容易验证平面上的任意两点E, F E,F，满足细致平稳条件：

$$\pi(E)P(E \to F) = \pi(F)P(F \to E)$$
π(E)P(E→F)=π(F)P(F→E)

2. 二维Gibbs采样

利用上一节找到的状态转移矩阵，我们就得到了二维Gibbs采样，这个采样需要两个维度之间的条件概率。具体过程如下：

1）输入平稳分布$\pi(x_1, x_2)$π(x1,x2)，设定状态转移次数阈值n_1n1，需要的样本个数n_2n2

2）随机初始化初始状态值$x_1^{(0)}$x1(0)和$x_2^{(0)}$x2(0)

3）for $t = 0$t=0 to $n_1 + n_2 - 1$n1+n2−1:

a) 从条件概率分布$P(x_2|x_1^{(t)})$ P(x2|x1(t))中采样得到样本x_2^{t+1}x2t+1

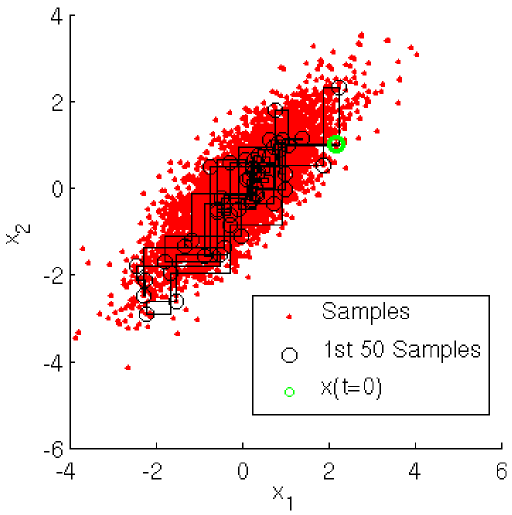
b) 从条件概率分布$P(x_1|x_2^{(t+1)})$ P(x1|x2(t+1))中采样得到样本x_1^{t+1}x1t+1

样本集$\{(x_1^{(n_1)}, x_2^{(n_1)}), (x_1^{(n_1+1)}, x_2^{(n_1+1)}), \ldots, (x_1^{(n_1+n_2-1)}, x_2^{(n_1+n_2-1)})\}$ {(x1(n1),x2(n1)),(x1(n1+1),x2(n1+1)),...,(x1(n1+n2−1),x2(n1+n2−1))}即为我们需要的平稳分布对应的样本集。

整个采样过程中，我们通过轮换坐标轴，采样的过程为：

$$(x_1^{(1)}, x_2^{(1)}) \to (x_1^{(1)}, x_2^{(2)}) \to (x_1^{(2)}, x_2^{(2)}) \to \ldots \to (x_1^{(n_1+n_2-1)}, x_2^{(n_1+n_2-1)})$$
(x1(1),x2(1))→(x1(1),x2(2))→(x1(2),x2(2))→...→(x1(n1+n2−1),x2(n1+n2−1))

用下图可以很直观的看出，采样是在两个坐标轴上不停的轮换的。当然，坐标轴轮换不是必须的，我们也可以每次随机选择一个坐标轴进行采样。不过常用的Gibbs采样的实现都是基于坐标轴轮换的。



3. 多维Gibbs采样

上面的这个算法推广到多维的时候也是成立的。比如一个n维的概率分布$\pi(x_1, x_2, \ldots x_n)$ π(x1,x2,...xn)，我们可以通过在n个坐标轴上轮换采样，来得到新的样本。对于轮换到的任意一个坐标轴$x_i$xi上的转移，马尔科夫链的状态转移概率为$P(x_i|x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$ P(xi|x1,x2,...,xi−1,xi+1,...,xn)，即固定$n-1$n−1个坐标轴，在某一个坐标轴上移动。

具体的算法过程如下：

1）输入平稳分布$\pi(x_1, x_2，\ldots, x_n)$ π(x1,x2，...,xn)或者对应的所有特征的条件概率分布，设定状态转移次数阈值n_1n1，需要的样本个数n_2n2

2）随机初始化初始状态值$(x_1^{(0)}, x_2^{(0)}, \ldots, x_n^{(0)}$ (x1(0),x2(0),...,xn(0)

3）for $t=0$t=0 to n_1+n_2-1n1+n2−1:

a) 从条件概率分布$P(x_1|x_2^{(t)}, x_3^{(t)}, \ldots, x_n^{(t)})$ P(x1|x2(t),x3(t),...,xn(t))中采样得到样本x_1^{t+1}x1t+1

b) 从条件概率分布$P(x_2|x_1^{(t+1)}, x_3^{(t)}, x_4^{(t)}, \ldots, x_n^{(t)})$ P(x2|x1(t+1),x3(t),x4(t),...,xn(t))中采样得到样本x_2^{t+1}x2t+1

c)...

d) 从条件概率分布$P(x_j|x_1^{(t+1)}, x_2^{(t+1)}, \ldots, x_{j-1}^{(t+1)}, x_{j+1}^{(t)} \ldots, x_n^{(t)})$ P(xj|x1(t+1),x2(t+1),...,xj−1(t+1),xj+1(t)...,xn(t))中采样得到样本x_j^{t+1}xjt+1

e)...

f) 从条件概率分布$P(x_n|x_1^{(t+1)}, x_2^{(t+1)}, \ldots, x_{n-1}^{(t+1)})$ P(xn|x1(t+1),x2(t+1),...,xn−1(t+1))中采样得到样本x_n^{t+1}xnt+1

样本集$\{(x_1^{(n_1)}, x_2^{(n_1)}, \ldots, x_n^{(n_1)}), \ldots, (x_1^{(n_1+n_2-1)}, x_2^{(n_1+n_2-1)}, \ldots, x_n^{(n_1+n_2-1)})\}$ {(x1(n1),x2(n1),...,xn(n1)),..., (x1(n1+n2−1),x2(n1+n2−1),...,xn(n1+n2−1))}即为我们需要的平稳分布对应的样本集。

整个采样过程和Lasso回归的坐标轴下降法算法非常类似，只不过Lasso回归是固定$n-1$个特征，对某一个特征求极值。而Gibbs采样是固定$n-1$个特征在某一个特征采样。

同样的，轮换坐标轴不是必须的，我们可以随机选择某一个坐标轴进行状态转移，只不过常用的Gibbs采样的实现都是基于坐标轴轮换的。

4. 二维Gibbs采样实例

这里给出一个Gibbs采样的例子。完整代码参见我的github: https://github.com/ljpzzz/machinelearning/blob/master/mathematics/mcmc_3_4.ipynb

假设我们要采样的是一个二维正态分布$Norm(\mu, \Sigma)$,其中：

$$\mu = (\mu_1, \mu_2) = (5, -1)$$

μ=(μ1,μ2)=(5,−1)

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 1 & 4 \end{pmatrix}$$

Σ=(σ12ρσ1σ2ρσ1σ2σ22)=(1114)

而采样过程中的需要的状态转移条件分布为：

$$P(x_1|x_2) = Norm\left(\mu_1 + \rho\sigma_1/\sigma_2(x_2 - \mu_2), (1-\rho^2)\sigma_1^2\right)$$

P(x1|x2)=Norm(μ1+ρσ1/σ2(x2−μ2),(1−ρ2)σ12)

$$P(x_2|x_1) = Norm\left(\mu_2 + \rho\sigma_2/\sigma_1(x_1 - \mu_1), (1-\rho^2)\sigma_2^2\right)$$

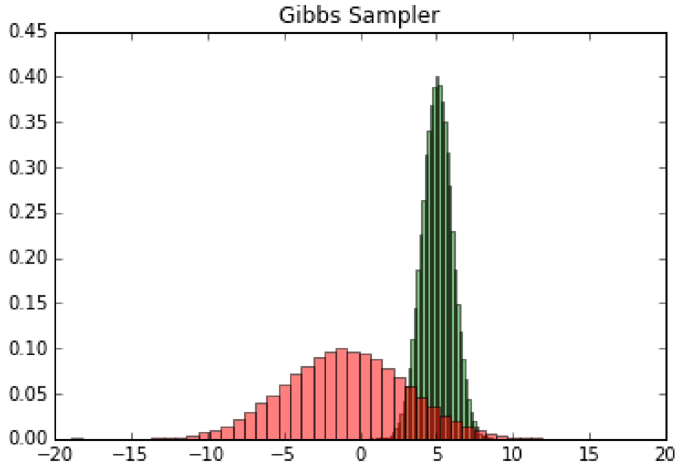
P(x2|x1)=Norm(μ2+ρσ2/σ1(x1−μ1),(1−ρ2)σ22)

具体的代码如下：

```
from mpl_toolkits.mplot3d import Axes3D
from scipy.stats import multivariate_normal
samplesource = multivariate_normal(mean=[5,-1], cov=[[1,0.5],[0.5,2]])

def p_ygivenx(x, m1, m2, s1, s2):
    return (random.normalvariate(m2 + rho * s2 / s1 * (x - m1), math.sqrt(1 - rho ** 2) * s2))

def p_xgiveny(y, m1, m2, s1, s2):
    return (random.normalvariate(m1 + rho * s1 / s2 * (y - m2), math.sqrt(1 - rho ** 2) * s1))

N = 5000
K = 20
x_res = []
y_res = []
z_res = []
m1 = 5
m2 = -1
s1 = 1
s2 = 2

rho = 0.5
y = m2

for i in xrange(N):
    for j in xrange(K):
        x = p_xgiveny(y, m1, m2, s1, s2)
        y = p_ygivenx(x, m1, m2, s1, s2)
        z = samplesource.pdf([x,y])
        x_res.append(x)
        y_res.append(y)
        z_res.append(z)
```

```
num_bins = 50
plt.hist(x_res, num_bins, normed=1, facecolor='green', alpha=0.5)
plt.hist(y_res, num_bins, normed=1, facecolor='red', alpha=0.5)
plt.title('Histogram')
plt.show()
```

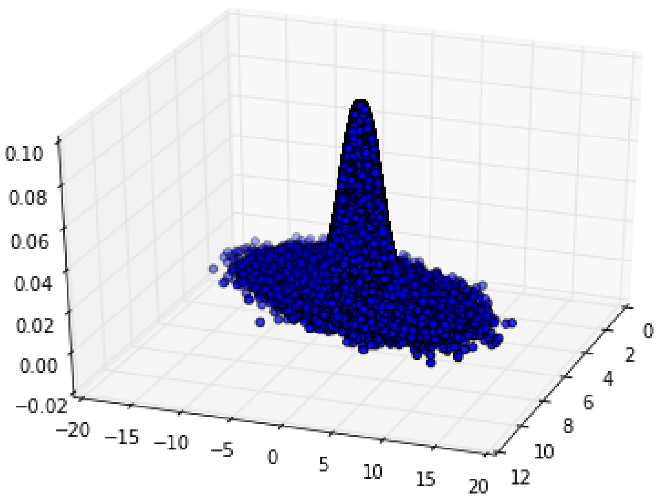
输出的两个特征各自的分布如下：



然后我们看看样本集生成的二维正态分布，代码如下：

```
fig = plt.figure()
ax = Axes3D(fig, rect=[0, 0, 1, 1], elev=30, azim=20)
ax.scatter(x_res, y_res, z_res,marker='o')
plt.show()
```

输出的正态分布图如下：



5. Gibbs采样小结

由于Gibbs采样在高维特征时的优势，目前我们通常意义上的MCMC采样都是用的Gibbs采样。当然Gibbs采样是从M-H采样的基础上的进化而来的，同时Gibbs采样要求数据至少有两个维度，一维概率分布的采样是没法用Gibbs采样的,这时M-H采样仍然成立。

有了Gibbs采样来获取概率分布的样本集，有了蒙特卡罗方法来用样本集模拟求和，他们一起就奠定了MCMC算法在大数据时代高维数据模拟求和时的作用。MCMC系列就在这里结束吧。

（欢迎转载，转载请注明出处。欢迎沟通交流： liujianping-ok@163.com）

« 上一篇：MCMC(三)MCMC采样和M-H采样
» 下一篇：文本挖掘的分词原理

评论列表

#1楼 2017-07-31 15:58 jingchunzhen
请问博主在二维Gibbs采样实例中，为什么设置了双重for循环for i in xrange(N):
for j in xrange(K):，这样做的意义是什么呢？参照伪代码是for i in xrange(n1+n2):
支持(0)反对(0)

#2楼[楼主] 2017-07-31 16:19 刘建平Pinard
@ jingchunzhen

你好，这里代码里的用法的确和算法描述稍有不同：

在代码里外层循环是得到N个样本，对于每个样本，我们要在内层循环中分别进行马尔科夫状态转移K次得到对应的样本。这样的好处是得到的样本独立性好一点，毕竟我们的条件概率计算量不大。

在算法里描述的的确如你所说，先进行若干次马尔科夫状态转移，到了我们认为已经收敛到目标分布时，就进行连续的采样，得到N个样本。这样的好处是采样的计算量小很多，尤其是对于复杂分布的计算，一次马尔科夫过程后就可以得到所有想要个数的样本，但是样本之间的独立性会稍差。
支持(5)反对(0)

#3楼 2017-12-29 18:35 忆梦涟
你好，请问状态转移次数阈值，也就是马尔科夫链达到稳定的转移次数n1，这个值如何确定，
支持(0)反对(0)

#4楼[楼主] 2018-01-02 11:03 刘建平Pinard
@ 忆梦涟
你好，实际应用中都是排脑袋拍出来的，如果计算能力足够，n1可以比较大一些，这样保险，否则就没有办法了，甚至可能只能取n1为个位数，当然这样有可能马尔科夫链并没有真正得到稳态，但是也算是一种近似了。
支持(0)反对(0)

#5楼 2018-01-11 09:25 CuiPeng
你好，我看到LDA中有用gibbs采样训练参数，我看文章中只讲到gibbs采样是用来生成稳定的指定分布的数据，请问怎样用gibbs采样去更新参数呢？比如我给定了一组参数，如何根据样本去训练参数呢？
支持(0)反对(0)

#6楼[楼主] 2018-01-11 11:45 刘建平Pinard
@ CuiPeng
你好，Gibbs采样只是一个数学方法，不是一个机器学习算法，所以本身不能帮你训练模型参数，更新模型参数。
在LDA中，Gibbs采样也只是为了采样分布数据然后做近似计算的，也就是为LDA求解服务的。
支持(1)反对(0)

#7楼 2018-04-15 11:21 小小老实人
楼主你好，文章写得很好，对我帮助很大。一个无关紧要得问题，请问算法里随机初始化状态值，角标是不是该为0呢？
支持(0)反对(0)

#8楼[楼主] 2018-04-16 20:35 刘建平Pinard
@ 小小老实人
你好，那里的确是我写错，已经修改，感谢指出错误。
支持(0)反对(0)

#9楼 2018-05-12 14:06 推荐算法萌新
博主您好！看了这篇文章我有一点疑问：
1、在每一节里面，采样过程的第一步都写了，输入平稳分布π(x)，那么这个平稳分布在采样过程中的作用是什么？
支持(0)反对(0)

#10楼 2018-05-12 14:15 推荐算法萌新
或者我是这样理解，多维Gibbs采样时，不管初始状态如何，只要依据当前状态下计算的单轴条件概率进行转移，最终都会收敛，然后的得到一个分布P，而这个分布P，是满足那个细致平稳条件的，所以分布P是一个平稳分布。
支持(0)反对(0)

#11楼[楼主] 2018-05-12 22:11 刘建平Pinard
@ 推荐算法萌新
你好，是的，这个平稳分布就是我们需要采样的分布，而Gibbs采样就是帮助我们采样得到改分布样本的方法。
支持(0)反对(0)

#12楼 2018-06-09 22:29 zhuyunxiu
有个问题想请教一下，样本是根据任意的Q概率分布采样得到的，然后是Q开始改变了对吧，变成了Q（x*|xt），再根据这个概率进行采样的。那到底这个概率是什么呢？我看到您的实例中用高斯分布先采一个，再用采到的样品作为新的高斯分布的均值进行采样，那要是不是高斯分布，假设是多项分布作为起始采样概率，那这个Q（x*|xt）该怎么求？
还有，同样的，吉布斯中的条件概率如果不是高斯要怎么进行采样？
支持(0)反对(0)

#13楼[楼主] 2018-06-10 12:43 刘建平Pinard
@ zhuyunxiu
你好，我们选择Q的原则肯定是好采样的常用概率分布。Q的联合分布和条件分布函数也都是已知很好采样的概率分布。所以你说的Q（x*|xt）也是一个已知的好采样的概率分布。
回到你说的：“假设是多项分布作为起始采样概率，那这个Q（x*|xt）该怎么求？”，选择Q的原则是好采样的条件概率。所以如果你这个Q（x*|xt）不好求，那可以换一个Q了。
最后一个问题，“条件概率如果不是高斯要怎么进行采样？”，Gibbs采样比较特殊，它是基于坐标轴轮转迭代的，所以需要任意一个坐标轴基于其他所有坐标轴的条件概率，这个条件概率如果不是常见的好采样的分布，那么也是不好采样求解的。也就是说Gibbs采样也不是万能的。
支持(2)反对(0)

#14楼 2018-08-19 12:45 Lszx
老师您好，看了您的文章受益匪浅
请问:
a) 从条件概率分布 P(x2|x1(t))中采样得到样本 x2 (t+1)
b) 从条件概率分布P(x1|x2(t+1))中采样得到样本 x1 (t+1)
根据原理 任意两点E F之间的转移 都服从细致平稳分布
这里的第二步采样 为什么不是从P(x1|x2(t))中采样得到样本 x1 (t+1)
支持(0)反对(0)

#15楼[楼主] 2018-08-19 23:02 刘建平Pinard
@ Lszx
你好，你这样采样也是可以的。我采用坐标轴轮换只是为了符合大多数Gibbs采样算法使用的的思路。第三节的末尾也讲到了原因:
用下图可以很直观的看出，采样是在两个坐标轴上不停的轮换的。当然，坐标轴轮换不是必须的，我们也可以每次随机选择一个坐标轴进行采样。不过常用的Gibbs采样的实现都是基于坐标轴轮换的。
支持(3)反对(0)

#16楼 2018-08-21 11:12 Lszx
@ 刘建平Pinard
老师 我有个问题想请您指导一下 具体发您163邮箱了
期待您百忙之中的回复
支持(0)反对(0)

#17楼 2018-09-13 10:03 Process520
老师，您好，关于下面两步算法描述中，有三个疑问:
a) 从条件概率分布 P(x2|x1(t))中采样得到样本 x2 (t+1)
b) 从条件概率分布P(x1|x2(t+1))中采样得到样本 x1 (t+1)
1. 这里说“得到样本x2(t+1)”,我理解是得到样本为(x1(t), x2(t+1)),
x2(t+1)是它在属性x2处的取值，这样理解对吗?
2.如果上面对的话，那我们取的样本(x1(t+1), x2(t+1))是由(x1(t), x2(t+1))转移来的吗?
3.如果第二问对的话，能把(x1(t), x2(t+1))作为一个采样样本吗？不能或不好的原因是因为样本(x1(t+1), x2(t+1))与(x1(t), x2(t))更加独立吗?
支持(0)反对(0)

#18楼[楼主] 2018-09-13 10:25 刘建平Pinard
@ Process520
你好!
1） 的确这里采样得到的是某一个维度的数据，而不是整个数据。
2） 是的，这就是坐标轴轮转。参看第二节图中的黑色采样路径。
3） 也是可以的，但是如你所说，我们期望两个样本的独立性稍微强一些，所以一般不会把相邻采样的数据的弄的某个维度总是一样。
支持(0)反对(0)

#19楼 2018-09-13 11:25 Process520
@ 刘建平Pinard
明白了，感谢您的解答。
祝您生活愉快，身体健康。
支持(0)反对(0)

#20楼 2018-09-19 16:38 aaronwang123
你好，博主:

“MCMC算法在大数据时代高维数据模拟求和时的作用。"

看了这个mcmc系列后，想问下，我的理解是否是对的？

1.整个系列从求不规则面积出发。引出了不同概率分布如何采样问题。
2.那么我们最后有价值的东西是，得到了如何正确采样的方法论。而不是求和问题的解决方法。
3.请问我们大数据处理，为什么会有采样问题
支持(0)反对(0)

#21楼[楼主] 2018-09-20 13:41 刘建平Pinard
@ aaronwang123
你好！

1. 这个没问题，其实最初的原理就是第一篇的接受-拒绝采样。

2. 有价值的东西的确是正确采样的方法论，当有了这个方法后，一些复杂的问题比如基于概率分布的积分求和问题可以使用采样的方法近似求解。

3. 主要是有些计算式太复杂，比如包含复杂的概率分布积分等，没有办法直接求得代数解，但是我们又需要这个结果，那么可以通过采样的方法得到对应分布样本，然后求对应的近似解。

比如在LDA主题模型中，我们可以用MCMC来采样求主题分布和词分布的近似解。
文本主题模型之LDA(二) LDA求解之Gibbs采样算法
https://www.cnblogs.com/pinard/p/6867828.html

此外还有分解机模型里的积分求解。
分解机(Factorization Machines)推荐算法原理
https://www.cnblogs.com/pinard/p/6370127.html
支持(0)反对(0)

#22楼 2018-10-06 18:05 hui0618
博主你好, 请问状态转移条件是如何得到的? 看表达式我的理解是, 当前采集到的x1如果偏移其均值较大, 那么当前采集到的x1对应其小概率分布, 于是调整x2所属的正态分布的均值和方差, 使得采样x2时也会偏移x2的均值更可能得到概率密度较小的x2样本.
支持(0)反对(0)
刷新评论刷新页面返回顶部
注册用户登录后才能发表评论，请 登录 或 注册，访问网站首页。

最新IT新闻:
· SwiftKey Android版即将支持Windows 10云剪贴板
· 福布斯专访贝索斯：他的字典里没有“边界”这个词