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。

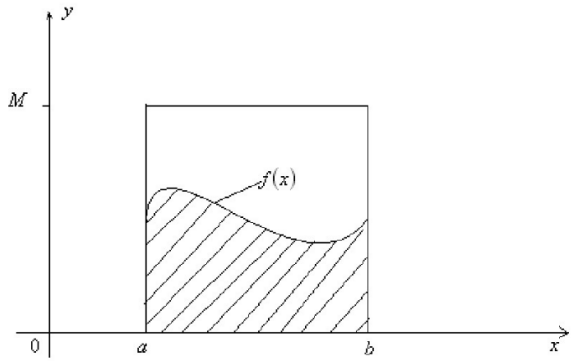
2. 蒙特卡罗方法引入

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 x在[a,b]之间是均匀分布的，而绝大部分情况，x 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 x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：

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

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 x在[a,b]之间是均匀分布的时候，$p(x_i) = 1/(b-a)$p(xi)=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$p(x)$ 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 x的分布$p(x)$ 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 x的概率分布。如果求出了x 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 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 x的样本集。

对于常见的均匀分布$uniform(0, 1)$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$uniform(0, 1)$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) (Z1,Z2)可以通过通过独立采样得到的$uniform(0, 1)$uniform(0,1)样本对(X_1, X_2) (X1,X2)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)

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

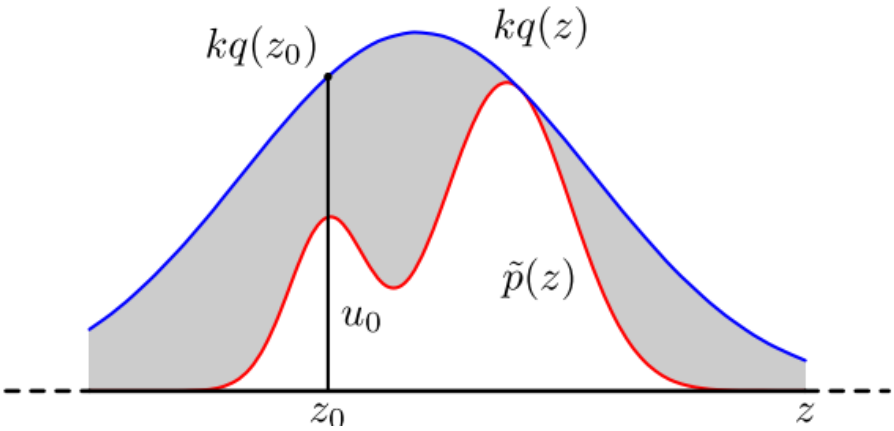
Z2=−2lnX1sin(2πX2)

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$uniform(0, 1)$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 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 $p(x)$ 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 $q(x)$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 $p(x)$ p(x) 分布的目的，其中$q(x)$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。

（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
粉丝 - 2234
+加关注
13
0
« 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
posted @ 2017-03-27 15:08 刘建平Pinard 阅读(12982) 评论(30) 编辑 收藏

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xx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)