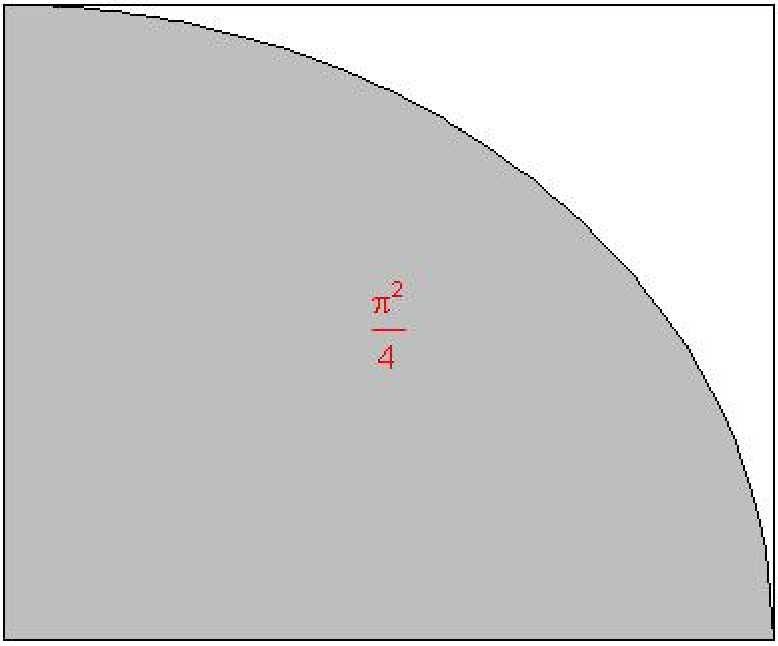
统计计算 张楠 2019秋: Monte Carlo Integration

(返回 统计计算 张楠 2019秋)

引子

蒙特卡罗(Monte Carlo)积分是一种基于随机抽样的统计方法. 蒙特卡罗方法其实 也只是对一种思想的泛称, 只要在解决问题时, 利用产生大量随机样本，然后对 这些样本结果进行概率分析, 从而来预测结果的方法, 都可以称为蒙特卡罗方法.

比如要求圆周率π的值, 最著名的就是中学时学过的“割圆法”(刘徽(魏晋, 3.1416), 祖冲之(南北朝, 3.1415926 $<\pi<$3.1415927)). 现在也可以使用蒙特卡罗积分方法：由概率论知若 $r.v.\ X,Y\ i.i.d\ U(0,1)$, 则 $P(X^2+Y^2\leq 1)=\frac{\pi}{4}$



因此, $\pi=4P(X^2+Y^2\leq 1)\approx 4\#\{x^2+y^2\leq 1\}/n$.

- $n=1000$, $\hat{\pi}=3.168$.
- $n=100,000$, $\hat{\pi}=3.14312$.
- $n=10^7$, $\hat{\pi}=3.141356$.

假设g是一个可积函数, 我们要计算$\int_a^b g(x)dx$. 回忆在概率论中, 若随机变量X的密度为$f(x)$, 则随机变量$Y=g(X)$的期望为 $Eg(X)=\int_{-\infty}^{\infty} g(x)f(x)dx$ 如果从X的分布中产生了一些随机数, 则$Eg(X)$的无偏估计就是相应的样本平均值.

Simple Monte Carlo estimator

考虑估计$\theta=\int_0^1 g(x)dx$. 若$X_1,\cdots,X_m$为均匀分布$U(0,1)$总抽取的样本，则由强大数律知$\hat{\theta}=\overline{g_m(X)}=\frac{1}{m}\sum_{i=1}^m g(X_i)$以概率1收敛到期望$Eg(X)$. 因此$\int_0^1 g(x)dx$的简单的Monte Carlo估计量为$\overline{g_m(X)}$.

例1: (简单的Monte Carlo 积分) 计算$\theta=\int_0^1 e^{-x}dx$的简单Monte Carlo估计以及与积分值相比较.

```
m <- 10000
x <- runif(m)
theta.hat <- mean(exp(-x))
print(theta.hat)
print(1 - exp(-1))
#[1] 0.6355289
#[1] 0.6321206
```

估计为$\hat{\theta}\doteq 0.6355$, 而积分值为$\theta=1-e^{-1}\doteq 0.6321$.

若要计算$\int_a^b g(x)dx$, 此处$a<b$为有限数. 则作一积分变量代换使得积分限从0到1. 即作变换$y=(x-a)/(b-a)$, 因此

$$\int_a^b g(x)dx=\int_0^1 g(y(b-a)+a)(b-a)dy$$

另外一种做法就是利用均匀分布$U(a,b)$, 即$\int_a^b g(x)dx=(b-a)\int_a^b g(x)\frac{1}{b-a}dx$

例2: 简单Monte Carlo 积分(续) 计算$\theta=\int_2^4 e^{-x}dx$的Monte Carlo估计并和积分值相比较.

```
m <- 10000
x <- runif(m, min=2, max=4)
theta.hat <- mean(exp(-x)) * 2
print(theta.hat)
print(exp(-2) - exp(-4))
#[1] 0.1172158
#[1] 0.1170196
```

即, 估计的值为$\hat{\theta}\doteq 0.1172$, 而真值为$\theta=e^{-2}-e^{-4}\doteq 0.1170$.

总结一下, 积分$\int_a^b g(x)dx$的简单Monte Carlo估计方法为

1. 从均匀分布$U(a,b)$中产生$i.i.d$样本$X_1,\cdots,X_m$;
2. 计算$\overline{g_m(X)}=\frac{1}{m}g(X_i)$
3. $\hat{\theta}=(b-a)\overline{(g_m(X))}$.

例3: Monte Carlo积分, 无穷积分 比如使用Monte Carlo积分方法计算标准正态的分布函数$\Phi(x)=\int_{-\infty}^x \frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}dt$.

首先我们不能直接使用以前的方法(因为积分区间无界), 但是我们可以将此问题分为两种情形: $x>0$和$x\le 0$. 若$x>0$, 注意到对标准正态分布, 积分区间可以分为$(-\infty,0)$和$(0,x)$,因此只需要计算积分$\theta=\int_0^x e^{-\frac{t^2}{2}}dt$即可. 故可以使用之前的方法. 但是需要从均匀分布$U(0,x)$中抽取样本, 若x发生变化, 则均匀分布也就变化了. 若要求从$U(0,1)$中抽样, 则可以作变换$y=t/x$, 则

$$\theta = \int_0^1 xe^{-(xy)^2}dy$$

因此, $\theta = E_Y[xe^{-(xY)^2}]$, 其中$Y \sim U(0,1)$. 从而产生$U(0,1)$的$i.i.d$随机数$u_1,\cdots,u_m$, 则 θ的估计为

$$\hat{\theta} = \overline{g_m(u)} = \frac{1}{m}\sum_{i=1}^m xe^{-(xu)^2}.$$

此时对$x > 0$,$\Phi(x)$的估计为$0.5 + \hat{\theta}/\sqrt{2\pi}$; 对$x \le 0$, 计算$\Phi(x) = 1 - \Phi(-x)$.

```
x <- seq(.1, 2.5, length = 10)
m <- 10000
u <- runif(m)
cdf <- numeric(length(x))
for (i in 1:length(x)) {
    g <- x[i] * exp(-(u * x[i])^2 / 2)
    cdf[i] <- mean(g) / sqrt(2 * pi) + 0.5
}
Phi <- pnorm(x)
print(round(rbind(x, cdf, Phi), 3))
```

例4: Monte Carlo积分, 无穷积分(续) 对上例, 我们可以使用另外一种方式(hit-or-miss). 记$Z \sim N(0,1)$, 则对任何实数x有
$\Phi(x) = P(Z \le x) = EI(Z \le x)$.

从而从标准正态中产生随机样本$z_1,\cdots,z_m$后, 即可得到$\Phi(x)$的估计为

$$\hat{\Phi}(x) = \frac{1}{m}\sum_{i=1}^m I(z_i \le x).$$

其以概率1收敛到$\Phi(x)$.

```
x <- seq(.1, 2.5, length = 10)
m <- 10000
z <- rnorm(m)
dim(x) <- length(x)
p <- apply(x, MARGIN = 1,
         FUN = function(x, z) {mean(z <= x)}, z = z)
Phi <- pnorm(x)
print(round(rbind(x, p, Phi), 3))
```

总结如下: 欲估计积分

$$\theta = \int_A g(x)f(x)dx$$

其中f为以A为支撑的概率函数, 即$\int_A f(x)dx = 1$. 则产生f的$i.i.d$随机数$x_1,\cdots,x_m$后, 由大数律知θ的估计为

$$\hat{\theta} = \frac{1}{m}\sum_{i=1}^m g(x_i).$$

由于$\hat{\theta}$的方差为σ^2/m, 其中$\sigma^2 = Var_f(g(X))$. 当随机变量X的分布未知时, 我们可以使用样本$x_1,\cdots,x_m$的经验分布函数, 从而得到σ^2/m的估计为

$$\hat{\sigma}^2/m = \frac{1}{m^2}\sum_{i=1}^m[g(x_i) - \overline{g(x)}]^2.$$

再由中心极限定理知道当$m \to \infty$时

$$\frac{\hat{\theta}-E\hat{\theta}}{\sqrt{Var(\hat{\theta})}}$$

依分布收敛到标准正态分布. 因此对大样本, 渐近正态性可以给出积分的Monte Carlo估计的误差界, 以此可以来检查收敛性.

```
x <- 2
m <- 10000
z <- rnorm(m)
g <- (z <= x)  #the indicator function
v <- mean((g - mean(g))^2) / m
cdf <- mean(g)
c(cdf, v)
c(cdf - 1.96 * sqrt(v), cdf + 1.96 * sqrt(v))
#[1] 9.7800e-01 2.1516e-06
#[1] 0.975125 0.980875
```

随机变量$I(Z\leq 2)$取值1的概率为$\Phi(2)\approx 0.977$. 此处$g(X)\sim B(10000,\Phi(2))$, 因此 $g(X)$的方差为$0.977(1-0.977)/10000=2.223e-06$. Monte Carlo积分估计的方差为$2.1516e-06$, 已经非常接近了.

Variance and Efficiency

Monte Carlo 方法在估计一个积分$\int_a^b g(x)dx$时, 将其表示为一个均匀随机变量的期望, 从而

$$\theta=\int_a^b g(x)dx=(b-a)\int_a^b g(x)\frac{1}{b-a}dx=(b-a)E[g(X)],\quad X\sim U(a,b).$$

从而算法如下

1. 从$U(a,b)$中产生$i.i.d$样本$X_1,\cdots,X_m$.
2. 计算$\overline{g(X)}=\frac{1}{m}\sum_{i=1}^m g(X_i)$.
3. $\hat{\theta}=(b-a)\overline{g(X)}$.

易知,

$$E\hat{\theta}=\theta,\quad Var(\hat{\theta})=(b-a)^2Var(\overline{g(X)})=\frac{(b-a)^2}{m}Var(g(X)).$$

有中心极限定理, $\overline{g(X)}$依分布渐近到正态分布, 因此$\hat{\theta}$也渐近到正态分布.

Hit-or-miss Monte Carlo 方法则使用了另外一种估计积分的方式, 其方差和上面说的方法不同. 表述如下: 假设f为随机变量X的概率函数, 使用``hit-or-miss"方法估计积分$F(x)=\int_{-\infty}^x f(t)dt$:

1. 从X的分布中产生$i.i.d$样本$X_1,\cdots,X_m$.
2. 计算$\overline{g(X)}=\frac{1}{m}\sum_{i=1}^m I(X_i\leq x)$.
3. $F\hat{(x)}=\overline{g(X)}$.

显然对每个有限的x, $Y=g(X)=I(X\leq x)\sim B(1,p),\quad p=F(x)$. 因此

$$E[F\hat{(x)}]=F(x),\quad Var[F\hat{(x)}]=F(x)(1-F(x))/m.$$

$F\hat{(x)}$的方差可以通过$F\hat{(x)}(1-F\hat{(x)})/m$来估计.

这两种方法的方差不同, 自然会问哪种优一些, 即更有效率.

设$\hat{\theta}_1$和$\hat{\theta}_2$是θ的两个无偏估计量, 则$\hat{\theta}_1$比$\hat{\theta}_2$更有效, 如果

$$Var(\hat{\theta}_1) \le Var(\hat{\theta}_2).$$

如果一个估计量的方差未知, 则我们通过样本将其估计出来. 另外, 估计量的方差总是可以通过增加样本量来减少的.
