统计计算 张楠 2019秋: Variance Reduction: Stratified Sampling and Stratified Importance Sampling

(返回 统计计算 张楠 2019秋)

Stratified Sampling

利用积分的线性性和和强大数律, 可以知道要求的积分$\int g(x)dx$可以分为几个积分之和, 对每个积分可以单独使用Monte Carlo积分方法. 不妨设将此积分分为k个积分之和, 对每个积分 抽样m_i次, $m=m_1+\cdots+m_k$, 满足目标

$$Var(\hat{\theta}_k(m_1,\cdots,m_k)) < Var(\hat{\theta}).$$

例11: 考虑之前的例子 将积分区间$(0,1)$比如分为四个子区间, 在每个子区间上抽样$m/4$次, 这里m为总的抽样次数. 然后将这些估计加起来得到积分$\int_0^1 e^{-x}(1+x^2)^{-1}dx$的估计.

```
M <- 20    #number of replicates
T2 <- numeric(4)
estimates <- matrix(0, 10, 2)

g <- function(x) {
    exp(-x - log(1+x^2)) * (x > 0) * (x < 1) }

for (i in 1:10) {
    estimates[i, 1] <- mean(g(runif(M)))
    T2[1] <- mean(g(runif(M/4, 0, .25)))
    T2[2] <- mean(g(runif(M/4, .25, .5)))
    T2[3] <- mean(g(runif(M/4, .5, .75)))
    T2[4] <- mean(g(runif(M/4, .75, 1)))
    estimates[i, 2] <- mean(T2)
}

estimates
apply(estimates, 2, mean)
apply(estimates, 2, var)
```

定理2.记M次抽样下的简单Monte Carlo积分估计量(即从均匀分布中抽样)为$\hat{\theta}^M$, 以及

$$\hat{\theta}^S=\frac{1}{k}\sum_{i=1}^k\hat{\theta}_i$$

表示分层的估计量, 每个层的抽样次数为m/k. 在第j个层上, $g(U)$的均值和方差分别 记为θ_j和σ_j^2, $j=1,\cdots,k$. 则$Var(\hat{\theta}^M)\geq Var(\hat{\theta}^S)$.

证: [折叠]

用J表示随机选择的层, $P(J=j)=1/k, j=1,\cdots,k$. 则

$$
\begin{aligned}
Var(\hat{\theta}^M) &= \frac{Var(g(U))}{M} = \frac{1}{M}[Var(E(g(U)|J)) + E(Var(g(U)|J))] \\
&= \frac{1}{M}(Var(\theta_J) + E\sigma_J^2) \\
&= \frac{1}{M}(Var(\theta_J) + \frac{1}{k}\sum_{i=1}^{k}\sigma_j^2) \\
&= \frac{1}{M}Var(\theta_J) + Var(\hat{\theta}^S) \geq Var(\hat{\theta}^S).
\end{aligned}
$$

□

例12: 使用分层抽样方法估计前面例子中的积分 对积分$\int_0^1 e^{-x}(1+x^2)^{-1}dx$应用 k层抽样方法估计, 并和简单的Monte Carlo积分方法比较.

```
M <- 10000  #number of replicates
   k <- 10     #number of strata
   r <- M / k  #replicates per stratum
   N <- 50     #number of times to repeat the estimation
   T2 <- numeric(k)
   estimates <- matrix(0, N, 2)

   g <- function(x) {
       exp(-x - log(1+x^2)) * (x > 0) * (x < 1)
       }

   for (i in 1:N) {
       estimates[i, 1] <- mean(g(runif(M)))
       for (j in 1:k)
           T2[j] <- mean(g(runif(M/k, (j-1)/k, j/k)))
       estimates[i, 2] <- mean(T2)
   }

   apply(estimates, 2, mean)
   apply(estimates, 2, var)
```

Stratified Importance Sampling

分层抽样的思想可以用在重要性抽样方法中. 在重要性抽样方法中, $\theta = \int g(x)dx$ 的估计量方差为σ^2/M, 其中 $\sigma^2 = Var(g(X)/f(X)), X \sim f$. 应用分层抽样方法, 将直线分为$k$个子区间 $I_j = \{x : a_{j-1} \leq x < a_j\}$, 其中 $a_0 = -\infty, a_j = F^{-1}(j/k), j = 1, \cdots, k-1, a_k = \infty$. 记$g_j(x) = g(x)I(a_{j-1} \leq x < a_j)$, 以及

$$\theta_j = \int_{a_{j-1}}^{a_j} g_j(x)dx, j = 1, \cdots, k.$$

则$\theta = \theta_1 + \cdots + \theta_k$. 对每个子区间$I_j$, 重要性函数可以取为条件密度:

$$f_j(x) = f(x|I_j) = \frac{f(x)I(a_{j-1} \leq x < a_j)}{P(a_{j-1} \leq X < a_j)} = kf(x)I(a_{j-1} \leq x < a_j)$$

再记$\sigma_j^2 = Var(g_j(X)/f_j(X)), j = 1, \cdots, k$. 然后得到$\theta$的分层重要性抽样 下的估计为

$$\hat{\theta}^{SI} = \sum_{i=1}^{k} \hat{\theta}_j$$

其方差为

$$Var(\hat{\theta}^{SI}) = \sum_{i=1}^{k} Var(\hat{\theta}_j) = \frac{1}{m}\sum_{i=1}^{k}\sigma_j^2.$$

其中$m = M/k$. 则希望

$$\sigma^2/M > \frac{k}{M}\sum_{i=1}^{k}\sigma_j^2.$$

我们可以证明如下结论:

定理3：假设$M = km$为重要性抽样下的估计量$\hat{\theta}^{I}$的抽样个数, $\hat{\theta}^{SI} = \sum_{i=1}^{k}\hat{\theta}_j$为分层重要性抽样下的估计量, 这里$\hat{\theta}_j$为第$j$层$\theta_j$的重要性抽样下的估计量, 抽样个数为$m$. 若$Var(\hat{\theta}^{I}) = \sigma^2/M$, 以及$Var(\hat{\theta}_j) = \sigma_j^2/m$,则

$$\sigma^2 - k\sum_{j=1}^{k}\sigma_j^2 \geq 0,$$

等号成立当且仅当$\theta_1 = \cdots = \theta_k$.

此结论说明分层抽样绝不会扩大方差, 在g非常数的场合总是存在一个可以减少方差的分层.

例13: 在前面的例子中, 试使用$f_3(x)$作为重要性函数, 将积分区间分为$(j/5, (j+1)/5), j = 0, 1, \cdots, 4$这 5个子区间, 对每个子区间应用重要性抽样方法, 计算此时积分的估计量及其经验方差.
