泊松随机数的生成算法：数学推导和程序实现

拉普拉斯算符
数学话题下的优秀答主

最近在做一个机器学习的项目，其中用到了泊松随机数。查维基百科 Poisson distribution 发现了一个算法，可以生成泊松随机数：

```
algorithm poisson random number (Knuth):
    init:
         Let L ← e^{−λ}, k ← 0 and p ← 1.
    do:
         k ← k + 1.
         Generate uniform random number u in [0,1] and let p ← p × u.
    while p > L.
    return k − 1.
```

词条里面没有解释为什么这个算法可以生成泊松随机数，我在此给出证明。

第一节：算法描述

上面的这个算法可以描述为：

第一步：给定一个参数 $\lambda>0$ ，生成一系列的随机数，这些随机数服从 $\mathrm{Uniform}(0,1)$ 分布，也就是这些随机数在开区间 (0, 1) 之间均匀分布。

第二步：求这些随机数的乘积，当乘积小于或者等于 $e^{-\lambda}$ 时，程序停止。记下此时参与求乘积的随机数的个数。

第三步：程序终止时参与乘积的随机数的个数减一服从参数为 λ 的泊松分布。

第二节：算法的数学表达

为了证明这个算法确实可以生成泊松随机数，我们记

$$P=\prod_{i=1}^{n}X_i, X_i\sim \mathrm{Uniform}(0,1)$$

这就等价于

$$\log P=\sum_{i=1}^{n}\log X_i$$

已知随机变量 X 的概率密度为 $f_X(x)$ ，令 $Y=\log X$.

$$p(Y\le y)=\int_{-\infty}^{y}f_Y(y')dy'=p(\log X\le y)=p(X\le e^y)=\int_{-\infty}^{e^y}f_X(x)dx$$

所以变量 Y 的概率密度为

$$f_Y(y)=\frac{d}{dy}\int_{-\infty}^{e^y}f_X(x)dx=f_X(e^y)e^y$$

已知

$$f_X(x)=\begin{cases}1 & 0\le x\le 1\\ 0 & \text{otherwise}\end{cases}$$

所以

$$f_Y(y)=\begin{cases}e^y & -\infty<y\le 0\\ 0 & \text{otherwise}\end{cases}$$

这是一个指数分布。

因为随机变量 $\log P=\sum_{i=1}^{n}\log X_i:=\sum_{i=1}^{n}Y_i$ ，所以我们要计算一系列服从指数分布的随机变量的和。已知，对于独立随机变量 X,Y ，它们的和 $Z=X+Y$ 的概率密度为

$$f_Z(z)=\int_{-\infty}^{\infty}f_X(\xi)f_Y(z-\xi)d\xi$$

这是两个概率密度函数的卷积。做傅里叶变换，得到 Z 的概率分布的特征函数是 X,Y 两个随机变量的概率密度分布的特征函数的乘积。为了计算 $\log P$ 的概率分布，我们先要计算指数分布的特征函数。根据特征函数的定义，我们有

$$\hat{f}_Y(\eta)=\int_{-\infty}^{\infty}f_Y(y)e^{i\eta y}dy=\frac{1}{i\eta+1}$$

所以变量 $\log P$ 的概率密度的特征函数为 $\frac{1}{(i\eta+1)^n}$.

第三节：根据概率密度的特征函数计算所对应的概率密度

现在我们已经知道了概率密度的特征函数，接下来我们要根据这个特征函数计算所对应的概率密度。做傅里叶逆变换可以得到所对应的概率密度分布:

$$I(y)=\frac{1}{2\pi}\int_{-\infty}^{\infty}\frac{1}{(i\eta+1)^n}e^{-i\eta y}d\eta$$

因为变量 $\log P$ 是一系列负数的求和，所以上面的积分中， $y<0$.

选择如下图所示的一个积分围道：

计算在这个围道上的积分：

$$\frac{1}{2\pi}\oint_{\gamma_R}\frac{1}{(iz+1)^n}e^{-izy}dz$$

这个积分可以分为两部分，第一部分是沿着横轴求积分，第二部分是沿着外面的大圆求积分。可以证明沿着大圆的积分为零，因为

$$\left|\frac{1}{2\pi}\int_{z=Re^{i\theta},\sin\theta>0}\frac{1}{(iz+1)^n}e^{-iyR(\cos\theta+i\sin\theta)}dz\right|\le\frac{1}{2\pi}\int_{z=Re^{i\theta}}\frac{1}{(R+1)^n}e^{yR\sin\theta}Rd\theta\to 0$$

当大圆半径为无穷大的时候该积分趋近于零，因为当 $y<0,\sin\theta>0$ 时， $e^{yR\sin\theta}$ 以指数速度衰减到零。

所以我们就有

$$\frac{1}{2\pi}\oint_{\gamma_R}\frac{1}{(iz+1)^n}e^{-izy}dz=\frac{1}{2\pi}\int_{-\infty}^{\infty}\frac{1}{(i\eta+1)^n}e^{-i\eta y}d\eta$$

根据柯西积分定理，左边的积分为

$$\frac{1}{2\pi}\oint_{\gamma_R}\frac{1}{(iz+1)^n}e^{-izy}dz=\frac{1}{2\pi}\oint_{z=i+\delta e^{i\theta}}\frac{1}{(i\delta e^{i\theta})^n}e^{-iy(i+\delta e^{i\theta})}\delta e^{i\theta}id\theta=\frac{1}{2\pi}e^y\oint\frac{e^{-iy\delta e^{i\theta}}}{(i\delta e^{i\theta})^{n-1}}d\theta$$

上面式子最右边的积分为

$$\oint\frac{e^{-iy\delta e^{i\theta}}}{(i\delta e^{i\theta})^{n-1}}d\theta=\oint\sum_{m=0}^{\infty}\frac{(-y)^m(i\delta e^{i\theta})^{m-n+1}}{m!}d\theta=\sum_{m=0}^{\infty}\frac{(-y)^m}{m!}2\pi\delta_{m,n-1}=2\pi\frac{(-y)^{n-1}}{(n-1)!}$$

所以围道积分为

$$\frac{1}{2\pi}\oint_{\gamma_R}\frac{1}{(iz+1)^n}e^{-izy}dz=e^y\frac{(-y)^{n-1}}{(n-1)!}, y<0$$

最终我们得到随机变量 $\log P$ 所服从的概率密度函数为

$$f_{\log P}(y)=\begin{cases}\frac{(-y)^{n-1}}{(n-1)!}e^y & -\infty<y\le 0\\ 0 & \text{otherwise}\end{cases}$$

这个分布的名字叫做 Γ 分布。显然，根据上式，当 $n=1$ 的时候，上面的分布退化为指数分布。

第四节：计算随机变量 P 的概率密度函数

已经知道了 $\log P$ 服从 Γ 分布，那么计算 P 的分布也很简单了。已知

$$p(\log P\le y)=\int_{-\infty}^{y}f_{\log P}(y')dy'=p(P\le e^y)=\int_{-\infty}^{e^y}f_P(p)dp$$

所以

$$f_P(p)=p^{-1}f_{\log P}(\log p)=\begin{cases}\frac{(-\log p)^{n-1}}{(n-1)!} & 0<p\le 1\\ 0 & \text{otherwise}\end{cases}$$

第五节：计算 $p<e^{-\lambda},\lambda>0$ 的概率

我们已经知道了变量 P 的分布函数，那么就可以计算 $p<e^{-\lambda}$ 的概率为

$$p(P<e^{-\lambda})=\int_{0}^{e^{-\lambda}}\frac{(-\log p)^{n-1}}{(n-1)!}dp=\frac{1}{(n-1)!}\int_{\lambda}^{\infty}e^{-t}t^{n-1}dt$$

因为这个概率依赖于 n ，所以可以将这个概率重新写作

$$p_n(P<e^{-\lambda})=\frac{1}{(n-1)!}\int_{\lambda}^{\infty}e^{-t}t^{n-1}dt$$

利用分部积分，可以得到概率的递归关系为

$$p_n(P<e^{-\lambda})=\frac{\lambda^{n-1}}{(n-1)!}e^{-\lambda}+p_{n-1}(P<e^{-\lambda}),n>1$$

因为 $p_1(P<e^{-\lambda})=e^{-\lambda}$ ，所以我们有

$$p_n(P<e^{-\lambda})=\sum_{k=0}^{n-1}\frac{\lambda^k}{k!}e^{-\lambda}$$

第六节：根据对概率的两种等价解释得到泊松分布

现在我们已经算出来了当我们用 n 个 [0, 1] 均匀分布的随机数连乘时，所得到的乘积小于 $e^{-\lambda},\lambda>0$ 的概率。这里，我们相当于是固定了 n ，扫描不同的参数 λ ，得到了概率。我们可以换一个角度。这个概率也可以看作是我们固定了参数 λ ，计算需要多少个 [0, 1] 之间均匀分布的随机数连乘才能让最后的乘积小于 $e^{-\lambda}$. 也就是，

$$p(P<e^{-\lambda})=p(N\le n)=\sum_{k=1}^{n}p(N=k)$$

根据第五节的结果，我们知道

$$p_n(P<e^{-\lambda})=\sum_{k=0}^{n-1}\frac{\lambda^k}{k!}e^{-\lambda}$$

所以，假设 n 个 [0, 1] 之间均匀分布的随机数连乘后刚好小于 $e^{-\lambda}$ ， 那么 n 服从这样的概率分布：

$$p(N=n)=\frac{\lambda^{n-1}}{(n-1)!}e^{-\lambda}$$

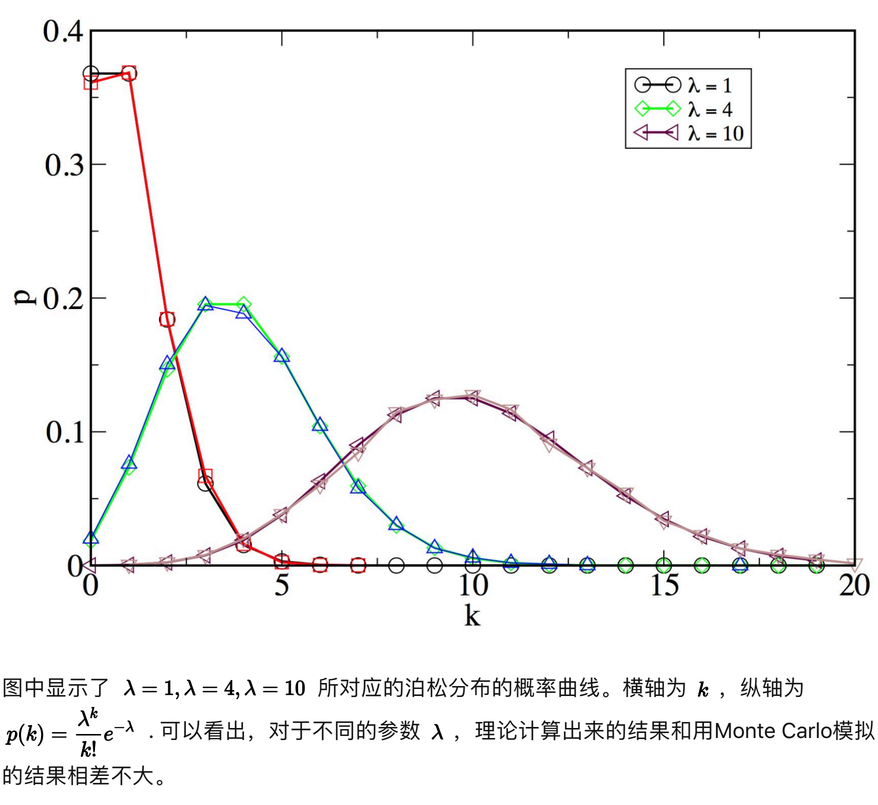
这就是泊松分布。

第七节：程序实现

我已经写了一个程序来实现这个算法，并且得到了测试结果。程序GitHub地址为

PrimerLi/Poisson

第八节：实验结果



图中显示了 $\lambda=1,\lambda=4,\lambda=10$ 所对应的泊松分布的概率曲线。横轴为 k ， 纵轴为 $p(k)=\frac{\lambda^k}{k!}e^{-\lambda}$. 可以看出，对于不同的参数 λ ， 理论计算出来的结果和用Monte Carlo模拟出来的结果相差不大。

编辑于 2018-05-14

4 条评论

lixin liu 2017-10-31

我猜这个算法提出的背景可能是这样的。（1）一个时间上的泊松点过程，如果强度参数为1，则lambda时间内，事件点出现的个数服从泊松分布，参数为lambda。（2）相邻事件点之间的时间长度互相独立，且服从参数为1的指数分布。（3）因此重复生成服从指数分布的随机变量，代表时间区间长度，把他们累加起来，直到总和恰好超过时间段总长lambda，累计的个数减去1就是这段时间“内”事件点的个数，即是一个服从泊松分布的随机变量。（4）[0,1]均匀随机变量的负对数服从参数为1的指数分布，因此做一下变量代换就可以得到原算法。

（1）算是泊松过程和泊松分布的定义

（2）是个性质（多次间隔即是多个指数分布的独立和，服从gamma分布，这也是个性质，你也证明了）

（4）就是第二节证明的

6

田永龙 回复 lixin liu 2019-12-02

晖哥解释得很透彻啊，666

赞

Lee Sam 2019-12-02

这研究流程记录的，真标准

赞

Rainhow Zhang 2017-11-04

第三节用Moment Generating Function会不会简化很多。 当然如果这里最后logP 不是一些已知的分布，就不行了。

赞