Bayesian规则下的Go/Nogo概率

Chao Cheng

2023 年 7 月 20 日

目录

1 引言

事件时间$T_e \sim Exp(\lambda_1)$, $f(t;\lambda_1) = \lambda_1 \exp(-\lambda_1 t)$，删失时间$T_c \sim Exp(\lambda_2)$, $f(t;\lambda_2) = \lambda_2 \exp(-\lambda_2 t)$. 则观察到的数据的联合密度/质量

$$f(t,\delta) = \begin{cases} \lambda_1 \exp\left(-(\lambda_1+\lambda_2)t\right) & \delta = 1 \\ \lambda_2 \exp\left(-(\lambda_1+\lambda_2)t\right) & \delta = 0 \end{cases} \tag{1}$$

其中$\delta = 1$表示观察到事件，$\delta = 0$表示观察到删失。

2 数据似然

记实际观察到的数据为(T,δ)，其中$T=\min(T_e,T_c)$；$\delta=1$表示观察到事件，$\delta=0$表示观察到删失。那么对于i.i.d.样本(t_i,δ_i), $i=1,\cdots,n$，数据的likelihood为

$$\begin{aligned}L(\lambda_1,\lambda_2;Data)&=\prod_{i=1}^n f(t_i,\delta_i)\\&=\lambda_1^{\sum_{i=1}^n\delta_i}\lambda_2^{n-\sum_{i=1}^n\delta_i}\exp\left(-(\lambda_1+\lambda_2)\sum_{i=1}^n t_i\right)\\&=\lambda_1^{\sum_{i=1}^n\delta_i}\exp\left(-\lambda_1\sum_{i=1}^n t_i\right)\cdot\lambda_2^{n-\sum_{i=1}^n\delta_i}\exp\left(-\lambda_2\sum_{i=1}^n t_i\right)\end{aligned}\tag{2}$$

从(2)中可以看出最大化likelihood时，λ_1和λ_2相互独立。若要关于λ_1来最大化likelihood，只需要最大化

$$\lambda_1^{\sum_{i=1}^n\delta_i}\exp\left(-\lambda_1\sum_{i=1}^n t_i\right)$$

若要关于λ_2来最大化likelihood，只需要最大化

$$\lambda_2^{n-\sum_{i=1}^n\delta_i}\exp\left(-\lambda_2\sum_{i=1}^n t_i\right)$$

因此MLE分别是

$$\hat{\lambda_1}=\frac{\sum_{i=1}^n\delta_i}{\sum_{i=1}^n t_i},\quad\hat{\lambda_2}=\frac{n-\sum_{i=1}^n\delta_i}{\sum_{i=1}^n t_i}$$

3 先验和后验分布

一般对λ采用共轭先验$Gamma(\alpha,\beta)$分布，密度为

$$\pi(\lambda|\alpha,\beta)=\frac{\beta^\alpha}{\Gamma(\alpha)}\lambda^{\alpha-1}\exp(-\beta\lambda)\tag{3}$$

取决于参数的定义和记号的设置，有时会采用Inverse-Gamma分布，或参数β设置为$1/\beta$，因此在实际使用中，请确认到底使用的何种形式的密度和先验。本文档中的写法，α和β对应着`R`中`dgamma`函数中的`shape`和`rate`参数。

结合(2)和(3)可知，λ_1的后验分布核心为

$$f(\lambda_1|Data)\propto\lambda_1^{\alpha+\sum_{i=1}^n\delta_i-1}\exp\left(-\left(\beta+\sum_{i=1}^n t_i\right)\lambda_1\right),$$

因此λ_1的后验分布为$Gamma\left(\alpha+\sum\delta_i,\beta+\sum t_i\right)$。

同理，λ_2的后验分布为$Gamma\left(\alpha+n-\sum\delta_i,\beta+\sum t_i\right)$。

4 操作特性的计算

4.1 理论计算

观察到事件的概率：

$$P\left(\delta=1\right)=\int_0^\infty f\left(t,\delta=1\right)\mathrm{d}t=\frac{\lambda_1}{\lambda_1+\lambda_2}.$$

那么观察时间的条件分布

$$P\left(T\leq t|\delta\right)=\frac{P\left(T\leq t,\delta\right)}{P\left(\delta\right)}=\begin{cases}\dfrac{P\left(T\leq t,\delta=1\right)}{P\left(\delta=1\right)}\sim Exp\left(\lambda_1+\lambda_2\right) & \text{if } \delta=1\\ \dfrac{P\left(T\leq t,\delta=0\right)}{P\left(\delta=0\right)}\sim Exp\left(\lambda_1+\lambda_2\right) & \text{if } \delta=0\end{cases}$$

可见该条件分布与δ取值无关，因此T与δ独立。（一般这样的结论是不会成立的，这里依赖于两者都是指数分布）。

由此可得

$$\sum_{i=1}^n T_i\sim Gamma\left(n,\lambda_1+\lambda_2\right)$$
$$\sum_{i=1}^n \delta_i\sim Binomial(n,\frac{\lambda_1}{\lambda_1+\lambda_2})$$

且$\sum\limits_{i=1}^n T_i$和$\sum\limits_{i=1}^n \delta_i$独立。

在Bayesian角度下，风险的后验分布服从$Gamma\left(\text{prior_shape}+\sum\limits_{i=1}^n\delta_i,\text{prior_rate}+\sum\limits_{i=1}^n T_i\right)$对于给出某一个Go/Nogo的规则，例如$P\left(\lambda<\text{eff_cut}\right)\geq\text{go_prob}$后，我们可以将其转化为在所有可能的观察到的事件数的情况下，需要的总观察时间的要求，因此某一个潜在的风险λ_1给出Go信号的概率为

$$\begin{aligned}&P\left(\text{Declare Go}\right)\\&=\sum_{evt_{num}=0}^n P\left(\text{Declare Go},\sum_{i=0}^n\delta_i=evt_{num}\right)\\&=\sum_{evt_{num}=0}^n P\left(\text{Declare Go}\middle|\sum_{i=0}^n\delta_i=evt_{num}\right)P\left(\sum_{i=0}^n\delta_i=evt_{num}\right)\\&=\sum_{evt_{num}=0}^n P\left(\sum_{i=1}^n T_i\geq t_{\text{eff cut for }evt_{num}}\middle|\sum_{i=0}^n\delta_i=evt_{num}\right)P\left(\sum_{i=0}^n\delta_i=evt_{num}\right)\\&=\sum_{evt_{num}=0}^n P\left(\sum_{i=1}^n T_i\geq t_{\text{eff cut for }evt_{num}}\right)P\left(\sum_{i=0}^n\delta_i=evt_{num}\right).\end{aligned}$$

同理我们可以算出某个λ_1给出Nogo信号的概率为

$$P\left(\text{Declare Nogo}\right)=\sum_{evt_{num}=0}^n P\left(\sum_{i=1}^n T_i\leq t_{\text{fut cut for }evt_{num}}\right)P\left(\sum_{i=0}^n\delta_i=evt_{num}\right).$$

4.2 考虑data cutoff的OC计算

待添加，当前基于模拟计算。