Bayes Statistics

A Basic Lecture Note

作者：Jianqi Huang

组织：Central University of Finance and Economics

时间：Mar, 2023

版本：2.0

ElegantLaTeX Program

这本书所使用的模版是在 elegant book

目录

第 1 章 Introduction

1.1 贝叶斯统计的若干概念

定义 1.1

在参数空间上的 Θ 上的任意概率分布都称为是先验分布。通常以 $\pi(\theta)$ 来表示随机变量 θ 的概率函数。

定义 1.2

获得样本 X 后的分布称为后验分布。在给定 $X = x$ 下的 θ 条件分布记为 $\pi(\theta|x)$，在有密度下，密度函数为

$$\pi(\theta|x) = \frac{h(x,\theta)}{m(x)} = \frac{f(x|\theta)\pi(\theta)}{\int_\Theta \pi(\theta)d\theta}$$

其中 $h(x,\theta) = f(x|\theta)\pi(\theta)$，为 X 和 θ 的联合密度。

$$m(x) = \int_\Theta h(x,\theta) = \int_\Theta f(x|\theta)\pi(\theta)d\theta$$

称为是 X 的边缘概率密度。

定义 1.3

似然函数：在给定联合样本值 X 下关于未知参数 θ 的函数：$L(\theta|x) = f(x|\theta)$。

这里的 $f(x\theta)$ 就是一个密度函数，也就是在给定 θ 的联合样本值。因此从定义上来看，似然函数与密度函数是两个完全不同的数学对象。前者是关于的函数，后者是关于 $\boldsymbol{x}$ 的函数。这里的等号只能理解为数值上的等价，不能在定义上划等。

笔记 竖线"|" 表示的是条件分布。";" 将两个参数分割开。一般情况下 $f(x|\theta)$ 写为 $f(x;\theta)$.

1.2 选择先验

1.2.1 直方图法

将参数空间划分为一些小区间，在每一个小区间上决定主观概率或按照历史数据计算频率；绘制直方图，再画一条光滑曲线。这个曲线就是先验 $\pi(\theta)$ 的一个近似估计。

1.2.2 相对似然法

定义 1.4 (超参数)

先验分布中的参数就是超参数。

若我们确定了先验分布的形式，进一步来确定参数的大小，可以使用的方法有

- 矩估计法
- 分位数法

1.3 边缘分布确定先验分布

定义 1.5

假设随机变量 X 有概率函数 $f(x|\theta)$，θ 有先验分布 $F^{\pi}(\theta)$ 其中的概率函数为 $\pi(\theta)$，定义随机变量 X 的边缘分布

$$m(x)=\int_{\Theta} f(x|\theta)dF^{\pi}(\theta)$$

♣

当先验中有未知超参数 λ 时，记 $\pi(\theta)=\pi(\theta|\lambda)$ 则边缘分布依赖于 λ，此时可记 $m(x)=m(x|\lambda)$。

1.3.1 边缘分布的统计意义解释

定义 1.6

混合分布，一个随机变量 X 分别从 F_1,F_2 中抽取样本，其概率为 $p,1-p$。这个混合分布函数为 $F(x)=pF(x|\theta_1)+(1-p)F(x|\theta_2)$，用概率来表示

$$f(x)=pf(x|\theta_1)+(1-p)f(x|\theta_2)$$

称 $F(x)$ 为 $F(x|\theta_1)$ 和 $F(x|\theta_2)$ 的混合分布。p 和 1-p 可认为是随机变量 θ 的分布。即 $P(\theta=\theta_1)=p=\pi(\theta_1)$，$P(\theta=\theta_2)=1-p=\pi(\theta_2)$。从混合中抽取容量 n 的样本，那么将约有 $n\pi(\theta_1)$ 个样本抽自于总体。

♣

从这个定义中，我们可看见，边缘分布 $m(x)$ 是混合分布的推广。

1.3.2 选择先验分布的 ML-II 方法

当我们观测到先验分布 π_1 和 π_2 的时候，使得

$$m(x|\pi_1)>m(x|\pi_2)$$

我们可以认为选择 π_1 的似然比会高于 π_2。更有可能从 π_2 中产生。

定义 1.7

设 Γ 为所考虑的先验类，若存在 $\hat{\pi}\in\Gamma$，有样本 $\boldsymbol{x}=(x_1,x_2,\cdots,x_n)$ 后，使得

$$m(x|\hat{\pi})=\sup_{\pi\in\Gamma} m(\boldsymbol{x}|\pi)=\sup_{\pi\in\Gamma}\prod_{i=1}^{n} m(x_i|\pi)$$

则称其为 Γ 中的最大似然先验，或简称为 ML-II 先验。

♣

1.3.3 选择先验的矩估计法

当先验分布 $\pi(\theta|\lambda)$ 的形式已知，但含有的未知超参数 λ 时候，可以利用的先验分布的矩与边缘分布的矩之间的关系来寻求超参数 λ 的估计量 $\hat{\lambda}$。

步骤：计算样本分布 $f(x|\theta)$ 的期望和方差：

$$\mu(\theta)=E^{X|\theta}(X),\quad \sigma^2(\theta)=E^{X|\theta}[X-\mu(\theta)]^2$$

此处的 $E^{X|\theta}$ 表示为在给定 θ 的条件下关于样本分布的期望。第二步需要计算边缘密度 $m(x)=m(x|\lambda)$ 的期望 $\mu_m(\lambda)$ 和方差 $\sigma_m^2(\lambda)$，即

$$\mu_m=E^{X|\lambda}(X)=\int_{\mathfrak{X}} xm(x\lambda)dx=\int_{\mathfrak{X}}\int_{\Theta} xf(x|\theta)\pi(\theta|\lambda)d\theta dx=E^{\theta|\lambda}[\mu(\theta)] \tag{1.1}$$

1.4 无信息先验分布

贝叶斯分析中的重要特点在于统计推断时候利用先验信息，但可能常常出现没有先验信息的情况或者信息较少，这时候需要使用**无信息先验**，即对参数空间没有的先验信息。

1.4.1 Laplace 先验与广义先验分布

定义 1.8 (Laplace 先验)

假设随机变量 $X \sim f(x|\theta), \theta \in \Theta$，若 θ 的先验密度 $\pi(\theta)$ 满足条件：

- $\pi(\theta) \geq 0$ 且 $\int_\Theta \pi(\theta) = \infty$
- 后验密度 $\pi(\theta|x)$ 是正常的密度函数。

就称 $\pi(\theta)$ 为 θ 的广义先验密度。

笔记 即使给一个常数使得 $c\pi(\theta)$ 仍然是一个广义先验密度。

定义 1.9

假设检验：$H_0 : \theta \in \Theta \leftrightarrow H_1 : \theta \in \Theta$ 此时 $\theta_0 \cap \theta_1 = \Theta$ 获得 θ 的后验分布。

定义 1.10 (行为空间)

$d = d(x)$ 表示的是所采取的决策活动；检验中 $\mathcal{D} = \{d_0, d_1, \cdots\}$ d_0 表示的是接受原假设 H_0，d1 表示拒绝原假设。

定义 1.11

损失函数：$\theta \times \mathcal{D}$ 非负可测。损失函数 $L(\theta, d)$ 损失越小，表示决策越优。

1.4.2 位置参数的无信息先验

定义 1.12

假设总体 X 的密度函数有 $f(x-\theta)$，其样本空间 $\mathcal{X}$ 和参数空间 Θ 都是 R，此类称为位置参数。$\theta \in \Theta$

比如在 $X \sim N(\theta, \sigma^2)$，其中 σ^2 已知，则

$$\frac{1}{\sqrt{2\pi}\sigma} exp\{-\frac{1}{2\sigma^2}(x-\theta)^2\} = f(x-\theta)$$

属于位置参数族，其中的 θ 为位置参数。

位置参数具有平移变换群下的不变性。对 X 作平移得到 $Y = X + c$，同时对 θ 也作平移变换得到 $\eta = \theta + c$，显然 Y 的密度函数有形式 $f(y-\eta)$ 仍然是位置参数族中。此时对于此的先验密度 $\pi(\theta) = \equiv 0$，是一个广义的先验密度。

1.4.3 刻度参数无信息先验

定义 1.13 (刻度参数族)

假设总体 X 的密度函数有 $\sigma^{-1}\phi(x/\sigma)$ 的形式，其中 $\sigma > 0$ 为刻度参数。参数空间 $R^+ = (0, \infty)$ 则此类密度函数构成的分布称为刻度参数族 (scale parameter family)。

对于尺度参数，实际上不过是在图像上进行收缩变换。得到 $Y=CX$，参数 σ 也同样进行变换 $\eta=c\sigma$，变换系数 c>0。显然仍然是属于这个尺度参数族的。即对于任意的 $a,b,0<a<b,c>0$，σ 落在 $[a,b]$ 上的先验概率，应当等于 η 落在 $[ca,cb]$ 上的概率。不难看出的是，只有当先验为 $\pi=1/\sigma$ 时候成立。

1.4.4 杰弗里斯先验

对于一个分布族可能都不是尺度参数也不是位置参数。假设分布族 $\{f(x|\boldsymbol{\theta})\}$ 满足 Cramer-Rao 正则条件，其中 $\boldsymbol{\theta}=(\theta_1,\cdots,\theta_n)$ 是 p 维的参数向量。正则条件为五条。杰弗里斯先验：写出总体的自然对数

$$l(\boldsymbol{\theta},x)=l(\boldsymbol{\theta})=\ln[f(x|\boldsymbol{\theta})]$$

求出 Fisher 信息矩阵

$$I(\boldsymbol{\theta})=[I_{ij}(\boldsymbol{\theta})]_{p\times p},I_{ij}(\theta)=E^{x|\theta}(\frac{\partial l}{\partial\theta_i}\times\frac{\partial l}{\partial\theta_j})i,j=1,\cdots,p$$

这的表示对总体密度 $f(x|\theta)$ 求期望。参数向量 θ 的无信息先验密度：

$$\pi(\boldsymbol{\theta})=\sqrt{det[I(\boldsymbol{\theta})]}$$

1.5 共轭先验分布

1.5.1 概念

定义 1.14

假设 $\mathcal{F}$ 表示从 θ 的先验分布中构成的分布族。若对于任取的 $\pi\in\mathcal{F}$ 及样本值 x，后验分布 $\pi(\theta|x)$ 仍然属于 $\mathcal{F}$，则称这个是一个共轭先验分布。 ♣

也就是样本的先验分布与后验密度函数 $\pi(\theta|x)$ 同属于一个分布。也称 $\pi(\theta)$ 是参数 θ 的共轭先验分布。

例题 1.1 $N\sim(\theta,\sigma^2)$ 的方差为 σ^2 的共轭先验。假设 $\boldsymbol{x}=(x_1,x_2,\cdots,x_n)$ 来自正态分布，此样本密度为

$$p(\boldsymbol{x}|\sigma^2)=(\sqrt{2\pi\sigma})^{-n}exp\{-\frac{1}{2\sigma^2}\sum(x_i-\theta)^2\}\propto(\frac{1}{\sigma^2})^{n/2}exp\{-\frac{1}{2\sigma^2}\sum(x_i-\theta)^2\}$$

这里的密度显然可以看为是似然函数。其中的 σ^2 的因式决定了 σ^2 的共轭先验分布形式。若有一个分布的核有这种形式，这个分布与似然函数的乘积也有相同的形式，进而推导出这个分布是共轭先验。现假设随机变量 X 服从 $Gamma(\alpha,\lambda)$ 分布，其中 $\alpha>0$ 称为形状参数，$\lambda>0$ 称为位置参数。密度函数

$$p(x|\alpha,\lambda\frac{\lambda^\alpha}{\Gamma(\alpha)}x^{\alpha-1}e^{\lambda x},x>0$$

再令 $Y=X^{-1}$ 可得 Y 的密度函数

$$p(y|\alpha,\lambda)=\frac{\lambda^\alpha}{\Gamma(\alpha)}(\frac{1}{y})^{\alpha+1}exp(\frac{-\lambda}{y})$$

这个分布称为倒 Γ 分布。

1.5.2 后验分布计算实例

$$\pi(\theta|x)=\frac{f(\boldsymbol{x}\pi(\theta))\pi(\theta)}{m(\boldsymbol{x})}\propto f(\boldsymbol{x}|\theta)\pi(\theta)$$

正比于：可以使我们只关注等式左右两侧相关的因子，一些常数可直接省略。

比如 $X\sim B(n,\theta)$，若取 θ 的先验分布为 $Be(a,b)$，求 θ 的后验分布。似然函数的核是 $\theta^x(1-\theta)^{n-x}$，先验密度的核是 $\theta^{a-1}(1-\theta)^{b-1}$，因此有

$$\pi(\theta|x)\propto f(x|\theta)\pi(\theta)\propto\theta^{x+a-1}(1-\theta)^{n-x+b-1}$$

可清晰的看出右侧为贝塔分布 $Be(x+a,n-x+b)$ 的核。因此添加正则化因子得到的后验密度

$$\pi(\theta|x)=\frac{\Gamma(n+a+b)}{\Gamma(x+a)\Gamma(n-x+b)}\theta^{(x+a)-1}(1-\theta)^{(n-x+b)-1},0<\theta<1$$

1.5.3 共轭先验优点

方便计算、后验分布的某些参数可以很好的解释。

1.6 分层先验

定义 1.15

当给定的先验分布的超参数难以确定时候，可以对超参数再给出一个先验、第二个先验称为超先验。若超先验的超参数仍然难以确定，再给出一个先验。由先验和超先验决定的一个新先验称为分层先验。

$$X|\theta\sim f(x|\theta)$$
$$\theta|\lambda\sim\pi_1(\theta|\lambda)$$
$$\lambda\sim\pi_2(\lambda)$$

其中 $\mathcal{X}$ 表示的是样本空间，Θ 为参数空间，Λ 为超参数空间。$\pi_2(\lambda)$ 常取无信息先验。

其中 $\mathcal{X}$ 表示样本空间。

笔记 任何一个分层先验都可以写成一个规范的先验，一个二层先验，这个规范先验是

$$\pi(\theta)=\int_v arLambda\pi_1(\theta|\lambda)\pi_2(\lambda)d\lambda=\int_\Lambda\pi(\theta,\lambda)d\lambda$$

第 2 章 常见的统计模型参数的后验分布

一切统计推断都是从后验分布出发，后验的计算较为关键。

2.1 后验分布与充分性

当先验分布有密度时候的后验分布的计算公式

$$h(x,\theta)=f(x|\theta)\pi(\theta)$$

样本的边缘密度为:

$$m(x)=m(x|\pi)=\int_{\Theta}f(x|\theta)\pi(\theta)d\theta$$

以公式可知

$$\pi(\theta|x)=\frac{h(x,\theta)}{m(x)}=\frac{f(x|\theta)\pi(\theta)}{m(x)}=\frac{f(x|\theta)\pi(\theta)}{\int_{\Theta}f(x|\theta)\pi(\theta)d\theta}$$

因为 $m(x)$ 仅为与 x 相关的函数，因此可认为其为一个常数，将上述公式转化为

$$\pi(\theta|x)$$

定义 2.1 (充分统计量)

$X\sim f(x|\theta),\theta\in\Theta,X=(X_1,X_2,\cdots,X_n)$ 是从总体中获得的独立同分布的样本。$T=T(X)$ 是一个统计量，若在 $T=t$ 给定的情况下，X 的条件分布于 θ 无关，那么就认为其是一个充分统计量。一个比较好的判断充分统计量的方式是因子分解定理。

♣

引理 2.1

假设 $X\sim f(x|\theta),\theta\in\Theta$，此时的 $f(x|\theta)$ 为随机变量 X 的概率密度函数，$X=(X_1,X_2,\cdots,X_n)$ 是从总体抽取的 iid 样本。$T=T(X)$ 是统计量，其密度函数为 $q(t|\theta)$，若充分，则 $\forall\pi\in\Gamma$ 有

$$\pi(\theta|x)=\widetilde{\pi}(\theta|t)$$

也就是对于样本的 X 分布，与基于充分统计量的后验得到的估计是相同的。

♡

2.2 正态总体参数的后验分布

假设 $X_1,X_2,\cdots,X_n i.i.d\sim N(\theta,\sigma^2)$，记 $X=(X_1,X_2\cdots,X_n)$，若给定的 $\varphi=(\theta,\sigma^2)$ 时样本 X 的联合概率分布为

$$\begin{aligned}f(x|varphi)&=(2\pi\sigma^2)^{-\frac{n}{2}}exp\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\theta)^2\}\\&=(2\pi\sigma^2)^{-\frac{n}{2}}exp\{-\frac{1}{2\sigma^2}[\sum_{i=1}^{n}(x_i-\overline{x})^2+n(\bar{x}-\theta)^2]\}\end{aligned}$$

定义 2.2

若随机变量 Y 具有下的概率密度

$$p(y|v,\mu,\tau^2)=\frac{\Gamma(\frac{v+1}{2})}{\Gamma(\frac{v}{2})\sqrt{v\pi}}\cdot\frac{1}{\tau}\cdot[1+\frac{1}{v}(\frac{y-\mu}{\tau})^2]^{-\frac{v+1}{2}}$$

则称其为广义一元 t 分布。

♣

2.2.1 无信息先验下的后验分布

1. 当 σ^2 已知时，均值参数的后验分布 $\bar{X}=\frac{1}{n}\sum_{i=1}^{n}X_i$ 在总体 $N(\mu,\sigma)$ 中可以得到 θ 的似然函数为

$$l(\theta|\bar{x})=\sqrt{\frac{n}{2\pi\sigma^2}}exp\{-\frac{n}{2\sigma^2}(\bar{x}-\theta)^2\}\propto exp\{-\frac{n}{2\sigma^2}(\bar{x}-\theta)^2\}$$

由 $\pi(\theta)\equiv 1,\theta\in R$ 可得到 θ 的后验密度为

$$\pi(\theta|x)\propto l(\theta|\bar{x})\pi(\theta)\propto exp\{-\frac{n}{2\sigma^2}(\theta-\bar{x})^2\}$$

这个是非标准化的，属于一个核。因此需要添加正则化常数因子得到即 θ 的后验分布为 $N(\bar{x},\sigma^2/n)$

当 σ^2 和 μ 都未知时候，给定 x，(θ,σ^2) 的似然函数：

$$l(\theta,\sigma^2|x)\propto(\sigma^2)6-\frac{n}{2}exp\{-\frac{1}{2\sigma^2}[vs^2+n(\theta-\bar{x})^2]\}$$

此处的 $v=n-1,s^2=\frac{1}{v}\sum_{i=1}^{n}(x_i-\bar{x})^2$

假设 θ 和 σ 的无信息先验分别为 $\pi_1(\theta)\equiv 1$ 和 $\pi_2(\sigma^2)=1/\sigma^2$

2.3 共轭先验下的后验分布

1. 当 σ^2 已知，均值参数 θ 的后验分布

$$l(\theta|\bar{x})=\sqrt{\frac{n}{2\pi\sigma^2}}exp\{-\frac{n}{2\sigma^2}(\bar{x}-\theta)^2\}\propto exp\{-\frac{n}{2\sigma^2}(\bar{x}-\theta)^2\}$$

假设 θ 的共轭先验分布为 $N(\mu,\tau^2)$，其密度函数

$$\pi(\theta)=\frac{1}{\sqrt{2\pi}\tau}exp\{-\frac{(\theta-\mu)^2}{2\tau^2}\}\propto exp\{-\frac{(\theta-\mu)^2}{2\tau^2}\}$$

记 $\sigma_n^2=\sigma^2/n$ 则我们就有

其中

$$A=\frac{1}{\sigma^2}+\frac{1}{\tau^2},B=\frac{\bar{x}}{\sigma_n^2}+\frac{\mu}{\tau^2},C=\frac{\bar{x}^2}{\sigma_n}+\frac{\mu^2}{\tau^2}$$

2. 当 θ 已知时候，参数 σ^2 的后验分布

令 $T=\sum_{i=1}^{n}(X_i-\theta)^2$，则给定 θ 时候，T 为 σ 的充分统计量。$T/\sigma^2\sim\chi_n^2$，给定 $T=t$，由其可得到

$$l(\sigma^2|t)\propto(\sigma^2)^{-n/2}exp\{-\frac{t}{2\sigma^2}\}$$

根据前给出的参数 σ^2 的无信息先验

$$\pi(\sigma^2)\propto\frac{1}{\sigma^2}$$

因此 σ^2 的后验分布

$$\pi(\sigma^2|t)\propto l(\sigma^2|t)\pi(\sigma^2)\propto(\sigma^2)^{-(\frac{n}{2}+1)}exp\{-\frac{t}{2\sigma^2}\}$$

2.4 一类离散分布和多项式分布

假设离散随机变量 X 的概率分布有以下形式：

$$f(x|\theta)=P(X=x|\theta)=h(x)\theta^{b(x)}(1-\theta)^{d(x)}\tag{2.1}$$

其中的 $b(x)$ 和 $d(x)$ 分别取非负整数，此类包含几个常见分布

1. 0-1 分布 $B(1,\theta)$，$h(x)=1,b(x)=x,d(x)=1-x$，即 $f(x|\theta)=P(X=x|\theta)=\theta^x(1-\theta)^{1-x}$
2. 二项分布 $B(n,\theta):h(x)=(nx),b(x)=x,d(x)=n-x$
3. 几何分布

4. 负二项分布

2.4.1 参数的先验为无信息先验的后验分布

由式 (2.1) 可知 θ 的似然函数

$$l(\theta|x) \propto \theta^{b(x)}(1-\theta)^{d(x)}$$

先验为 $\pi(\theta) \equiv 1$ 时候，θ 的后验分布为：

$$\pi(\theta|x) \propto \theta^{b(x)}(1-\theta)^{d(x)}$$

上式中为 Beta 分布的核，添加正则化因子后可以得到：

$$\pi(\theta|x) = \frac{\Gamma(b(x)+d(x)+2)}{\Gamma(b(x)+1)\Gamma(d(x)+1)}\theta^{b(x)+1}(1-\theta)^{(d(x)+1)-1} \tag{2.2}$$

其中 $0 < \theta < 1$，θ 的后验分布式贝塔分布 $Be(b(x)+1, d(x)+1)$。

2.4.2 参数的先验分布为共轭先验的后验分布

令 θ 的共轭分布为贝塔分布 $Be(\alpha, \beta)$，其密度函数为

$$\pi(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1} \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

其后验分布为：

$$\pi(\theta|x) \propto l(\theta|x)\pi(\theta)$$

因此通过上述的推导我们可得到

- θ 的无信息先验为 $\pi(\theta) \equiv 1$，则 θ 的后验分布为 $Be(b(x)+1, d(x)+1)$
- 若 θ 的共轭先验由 $Be(\alpha, \beta)$ 给出，则 θ 的后验为 $Be(b(x)+\alpha, d(x)+\beta)$

2.5 多项分布的后验分布

假设 $X = (X_1, X_2, \cdots, X_k)$ 服从多项分布 $M(n, \theta)$，其中的 $X_i \geq 0$，$\sum_{i=1}^{k} X_i = n, \theta = (\theta_1, \theta_2, \cdots, \theta_k), \theta_i \geq 0, \sum_{i=1}^{k}\theta_i = 1$ 独立参数只有 k-1，因此 X 的概率分布为

$$p(x|\theta) = P(X_1 = x_1, X_2 = x_2, \cdots, X_k = x_k|\theta) \tag{2.3}$$

$$= \frac{n!}{x_1!x_2!\cdots x_k!}\left(\prod_{i=1}^{k-1}\theta_i^{x_i}\right)\left(1-\sum_{i=1}^{k-1}\right)^{x_k} \tag{2.4}$$

2.6 无信息下的后验分布

当 σ 已知下的均值参数 θ 的后验分布记 $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$，在 $N(\theta, \sigma^2)$ 总体中，当 σ 已知，则 $T = \bar{X}$ 为充分统计量。可得 $\bar{X} \sim N(\theta, \sigma^2/n)$，故 θ 的似然函数为

$$l(\theta|\bar{x}) = \sqrt{\frac{n}{2\pi\sigma^2}}exp\{-\frac{n}{2\sigma^2}\} \propto exp\{-\frac{n}{2\sigma^2}(\bar{x}-\theta)\}$$

第 3 章 贝叶斯推断

从统计模型中参数的贝叶斯分析的两种方式：

1. 从 θ 后验出发，考虑 θ 的贝叶斯推断问题，不考虑损失
2. 考虑损失，用统计决策方法来考虑 θ 的贝叶斯分析问题。

3.1 条件方法和原理

3.1.1 条件方法

后验分布式在给定样本 x 下的 θ 的条件分布，基于后验分布的统计推断只有考虑了已经出现的数据，认为未出现的数据并没有做出贡献。经典统计学家认为参数 θ 的无偏估计 $\hat{\theta}(X)$ 应满足

$$E[\theta(\hat{X})] = \int_{\mathfrak{X}} \hat{\theta}(x)p(x|\theta)dx = \theta \tag{3.1}$$

求平均是对于所有的样本空间出现的样本求得的。实际上在不少的估计量只使用 1 次或几次，多数为出现的样本也要参与平均就较为难以理解。

3.1.2 似然原理

似然原理的核心是似然函数，在从总体中抽取独立同分布的样本之后，其联合分布 $f(x|\theta) = f(x_1, x_2, \cdots, x_n|\theta)$ 。当我们固定 x 时候，把 $f(x|\theta)$ 看为是一个 θ 的函数，称为似然函数，记

$$L(\theta|x) = f(x|\theta) = \prod_{i=1}^{n} f(x_i|\theta) \tag{3.2}$$

似然函数强调的是 θ 的函数，样本 x 是在似然函数中给出了一组观测值。在这组观测值之下使得似然函数取值更大的就更为可能是 θ 的真值。特别的，当使得 $L(\theta|x)$ 在参数空间 Θ 取值达到最大的 θ 之后的 $\hat{\theta}(x)$ 称为是最大似然估计。假如两个似然函数成比例，比例因子又并不依赖于 θ，则他们的 MLE 是相同的。

3.2 贝叶斯点估计

从不同的统计量中考虑：一个后验概率 $\pi(\theta|\boldsymbol{x})$ 的众数称为参数 θ 的后验众数估计。中位数也就对应于后验中位数估计；期望也即后验期望估计，这些都是称为点估计 $\hat{\theta_\beta}$

3.2.1 贝叶斯点估计的误差

定义 3.1

假设参数 θ 的后验分布为，其中是已知样本，则 $(\theta - \hat{\theta})^2$ 后验期望

$$PMSE(\hat{\theta}) = E^{\theta|x}(\theta - \hat{\theta})^2 = E[(\theta - \hat{\theta})^2|\boldsymbol{x}]$$

这个就称为 $\hat{\theta}$ 的后验均方差。其平方根称为是 $\hat{\theta}$ 的标准误。 ♣

其中的 PMSE 越小越好，若 $\mu^\pi(x)$ 为 θ 的后验均值，特别当 $\delta(x) = E(\theta|x) = \mu^\pi(x)$ 时候，则 $\delta(x)$ 的 PMSE 为后验方差，即

$$PMSE(\delta(x)) = E^{\theta|x}[(\theta - \mu^\pi(x))^2] = V^\pi(x) \tag{3.3}$$

其中的 $V^{\pi}(x)$ 是 θ 的后验方差。对 θ 的任一估计 $\delta(x)$，其后验均方误差 $PMSE(\delta(x))$ 与它的后验方差 $V^{\pi}(x)$ 的关系

$$\begin{aligned} PMSE(\delta(x)) &= E^{\theta|x}[(\theta-\delta(x))^2] \\ &= E^{\theta|x}[(\theta-\mu^{\pi}(x))+(\mu^{\pi}(x)-\delta(x))]^2 \\ &= V^{\pi}(x)+[\mu^{\pi}(x)-\delta(x)]^2 \geq V^{\pi}(x) \end{aligned} \tag{3.4}$$

其中等号成立的充分条件为 $\delta(x)=\mu^{\pi}(x)$，即 θ 的后验期望估计使得 $PMSE$ 达到最小。

3.2.2 多参数情形

若 $\theta=(\theta_1,\theta_2,\cdots,\theta_n)$，则后验众数估计：从后验分布中用广义最大似然估计来获得的后验密度作为后验众数估计；后验期望估计：$\mu^{\pi}(x)=E^{\theta|x}(\theta)$，估计量的精度用后验协方差举证来衡量

$$COve^{\pi}(x)=E^{\theta|x}[(\theta-\mu^{\pi}(x))(\theta-\mu^{\pi}(x))]$$

3.3 区间估计

对于区间估计，当 θ 的后验分布获得后，就可以求出其落入某个区间 [a,b] 的概率 $1-\alpha$ 的估计。

$$P(a\leq\theta\leq b|x)=\int +a^{b}\pi(\theta|x)d\theta=1-\alpha$$

定义 3.2 (可信区间)

θ 的后验分布 $\pi(\theta|x)$，对于给定的概率 $1-\alpha$，集合 C 满足如下的条件

$$P(\theta\in C|x)=\int_{C}\pi(\theta|x)d\theta=1-\alpha$$

对于任意给定的 $\theta_1\in C$ 和 $\theta_2\notin C$，总会存在 $\pi(\theta_1|x)>\pi(\theta_2|x)$ 称 C 为 θ 下的最大后验密度可信集。$1-\alpha$HPD 可信集。

3.3.1 泊松分布参数的估计

$$p(x|\lambda)=\frac{\lambda^{x}}{x!}e^{-\lambda}x=0,1,2\cdots$$

后验分布

$$\pi(\lambda|x)\propto p(x|\lambda)\pi(\lambda)\propto$$

参数估计当柏松分布的均值取 $Gamma(\alpha,\beta)$ 作为共轭先验。

3.4 假设检验

假设检验是统计推断的重要问题，其具体步骤如下：

对于问题来给出原假设 H_0，和备择假设 H_1，将假设检验问题写成

$$H_0:\theta\in\Theta_0\leftrightarrow H_1:\theta\in\Theta_1$$

其中的 Θ_0 为参数空间 Θ 的非空真子集。同时满足互斥性：$\Theta_1=\Theta-\Theta_0$。

3.4.1 贝叶斯因子

定义 3.3

假设两个的先验概率 π_0 和 π_1。后验概率分别为 α_0 和 α_1。比例 α_0/α_1 表示的含义是后验机会比。而 π_0/π_1 表示的是先验机会比。则贝叶斯因子 (Baysian factor):

$$B^{\pi}(x) = \frac{\alpha_0/\alpha_1}{\pi_0/\pi_1}$$

$B^{\pi}(x)$ 的取值越高，表示对 H_0 的支出度越高。

$$\frac{\alpha_0}{\alpha_1} = \frac{\pi_0 f(x|\theta_0)}{\pi_1 f(x|\theta_1)} \tag{3.5}$$

若要拒绝原假设，也就是要求 $\alpha_1/\alpha_1 < 1$ 成立，由 (3.5) 可以得到

$$\frac{f(x|\theta_1)}{f(x|\theta_0)} > \frac{\pi_0}{\pi_1}$$

也即要求两个密度函数值之比要大于临界值，与 NP 引理的基本结果类似。

3.4.2 简单假设对复杂假设

$$H_0 : \theta = \theta_0 \leftrightarrow H_1 : \theta \neq \theta_0 \tag{3.6}$$

检验问题 (3.6) 是经典统计的一类常见问题，当参数 θ 为连续变量时候，用简单假设是不够合理的。因此往往是用一个检验区间来考虑。

假设样本分布为 $f(x|\theta)$，容易求出边缘分布

$$m(x) = \int_{\Theta} f(x|\theta)\pi(\theta)d\theta = \pi_0 f(x|\theta_0) + \pi_1 m_1(x)$$

3.4.3 多重假设检验

$$H_i : \theta \in \Theta_i, i = 1, 2, \cdots, k \tag{3.7}$$

其中 $\Theta_1 \cap \Theta_2 \cdots \cap_k = \Theta$，每一个都是非空子集。计算后验概率

$$\alpha_i = P(\Theta|x), i = 1, 2, \cdots, k$$

若存在一个 α_{i0} 最大，则接受假设 $i0$。

3.5 预测推断

3.5.1 贝叶斯预测分布

定义 3.4

假设 $X \sim f(x|\theta)$，$X = (X_1, X_2, \cdots, X_n)$，从总体中抽样获得的历史数据。Z 的未来预测值 Z_0 的后验预测密度 (posterior predictive density) 定义为

$$p(z_0|x) = \int_{\Theta} g(z_0|\theta)\pi(\theta|x)d\theta$$

3.6 模型选择

$M_0 : X$ 有密度 $f(x|\theta)$，其中 $\theta \in \Theta_0$
$M_1 : X$ 有密度 $f(x|\theta)$，其中 $\theta \in \Theta_1$

3.6.1 贝叶斯模型评价

1. 评价的重要性贝叶斯模型的基本想法是指定抽样分布和所有未知的参数先验，贝叶斯模型的任何推断是基于后验分布所进行的。贝叶斯推断的结果依赖于指定的模型。

2.AIC 和 BIC 准则最大似然原理：AIC 准则

$$AIC = -2\ln f(x_n|\hat{\theta_{MLE}}) + 2p$$

其中的 $\hat{\theta_{MLE}}$ 为 θ 的最大似然估计。p 是参数向量的维数，2p 为惩罚项。最优模型可通过最小化 AIC 得到。

3. 贝叶斯预测信息准则 (BPIC)

$$BPIC = -2\int_{\Theta} \ln f(x_n|\theta)\pi(\theta|x_n)d\theta + 2p$$

p 为模型中的参数个数。

4. 偏差信息准则 (DIC) 令 $D(\theta) = -2\ln f(x_n|\theta)$，常用于度量偏差的方式

$$p_D = \bar{D} - D(\bar{\theta}) = 2\ln f(x|\bar{\theta_n}) - 2\int_{\Theta} \ln f(x_n|\theta)\pi(\theta|x_n)d\theta$$

第 4 章 贝叶斯统计决策

决策实际上是一个过程，可划分为两个部分，第一个部分是将决策问题进行描述。第二个部分是将如何决策使得收益最大化的过程进行表示。

4.1 后验风险最小原则

假设 $L(\theta,\delta(x))$ 为损失函数，损失函数进行样本求分布求平均就可得到风险函数。（也就是用期望来定义风险函数）损失函数按照后验分布 $\pi(\theta|x)$ 求期望就可以得到后验风险。

定义 4.1

假设 $\pi(\theta|x)$ 为 θ 的后验分布，$L(\theta,\delta(x))$ 为损失函数，则

$$R(\delta(x)|x)=E[L(\theta,\delta(x))]=\begin{cases}\int_{\Theta}L(\theta,\delta(x))\pi(\theta|x)d\theta,\theta\text{为连续型变量}\\ \sum_{i}L(\theta_i,\delta(x))\pi(\theta_i|x),\theta\text{为离散型随机变量}\end{cases}$$

称为决策函数 $\delta(x)$ 的后验风险。

若存在一个决策函数 $\delta^*(x)\in\mathcal{D}$，使得对任意决策函数 $\delta(x)\in\mathcal{D}$，有

$$R(\delta^*(x)|x)=\min_{\delta\in\mathcal{D}}R(\delta(x)|x)$$

则称这个是后验风险最小准则下的最优贝叶斯决策函数。

4.1.1 后验风险与贝叶斯风险关系

$f(x,\theta)=f(x|\theta)\pi(\theta)=\pi(\theta|x)m(x)$ 其中的 $R_\pi(\delta(x))$ 改写为：

$$\begin{aligned}R_\pi(\delta(x))&=E^\theta[R(\theta,\delta(x))]\\&=\int_\Theta R(\theta,\delta(x))\pi(\theta)d\theta\\&=\int_\Theta\left[\int_{\mathfrak{X}^n}L(\theta,\delta(x))f(x|\theta)dx\right]\pi(\theta)d\theta\\&=E^X[R(\delta(X)|X)]\end{aligned}$$

从该公式可看出，贝叶斯风险决策的两种表达式：

$$R_\pi(\delta(x))=E^\theta[R(\theta,\delta(x))]=E^X[R(\delta(\boldsymbol{X})|\boldsymbol{X})]$$

也就是两种积分顺序：先对 θ 求积分，再对 X 的绝对分布 $m(x)$ 求均值（积分）。

4.1.2 后验风险最小的原则

定理 4.1

假设存在决策函数 $\delta(x)\in\mathfrak{D}$ 对于任意的决策函数 $\delta(x)\in\mathfrak{D}$ 使得

$$R(\delta(x)|x)=\inf_{\delta\in\mathfrak{D}}R(\delta(x)|x)=\inf\int_\Theta L(\theta,\delta(x))\pi(\theta|x)d\theta$$

则 $\delta_\pi(x)$ 为先验分布下的贝叶斯解。

证明 假设 $\delta \in \mathfrak{D}$ 为任一决策函数，由已知条件可知：

$$R(\delta(x)|x) = \int_{\Theta} L(\theta, \delta(x))\pi(\theta|x)d\theta \geq \int_{\Theta} L(\theta, \delta_\pi(x))\pi(\theta|x)d\theta = R(\delta_\pi(x)|x)$$

对两侧求积分可得到：

$$R_\pi(\delta(x)) = \int_{\mathfrak{X}} R(\delta(x)|x)m(\boldsymbol{x})dx \geq R_\pi(\delta_\pi(x))$$

定义 4.2

若 $\pi(\theta)$ 为广义先验，所求得的最优决策函数称为广义的贝叶斯解。 ♣

例题 4.1 通过血液检测的例子，假设 θ_1 患者患病，θ_2 患者不患病，$X = 1, 0$ 分别表示化验结果为阳性和阴性。

$$p(X = 1|\theta_1) = 0.8, p(X = 0|\theta_1) = 0.2$$
$$p(X = 1|\theta_2) = 0.1, p(X = 0|\theta_1) = 0.9$$

损失函数 $L(\theta, a)$ 如下：

由上述说明可得到我们参数 θ 的后验分布

$$\pi(\theta_1|x = 0) = 0.012, \pi(\theta_2|x = 0) = 0.988$$
$$\pi(\theta_1|x = 1) = 0.296, \pi(\theta_2|x = 1) = 0.704$$

进一步对后验风险进行刻画：

$$\begin{aligned} R(a_1|x = 0) &= E^{\theta|x}[L(a_1, \theta)] \\ &= L(a_1, \theta_1) \times \pi(\theta_1|x = 0) + L(a_1, \theta_2) \times \pi(\theta_2|x = 0) \\ &= 0 \times 0.012 + 4 \times 0.988 = 3.952 \end{aligned}$$

类似的可得到

$$R(a_2|x = 0) = 10 \times 0.012 + 0 \times 0.988 = 0.12,$$
$$R(a_3|x = 0) =$$

同理：当 $x = 1$ 时，可计算得到

$$R(a_1|x = 1) = 2.816$$
$$R(a_2|x = 1) = 2.96$$
$$R(a_3|x = 1) = 3.184$$

X 所表达的是后验所得到的类别能对先验的修正，同样还是一个随机事件，而我们想要知道的是在这个随机事件给定的情况下得到决策函数的后验风险的期望值。

4.2 一般损失函数的贝叶斯估计

在先前所讨论的损失函数是一个线性的函数，在这里进一步考虑一个连续的损失函数。

定理 4.2

在平方损失函数下，$L(\theta, a) = (\theta - a)^2$,$theta$ 的贝叶斯估计为后验期望值。

$$\hat{\theta}_B(\boldsymbol{x}) = E(\theta|x) = \int_{\Theta} \theta\pi(\theta|x)d\theta$$

♡

证明略。

加权的平方损失函数:

$$\widehat{\theta_B} = \frac{E(\theta w(\theta)|x)}{E(w(\theta)|x)}$$

绝对值的损失函数 (abosolute error loss function)

$$L(\theta, a) = |a - \theta|$$

下 θ 为贝叶斯估计的后验中位数。

定理 4.3 (线性损失函数)

$$L(\theta, a) = \begin{cases} k_0(\theta - a) \\ k_1(a - \theta) \end{cases}$$

后验风险最小准则的贝叶斯估计为 $k_0/(k_0 + k_1)$ 的分位数。 ♡

第 5 章 贝叶斯统计计算方法

在贝叶斯统计方法中，常常需要计算后验期望分布的期望、方差、分位数等数字特征。常用的后验均值，是平方损失下的贝叶斯估计，此估计量的精度是在后验方差来度量的。后验众数、中位数等常常作为贝叶斯估计的可信区间。一些后验分布并没有很好的数字特征显示表达，需要一些特殊的计算方法。

5.1 蒙特卡洛抽样方法

MCMC 方法是常常需要计算一些后验期望等数字特征时候。

$$\pi(\theta|x) = \frac{p(x|\theta)\pi(\theta)}{\int p(x|\theta)\pi(\theta)\pi(\theta)d\theta}$$

我们所感兴趣的是函数 $h(\theta)$ 的后验期望：

$$E[(h(\theta))|x] = \int h(\theta)\pi(\theta|x)d\theta = \frac{\int h(\theta)p(x|\theta)\pi(\theta)d\theta}{\int p(x|\theta)\pi(\theta)d\theta} \tag{5.1}$$

5.1.1 蒙特卡洛抽样

若 (5.1) 没有显式表达，除了可以使用分析学的逼近或数值积分的方法外，还可以考虑使用蒙特卡洛抽样方法。若可以从后验分布 $\pi(\theta|x)$ 中产生独立同分布的观测值 $\theta_1, \theta_2, \cdots, \theta_m$，由大数定理可得到：

$$\bar{h_m} = \frac{1}{m}\sum_{i=1}^{m}(h(\theta_i)) \tag{5.2}$$

几乎处处收敛到 $E[h(\theta)|x]$，这一结果保证了样本量 m 足够大的时候可以使用 $\bar{h_m}$ 来作为 $E[h(\theta)|x]$ 的估计。而估计量 (5.2) 被称为积分的蒙特卡洛逼近 (Monte Carlo approximation)。这种估计量 (5.2) 取毕节积分的方法称为是蒙特卡洛抽样 (Monte Carlo sampling) 方法。

5.1.2 蒙特卡洛重要性抽样方法

5.2 MCMC

5.3 Metropolis-Hastings 算法

5.4 Gibbs 抽样方法