第九章 概率模型

第 27 讲 统计决策理论

黄定江

DaSE @ ECNU
djhuang@dase.ecnu.edu.cn

27.1.1 比较风险函数

一般使用风险函数比较两个估计。然而，这并不能提供一个明确的答案说哪一个估计更好。考虑下面的例子。

例 1

令 $X \sim N(0,1)$，假设使用平方损失函数。考虑两个估计 $\widehat{\theta}_1 = X$ 和 $\widehat{\theta}_2 = 3$ 。风险函数为 $R\left(\theta, \widehat{\theta}_1\right) = \mathbb{E}_\theta(X-\theta)^2 = 1$ 和 $R\left(\theta, \widehat{\theta}_2\right) = \mathbb{E}_\theta(3-\theta)^2 = (3-\theta)^2$。如果 $2 < \theta < 4$，则 $R\left(\theta, \widehat{\theta}_2\right) < R\left(\theta, \hat{\theta}_1\right)$，否则，$R\left(\theta, \widehat{\theta}_1\right) < R\left(\theta, \hat{\theta}_2\right)$ 。没有哪一个估计一定比另一个好，见图*1*。

27.1.1 比较风险函数举例

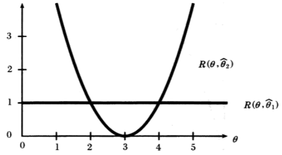

图 1: 比较两个风险函数

例 2

令 $X_1,\cdots,X_n\sim Bernoulli(p)$。考虑平方损失函数，令 $\widehat{p}_1=\bar{X}$。由于这是无偏的，就有

$$R\left(p,\widehat{p}_1\right)=\mathbb{V}(\bar{X})=\frac{p(1-p)}{n}$$

另一个估计为

$$\widehat{p}_2=\frac{Y+\alpha}{\alpha+\beta+n}$$

27.1.1 比较风险函数举例

其中，$Y=\sum_{i=1}^{n} X_i, \alpha, \beta$ 为正常数。这是使用先验 Beta (α, β) 的后验均值。现在，

$$\begin{aligned}
R(p, \widehat{p}_2) &= \mathbb{V}_p(\widehat{p}_2) + (\mathrm{bias}_p(\widehat{p}_2))^2 \\
&= \mathbb{V}_p\left(\frac{Y+\alpha}{\alpha+\beta+n}\right) + \left(\mathbb{E}_p\left(\frac{Y+\alpha}{\alpha+\beta+n}\right) - p\right)^2 \\
&= \frac{np(1-p)}{(\alpha+\beta+n)^2} + \left(\frac{np+\alpha}{\alpha+\beta+n} - p\right)^2
\end{aligned}$$

令 $\alpha=\beta=\sqrt{n/4}$ 得到的估计为：

$$\widehat{p}_2 = \frac{Y+\sqrt{n/4}}{n+\sqrt{n}}$$

风险函数为：

$$R(p, \widehat{p}_2) = \frac{n}{4(n+\sqrt{n})^2}$$

27.1.1 贝叶斯风险

这些例子说明了风险函数需要进行比较。为此，需要用一个数来描述这个风险函数。最大风险和贝叶斯风险就是采用这种形式定义的。

定义 1

最大风险为

$$\bar{R}(\widehat{\theta}) = \sup_{\theta} R(\theta, \widehat{\theta})$$

贝叶斯风险为

$$r(f, \widehat{\theta}) = \int R(\theta, \widehat{\theta}) f(\theta) \mathrm{d}\theta$$

其中，$f(\theta)$ 是 θ 的先验。

27.1.1 贝叶斯风险举例

例 3

再次考虑例2中的两个估计。得出

$$\bar{R}\left(\widehat{p}_1\right)=\max_{0\leqslant p\leqslant 1}\frac{p(1-p)}{n}=\frac{1}{4n}$$

和

$$\bar{R}\left(\widehat{p}_2\right)=\max_{p}\frac{n}{4(n+\sqrt{n})^2}=\frac{n}{4(n+\sqrt{n})^2}$$

因为 $\bar{R}\left(\widehat{p}_2\right)<\bar{R}\left(\widehat{p}_1\right)$，根据最大风险，$\hat{p}_2$ 是更好的估计。然而，当 n 很大时，除了在接近 $p=\frac{1}{2}$ 的参数空间对应的小区域内，$\bar{R}\left(\widehat{p}_1\right)$ 的风险要比 $\bar{R}\left(\widehat{p}_2\right)$ 小。因此，许多人宁愿选择 $\hat{p}_1$ 而不是 $\hat{p}_2$。这说明了像最大风险这种单个数对风险函数的描述并不是完美的。现在考虑贝叶斯风险。为了说明，令 $f(p)=1$，则

27.1.1 贝叶斯风险举例

$$r(f, \widehat{p}_1) = \int R(p, \widehat{p}_1)\,\mathrm{d}p = \int \frac{p(1-p)}{n}\,\mathrm{d}p = \frac{1}{6n},$$

并且

$$r(f, \widehat{p}_2) = \int R(p, \widehat{p}_2)\,\mathrm{d}p = \frac{n}{4(n+\sqrt{n})^2}$$

对于 $n \geq 20$，有 $r(f, \widehat{p}_2) > r(f, \widehat{p}_1)$，这表明了 $\hat{p}_1$ 是一个比较好的估计。从直觉上看比较合理，但是这个答案取决于先验的选择。尽管最大风险也有不足，但它的优点是不需要选择先验。

这两种风险函数的描述表明了设计估计的两种不同方法：选择使最大风险最小的 $\hat{\theta}$ 得到最小最大估计；选择使贝叶斯风险最小的 $\hat{\theta}$ 得到贝叶斯估计。

27.2.1 引言

令 f 是一先验。根据贝叶斯定理，后验密度为

$$f(\theta \mid x) = \frac{f(x \mid \theta) f(\theta)}{m(x)} = \frac{f(x \mid \theta) f(\theta)}{\int f(x \mid \theta) f(\theta) \mathrm{d}\theta}$$

其中，$m(x) = \int f(x, \theta) \mathrm{d}\theta = \int f(x \mid \theta) f(\theta) \mathrm{d}\theta$ 是 X 的边际分布。定义估计 $\hat{\theta}(x)$ 的后验风险为

$$r(\widehat{\theta} \mid x) = \int L(\theta, \widehat{\theta}(x)) f(\theta \mid x) \mathrm{d}\theta$$

27.2.1 贝叶斯估计

定理 1

贝叶斯风险 $r(f,\widehat{\theta})$ 满足

$$r(f,\widehat{\theta}) = \int r(\widehat{\theta} \mid x)m(x)\mathrm{d}x$$

令 $\hat{\theta}(x)$ 是使得 $r(\widehat{\theta} \mid x)$ 最小的 θ 值，则 $\hat{\theta}$ 是贝叶斯估计。

27.2.1 贝叶斯估计

证明.

可以把贝叶斯风险改写为

$$\begin{aligned} r(f,\widehat{\theta}) &= \int R(\theta,\widehat{\theta})f(\theta)\mathrm{d}\theta = \int\left(\int L(\theta,\widehat{\theta}(x))f(x\mid\theta)\mathrm{d}x\right)f(\theta)\mathrm{d}\theta \\ &= \iint L(\theta,\widehat{\theta}(x))f(x,\theta)\mathrm{d}x\,\mathrm{d}\theta = \iint L(\theta,\widehat{\theta}(x))f(\theta\mid x)m(x)\mathrm{d}x\,\mathrm{d}\theta \\ &= \int\left(\int L(\theta,\widehat{\theta}(x))f(\theta\mid x)\mathrm{d}\theta\right)m(x)\mathrm{d}x = \int r(\widehat{\theta}\mid x)m(x)\mathrm{d}x \end{aligned}$$

如果选择 $\hat{\theta}(x)$ 为使得 $r(\widehat{\theta}\mid x)$ 最小的 θ 值，那么就能使被积函数在每一个 x 都最小，因此使得积分 $r(\widehat{\theta}\mid x)m(x)\mathrm{d}x$ 最小。 □

27.2.1 贝叶斯估计

定理 2

如果 $L(\theta,\widehat{\theta})=(\theta-\widehat{\theta})^2$，则贝叶斯估计为

$$\widehat{\theta}(x)=\int \theta f(\theta \mid x)\mathrm{d}\theta=\mathbb{E}(\theta \mid X=x)$$

如果 $L(\theta,\widehat{\theta})=|\theta-\widehat{\theta}|$，则贝叶斯估计为后验 $f(\theta|x)$ 的中位数。如果 $L(\theta,\widehat{\theta})$ 是 *0-1* 损失，贝叶斯估计为后验 $f(\theta|x)$ 的众数。

证明.

下面将证明这个定理中损失函数为平方损失的情况。贝叶斯规则 $\hat{\theta}(x)$ 使得 $r(\widehat{\theta} \mid x)=\int(\theta-\widehat{\theta}(x))^2 f(\theta \mid x)\mathrm{d}\theta$ 最小。对 $r(\widehat{\theta} \mid x)$ 关于 $\hat{\theta}(x)$ 求导，并让它等于 0，得到 $2\int(\theta-\widehat{\theta}(x))f(\theta \mid x)\mathrm{d}\theta=0$。解方程得到定理2。 □

27.2.1 贝叶斯估计举例

例 4

令 $X_1, \cdots, X_n \sim N\left(\mu, \sigma^2\right)$。这里 σ^2 已知，假设用 $N(a, b^2)$ 作为 μ 的先验。根据平方损失的贝叶斯估计为后验均值，即

$$\widehat{\theta}\left(X_1, \cdots, X_n\right) = \frac{b^2}{b^2+\sigma^2/n}\bar{X} + \frac{\sigma^2/n}{b^2+\sigma^2/n}a$$

27.3.1 最小最大规则

求最小最大规则比较复杂，在这里并不能全面讲述这一理论，但会提到几个关键结果。这一节传达的主要信息就是：常数风险函数的贝叶斯估计是最小最大估计。

定理 3

令 $\widehat{\theta}^f$ 是某一先验 f 的贝叶斯规则，

$$r\left(f, \widehat{\theta}^f\right) = \inf_{\widehat{\theta}} r(f, \widehat{\theta})$$

假设对所有的 θ，有

$$R\left(f, \widehat{\theta}^f\right) \leqslant r\left(f, \widehat{\theta}^f\right)$$

则 $\widehat{\theta}^f$ 是最小最大估计，f 称为最不利先验。

27.3.1 最小最大规则

定理 4

假设 $\hat{\theta}$ 是基于先验 f 的贝叶斯估计。进一步假设 $\hat{\theta}$ 的风险为常数 c:$R(\theta, \widehat{\theta}) = c$，则 $\hat{\theta}$ 是最小最大的。

证明略。

27.3.1 最小最大规则

例 5

再次考虑 *Bernoulli* 模型，但是它的损失函数为

$$L(p,\hat{p}) = \frac{(p-\hat{p})^2}{p(1-p)}$$

令

$$\widehat{p}\left(X^n\right) = \widehat{p} = \frac{\sum_{i=1}^n X_i}{n}$$

风险为

$$R(p,\hat{p}) = \mathbb{E}\left(\frac{(p-\hat{p})^2}{p(1-p)}\right) = \frac{1}{p(1-p)} \frac{p(1-p)}{n} = \frac{1}{n}$$

这里，它作为 p 的函数，是一个常数。可以证明，对于这个损失函数，$\widehat{p}\left(X^n\right)$ 是在先验 $f(p)=1$ 下的贝叶斯估计。因此，$\hat{p}$ 是最小最大的。

27.3.1 最小最大规则

定理 5

令 $X_1, \cdots, X_n \sim N(0,1)$，且令 $\hat{\theta} = \bar{X}$，则 $\hat{\theta}$ 是关于任意优良的损失函数的最小最大规则。它是具有这种性质的唯一估计。

对于满足弱正则性条件的参数模型，极大似然估计近似最小最大估计。考虑平方损失函数，它是偏差的平方加上方差。在大样本的参数模型中，可以证明方差项远远大于偏差项，所有极大似然估计 $\hat{\theta}$ 约等于方差

$$R(\theta, \widehat{\theta}) = \mathbb{V}_{\theta}(\hat{\theta}) + \mathrm{bias}^2 \approx \mathbb{V}_{\theta}$$

极大似然估计的方差近似为

$$\mathbb{V}(\hat{\theta}) \approx \frac{1}{nI(\theta)}$$

其中,$I(\theta)$ 是 Fisher 信息量。因此，

$$nR(\theta, \widehat{\theta}) \approx \frac{1}{I(\theta)}$$

对于任意其他估计 $\theta^{'}$，可以证明对于足够大的 n，有 $R\left(\theta, \theta'\right) \geqslant R(\theta, \widehat{\theta})$。更精确地，

$$\lim_{\epsilon \to 0} \limsup_{n \to \infty} \sup_{|\theta - \theta'| < \epsilon} nR\left(\theta', \widehat{\theta}\right) \geqslant \frac{1}{I(\theta)}$$

这说明在局部大样本的情况下，极大似然 MLE 是最小最大的。可以证明 MLE 近似是贝叶斯规则。

总之，在绝大多数大样本参数模型中，MLE 是近似最小最大的和贝叶斯规则。