STAT201B Lecture 4 Bootstrap (2)

1 Bootstrap

> **Logic**
>
> Bootstrap 是一种 computer-intensive method, 当无法求出问题的 analytical solution 时, 可以用 Bootstrap 估计 measures of uncertainty
>
> 本章节将介绍 Bootstrap 的基本思路和几个应用
>
> 关于 Bootstrap 的详细论述, 见 STA3020 Lecture 12-13 与 STA4100 Lecture 12-13

1.1 Bootstrap 的用法一: 估计统计量的方差

情境设置:

- 我们有数据 $X_1, \dots, X_n$, 可以计算统计量 $T_n = g(X_1, \dots, X_n)$
- 我们希望估计 T_n 的方差 $V_F[T_n]$ (无法得到解析解)

思路:

- 若 F 已知, 则可以使用 Monte Carlo Integration 来估计 $V_F[T_n]$
- 但现实中 F 未知, 因此我们可以**先使用 empirical CDF $\hat{F}_n$ 来估计 F, 随后再使用 Monte Carlo Integration 估计 $V_{\hat{F}_n}[T_n]$** (需要对 $\hat{F}_n$ 进行 sampling)

对 $\hat{F}_n$ 进行 (re)sampling:

- 对 observations $X_1, \dots, X_n$ 进行随机取样
- 采用 sampling **with replacement**

具体算法:

1. 重复以下步骤 B 次, 以得到 $T^*_{n,1}, \dots, T^*_{n,B}$:
   - (Re)sampling $X^*_1, \dots, X^*_n \sim \hat{F}_n$
   - 计算 $T^*_n = g(X^*_1, \dots, X^*_n)$
2. 使用 MC integration 来估计 $V_{\hat{F}_n}(T_n)$, 即:

$$v_{bootstrap} = \hat{V}_{\hat{F}_n}(T_n) = \frac{1}{B}\sum_{j=1}^{B}\left(T^*_{n,j} - \frac{1}{B}\sum_{k=1}^{B}T^*_{n,k}\right)^2$$

构建 T_n 的置信区间:

1. 方法一: 使用 **Normal-based interval**:

$$C_n = T_n \pm z_{\alpha/2} \cdot \hat{se}_{bootstrap} = T_n \pm z_{\alpha/2} \cdot \sqrt{v_{bootstrap}}$$

> **Remark**
>
> 仅当 T_n 的分布接近 Normal 时有效, 需要注意 T_n 的 asymptotic normality 是一个关于 n 的性质, 无法通过控制 B 的大小来保证

2. 方法二: 使用 **Quantile interval**:

$$C_n = (T^*_{\alpha/2}, T^*_{1-\alpha/2})$$

其中 T^*_β 是 bootstrap sample $T^*_{n,1}, \dots, T^*_{n,B}$ 的 β quantile

1.2 Bootstrap 的用法二: 估计 bias

情境设置:

- 我们有数据 $X_1, \dots, X_n \sim F_0$; $F_1 := \hat{F}_n$ 为对应的 empirical distribution; $\theta(F_1)$ 为 $\theta(F_0)$ 的 plug-in estimator
- 我们希望估计 plug-in estimator 的 bias $t_0 = \mathbb{E}_{F_0}[\theta(F_0) - \theta(F_1)]$

思路与方法:

- 先从 F_1 中 (即 $X_1, \dots, X_n$ 中) 进行 (re)sampling, 得到 $Y_1, \dots, Y_n$ 与对应的 empirical CDF F_2
- 注意到 bias t_0 的表达式中 F_0 未知, 因此我们可以考虑用 F_1 来估计 F_0, 用 F_2 来估计 F_1, 即

$$\hat{t}_0 = \mathbb{E}_{F_1}[\theta(F_1) - \theta(F_2)]$$

Example: μ^2 的估计量的 debias

问题设置:

- 我们有数据 $X_1, \dots, X_n \overset{i.i.d.}{\sim} F_0$ with mean μ and variance σ^2
- 我们希望估计 $\theta(F_0) = (\mathbb{E}_{F_0}[X])^2 = \mu^2$
- 我们使用 empirical plug-in estimator $\theta(F_1) = (\mathbb{E}_{F_1}[Y])^2 = \bar{X}^2$, 其中 $Y \sim F_1$

解题过程:

注意到 plug-in estimator $\theta(F_1)$ 为 biased, 且 bias 为

$$\begin{aligned} t_0 &= \mathbb{E}_{F_0}[\theta(F_0) - \theta(F_1)] \\ &= \theta(F_0) - \mathbb{E}_{F_0}[\theta(F_1)] \\ &= \mu^2 - \left(\mu^2 + \frac{\sigma^2}{n}\right) \\ &= -\frac{\sigma^2}{n} \end{aligned}$$

为了 debias, 我们考虑以下 estimator (为了估计 bias t_0, 我们用 F_1 来估计 F_0, 用 F_2 来估计 F_1):

$$\begin{aligned} \tilde{\theta} &= \theta(F_1) + \hat{t}_0 \\ &= \theta(F_1) + [\theta(F_1) - \mathbb{E}_{F_1}[\theta(F_2)]] \\ &= 2\theta(F_1) - \mathbb{E}_{F_1}[\theta(F_2)] \end{aligned}$$

注意到此处 $\mathbb{E}_{F_1}[\theta(F_2)]$ 可以直接求出 (无需借助 Monte Carlo integration):
由于 $Y_1, \dots, Y_n \overset{i.i.d.}{\sim} F_1 = \hat{F}_0$ (由 $X_1, \dots, X_n$ 构成的分布), 因此满足: $\mathbb{E}[Y_i] = \bar{X}$, $V(Y_i) = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2$;
由于 $\theta(F_2) = (\mathbb{E}_{F_2}[Z])^2 = (\bar{Y})^2$ (令 $Z \sim F_2$), 因此有

$$\mathbb{E}_{F_1}[\theta(F_2)] = (\mathbb{E}_{F_1}[(\bar{Y})^2]) = (\bar{X})^2 + \frac{1}{n}\left(\sum_{i=1}^n \frac{(X_i - \bar{X})^2}{n}\right)$$

因此:

$$\tilde{\theta} = 2\theta(F_1) - \mathbb{E}_{F_1}[\theta(F_2)] = (\bar{X})^2 - \frac{1}{n}\left(\sum_{i=1}^n \frac{(X_i - \bar{X})^2}{n}\right)$$

Remark

可以验证该 estimator 的 bias 是否减少:

$$\begin{aligned} E_{F_0}(\tilde{\theta}) &= \left(\mu^2 + \frac{\sigma^2}{n}\right) - \mathbb{E}_{F_0}\left[\sum_{i=1}^n \frac{(X_i - \bar{X})^2}{n^2}\right] \\ &= \mu^2 + \frac{\sigma^2}{n} - \frac{n-1}{n^2}\sigma^2 \\ &= \mu^2 + \frac{\sigma^2}{n^2} \end{aligned}$$

注意到 $\mathbb{E}_{F_0}[\theta(F_1)] = \mathbb{E}_{F_0}[\bar{X}^2] = \mu^2 + \sigma^2/n$, 因此 bias 确实减少了

1.3 Bootstrap 的用法三: Pivotal intervals

Parametric statistics 中的 pivot:

- Pivot 为一个 function $R(X_1, \dots, X_n, \theta)$, 其**分布与 θ 无关**
- 它使得我们可以在不知道 θ 的情况下构建 $R_n = R(X_1, \dots, X_n, \theta)$ 的置信区间, 并通过简单的变换得到 θ 的置信区间

Nonparametric statistics 中的 pivot:

- 在 nonparametric statistics 设定下, 我们通常无法得到一个 exactly pivotal 的 quantity (完全不取决于任何可能的未知 F)

Nonparametric statistics 中 location parameter 的 confidence interval:

若 $\theta = T(F)$ 为一个 location parameter, 则 $R_n = \hat{\theta}_n - \theta$ 通常会是 approximately pivotal

> ⚠ **Remark**
>
> - 这主要是由于在大样本下, location parameter 的 estimator 通常会趋向正态分布
> - 对于 scale parameter / shape parameter, 它们的分布会更加复杂

若 R_n 的 CDF H, 则我们可以构建一个 $1-\alpha$ confidence interval for θ of (a, b), 其中

$$
\begin{aligned}
a &= \hat{\theta}_n - H^{-1}(1-\alpha/2) \\
b &= \hat{\theta}_n - H^{-1}(\alpha/2)
\end{aligned}
$$

但现实中我们不知道 H, 因此我们考虑使用 bootstrap samples 来构建 empirical CDF

$$
\hat{H}(r) = \frac{1}{B} \sum_{i=1}^{n} \mathbf{1}(R_{n,j}^* \le r)
$$

其中 $R_{n,j}^* = \hat{\theta}_{n,j}^* - \hat{\theta}_n$, 因此, $H^{-1}(1-\alpha/2)$ 和 $H^{-1}(\alpha/2)$ 的 estimator 即为这些 samples 的 $1-\alpha/2$ 和 $\alpha/2$ quantiles, 因此 $1-\alpha$ bootstrap pivotal interval 为

$$
C_n = (2\hat{\theta} - \hat{\theta}_{1-\alpha/2}^*, 2\hat{\theta} - \hat{\theta}_{\alpha/2}^*)
$$