STAT201B Lecture 3 Bootstrap (1)

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

1.1 Bootstrap 的类型

1. Nonparametric bootstrap
   主要使用以下两个重要概念
   - Monte Carlo integration
   - Empirical CDF
2. Parametric bootstrap

1.2 Monte Carlo Integration

Monte Carlo integration 的原理:

Monte Carlo integration 基于以下 approximation (见 STAT201B Lecture 2 Non-Parametric Inference 中 linear functional 的 plug-in estimator):

$$\begin{aligned}\mathbb{E}[h(Y)] &= \int h(y)dF_Y(y) \\ &\approx \frac{1}{B}\sum_{j=1}^{B} h(Y_j)\end{aligned}$$

其中 $Y_1, \ldots, Y_B \overset{i.i.d.}{\sim} F_Y$

Monte Carlo integration 的效果

令 $\mathbb{E}[|h(Y)|] < \infty$, 则当 $B \to \infty$ 时, 有

$$\frac{1}{B}\sum_{j=1}^{B} h(Y_j) \overset{a.s.}{\to} \mathbb{E}[h(Y)]$$

Remark

- 通常我们可以控制 B 足够大, 使得 approximation 变得足够好

Example

A simple example: Use Monte Carlo integration to approximate

$$\int_{-\infty}^{\infty} \sin^2(x)e^{-x^2}dx$$

Solution: We can write this as $\sqrt{\pi}\int_{-\infty}^{\infty}\sin^2(x)f(x)dx$, where $f(x)$ is the PDF of a $N(0, 1/2)$ r.v. Therefore, we can

1. Draw $Y_1, \ldots, Y_B \overset{iid}{\sim} N(0, 1/2)$.

```
> B <- 10000; y <- 1/sqrt(2) * rnorm(B)
```

2. Approximate $\sqrt{\pi}\int_{-\infty}^{\infty}\sin^2(x)f(x)dx \approx \frac{\sqrt{\pi}}{B}\sum_{j=1}^{B}\sin^2(Y_j)$.

```
> sqrt(pi) * mean(sin(y)^2)
[1] 0.5509956
```

⚠ Remark

Normal distribution 的 PDF 为 $\frac{1}{\sqrt{2\pi\sigma^2}} exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$

1.3 Important Sampling

Logic

Important sampling 是 Monte Carlo sampling 的变种, 我们可以从一个 "importance function" g 而不是 target density h 进行采样, 使得估计的效率和准确性更高

场景

假设我们要估计下面这个积分:

$$I = \int f(x)\, h(x)\, dx$$

其中 $h(x)$ 是某个概率分布 (比如标准正态分布 $N(0,1)$，而 $f(x)$ 在某些极端值上特别大

比如:

$$I = \mathbb{E}_{X\sim N(0,1)}\left[e^X\right]$$

方法 1: 普通 Monte Carlo Sampling

- 我们从 $N(0,1)$ 抽样: $x_1, x_2, \ldots, x_n$
- 估计值为:

$$\hat{I}_{MC} = \frac{1}{n}\sum_{i=1}^{n} e^{x_i}$$

问题:

- e^x 在**大正数**时增长很快
- 但从标准正态 $\mathcal{N}(0,1)$ 抽样时, 大部分样本都在 $[-2,2]$ 间, 几乎抽不到大于 5 的数
- 结果就是: 估计会偏向于 "低估", 要补偿只能疯狂增加样本量

方法 2: Importance Sampling

思路: 与其从 $\mathcal{N}(0,1)$ 抽样, 不如换一个更容易抽到大数的分布, 比如 **右偏的正态分布 $\mathcal{N}(3,1)$**

- 我们先从 $g(x) = \mathcal{N}(3,1)$ 抽样 $x_1, x_2, \ldots, x_n$ 。
- 但因为目标分布是 $h(x) = \mathcal{N}(0,1)$，所以需要加一个**权重修正**:

$$w(x) = \frac{h(x)}{g(x)}$$

- Importance Sampling 的估计变成:

$$\hat{I}_{IS} = \frac{1}{n}\sum_{i=1}^{n} e^{x_i} \cdot w(x_i)$$

效果:

- 因为 $g(x)$ 会经常给我们 "大数" (比如 4, 5, 6)，在这些区域 e^x 特别重要。
- 权重修正后，这些大数的贡献被准确计入，而不需要像 Monte Carlo 那样依赖运气去抽到。

直观类比

- Monte Carlo: 去北京三里屯随便问人年薪, 想估计 "中国平均年薪", 会发现高薪人很少被抽到, 所以估计结果偏低
- Importance Sampling：你专门去金融街抽样 (那里高薪人比较多)，用 "北京 vs 金融街人口比例" 来调整权重，这样能更快更准确地估计平均水平

Importance sampling 的原理:

$$
\begin{aligned}
\mathbb{E}_h[q(\theta)] &= \int q(\theta)h(\theta)d\theta \\
&= \int q(\theta)\frac{h(\theta)}{g(\theta)}g(\theta)d\theta \\
&\approx \frac{1}{B}\sum_{i=1}^{B} q(\theta_i)\frac{h(\theta_i)}{g(\theta_i)}
\end{aligned}
$$

其中 $\theta_1, \dots, \theta_B \overset{i.i.d.}{\sim} g(\theta)$

Remark

关于 expectation 的 notations 可能有一些 confusing, 此处 expectation 的下标 h 表示 θ 的 pdf 为 h

Remark

- 关于 Monte Carlo 和 importance sampling, 更详细的论述见 STA4042 Statistical Learning Lecture 6 Basic Probability Tools
- 关于 $g(x)$ 的选取, 可以遵循 **Rubenstein Theorem** (见 STA4042 Statistical Learning Lecture 6 Basic Probability Tools):
  对于 measurable mapping $q : \mathbb{R} \to \mathbb{R}$ 和 density $f : R \to R_+$, 可以 minimize $\frac{q(X)h(X)}{g(X)}$ 的 variance 的 density $g : R \to R_+$ 为:

  $$\frac{q(X)h(X)}{\int_{\mathbb{R}} |q(t)|h(t)dt}$$

  使用 Jensen inequality 易证
- 对于任意 integration, 都可以使用 gaussian kernel 进行 approximation:

  $$
  \begin{aligned}
  \int_{\Omega} r(x)dx &= \int_{-\infty}^{\infty} \frac{I(x \in \Omega)\cdot r(x)}{\frac{1}{\sqrt{2\pi}}e^{-x^2/2}} \frac{1}{\sqrt{2\pi}}e^{x^2/2} \\
  &\approx \frac{1}{B}\sum_{i=1}^{B} \frac{I(X_i \in \Omega)r(X_i)}{\frac{1}{\sqrt{2\pi}}e^{-x^2/2}}
  \end{aligned}
  $$

  其中 $X_1, X_2, \dots, X_B \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$