Johns Hopkins Engineering

625.464 Computational Statistics

Introduction to Bootstrapping

Module 9 Lecture 9A

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

Motivation for the Bootstrap

T estimates θ a parameter of distribution F

- want characteristics of dist of T
- ideally have multiple samples from $F \Rightarrow$ multiple T_i to allow us to obtain this information
- unrealistic $\rightarrow$ so how can we get more information from our single sample?

Bootstrap's Main Idea - Resampling

Resampling methods generally involve the use of many samples each drawn from a single sample from F

For bootstrap — the basic idea is that the observed sample should contain all of the information about the underlying population (F) & so the observed sample is considered to be the population.

Bootstrapping - More Formally

- Let $\theta = T(F)$ - can be expressed as a functional of F. Usually

Ex

$$T(F) = \int g(z)\, dF(z)$$

$$T(F) = \int z\, dF(z) \leftarrow \text{mean}$$

F is a distribution

- Let $x_1, \dots, x_n$ $X_1, \dots, X_n \sim \text{iid } F$

observed data from density f $\mathcal{X} = \{x_1, \dots, x_n\}$ CDF

Bootstrapping More Fomally

- $\theta = T(F)$ parameter of interest
- $\mathcal{X} = \{x_1, \ldots, x_n\}$ observed data $\sim$ iid F
- Let $\hat{F}$ be the empirical distribution of the observed data. (weight each obs. x_i with probability $1/n$)
- An estimator of θ, $\hat{\theta} = T(\hat{F})$

Ex/ θ is pop mean $\hat{\theta} = \int z \, dF(\hat{z}) = \sum_{i=1}^{n} \frac{x_i}{n}$

Bootstrapping More Formally

Let $\theta = T(F)$. $\mathcal{X} = \{x_1, \ldots, x_n\}$. $\hat{F}$ be empirical dist. and est $\hat{\theta} = T(\hat{F})$.

$$R(\mathcal{X}, F)$$

$R(\mathcal{X}, F)$ could be the bias of $T(\hat{F})$ or the var, etc.

~~Ex~~ $$R(\mathcal{X}, F) = \frac{T(\hat{F}) - T(F)}{S(\hat{F})}$$

where $S(\hat{F})$ is s.d. of $T(\hat{F})$

The Bootstrap Method

Let $\mathcal{X}^* = \{X_1^*, \dots, X_n^*\}$ be iid r.v. drawn from the dist $\hat{F}$. Then $\mathcal{X}^*$ is the bootstrap sample of pseudodata or pseudodata set.

The bootstrap strategy is to examine the dist of $R(\mathcal{X}^*, \hat{F})$ and use it to make inf. about $R(\mathcal{X}, F)$.

sample (X, replacement)

Bootstrap Example

$\mathcal{X} = \{x_1, x_2, x_3\} = \{1, 2, 6\}$ is observed from F. and we wish to est. the mean of F, θ.

$$\boxed{\begin{array}{l} \theta = T(F) \\ \hat{\theta} = T(\hat{F}) \\ \hat{\theta}^* = T(\hat{F}^*) \end{array}}$$

$$\hat{\theta} = T(\hat{F}) = 9/3 = R(\underline{\mathcal{X}}, \underline{F})$$

A bootstrap sample $\mathcal{X}^* = \{x_1^*, x_2^*, x_3^*\}$ will be drawn iid from $\hat{F}$.

$3^3 = 27$ possible $\mathcal{X}^*$

Ex/ $\mathcal{X}^* = \{1, 1, 6\}$ $\quad \hat{F}^*$ put $2/3$ at 1, $1/3$ at 6

$$\hat{\theta}^* = T(\hat{F}^*) = 8/3$$

$\mathcal{X} = \{1, 2, 6\}$

Bootstrap Example

$P^*[\hat{\theta}^* \le 6/3] = 8/27$

$\mathcal{X}^*$	$\hat{\theta}^*$	$P^*[\hat{\theta}^*]$
1,1,1	3/3	1/27
1,1,2	4/3	3/27
1,2,2	5/3	3/27
2,2,2	6/3	1/27
1,1,6	8/3	3/27
1,2,6	9/3	6/27
2,2,6	10/3	3/27
1,6,6	13/3	3/27
2,6,6	14/3	3/27
6,6,6	18/3	1/27

- For each $\mathcal{X}^*$ we calculate $\hat{\theta}^*$
- For each $\hat{\theta}^*$ we calculate the $P^*[\hat{\theta}^*]$ w.r.t the bootstrap experiment of drawing $\mathcal{X}^*$ conditional on $\mathcal{X}$.

$R(\mathcal{X}^*, \hat{F}) = \hat{\theta}^*$

$\hat{\theta} = R(\mathcal{X}, F)$