B O O T S T R A P (Clássico e Bayesiano)

Nívea Bispo

06 de maio de 2016

Sumário

- Efron, B. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics. vol. 7, p. 1-26, 1979*.
- Rubin, D. B. The Bayesian Bootstrap. *The Annals of Statistics, vol. 9, p. 130-134, 1981.*
- Efron B, Tibshirani R. An introduction to the Bootstrap. New York: Chapman & Hall, 1993.
- Davison, A. C., Hinkley, D. V. Bootstrap Methods and their Application. Cambridge Series in Statistical and Probabilistic Mathematics, 1997.
- Chernick, M. R. Bootstrap Methods: A Guide for Practitioners and Researchers. John Wiley & Sons, 2nd Edition, 2007.
- Chernick, M. R., Labudde, R. A. An Introduction to Bootstrap Methods with Applications to R. John Wiley & Sons, 2011.

- Método não-paramétrico de reamostragem que se baseia no pressuposto de que a amostra é uma boa representação da população;
- Tem como objetivo determinar as propriedades da distribuição do estimador de certo parâmetro sem fazer muitas suposições sobre a forma da distribuição que gerou os dados;
- É um método para melhorar estimadores, sendo frequentemente utilizado na estimação de erros-padrão, viés e na construção de IC's;
- Há dois tipos: Não-Paramétrico e o Paramétrico.

Ideia:

Suponha que F é a fda da população de interesse. Como fazer inferência para F?
No caso mais simples os dados observados são tratados como uma realização de uma a.a. $X_1, \ldots, X_n$ da distribuição F, e uma estimativa de F, baseada na amostra, é usada para inferir sobre F. O mais usual é utilizar a distribuição empírica como estimativa:

$$\hat{F} = \frac{\sharp\{X_i \leq X\}}{n}$$

$\hat{F}$ atribui pesos iguais para cada ponto amostral, e assume valores no conjunto $\{0, 1/n, 2/n, \ldots, n/n\}$.

Quando utilizá-lo:

- Quando existe, ou não, um modelo probabilístico bem definido para os dados;
- Quando a distribuição teórica de determinada estatística é complicada ou desconhecida;
- O tamanho amostral é insuficiente para fazer inferência sobre a quantidade de interesse;

Visão Geral

- Seja $X_1, \ldots, X_n$ uma a.a. de X, onde X possui f.d.a. F. Uma estatística T(X) é escolhida para estimar o parâmetro θ que indexa a distribuição F;
- A distribuição bootstrap de $T(X)$ é gerada a partir de B replicações de X_i.
- Existem duas maneiras de realizar o bootstrap:
  - Não-Parametricamente: Não conhecemos a forma de F, e estimamos-a a partir da distribuição empírica obtida com os dados.
  - Parametricamente: F pertence a uma família paramétrica de distribuições, e seus parâmetros são estimados a partir da amostra original.
- O bootstrap não-paramétrico é mais robusto com relação às hipóteses distribucionais, enquanto que o bootstrap paramétrico é esperado ser mais eficiente quando as suposições paramétricas são verdadeiras.

"Estimador" Bootstrap do erro-padrão:

O erro-padrão estimado pelas réplicas bootstrap é dado por:

$$\widehat{ep}_B = \left[\frac{\sum_{b=1}^{B} (\theta^{*(b)} - \theta^*)^2}{B - 1} \right]^{1/2}$$

Correção Bootstrap do vício:

A correção bootstrap para o vício é dada por:

$$\tilde{\theta} = 2\hat{\theta} - \theta^*$$

obs.: Na prática a correção de viés pode ser perigosa. Mesmo que $\tilde{\theta}$ seja menos viesado que $\hat{\theta}$, seu erro-padrão pode ser maior. Por isso, vale a pena checar a estimativa de ambos os erros-padrão.

IC's via bootstrap:

- Percentílico;
- BC, BCa, ABC (são intervalos que corrigem o viés);
- Básico, Bootstrap-t.

Algoritmo Bootstrap:

1. Gere uma amostra X^*, com reposição, da distribuição empírica dos dados;
2. Calcule a estimativa $\theta^* = T(X^*)$ para a amostra bootstrap gerada;
3. Repita os passos 1 e 2 B vezes.

Algoritmo Bootstrap:

1. Gere uma amostra X^*, com reposição, da distribuição empírica dos dados;
2. Calcule a estimativa $\theta^* = T(X^*)$ para a amostra bootstrap gerada;
3. Repita os passos 1 e 2 B vezes.

Onde o método falha?

- Estruturas de dependência (o Bootstrap se baseia na suposição de independência);
- Na presença de *outliers*: Nesse caso não iremos amostrar de uma boa estimativa de F [+ variabilidade seria adicionada às estimativas];
- Dados multivariados: Quando a dimensão dos problema é grande, $\hat{F}_n$ deixa de ser uma boa aproximação para F.
- Big Data: Nesse caso o bootstrap torna-se 'inviável' computacionalmente. (Possível solução: 'Parallel Bootstrap' e/ou 'Bags of Little Bootstrap').

Ideia:

Operacionalmente o que difere o BB do método usual é como as probabilidades (pesos) atribuídas a cada X_i são definidas. No BB esses 'pesos' serão assumidos desconhecidos, e uma distribuição a priori será atribuída a eles.

Funcionamento:

Suponha que estamos interessados em inferir sobre a média de certa população. Em cada réplica do BB $T(X) = \overline{X}$ será calculada de maneira analóga ao método usual ($\widehat{T(X)} = \sum_i \omega_i X_i$), contudo, novos ω_i serão gerados a cada nova reamostragem:

$$
\begin{aligned}
b &= 1 \Rightarrow \widehat{T_1(X)} = \sum_i \omega_i^{(1)} X_i^* \\
&\vdots \\
b &= B \Rightarrow \widehat{T_B(X)} = \sum_i \omega_i^{(B)} X_i^*
\end{aligned}
$$

Visão Geral

Seja $X_1, \ldots, X_n$ uma a.a. de X. Considere $d = (d_1, \ldots, d_K)$ como sendo o vetor de todos os possíveis valores distintos de X, e $\theta = (\theta_1, \ldots, \theta_K)$ um vetor de probabilidades, onde $P(X = d_k|\theta) = \theta_k$.

Se n_k é o número de X_i iguais a d_k, e a priori para $\pi(\theta) = \prod_k \theta_k^l$, a densidade a posteriori para θ será uma Dirichlet$(n_1 + l, \ldots, n_K + l)$.

- O BB gera da distribuição a posteriori do parâmetro θ, reamostrando os X_i com os pesos gerados a partir da Dirichlet;
- Assim como no método usual, o BB é sensível aos pressupostos assumidos para determinado modelo.

Algoritmo BB:

1. Gere os pesos ω_i de uma Dirichlet uniforme (Dir (1,1,. . . 1) com a mesma dimensão de X;
2. Gere uma amostra X^*, assumindo os pesos gerados em 1;
3. Calcule a estimativa $\widehat{T(X)}$ para a amostra bootstrap gerada em 2;
4. Repita os passos 1, 2 e 3 B vezes.

Algoritmo BB:

1. Gere os pesos ω_i de uma Dirichlet uniforme (Dir (1,1,... 1) com a mesma dimensão de X;
2. Gere uma amostra X^*, assumindo os pesos gerados em 1;
3. Calcule a estimativa $\widehat{T(X)}$ para a amostra bootstrap gerada em 2;
4. Repita os passos 1, 2 e 3 B vezes.