Introduction to Bootstrap

Preetham Reddy Armoor

Department of Mathematics
University of Applied Sciences, Mittweida

28 November 2016

Outline

Introduction

- Bootstrap is a computer based method for assigning measure of accuracy to statistical estimates.
- Bootstrap was introduced by B.Efron in 1979.

Bootstrap

Bootstrap Sample

A bootstrap sample $x^* = (x_1^*, x_2^*, x_3^* .., x_n^*)$ is obtained by random sampling n times, with replacement,from the original data points $x = (x_1, x_2, x_3, ..., x_n)$.

Bootstrap

Bootstrap Sample

A bootstrap sample $x^* = (x_1^*, x_2^*, x_3^*.., x_n^*)$ is obtained by random sampling n times, with replacement,from the original data points $x = (x_1, x_2, x_3, ..., x_n)$.

Example

Consider a sample $x = (x_1, x_2, x_3, x_4, x_5)$some bootstrap samples can be:

$$x_1^* = (x_2, x_3, x_2, x_4, x_1)$$

$$x_2^* = (x_2, x_1, x_3, x_3, x_1)$$

$$x_3^* = (x_4, x_1, x_2, x_3, x_4)$$

Definition

The Plug-in estimate of a parameter $\theta = t(F)$ is defined to be:

$$\hat{\theta} = t(\hat{F})$$

the function $\theta = t(F)$ of the probability distribution function F is estimated by the same function $t(.)$ of the empirical density $\hat{F}$.

Definition

The Plug-in estimate of a parameter $\theta = t(F)$ is defined to be:

$$\hat{\theta} = t(\hat{F})$$

the function $\theta = t(F)$ of the probability distribution function F is estimated by the same function $t(.)$ of the empirical density $\hat{F}$.

Bootstrap Replication

with each bootstrap sample $x^{*(1)}$ to $x^{(B)}$, we can compute a bootstrap replication $\hat{\theta}^*(b) = s(x^{*(b)})$ using the plug-in principle.

Accuracy of sample estimate

how accurate is $\hat{\theta}$ compared to the real value θ ?

- Standard error.
- Bias.
- Confidence interval.
- etc.

Preliminaries

Standard Error

The standard error is the standard deviation of sampling distribution of statistic $\hat{\theta}$. As such, it measures the precision of an estimate of the statistic of a population distribution.

$$se(\hat{\theta}) = \sqrt{var_F(\hat{F})}$$

Preliminaries

Standard Error

The standard error is the standard deviation of sampling distribution of statistic $\hat{\theta}$. As such, it measures the precision of an estimate of the statistic of a population distribution.

$$se(\hat{\theta}) = \sqrt{var_F(\hat{F})}$$

Standard Error of $\overline{X}$

$$se(\overline{X}) = \sqrt{var_F(\overline{X})} = \frac{\sigma_F}{\sqrt{n}}$$

Preliminaries

Standard Error

The standard error is the standard deviation of sampling distribution of statistic $\hat{\theta}$. As such, it measures the precision of an estimate of the statistic of a population distribution.

$$se(\hat{\theta}) = \sqrt{var_F(\hat{F})}$$

Standard Error of $\overline{X}$

$$se(\overline{X}) = \sqrt{var_F(\overline{X})} = \frac{\sigma_F}{\sqrt{n}}$$

Estimated Standard Error of $\overline{X}$

$$\hat{s}e(\overline{X}) = \frac{\hat{\sigma}}{\sqrt{n}}$$

Preliminaries

Bootstrap Estimated Standard Error

Estimate the standard error $se_F(\hat{\theta})$ by the standard deviation of the B replication:

$$\hat{se}_B = [\frac{\sum_{b=1}^{B}[\hat{\theta}^*(b) - \hat{\theta}^*(.)]^2}{B-1}]$$

where $\hat{\theta}^*(.) = \frac{\sum_{b=1}^{B}[\hat{\theta}^*(b)]}{B}$

Preliminaries

Bias

The Bias is the difference between the expectation of an estimator $\hat{\theta}$ and the quantity θ being estimated:

$$Bias_F(\hat{\theta}, \theta) = E_F(\hat{\theta}) - \theta$$

Preliminaries

Bias

The Bias is the difference between the expectation of an estimator $\hat{\theta}$ and the quantity θ being estimated:

$$Bias_F(\hat{\theta}, \theta) = E_F(\hat{\theta}) - \theta$$

Bootstrap Estimate of Bias

the bootstrap estimate of bias is defined to be the estimate:

$$Bias_{\hat{F}}(\hat{\theta}) = E_{\hat{F}}[S(\mathbf{x}^*)] - t(\hat{F}) = \theta^*(.) - \hat{\theta}$$

Non-Parametric Bootstrap

Algorithm

1. Assume a data set $x = (x_1, x_2, ..., x_n)$ is available.
2. Fix the number of bootstrap re-samples B.
3. Sample a new data set x^* set of size n from x with replacement.
4. Estimate θ from x^*.call the estimate $\hat{\theta}_i^*$,for $i = 1.....N$.
5. Repeat step 3 and 4 B times.
6. Consider the emperical distribution of $(\hat{\theta}_1^*, ..., \hat{\theta}_N^*)$ as an approximation of the true distribution of $\hat{\theta}$.

1. we assume data set $x = (x_1, x_2, ..., x_n)$ has a known distribution F_ψ
2. The data comes from a known distribution family F_ψ has a set of parameters.
3. Estimate parameters of ψ.
4. Fix the number of bootstrap re-samples B.
5. Sample a new data set x^* set of size n from x with replacement.
6. Estimate θ from x^*.call the estimate $\hat{\theta}^*_i$,for $i = 1.....N$.
7. Repeat step 5 and 6 B times.
8. Consider the emperical distribution of $(\hat{\theta}^*_1, ..., \hat{\theta}^*_N)$ as an approximation of the true distribution of $\hat{\theta}$.

Conclusion

- In Parametric bootstrap,$\hat{F}_{par}$ is not anymore the emperical density function.
- If the prior information on F is accurate,then $\hat{F}_{par}$ estimates better F than the empirical p.d.f .In this case the parametric bootstrap gives better estimation for the standard errors.
- If the parametric model is mis-specified then it rapidly converges to the wrong distribution.

Conclusion

when might bootstrap fail?

- Incomplete data.
- Dependent data.
- Noisy data.

Reference I

B.Efron,R.J.Tibshirani.
An Introduction to Bootstrap.
Chapman and hall, 1998.

Tim Hesterberg,Shaun Monaghan,David S.Moore.
Bootstrap Methods And Permutation Test.
W.H.Freeman and company, New York, 2003.