The Bootstrap - Basics and Applications

RTG Seminar Series

Dmitriy Izyumin

Overview

1. Basic idea of bootstrap
2. Bag of little bootstraps
3. Bootstrap aggregating

What is Bootstrapping?

- Data-driven resampling procedure
- Useful for obtaining measures of estimator quality (bias, variance, CI, PI).
- Simple in practice, but theory is rather complicated
- Efron, B. 1979

Steps

Goal: Estimate the sampling distribution of an estimate of a population parameter θ.

Given: A random sample S of n observations.

Repeat the following B times:

1. Obtain a sample of size n by sampling with replacement from S.*
2. Compute $\hat{\theta}$ from the sample.

Use the B bootstrap estimates to estimate the sampling distribution.

* This step can be modified for *parametric* bootstrapping.

Setting and Notation

Sample $X_1, X_2, \cdots, X_n$ drawn i.i.d from and unknown distribution P.

P_n is an approximation of P computed from the sample.

The estimator $\hat{\theta}_n$ has true underlying distribution $Q_n(P)$.

Goal: Compute estimator quality assessment $\xi(Q_n(P), P)$.

Bootstrap: Compute the plug-in estimate $\xi(Q_n(P_n), P_n)$

Favorable Properties

- Simple to implement
- Data-driven
- Automatic
- Outperforms estimation using asymptotic theory under failry general condititions
- Many types of bootstrapping for different scenarios

Theoretic Properties

- Have been studied extensively
- Refer to references (Hall 1992)

Bag of Little Bootstraps

- Scalable extension of the bootstrap to massive data
- Retains favorable properties of the bootstrap
- Less demanding computationally
- Designed with modern computing in mind
- Kleiner, A. et al, 2012

Motivation

When bootstrapping, we

- Start with sample of size n
- Obtain many resamples of size n
- Carry out calculations on each resample

If n is large, this becomes

- Computationally expensive or unfeasible
- Difficult to parallelize

Alternatives for Large Data

Some proposed alternatives:

- Subsampling
- *m* out of *n* bootstrap

Complications:

- Not robust to changes of hyperparameters
- Convergence computations require rescaling

New Idea

1. Start with sample of size n
2. Obtain s subsamples of size $b < n$ without replacement
3. Carry out bootstrap on each subsample using r bootstrap samples of **size** n
4. Compute bootstrap results for each subsample
5. Average over the results from the subsamples

Notation

Obtain subsamples of $X_1, X_2, \cdots, X_n$ of size b.

$P_{n,b}^{(j)}$ is an approximation of P computed from the j^{th} subsample.

Goal: Compute $\xi(Q_n(P), P)$.

BLB estimate: $\frac{1}{s}\sum_{i=1}^{s} \xi(Q_n(P_{n,b}^{(j)}), P_{n,b}^{(j)})$

Why Size b and not n?

- Each resample has *at most b* unique points
- Storage and computation requirements are now in $O(b)$
- Scales in b with respect to computation time and storage space
- Authors recommend $b = n^\gamma$, where $0.5 \geq \gamma < 1$

Practical Benefits

- Simple to implement
- Automatic
- Easily parallelizeable

Theoretical Properties

- Under standard assumptions, BLB is consistent.
- If b and s increase reasonably fast with n, then BLB has the same higher-order correctness as the bootstrap; i.e. same convergence rate!

Scalability

- Well-suited for parallel computing
- Resamples and subsamples are much smaller than original sample
- Bootstrap resampling is much faster
- Even though convergence rates are the same, the same relative error threshold can be achieved in less time than with bootstrap

Further Developments

- BOOT-TS for Massive Time Series Data
- Subsampled Double Bootstrap for Massive Data

Bagging - Bootstrap Aggregating

- Machine learning ensemble approach
- Average predictions of models trained on bootstrap resamples
- Breiman, L., 1994

Motivation

- Some predictors (e.g. CART, neural nets) are inherently unstable, and sensitive to perturbations in training data
- Fitting a predictor to different arrangements of training data can help us understand and correct the instability
- "The vital element is the instability of the prediction method. If perturbing the learning set can cause significant changes in the predictor constructed, then bagging can improve accuracy." - Leo Breiman, 1994

Notation

- Learning set $\mathcal{L} = \{(\mathbf{x_i}, y_i), i = 1, 2, \cdots, N\}$
  - Predict y from x
  - Regression or classification
- Predictor $\phi(\mathbf{x}, \mathcal{L})$ of y based on $\mathbf{x}$
- Take resamples $\mathcal{L}^{(|)}$ of $\mathcal{L}$ for $j = 1, 2, \cdots, B$
- Form $\phi_B(\mathbf{x}, \mathcal{L})$ by aggregating $\phi(\mathbf{x}, \mathcal{L}^{(|)})$ for $j = 1, 2, \cdots, B$
  - Regression: predict y by an average of the bootstrap predictions
  - Classification: predict y by a vote of the bootstrap classifiers

Benefits

- Improves stability
- Reduces variance
- Reduces overfitting

Theory

- Theory only recently developed
- Reduces variance and MSE for regression trees
- The above is not always true in general
- Refer to references (Buja 2006, Bühlmann 2003)

Further Developments

- Random forest (more flexible version of bagging for decision trees)
- Bragging - Bootstrap **robust** aggregating
  - Use a robust location estimator (e.g. median) when aggregating

References

Efron, B. (1979). "Bootstrap methods: Another look at the jackknife". The Annals of Statistics. 7 (1): 126.

Hall, Peter. "Rate of convergence in bootstrap approximations." The Annals of Probability (1988): 1665-1684.

Kleiner, Ariel, et al. "A scalable bootstrap for massive data." Journal of the Royal Statistical Society: Series B (Statistical Methodology) 76.4 (2014): 795-816.

Laptev, Nikolay, Carlo Zaniolo, and Tsai-Ching Lu. "BOOT-TS: A Scalable Bootstrap for Massive Time-Series Data." (2012).

References

Sengupta, Srijan, Stanislav Volgushev, and Xiaofeng Shao. "A subsampled double bootstrap for massive data." Journal of the American Statistical Association just-accepted (2015).

Breiman, Leo. "Bagging predictors." Machine learning 24.2 (1996): 123-140.

Buja, Andreas, and Werner Stuetzle. "Observations on bagging." Statistica Sinica (2006): 323-351.

Bühlmann, Peter. "Bagging, subagging and bragging for improving some prediction algorithms." (2003).