Insights, resources, and application of “Big Data"

BY YALIN LIAO

In the setting of large datasets, the application of the bootrap is limited due to its high computation cost. Authors in “The Big Data Bootstrap” present an alternative to the bootstrap, called Bag of Little Bootstraps (BLB).

In the outer loop of BLP, s index sets $I_j = \{i_1^j, i_2^j \cdots, i_b^j\}$ for $j = 1, 2, \cdots, s$ are sampled from $\{1, 2, \cdots, n\}$ where $i_m^j \neq i_n^j$ for ant $m \neq n$ (sample without replacement) and then get s bags $\text{Bag}_j = \{X_{i_1^j}, \cdots, X_{i_b^j}\}$; in the inner loop of the algorithm, the bootstrap is applied, that is, resample data points from the empirical distribution $P_{n,b}^{(j)}$ derived from bags $\text{Bag}_j = \{X_{i_1^j}, \cdots, X_{i_b^j}\}$ and average the estimator. Finally average the estimators from all bags.

BLB has the computational advantages over the Bootstraps and m out of n Bootstrap meanwhile it share the statistical consitency with the Bootstraps under the same assumptions. The computation cost is lower due to smaller bag size b from the high level analysis of computation complexity. The conclusion of the statistical consistency is

$$s^{-1}\sum_{j=1}^{s} \xi\left(Q_n\left(P_{n,b}^{(j)}\right)\right) - \xi\left(Q_n(P)\right) \xrightarrow{p} 0 \text{ as } n \to \infty \text{ for any sequence } b \to \infty \text{ and for any fixed } s.$$

This is beautiful. This means even if the rate $\frac{b}{n} \to 0$ as $n \to \infty$ and $b \to \infty$ (b is much smaller than n) the estimator convergency is garanteed. Morover, the results hold for sequences $s \to \infty$ if $E\left|\xi\left(Q_n\left(P_{n,b}^{(j)}\right)\right)\right| < \infty$. This means we could choose more bags to get more accurate estimator using smaller bags (size of $b \ll n$).

I just have one question? In the setting of large datasets, most estimators based on large datasets should be reliable. Why do we need to use bootstrap?[1]

1. I think in most case of machine learning setting we apply the bootstrap strategy because we don’t have enough data points. If we have a large set of data points, do we need to use bootstrap? Maybe yes, like we want to estimate the uncertainty of the learned model.