# Bootstrap

Wei Li

Syracuse University

Spring 2021

Bootstrap methods

Bootstrap distribution

Nonparametric bootstrap, (semi)-parametric bootstrap

Bootstrap estimate of prediciton error

Some discussions

# Bootstrap methods

# Bootstrap methods

- The bootstrap is a general tool for assessing statistical accuracy.
- As with cross-validation, the bootstrap can be used to estimate prediction error.
  - typically estimates well only the expected prediction error $Err$ (but not $Err_{\boldsymbol{\tau}}$)

# General ideas

Data are realizations of

$$Z_1, \ldots, Z_n \quad \text{i.i.d.} \ \sim P$$

where $P$ denotes an unknown distribution.

We denote a statistical procedure or estimator by

$$\hat{\theta}_n = S\left(Z_1, \ldots, Z_n\right)$$

which is a (known) function $S$ of the data $Z_1, \ldots, Z_n$.

Statisticians typically would need to find out the

- sampling distribution of $\hat{\theta}_n$,
- the expectation $E(\hat{\theta}_n)$ or the variance $\mathrm{Var}(\hat{\theta}_\mathrm{n})$ .

Suppose we knew what the distribution $P$ is

- can simulate to obtain the distribution of any $\hat{\theta}_n$ with arbitrary accuracy

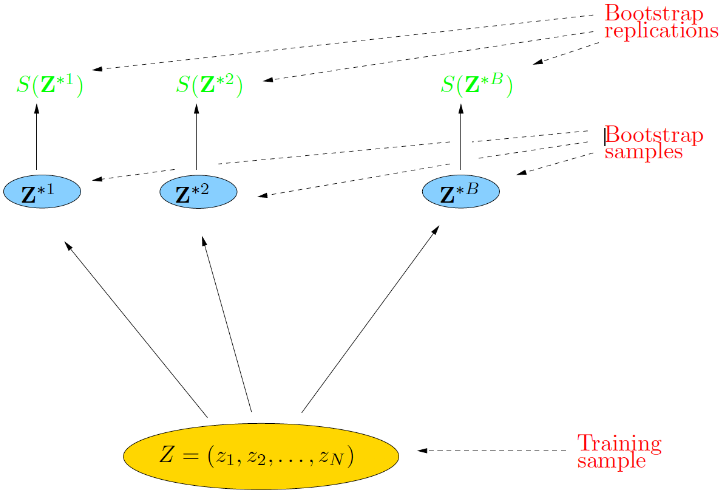
But we do not know what the distribution $P$.

Bootstrap:

- use the empirical distribution $\hat{P}_n$ which places probability mass $1/n$ on every data point $Z_i, i = 1, \ldots, n$.
- simulate from $\hat{P}_n$ : generate simulated data

$$Z_1^*, \ldots Z_n^* \quad \text{i.i.d.} \quad \sim \hat{P}_n$$

  i.e., generate $n$ random drawings with **replacement** from the original data set $\{Z_1, \ldots, Z_n\}$.
- Such a simulated new data set is called a **bootstrap sample**.
- compute our estimator $\hat{\theta}_n^* = S(Z_1^*, \ldots, Z_n^*)$ based on the bootstrap sample.
- We then repeat this many times (say obtain $B$ bootstrap samples, thus $B$ $\hat{\theta}_n^*$s )
- Get an approximate distribution for $\hat{\theta}_n$ by the histogram of $B$ $\hat{\theta}_n^*$s.



Bootstrap replications
$S(\mathbf{Z}^{*1})$
$S(\mathbf{Z}^{*2})$
$S(\mathbf{Z}^{*B})$
Bootstrap samples
$\mathbf{Z}^{*1}$
$\mathbf{Z}^{*2}$
$\mathbf{Z}^{*B}$
$Z = (z_1, z_2, \dots, z_N)$
Training sample


ESL. Fig. 7.12.

# Nonparametric bootstrap

The algorithm can be described as:

- 1. Generate a bootstrap sample

$$Z_1^*, \ldots, Z_n^* \quad \text{i.i.d.} \ \sim \hat{P}_n$$

  That is obtain $n$ random draws with replacement from the data set $\{Z_1, \ldots, Z_n\}$.

- 2. Compute the *bootstrapped estimator* based on the bootstrap sample

$$\hat{\theta}_n^* = S\left(Z_1^*, \ldots, Z_n^*\right)$$

- 3. Repeat steps 1 and 2 for $B$ times to obtain

$$\hat{\theta}_n^{*1}, \ldots, \hat{\theta}_n^{*B}$$

# Bootstrap distribution

# Bootstrap distribution

The **bootstrap distribution**, denoted here by $P^*$,

- the conditional probability distribution which is induced by i.i.d. resampling of the data given the original data.

The bootstrap distribution of $\theta_n^* = S\left(Z_1^*, \ldots, Z_n^*\right)$ is the distribution which arises when resampling with $\hat{P}_n$ and applying the function $S$ on such a bootstrap sample.

bootstrap expectation $\quad E^*\left[\hat{\theta}_n^*\right] \cong \frac{1}{B} \sum_{i=1}^{B} \hat{\theta}_n^{*i}$

bootstrap variance $\quad \operatorname{Var}^*\left(\hat{\theta}_n^*\right) \cong \frac{1}{B-1} \sum_{i=1}^{B}\left(\hat{\theta}_n^{*i} - \frac{1}{B} \sum_{j=1}^{B} \hat{\theta}_n^{*j}\right)^2$

$\alpha$ -quantile of the bootstrap distribution of $\hat{\theta}_n^*$:

empirical $\alpha$ -quantile of $\hat{\theta}_n^{*1}, \ldots, \hat{\theta}_n^{*B}$

If the empirical distribution $\hat{P}_n$ is “close” to the true data-generating probability $P$, the bootstrap values are “reasonable” estimates for the true quantities.

$$E\left[\hat{\theta}_n\right] \approx E^*(\hat{\theta}_n^*) \cong \frac{1}{B}\sum_{i=1}^{B}\hat{\theta}_n^{*i}$$

$$\operatorname{Var}\left(\hat{\theta}_n\right) \approx \operatorname{Var}^*\left(\hat{\theta}_n^*\right) \cong \frac{1}{B-1}\sum_{i=1}^{B}\left(\hat{\theta}_n^{*i} - \frac{1}{B}\sum_{j=1}^{B}\hat{\theta}_n^{*j}\right)^2.$$

# Bootstrap consistency

The bootstrap is called to be **consistent** for $\hat{\theta}_n$ if, for an increasing sequence $a_n$, for all $x$

$$P\left[a_n\left(\hat{\theta}_n-\theta\right)\leq x\right]-P^*\left[a_n\left(\hat{\theta}^*_n-\hat{\theta}_n\right)\leq x\right]\xrightarrow{P}0(n\to\infty)$$

In classical situations, $a_n=\sqrt{n}$.

# Estimating bias

Under bootstrap consistency, the bias of $\hat{\theta}_n$ may be approximated as

$$\begin{aligned} E(\hat{\theta}_n) - \theta &\approx E^*(\hat{\theta}_n^*) - \hat{\theta}_n \\ &\approx \frac{1}{B}\sum_{b=1}^{B} \hat{\theta}_n^{*b} - \hat{\theta}_n \end{aligned}$$

Also, $\text{Var}\left(\hat{\theta}_n\right) \approx \text{Var}^*\left(\hat{\theta}_n^*\right)$.

# Estimating confidence intervals

We can also construct the bootstrap confidence interval for $\theta$.

Recall that a $(1-\alpha)$ confidence interval for $\theta$, computed over $z_1, \ldots z_n$, is a random interval $[L, U]$ satisfying

$$P(L \leq \theta \leq U) = 1 - \alpha$$

The bootstrap confidence interval for $\theta$ is given by (why?)

$$\left[2\hat{\theta}_n - q^*_{1-\alpha/2}, 2\hat{\theta}_n - q^*_{\alpha/2}\right].$$

Here $q^*_{\alpha/2}$ and $q^*_{1-\alpha/2}$, are the $\alpha/2$ and $1-\alpha/2$ are the bootstrap quantiles of $\hat{\theta}_n^{*(1)}, \ldots \hat{\theta}_n^{*(B)}$.

# Studentized bootstrap confidence intervals

In some cases, the distributions of $(\hat{\theta}_n - \theta)/\widehat{\mathrm{SE}}(\hat{\theta}_n)$ and $(\hat{\theta}_n^* - \hat{\theta}_n)/\widehat{\mathrm{SE}}(\hat{\theta}_n^*)$ could be close, where $\widehat{\mathrm{SE}}(\cdot)$ denote estimated standard errors. Hence we could use what are called **studentized bootstrap confidence intervals**.

- repeat, for $b = 1, \dots B$ :
  - draw a bootstrap sample $z_1^{*(b)}, \dots z_n^{*(b)}$ from $\{z_1, \dots z_n\}$
  - recompute the statistic $\hat{\theta}_n^{*(b)}$ on $z_1^{*(b)}, \dots z_n^{*(b)}$
  - repeat, for $m = 1, \dots M$ :
    - draw a bootstrap sample $z_1^{*(b,m)}, \dots z_n^{*(b,m)}$ from$\{z_1^{*(b)}, \dots z_n^{*(b)}\}$
    - recompute the statistic $\hat{\theta}_n^{*(b,m)}$ from $\{z_1^{*(b,m)}, \dots z_n^{*(b,m)}\}$
  - compute the sample standard deviation $\hat{s}^{*(b)}$ of $\hat{\theta}_n^{*(b,1)}, \dots \hat{\theta}_n^{*(b,M)}$
  - compute $(\hat{\theta}_n^{*(b)} - \hat{\theta}_n)/\hat{s}^{*(b)}$.

From above we have a sample $\{(\hat{\theta}_n^{*(b)} - \hat{\theta}_n)/\hat{s}^{*(b)} : b = 1, \dots, B\}$, from which, we compute he quantiles $q_{\alpha/2}^*$ and $q_{1-\alpha/2}^*$.

The approximate $1 - \alpha$ bootstrap confidence interval for $\theta$ is given by

$$(\hat{\theta}_n - \widehat{\mathrm{SE}}(\hat{\theta}_n) q^*_{1-\alpha/2}, \hat{\theta}_n - \widehat{\mathrm{SE}}(\hat{\theta}_n) q^*_{\alpha/2}),$$

- $\widehat{\mathrm{SE}}(\hat{\theta}_n)$ can be approximated with $\mathrm{Var}^*(\hat{\theta}^*_n)$ using bootstrap samples $\{\hat{\theta}_n^{*(1)}, \ldots, \hat{\theta}_n^{*(B)}\}$.

# Nonparametric bootstrap, (semi)-parametric bootstrap

# Parametric bootstrap

Assume that the data are realizations from

$$Z_1, \ldots, Z_n \quad \text{i.i.d.} \quad \sim P_\theta$$

where $P_\theta$ is given up to an unknown parameter (vector) $\theta$.

- estimate the unknown parameter $\theta$ by $\hat{\theta}_n$
- draw

$$Z_1^*, \ldots, Z_n^* \quad \text{i.i.d.} \quad \sim P_{\hat{\theta}_n}$$

# Example

- the training data by $z=\{z_1, z_2, \ldots, z_n\}$, with $z_i=(x_i, y_i)$ $i=1,2,\ldots,n$.
- assume that $Y_i=\beta^{\top} x_i+\varepsilon_i, \quad (i=1,\ldots,n)$, $\varepsilon_1, \ldots, \varepsilon_n \stackrel{i.i.d.}{\sim} \mathcal{N}\left(0, \sigma^2\right)$, - $\theta=\left(\beta, \sigma^2\right)$. -$\hat{\beta}, \hat{\sigma}$ denote the MLE estimates based on original data.

1. Simulate $\varepsilon_1^*, \ldots, \varepsilon_n^* \quad$ i.i.d. $\quad \sim \mathcal{N}\left(0, \hat{\sigma}^2\right)$.
2. Construct

$$Y_i^*=\hat{\beta}^{\top} x_i+\varepsilon_i^*, i=1,\ldots,n$$

The parametric bootstrap regression sample is then

$$\left(x_1, Y_1^*\right), \ldots, \left(x_n, Y_n^*\right)$$

where the predictors $x_i$ are as for the original data.

# Nonparametric bootstrap

Denote the training data by $z = \{z_1, z_2, \ldots, z_n\}$, with $z_i = (x_i, y_i)$ $i = 1, 2, \ldots, n$. Here $x_i$ is the input, and $y_i$ the outcome. Suppose $E(Y|X = x) = \mu(x) = \sum_{j=1}^{M} \beta_j h_j(x)$ and $Y = \mu(X) + \epsilon$, where $var(\epsilon) = \sigma^2$.

$$
\begin{aligned}
\hat{\beta} &= \left(\mathbf{H}^T\mathbf{H}\right)^{-1}\mathbf{H}^T\mathbf{y} \\
\widehat{\mathrm{Var}}(\hat{\beta}) &= \left(\mathbf{H}^T\mathbf{H}\right)^{-1}\hat{\sigma}^2 \\
\hat{\sigma}^2 &= \sum_{i=1}^{n}\left(y_i - \hat{\mu}\left(x_i\right)\right)^2 / n
\end{aligned}
$$

- Let $h(x)^T = (h_1(x), h_2(x), \ldots, h_M(x))$.
- $\hat{\mu}(x) = h(x)^T\hat{\beta}$
- standard error $\widehat{\mathrm{se}}[\hat{\mu}(x)] = \left[h(x)^T\left(\mathbf{H}^T\mathbf{H}\right)^{-1}h(x)\right]^{\frac{1}{2}}\hat{\sigma}$.
- Then (biased) 95% confidence interval is $\hat{\mu}(x) \pm 1.96 \cdot \widehat{\mathrm{se}}[\hat{\mu}(x)]$.

Suppose we have $n = 50$. The nonparametric bootstrap works as in the following.

- We draw $B$ datasets each of size $n = 50$ with replacement from our training data, the sampling unit being the pair $z_i = (x_i, y_i)$.
- To each bootstrap dataset $\mathbf{Z}^*$ we fit a cubic spline $\hat{\mu}^*(x)$.
- Using B = 200 bootstrap samples, we can form a 95% pointwise confidence band from the percentiles at each x: we find the $2.5\% \times 200 =$ fifth largest and smallest values at each x.

# Semi-parametric bootstrap

Simulate new responses by adding Gaussian noise to the predicted values:

$$y_i^* = \hat{\mu}(x_i) + \varepsilon_i^*; \quad \varepsilon_i^* \sim N\left(0, \hat{\sigma}^2\right); \quad i = 1, 2, \ldots, n$$

This process is repeated $B$ times, where $B = 200$ say. The resulting bootstrap datasets have the form $(x_1, y_1^*), \ldots, (x_n, y_n^*)$ and we recompute the $B$ -spline smooth on each.

he confidence bands from this method will exactly equal the least squares bands, as the number of bootstrap samples goes to infinity.

Note that

- the estimate based on bootstrap sample $\hat{\mu}^*(x) = h(x)^T \left(\mathbf{H}^T \mathbf{H}\right)^{-1} \mathbf{H}^T \mathbf{y}^*$
- its distribution

$$\hat{\mu}^*(x) \sim N\left(\hat{\mu}(x), h(x)^T \left(\mathbf{H}^T \mathbf{H}\right)^{-1} h(x) \hat{\sigma}^2\right)$$

.

# Another version of bootstrap

Suppose

$$\begin{aligned} & Y_i = f(x_i) + \varepsilon_i \\ & \varepsilon_1, \ldots, \varepsilon_n \quad \text{i.i.d.} \quad \sim P_\varepsilon \end{aligned}$$

where $P_\varepsilon$ is unknown with expectation 0.

1. Estimate $\hat{f}$ from the original data and compute the residuals $r_i = Y_i - \hat{f}(x_i)$.
2. Consider the centered residuals $\tilde{r}_i = r_i - n^{-1} \sum_{i=1}^{n} r_i$. In case of linear regression with an intercept, the residuals are already centered. Denote the empirical distribution of the centered residuals by $\hat{P}_{\tilde{r}}$.
3. Generate

$$\varepsilon_1^*, \ldots, \varepsilon_n^* \quad \text{i.i.d.} \quad \sim \hat{P}_{\tilde{r}}$$

   Note that $\hat{P}_{\tilde{r}}$ is an estimate of $P_\varepsilon$.
4. Construct the bootstrap response variables

$$Y_i^* = \hat{m}(x_i) + \varepsilon_i^*, i = 1, \ldots, n$$

   and the bootstrap sample is then $(x_1, Y_1^*), \ldots, (x_n, Y_n^*)$.

# Bootstrap estimate of prediciton error

If $\hat{f}^{*b}(x_i)$ is the predicted value at $x_i$, from the model fitted to the $b$ th bootstrap dataset, our estimate is

$$\widehat{\mathrm{Err}}_{\mathrm{boot}} = \frac{1}{N} \sum_{i=1}^{N} \sum_{b=1}^{B} \frac{1}{B} L\left(y_i, \hat{f}^{*b}(x_i)\right)$$

- Repeat for $b = 1, \ldots, B$:
  - Generate $\left(X_1^{*b}, Y_1^{*b}\right), \ldots, \left(X_n^{*b}, Y_n^{*b}\right)$ by resampling with replacement from the original data.
  - Compute the bootstrapped estimator $\hat{f}^{*b}(\cdot)$ based on $\left(X_1^{*b}, Y_1^{*b}\right), \ldots, \left(X_n^{*b}, Y_n^{*b}\right)$
  - Evaluate $\mathrm{err}^{*b} = n^{-1} \sum_{i=1}^{n} L\left(Y_i, \hat{f}^{*b}(X_i)\right)$
- Approximate the bootstrap generalization error Err by

$$B^{-1} \sum_{i=1}^{B} err^{*i}$$

# Leave-one-out bootstrap estimate

Above estimate is not a good estimate in general.Tends to be overfitting.

The **leave-one-out bootstrap estimate** of prediction error is defined by

$$\widehat{\mathrm{Err}}^{(1)} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{|C^{-i}|}\sum_{b\in\{1,\ldots,B\}\cap C^{-i}} L\left(y_i, \hat{f}^{*b}(x_i)\right)$$

- $C^{-i}$ is the set of indices of the bootstrap samples $b$ that do not contain observation $i$,
- $\left|C^{-i}\right|$ is the number of such samples.

The leave-one out bootstrap solves the overfitting problem suffered by $\widehat{\mathrm{Err}}_{boot}$, but has the training-set-size bias mentioned in the discussion of cross-validation.

- Typically, the leave-one out bootstrap estimate will be biased upward.

# Bias

- Denote the bootstrap sample by $Z^* = \{Z_1^*, \ldots, Z_n^*\}$.
- an out-of-bootstrap sample

$$Z_{\text{out}}^* = \{Z_i; Z_i \notin Z^*\}$$

The out-of-bootstrap estimate above can be written as:

$$\widehat{\text{Err}}^{(1)} = \frac{1}{B} \sum_{b=1}^{B} \frac{1}{\left|Z_{\text{out}}^{*(b)}\right|} \sum_{i \in Z_{\text{out}}^{*(b)}} L\left(y_i, \hat{f}^{*(b)}(x_i)\right)$$

Note that $\hat{f}^{*(b)}(\cdot)$ involves only data from $Z^{*(b)}$, and $(X_i, Y_i) \in Z_{\text{out}}^*$.

- The expected size of the out-of-bootstrap sample: $E^*\left[|Z_{\text{out}}^*|\right] \approx 0.368n$.

Roughly speaking, $\widehat{\text{Err}}^{(1)}$ is like a CV estimate that uses about 36.8% data points as test data, or about 63.2% data points as training data.

# The .632 estimator

The ".632 estimator" is designed to alleviate this bias:

$$\widehat{\mathrm{Err}}^{(.632)} = .368 \cdot \overline{\mathrm{err}} + .632 \cdot \widehat{\mathrm{Err}}^{(1)}$$

The derivation of the .632 estimator is complex; intuitively it pulls the leave-one out bootstrap estimate down toward the training error rate, and hence reduces its upward bias.

Note that $\overline{\mathrm{err}} \leq \widehat{\mathrm{Err}}^{(.632)} \leq \widehat{\mathrm{Err}}^{(1)}$.

- The .632 estimator works well in "light fitting" situations
- In the heavily-overfitting situations, one can further improve the .632 estimator: the .632+ estimator

# Some discussions

# Some discussions

- One can use bootstrap to estimate effective degree of freedom
- The bootstrap distribution discussed above is a “poor man’s” Bayes posterior.

Bayesian bootstrap sample:

- Draw weights from a uniform Dirichlet distribution with the same dimension as the number of data points
- Sample from data accordingly to the probability defined by the Dirichlet weights
- Use the resampled data to calculate the statistics.