Machine Learning I

The bootstrap

Souhaib Ben Taieb

March 25, 2022

University of Mons

Table of contents

Table of contents

The bootstrap

- The bootstrap is a flexible and powerful statistical tool that can be used to *quantify the uncertainty* associated with a given (complex) estimator or machine learning method.
- For example, it can provide an estimate of the **standard error** of a coefficient, a **confidence interval** for that coefficient, or the **prediction error** of a machine learning method.
- The main idea is to obtain distinct data sets by repeatedly sampling observations from the original data set *with replacement*.

Resampling methods

Resampling methods are used in

1. **validating models** by using (random) subsets of the data (e.g cross-validation and bootstrap),
2. **estimating uncertainty** in sample statistics by drawing randomly with replacement from the data set (e.g. bootstrap),
3. performing **(non-parametric) significance tests** (permutation tests).
4. ...

Where does the name come from?

Pull yourself up by your bootstraps

It is not the same as the term “bootstrap” used in computer science meaning to “boot” a computer from a set of core instructions, though the derivation is similar.

Table of contents

A simple example

- Suppose that we wish to invest a fixed sum of money in two financial assets that yield returns of X and Y, respectively, where X and Y are random quantities.
- We will invest a fraction α of our money in X, and will invest the remaining $1 - \alpha$ in Y.
- We wish to choose α to minimize the total risk, or variance, of our investment. In other words, we want to minimize $\mathrm{Var}(\alpha X + (1 - \alpha)Y)$.
- One can show that the value that minimizes the risk is given by

$$\alpha = \frac{\sigma_Y^2 - \sigma_{XY}}{\sigma_X^2 + \sigma_Y^2 - 2\sigma_{XY}},$$

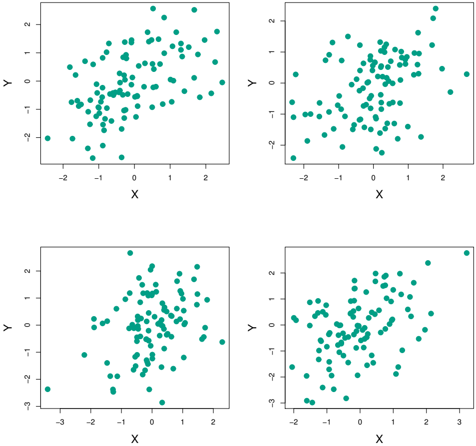
where $\sigma_X^2 = \mathrm{Var}(X), \sigma_Y^2 = \mathrm{Var}(Y)$, and $\sigma_{XY} = \mathrm{Cov}(X, Y)$.

Example continued

- But the values of σ_X^2, σ_Y^2, and σ_{XY} are unknown.
- We can compute estimates for these quantities, $\hat{\sigma}_X^2$, $\hat{\sigma}_Y^2$, and $\hat{\sigma}_{XY}$, using a data set that contains measurements for X and Y.
- We can then estimate the value of α that minimizes the variance of our investment using

$$\hat{\alpha} = \frac{\hat{\sigma}_Y^2 - \hat{\sigma}_{XY}}{\hat{\sigma}_X^2 + \hat{\sigma}_Y^2 - 2\hat{\sigma}_{XY}}.$$

Example continued



Each panel displays 100 *simulated returns for investments* X *and* Y*. From left to right and top to bottom, the resulting estimates for* α *are* 0.576, 0.532, 0.657, *and* 0.651.

Example continued

- To estimate the standard deviation of $\hat{\alpha}$, we repeated the process of simulating 100 paired observations of X and Y, and estimating α 1,000 times.
- We thereby obtained 1,000 estimates for α, which we can call $\hat{\alpha}_1, \hat{\alpha}_2, \ldots, \hat{\alpha}_{1000}$.

- For these simulations the parameters were set to $\sigma_X^2 = 1, \sigma_Y^2 = 1.25$, and $\sigma_{XY} = 0.5$, and so we know that the true value of α is 0.6 (indicated by the red line).

Example continued

- The mean over all 1,000 estimates for α is

$$\bar{\alpha} = \frac{1}{1000} \sum_{r=1}^{1000} \hat{\alpha}_r = 0.5996,$$

very close to $\alpha = 0.6$, and the standard deviation of the estimates is

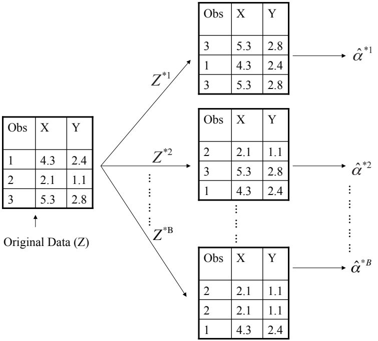
$$\sqrt{\frac{1}{1000-1} \sum_{r=1}^{1000} (\hat{\alpha}_r - \bar{\alpha})^2} = 0.083.$$

- This gives us a very good idea of the accuracy of $\hat{\alpha}$: $\mathrm{SE}(\hat{\alpha}) \approx 0.083$.
- So roughly speaking, for a random sample from the population, we would expect $\hat{\alpha}$ to differ from α by approximately 0.08, on average.

Now back to the real world

- The procedure outlined above cannot be applied, because for real data we cannot generate new samples from the original population.
- However, the bootstrap approach allows us to use a computer to mimic the process of obtaining new data sets, so that we can estimate the variability of our estimate without generating additional samples.
- Rather than repeatedly obtaining independent data sets from the population, we instead obtain distinct data sets by repeatedly sampling observations from the original data set *with replacement*.
- Each of these "bootstrap data sets" is created by sampling *with replacement*, and is the *same size* as our original dataset. As a result some observations may appear more than once in a given bootstrap data set and some not at all.

Example with just 3 observations



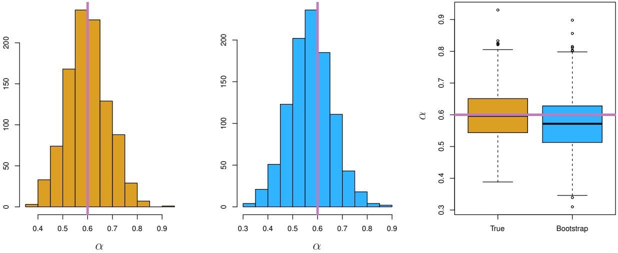
A graphical illustration of the bootstrap approach on a small sample containing $n = 3$ observations. Each bootstrap data set contains n observations, sampled with replacement from the original data set. Each bootstrap data set is used to obtain an estimate of α

- Denoting the first bootstrap data set by Z^{*1}, we use Z^{*1} to produce a new bootstrap estimate for α, which we call $\hat{\alpha}^{*1}$
- This procedure is repeated B times for some large value of B (say 100 or 1000), in order to produce B different bootstrap data sets, $Z^{*1}, Z^{*2}, \ldots, Z^{*B}$, and B corresponding α estimates, $\hat{\alpha}^{*1}, \hat{\alpha}^{*2}, \ldots, \hat{\alpha}^{*B}$.
- We estimate the standard error of these bootstrap estimates using the formula

$$\mathrm{SE}_B(\hat{\alpha}) = \sqrt{\frac{1}{B-1} \sum_{r=1}^{B} \left(\hat{\alpha}^{*r} - \bar{\hat{\alpha}}^{*}\right)^2}.$$

- This serves as an estimate of the standard error of $\hat{\alpha}$ estimated from the original data set. See center and right panels of Figure on slide 29. Bootstrap results are in blue. For this example $\mathrm{SE}_B(\hat{\alpha}) = 0.087$.

Results



Left: A histogram of the estimates of α obtained by generating 1,000 simulated data sets from the true population. *Center:* A histogram of the estimates of α obtained from 1,000 bootstrap samples from a single data set. *Right:* The estimates of α displayed in the left and center panels are shown as boxplots. In each panel, the pink line indicates the true value of α.

Table of contents

The bootstrap procedure

- Let $\hat{P}$ be an estimate of P, the population distribution.
- Draw B independent bootstrap samples/datasets from $\hat{P}$:

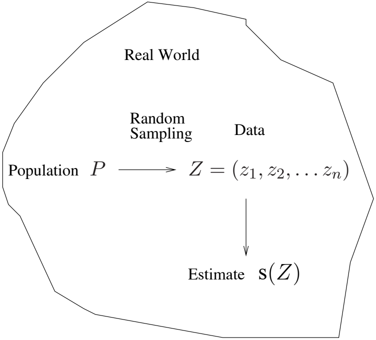
$$Z_1^{*(b)}, \ldots, Z_n^{*(b)} \sim \hat{P} \quad b = 1, \ldots, B.$$

- Evaluate the bootstrap replications:

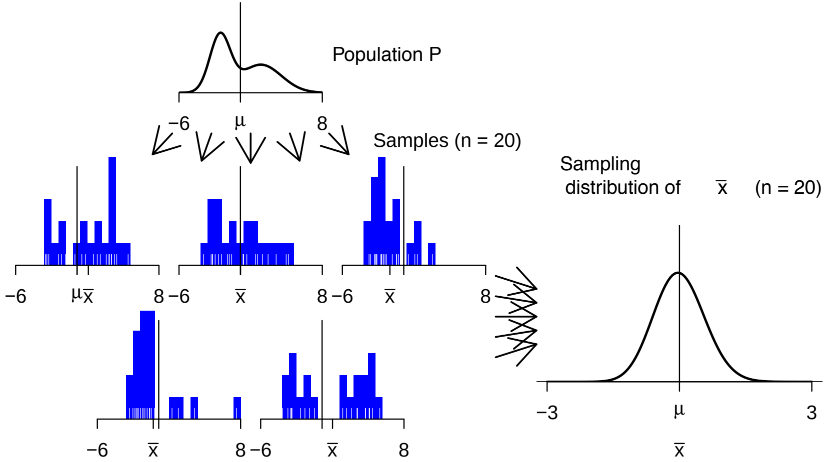
$$\hat{\theta}^{*(b)} = s(Z^{*(b)}) \quad b = 1, \ldots, B,$$

  where $s(\cdot)$ is the statistic of interest (e.g. mean, median, correlation coefficient, etc)
- Compute the sampling distribution of $\hat{\theta}^{*(b)}$ or any associted statistic of interest (standard deviation, confidence intervals, etc).

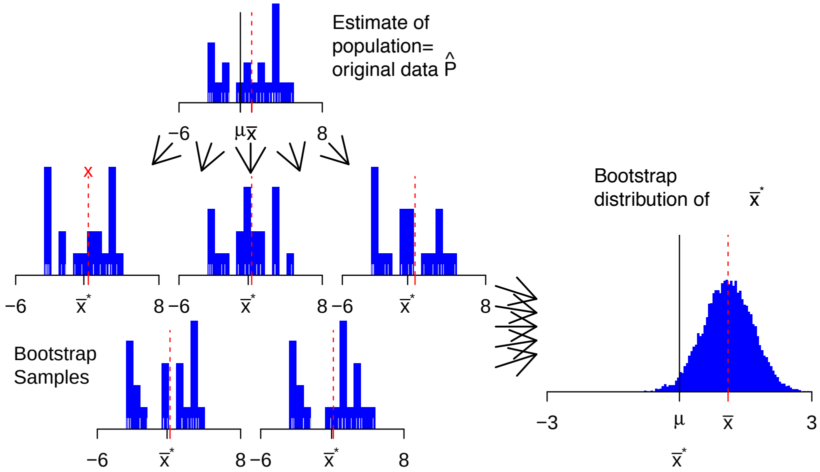
A general picture for the bootstrap



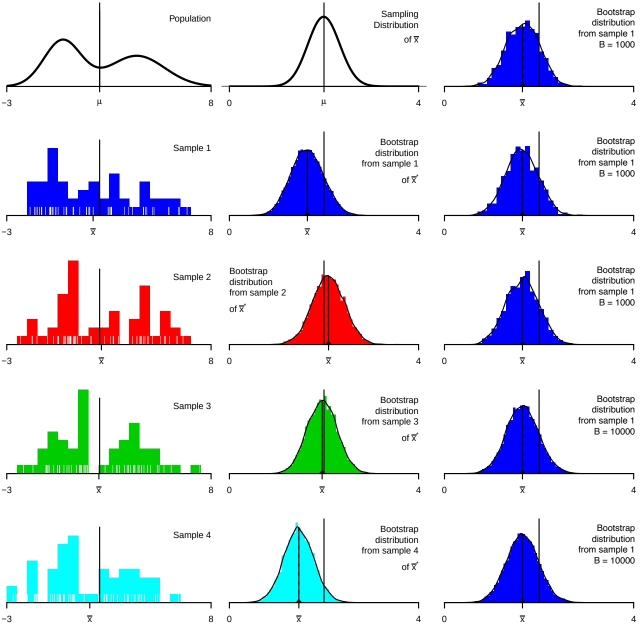
Bootstrapping: Ideal world



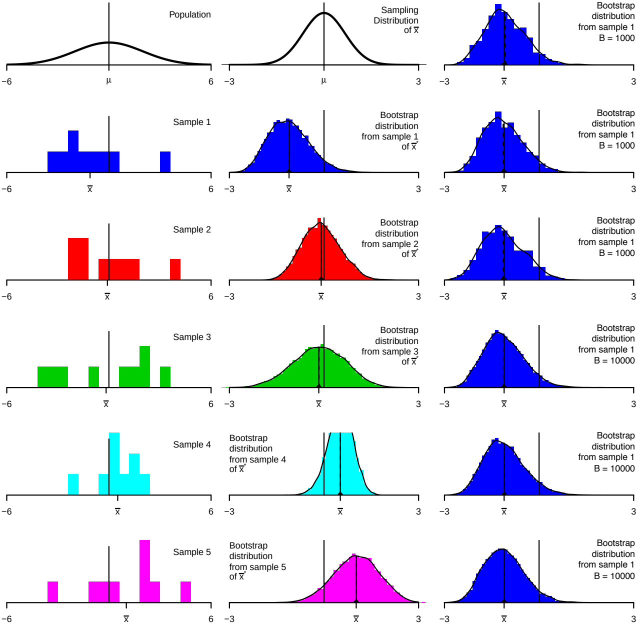
Bootstrapping: Bootstrap world



Sources of random variation - $n = 50$, $B = 10^3$ or 10^4



Sources of random variation - $n = 9$, $B = 10^3$ or 10^4



Other uses of the bootstrap

- Primarily used to obtain standard errors of an estimate.
- Also provides approximate confidence intervals for a population parameter. For example, looking at the histogram in the middle panel of the Figure on slide 29, the 5% and 95% quantiles of the 1000 values is $(.43, .72)$.
- This represents an approximate 90% confidence interval for the true α. *How do we interpret this confidence interval?*
- The above interval is called a *Bootstrap Percentile* confidence interval. It is the simplest method (among many approaches) for obtaining a confidence interval from the bootstrap.

Table of contents

Prediction error estimation

- In cross-validation, each of the K validation folds is **distinct** from the other $K-1$ folds used for training: there is **no overlap**. This is crucial for its success.
- To estimate prediction error using the bootstrap, we could think about using each bootstrap dataset as our training sample, and the original sample as our validation sample.
- In other words, we fit the model on a set of bootstrap samples, and then keep track of how well it predicts the original dataset

$$\mathrm{Err}_{\mathrm{boot}} = \frac{1}{B}\frac{1}{n}\sum_{b=1}^{B}\sum_{i=1}^{n} L(y_i, h^{*b}(x_i)),$$

where h^{*b} is fitted on the b-th bootstrap sample.
Does that work?

Probability that an observation belongs to a bootstrap sample

$$
\begin{aligned}
&\mathrm{P}(\text{observation } i \in \text{ bootstrap sample })\\
&= 1 - \mathrm{P}(\text{observation } i \notin \text{ bootstrap sample })\\
&= 1 - \Pi_{j=1}^{n}\mathrm{P}(\text{observation } i \text{ not in the j-th position in bootstrap sample})\\
&= 1 - \mathrm{P}(\text{observation } i \text{ not in the j-th position in bootstrap sample})^n\\
&= 1 - (1 - \mathrm{P}(\text{observation } i \text{ in the j-th position in bootstrap sample}))^n\\
&= 1 - \left(1 - \frac{1}{n}\right)^n\\
&\approx 1 - \frac{1}{e} \qquad\qquad \left(e^x = \lim_{n\to\infty}\left(1 + \frac{x}{n}\right)^n\right)\\
&= 0.632
\end{aligned}
$$

Prediction error estimation

- No. Each bootstrap sample has significant overlap with the original data. About **two-thirds** of the original data points appear in each bootstrap sample.
- In fact, each of these bootstrap data sets is created by **sampling with replacement**, and is the **same size as our original dataset**.
- As a result **some observations may appear more than once in a given bootstrap data set and some not at all**.
- Training and validation sets have observations in common! Overfit predictions will look very good.
- The other way around— with original sample = training sample, bootstrap dataset = validation sample— is worse!

Prediction error estimation

Better bootstrap version: we only keep track of predictions from bootstrap samples <u>not containing that observation</u>. The **leave-one-out bootstrap estimate of prediction error** can be defined as

$$\mathrm{Err}_{\text{loo-boot}} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{|S^{-i}|}\sum_{b\in S^{-i}} L(y_i, h^{*b}(x_i))$$

where S^{-i} is the set of indices of the bootstrap samples that do not contain observation i.

Problem of overfitting with $\mathrm{Err}_{\text{boot}}$ solved but **training-set-size bias as with cross-validation**.

Many applications

- Computing standard errors and confidence intervals for complex statistics
- Prediction error estimation
- Bagging (Bootstrap aggregating)
- ...

The bootstrap method we presented here is called the **non-parametric bootstrap**. There are other types of bootstrap methods based on different assumptions:

- parametric bootstrap
- block bootstrap
- smooth bootstrap
- residual bootstrap
- ...