Machine Learning I

The bootstrap

Souhaib Ben Taieb

University of Mons

UMONS Université de Mons

Faculté des Sciences

Table of contents

Table of contents

Resampling methods

Resampling methods are used in

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

The bootstrap

- The **bootstrap** is a flexible and powerful resampling method that can be used to *quantify the uncertainty* associated with almost any statistic or to estimate the prediction error of any learning model.
- By estimation the sampling distribution of any statistic, the bootstrap can provide for example an estimate of the **standard error** of a coefficient or its **confidence interval**. It can also be used to compute the **validation error** of any learning model.
- The main idea is to obtain distinct data sets by **repeatedly sampling** observations from the original data set **with replacement**.

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

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 $\alpha \in [0, 1]$ 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 compute:

$$\alpha^* = \underset{\alpha \in [0,1]}{\text{argmin}} \ \text{Var}(\alpha X + (1 - \alpha) Y),$$

  where $\text{Var}(X) = \sigma_X^2$, $\text{Var}(Y) = \sigma_Y^2$, and $\text{Cov}(X, Y) = \sigma_{X,Y}$.
- We can show that the solution is given by

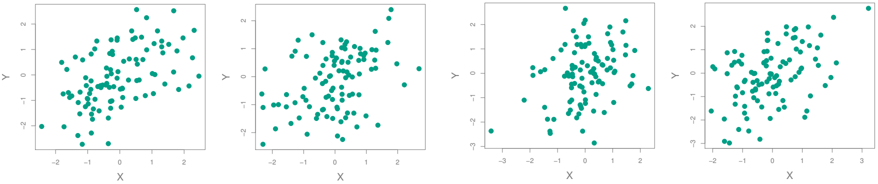
$$\alpha^* = \frac{\sigma_Y^2 - \sigma_{X,Y}}{\sigma_X^2 + \sigma_Y^2 - \sigma_{X,Y}}.$$

Example

- In practice, the values of σ_X^2, σ_Y^2, and $\sigma_{X,Y}$ are unknown.
- Given a dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$, we can compute estimates of these quantities $\hat{\sigma}_X^2$, $\hat{\sigma}_Y^2$, and $\hat{\sigma}_{X,Y}$.
- We can plug in these estimates in the previous formula and compute

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

- Consider $\sigma_X^2 = 1$, $\sigma_Y^2 = 1.25$, and $\sigma_{X,Y} = 0.5$ ($\alpha^* = 0.6$). Each panel displays 100 simulated returns for investments X and Y.



From left to right, the resulting values for $\hat{\alpha}$ are 0.576, 0.532, 0.657, and 0.651.

Sampling distribution

- To estimate the **sampling distribution**,, we repeated the process of simulating 100 paired observations of X and Y, and computing $\hat{\alpha}$ 1,000 times.
- We thereby obtained 1,000 estimates, which we call $\hat{\alpha}_1, \hat{\alpha}_2, \ldots, \hat{\alpha}_{1000}$.
- We can compute
  - The mean $\frac{1}{1000}\sum_{r=1}^{1000} \hat{\alpha}_r = 0.5996$
  - The standard deviation $\sqrt{\frac{1}{1000-1}\sum_{r=1}^{1000}(\hat{\alpha}_r - \bar{\alpha})^2} = 0.083$, also called the **standard errors**.

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** 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.
- 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.

Comparison of results

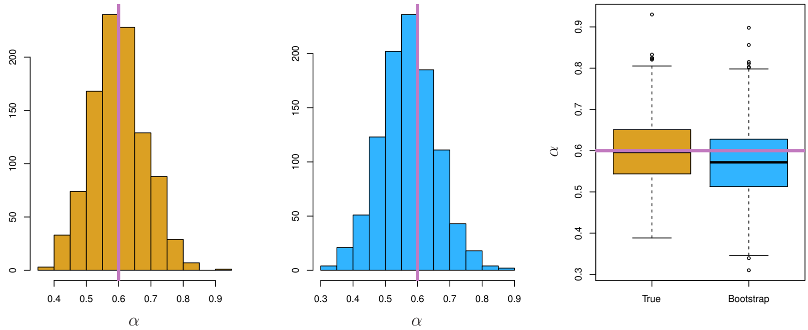


Table of contents

The (non-parametric) bootstrap procedure

Let $\mathcal{D} = \{z_i\}_{i=1}^n$, be a dataset with n observations, and $s(\cdot)$ a statistic of interest (e.g. mean, median, correlation coefficient, etc) for which we want to estimate the sampling distribution.

- Draw B **independent** bootstrap samples/datasets from $\mathcal{D}$:

$$\mathcal{D}^{*(b)} = \{z_1^{*(b)}, z_2^{*(b)}, \dots, z_n^{*(b)}\}, \quad b = 1, \dots, B,$$

  where $z_i^{*(b)}$ is **sampled** from $\mathcal{D}$ with **replacement**.
- Evaluate the bootstrap replications:

$$\hat{\theta}^{*(b)} = s(\mathcal{D}^{*(b)}) \quad b = 1, \dots, B,$$

- Compute the sampling distribution of $s()$ or any associated statistic of interest (standard deviation, confidence intervals, etc) using $\{\hat{\theta}^{*(1)}, \dots, \hat{\theta}^{*(B)}\}$.

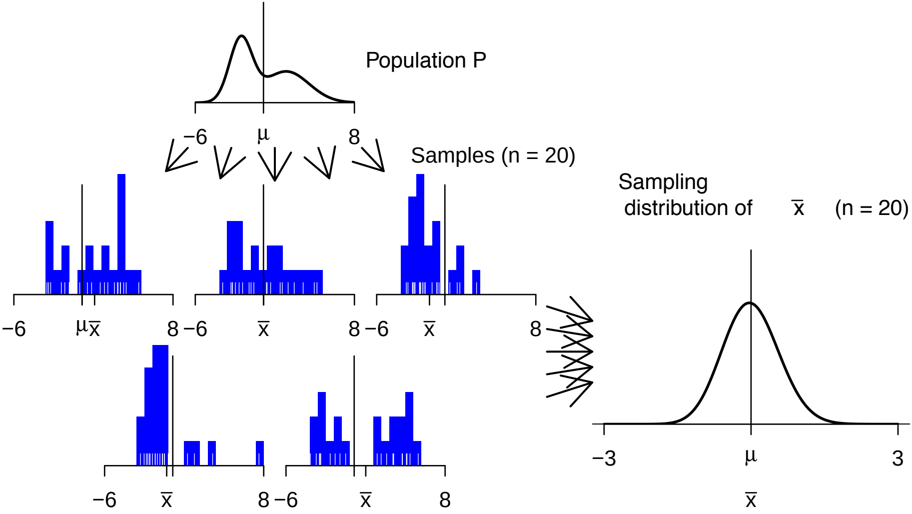
An equivalent description

- Let $\hat{P}$ be the empirical distribution function of the observed data, i.e. an **estimate** of P, the population distribution.
- Draw B **independent** bootstrap samples/datasets from $\hat{P}$:

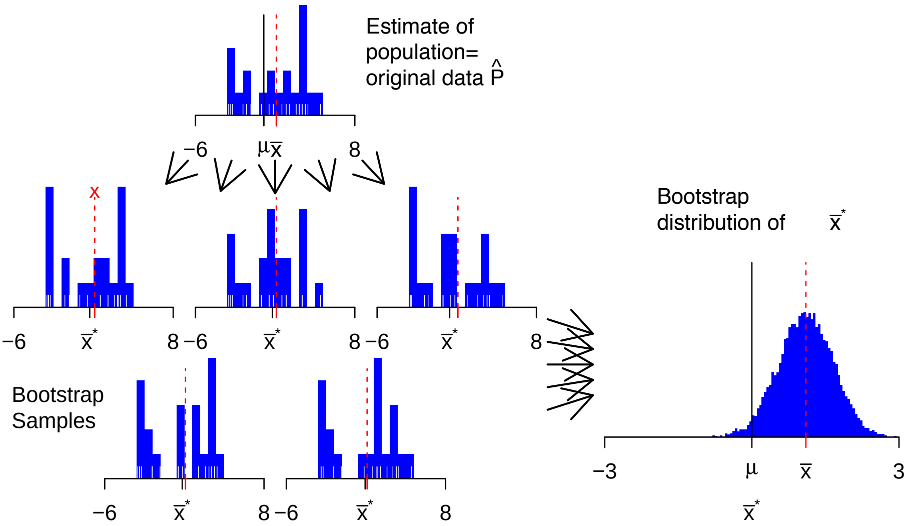
$$\mathcal{D}^{*(b)} = \{z_1^{*(b)}, z_2^{*(b)}, \ldots, z_n^{*(b)}\}, \quad b = 1, \ldots, B,$$

where $z_i^{*(b)}$ is **sampled** from $\hat{P}$.

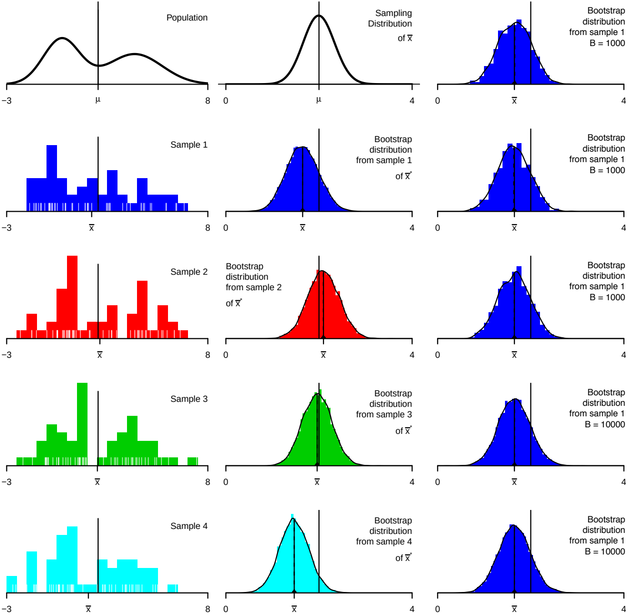
Bootstrapping: Ideal world



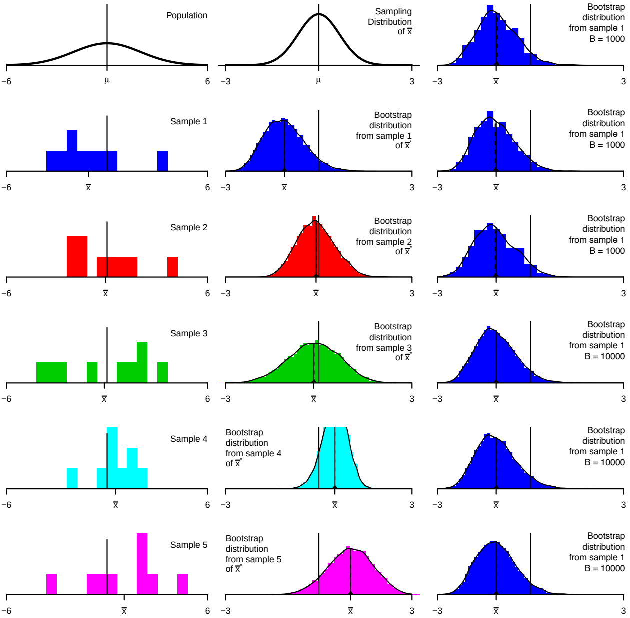
Bootstrapping: Bootstrap world



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



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



Probability that an observation belongs to a bootstrap sample

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

Probability that an observation belongs to a bootstrap sample

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

Probability that an observation belongs to a bootstrap sample

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

Probability that an observation belongs to a bootstrap sample

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

Probability that an observation belongs to a bootstrap sample

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

Probability that an observation belongs to a bootstrap sample

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

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?

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 not containing that observation. 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)
- ...

We presented the **non-parametric bootstrap**. There are other types of bootstrap methods based on different assumptions:

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