Master Program in *Data Science and Business Informatics*

Statistics for Data Science

Lesson 27 - Bootstrap and resampling methods

Salvatore Ruggieri

Department of Computer Science
University of Pisa, Italy
salvatore.ruggieri@unipi.it

Bootstrap principle

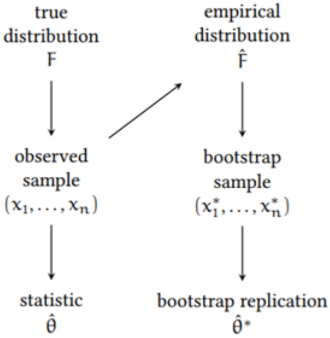
- Let $X_1, \ldots, X_n \sim F$ be a random sample
  - with *unknown distribution* F
- Estimator $T = h(X_1, \ldots, X_n)$, e.g., $\bar{X}_n = (X_1 + \ldots + X_n)/n$
  - with *unknown (sampling) distribution*
- From a dataset $x_1, \ldots, x_n$, we can derive a point estimate $\hat{\theta} = h(x_1, \ldots, x_n)$
- From many datasets $\{x_1^i, \ldots, x_n^i\}_{i=1}^m$, we can derive many point estimates $\hat{\theta}^i = h(x_1^i, \ldots, x_n^i)$
- By the **Glivenko-Cantelli Thm**, the empirical distribution of $\hat{\theta}^i$ approximates the distribution of T
- **Problem:** typically, we do not have many datasets, but only one!

See R script

Bootstrap principle

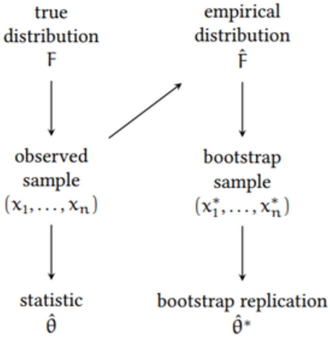
- Let $X_1, \ldots, X_n \sim F$ be a random sample
  - with *unknown distribution* F
- Estimator $T = h(X_1, \ldots, X_n)$, e.g., $\bar{X}_n = (X_1 + \ldots + X_n)/n$
- From a dataset $x_1, \ldots, x_n$, we can
  - derive a point estimate $\hat{\theta} = h(x_1, \ldots, x_n)$
  - or, derive an estimate $\hat{F}$ of F
- From $\hat{F}$ we can generate (a lot of) *bootstrap samples* $x_1^*, \ldots, x_n^*$
  - as realizations of $X_1^*, \ldots, X_n^* \sim \hat{F}$

and then (many) bootstrap point estimates $\hat{\theta}^* = h(x_1^*, \ldots, x_n^*)$



Bootstrap principle. Use the dataset $x_1, x_2, \ldots, x_n$ to compute an estimate $\hat{F}$ for the "true" distribution function F. Replace the random sample $X_1, X_2, \ldots, X_n$ from F by a random sample $X_1^*, X_2^*, \ldots, X_n^*$ from $\hat{F}$, and approximate the probability distribution of $h(X_1, X_2, \ldots, X_n)$ by that of $h(X_1^*, X_2^*, \ldots, X_n^*)$.

Empirical bootstrap



- How to derive $\hat{F}$ from $x_1, \ldots, x_n$?
- If we know nothing about F, use the empirical distribution:

  $\hat{F}(a) = F_n(a) = \frac{|\{i \in 1,\ldots,n \ | \ x_i \leq a\}|}{n}$
- How to generate a bootstrap sample $x_1^*, \ldots, x_n^*$?
  - x_i^* is chosen randomly from $\hat{F}$
  - i.e., x_i^* s chosen randomly from $x_1, \ldots, x_n$ (our dataset)
- Hence, a bootstrap dataset $x_1^*, \ldots, x_n^*$ is obtained by *random sampling with replacement*!
- Often the bootstrap approximation of the distribution of T will improve if we shift T by relating it to a corresponding feature of the "true" distribution.
  - rather than approximating the distribution of $\bar{X}_n$ by the one of $\bar{X}_n^*$, better to approximate $\Delta = \bar{X}_n - \mu$ by $\Delta^* = \bar{X}_n^* - \mu^*$, where $\mu^* = E[\hat{F}] = \bar{x}_n = (x_1 + \ldots + x_n)/n$

[See remarks 18.1 and 18.2 of textbook]

Empirical bootstrap

EMPIRICAL BOOTSTRAP SIMULATION (FOR $\bar{X}_n - \mu$). Given a dataset $x_1, x_2, \ldots, x_n$, determine its empirical distribution function F_n as an estimate of F, and compute the expectation

$$\mu^* = \bar{x}_n = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

corresponding to F_n.

1. Generate a bootstrap dataset $x_1^*, x_2^*, \ldots, x_n^*$ from F_n.
2. Compute the centered sample mean for the bootstrap dataset:

$$\bar{x}_n^* - \bar{x}_n,$$

where

$$\bar{x}_n^* = \frac{x_1^* + x_2^* + \cdots + x_n^*}{n}.$$

Repeat steps 1 and 2 many times.

- Use the empirical distribution of $\delta^* = \bar{x}_n^* - \bar{x}_n$ (realizations of $\Delta^* = \bar{X}_n^* - \bar{x}_n$)
  - for estimating the distribution of $\Delta = \bar{X}_n - \mu$, and in particular:

$$E[\Delta] = E[\bar{X}_n] - \mu \approx E[\Delta^*] \approx mean(\delta^*)$$

  - and then estimate μ as $\hat{\mu} = E[\bar{X}_n] - mean(\delta^*) \approx \bar{x}_n - mean(\delta^*)$

$mean(\delta^*)$ is the estimated bias

  - and $se(\bar{X}_n) = \sqrt{Var(\bar{X}_n)} = \sqrt{Var(\bar{X}_n - \mu)} \approx \sqrt{Var(\bar{X}_n^* - \bar{x}_n)} \approx sd(\delta^*)$

See R script

Empirical bootstrap

Empirical bootstrap simulation (for $\bar{X}_n - \mu$). Given a dataset $x_1, x_2, \ldots, x_n$, determine its empirical distribution function F_n as an estimate of F, and compute the expectation

$$\mu^* = \bar{x}_n = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

corresponding to F_n.

1. Generate a bootstrap dataset $x_1^*, x_2^*, \ldots, x_n^*$ from F_n.
2. Compute the centered sample mean for the bootstrap dataset:

$$\bar{x}_n^* - \bar{x}_n,$$

where

$$\bar{x}_n^* = \frac{x_1^* + x_2^* + \cdots + x_n^*}{n}.$$

Repeat steps 1 and 2 many times.

- Use the empirical distribution of $\delta^* = \bar{x}_n^* - \bar{x}_n$ (realizations of $\Delta^* = \bar{X}_n^* - \bar{x}_n$)
  - for estimating the distribution of $\Delta = \bar{X}_n - \mu$, and in particular:
  - confidence interval for $\delta = \bar{x}_n - \mu$ is $(q_{\alpha/2}, q_{1-\alpha/2})$ of δ^* empirical distribution
  - $q_{\alpha/2} \leq \delta = \bar{x}_n - \mu \leq q_{1-\alpha/2}$ implies c.i. for μ is $(\bar{x}_n - q_{1-\alpha/2}, \bar{x}_n - q_{\alpha/2})$

See R script

Empirical bootstrap

`boot.ci` method in R confidence intervals:

- `type='basic'`: $(\bar{x}_n - q_{1-\alpha/2}, \bar{x}_n - q_{\alpha/2})$ with quantiles over the distribution of δ^*
- `type='perc'`: $(q_{\alpha/2}, q_{1-\alpha/2})$ with quantiles over the distribution of $\bar{x}_n^*$ (without shift)
- `type='norm'`: $(\bar{x}_n - q_{1-\alpha/2}, \bar{x}_n - q_{\alpha/2})$ with quantiles over $N(mean(\delta^*), var(\delta^*))$
- `type='bca'`: bias (and skewness) correction and acceleration

See R script

Empirical bootstrap

`boot.ci` method in R confidence intervals:

- `type='stud'`: $(\bar{x}_n - q_{1-\alpha/2}\frac{s_n}{\sqrt{n}}, \bar{x}_n - q_{\alpha/2}\frac{s_n}{\sqrt{n}})$ with quantiles over the distribution of t^*

Empirical bootstrap simulation for the studentized mean.
Given a dataset $x_1, x_2, \ldots, x_n$, determine its empirical distribution function F_n as an estimate of F. The expectation corresponding to F_n is $\mu^* = \bar{x}_n$.

1. Generate a bootstrap dataset $x_1^*, x_2^*, \ldots, x_n^*$ from F_n.
2. Compute the studentized mean for the bootstrap dataset:

$$t^* = \frac{\bar{x}_n^* - \bar{x}_n}{s_n^*/\sqrt{n}},$$

   where $\bar{x}_n^*$ and s_n^* are the sample mean and sample standard deviation of $x_1^*, x_2^*, \ldots, x_n^*$.

Repeat steps 1 and 2 many times.

See R script

Empirical bootstrap

- Bootstrap approach applies to **any** estimator, not only the mean
- Example 1: the German Tank problem *[see Lesson 19]*

$$T_2 = \frac{n+1}{n} M_n - 1 \qquad\qquad E[T_2] = N$$

- Example 2: linear regression coefficients *[see Lesson 26]*
  - 95% confidence intervals (assuming $\underline{U_i \sim \mathcal{N}(0, \sigma^2)}$):

$$\hat{\beta} \pm t_{n-2,0.025} se(\hat{\beta}) \qquad\qquad \hat{\alpha} \pm t_{n-2,0.025} se(\hat{\alpha})$$

See R script

An application: probability of large errors

- Bootstrap principle: for $X \sim F$
  - the empirical distribution of $\Delta^* = \bar{X}_n^* - \bar{x}_n$ approximates the distribution of $\Delta = \bar{X}_n - \mu$
- Application: estimate $P_F(|\bar{X}_n - \mu| > 1)$ as
  - $P_{\hat{F}}(|\bar{X}_n^* - \bar{x}_n| > 1)$ and then by the fraction of $\delta^* = \bar{x}_n^* - \bar{x}_n$ such that $|\delta^*| > 1$

See R script

Wrap up on empirical bootstrap

- How many bootstrap samples?
  - There are $\binom{2n-1}{n-1}$ distinct bootstrap samples *[Why?]*
  - Suggested to use at least 1000 bootstrap samples
  - **Jackknife resampling**: bootstrap samples $x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n$, for $i = 1, \ldots, n$
- How good is the approximation by bootstrap?
  - Small perturbation to data-generating process should produce small perturbation of the parameter to estimate (θ)
  - Problems with extreme values, e.g., percentiles, maximum, etc.

Resampling methods for classifier performance estimation

- Decision rule $y^+_\theta(w)$ (classifier) or score function $s_\theta(w)$ (binary probabilistic classifier, Lesson 23)
- Loss function, e.g., 0-1 loss $\ell_\theta(c, w) = \mathbb{1}_{y^+_\theta(w) \neq c}$

Risk (or Expected Prediction Error EPE)

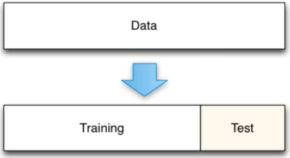
The risk w.r.t. a loss function ℓ_θ is $R(\theta_{TRUE}, \theta) = E_{(W,C) \sim f_{\theta_{TRUE}}}[\ell_\theta(C, W)]$.

Question: how to estimate risk given a dataset?

- **Holdout method:** split dataset into training and test, build $y^+_\theta()$ on training, estimate as the empirical risk on test set $(w_1, c_1), \ldots, (w_n, c_n)$:

$$\hat{r} = \frac{1}{n} \sum_{i=1}^{n} \ell_\theta(c_i, w_i) \qquad se = \sqrt{\frac{\hat{r}(1-\hat{r})}{n}}$$

[see Lesson 26 on CI for proportions]



  - Drawbacks: variability of training/test set, and then of empirical risk estimates

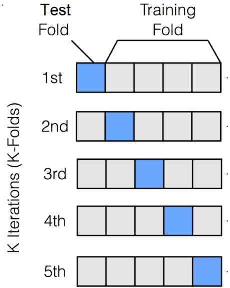
Resampling methods for classifier performance estimation

Question: how to estimate risk given a dataset?

- **Random sampling:** repeat holdout k times, and average the empirical risks:
  $\hat{r} = \frac{1}{k}\sum_{j=1}^{k}\hat{r}^j$ with $\hat{r}^j = \frac{1}{n_j}\sum_{i=1}^{n_j}\ell_\theta(c_i^j, w_i^j)$ is the error on j^{th} training-test split
- Standard error calculated as standard deviation over the k repetitions:
  $se = \sqrt{\frac{1}{k-1}\sum_{j=1}^{k}(\hat{r}^j - \hat{r})^2}$
  Wrong! As test sets (and then $\hat{r}^j$'s) are not independent!

Resampling methods for classifier performance estimation

Question: how to estimate risk given a dataset?



- k-**fold cross-validation:** average the empirical risks over k-fold splits:
  $$\hat{r} = \frac{1}{k}\sum_{j=1}^{k} \hat{r}^j \text{ with } \hat{r}^j = \frac{1}{n/k}\sum_{i=1}^{n/k} \ell_\theta(c_i^j, w_i^j)$$
- Standard deviation calculated over the k folds, with
  $$se = \sqrt{\frac{1}{k-1}\sum_j(\hat{r}^j - \hat{r})^2}$$
  Wrong!(⋆) Test sets are independent, but training sets (and then $\hat{r}^j$'s) are not!
- If classifier is stable over the folds (see [Kohavi, 1995]), use:
  $$se = \sqrt{\frac{\hat{r}(1-\hat{r})}{n}}$$
  [see Lesson 26 on CI for proportions]
  - Boils down to estimation as holdout but using all data instances (lower variability)!
  - This is the one implemented in R/caret
- Setting $k = n$ is the **leave-one out cross-validation** (LOOCV)

(⋆) CV should be treated as an estimator of the average prediction error across training sets!

Resampling methods for classifier performance estimation

Question: how to estimate risk given a dataset?

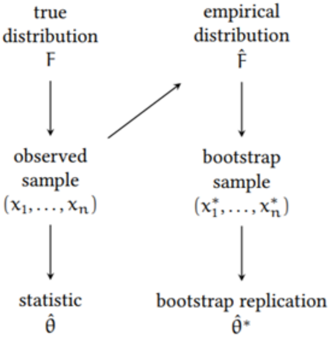
- training = bootstrap $x_1^*, \dots, x_n^*$, test = dataset $\setminus$ bootstrap = $\{x_1, \dots, x_n\} \setminus \{x_1^*, \dots, x_n^*\}$
  - .632 **bootstrap algorithm** for k bootstrap runs

$$\hat{r} = \frac{1}{k} \sum_j (0.632 \cdot \hat{r}^j + 0.368 \cdot \hat{r}_{tr})$$

  where $\hat{r}^j$ is the empirical risk on j^{th} bootstrap run, and $\hat{r}_{tr}$ is the empirical risk on the dataset
- [Kohavi, 1995, Kim, 2009] conclusions and recommendations:
  - Bootstrap has low variance, but it is extremely biased
  - k-fold cross-validation has low bias and variance can be controlled
    - by averaging multiple k-fold cross-validation
  - Recommendation: use **repeated (stratified) k-fold cross-validation**, with $k \approx 10$
- [Vanwinckelen, 2012] warns against "repeated", and it recommends k-**fold cross-validation**

See R script

Parametric bootstrap principle



- Let $X_1, \ldots, X_n \sim F(\gamma)$ be a random sample
  - with known family F but *unknown* parameter γ
- Estimator $T = h(X_1, \ldots, X_n)$, e.g., $\bar{X}_n = (X_1 + \ldots + X_n)/n$
- From a dataset $x_1, \ldots, x_n$, we can
  - derive an estimate $\hat{\gamma}$ of γ
- From $F(\hat{\gamma})$ we can generate (a lot of) *bootstrap samples* $x_1^*, \ldots, x_n^*$
  - as realizations of $X_1^*, \ldots, X_n^* \sim F(\hat{\gamma})$ *[a form of* **Monte Carlo simulation***]*

and then (many) bootstrap point estimates $\hat{\theta}^* = h(x_1^*, \ldots, x_n^*)$

Parametric bootstrap

PARAMETRIC BOOTSTRAP SIMULATION (FOR $\bar{X}_n - \mu$). Given a dataset $x_1, x_2, \ldots, x_n$, compute an estimate $\hat{\theta}$ for θ. Determine $F_{\hat{\theta}}$ as an estimate for F_θ, and compute the expectation $\mu^* = \mu_{\hat{\theta}}$ corresponding to $F_{\hat{\theta}}$.

1. Generate a bootstrap dataset $x_1^*, x_2^*, \ldots, x_n^*$ from $F_{\hat{\theta}}$.
2. Compute the centered sample mean for the bootstrap dataset:

$$\bar{x}_n^* - \mu_{\hat{\theta}},$$

   where

$$\bar{x}_n^* = \frac{x_1^* + x_2^* + \cdots + x_n^*}{n}.$$

Repeat steps 1 and 2 many times.

- Cfr with non-parametric bootstrap: use $\mu_{\hat{\theta}}$ instead of $\bar{x}_n$
- Use the empirical distribution of $\delta^* = \bar{x}_n^* - \mu_{\hat{\theta}}$ for estimating
  - confidence interval for $\delta = \bar{x}_n - \mu$ is $(q_{\alpha/2}, q_{1-\alpha/2})$ of δ^* empirical distribution
  - $q_{\alpha/2} \leq \delta = \bar{x}_n - \mu \leq q_{1-\alpha/2}$ implies c.i. for μ is $(\bar{x}_n - q_{1-\alpha/2}, \bar{x}_n - q_{\alpha/2})$

See R script

Application: distribution fitting

- Consider $x_1, \ldots, x_n$ realizations of a random sample $X_1, \ldots, X_n \sim F$
- Is the dataset from an $Exp(\lambda)$ for some λ? I.e., is it $F = Exp(\lambda)$?
- We estimate $\hat{\lambda} = 1/\bar{x}_n$ *[MLE estimation]*
- We measure how close is the dataset to the distribution as:

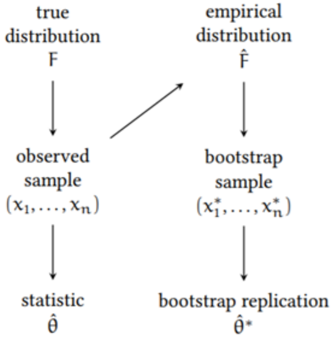
$$t_{ks} = \sup_{a \in \mathbb{R}} |F_n(a) - F_{\hat{\lambda}}(a)|$$

  where:
  - $F_n(a)$ is the empirical cumulative distribution function of $x_1, \ldots, x_n$
  - $F_{\hat{\lambda}}(a) = 1 - e^{\hat{\lambda} a}$, for $a \geq 0$, is the CDF of $Exp(\hat{\lambda})$
  - t_{ks} is the *Kolmogorov-Smirnov* distance *[See Lesson 11]*
- if $F = Exp(\lambda)$ then both $F_n \approx F$ and $F_{\hat{\lambda}} \approx F$, and then $F_n \approx F_{\hat{\lambda}}$, so that t_{ks} is small
- if $F \neq Exp(\lambda)$ then $F_n \approx F \neq Exp(\lambda) \approx F_{\hat{\lambda}}$, so that t_{ks} is large

Application: distribution fitting

- For the software dataset from the textbook
  - $\hat{\lambda} = 0.0015$ and $t_{ks} = 0.17$
- Is $t_{ks} = 0.17$ expected or an extreme value?
- Let's study the distribution of the bootstrap estimator:

$$T_{ks} = \sup_{a \in \mathbb{R}} |F_n^*(a) - F_{\hat{\Lambda}^*}(a)|$$



where:

  - $X_1^*, \ldots, X_n^* \sim Exp(\hat{\lambda})$ is a bootstrap sample
  - $F_n^*(a)$ is the empirical cumulative distribution of the bootstrap sample
  - $\hat{\Lambda}^* = 1/\bar{X}_n^*$
- It turns out $P(T_{ks} > 0.17) \approx 0$, unlikely that $Exp(\lambda)$ is the right model

See R script

Optional references

- Ji-HyunKim (2009)
  Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap.
  Computational Statistics & Data Analysis, 53 (11): 3735-3745

- Ron Kohavi (1995)
  A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection.
  Proc. of IJCAI 1995: 1137-1145

- Gitte Vanwinckelen and Hendrik Blockeel (2012)
  On Estimating Model Accuracy with Repeated Cross-Validation.
  Proc. of BeneLearn and PMLS 2012: 39 - 44

- Michael R. Chernick and Robert A. LaBudde (2011)
  An introduction to Bootstrap methods with applications to R.
  John Wiley & Sons, Inc.