Introduction to statistical learning

3.4 Supervised learning: Cross validation & Bootstrap

V. Lefieux

June 2018

Introduction
Cross validation
Bootstrap
References

Table of contents

Introduction

Cross validation

Bootstrap

Table of contents

Introduction

Cross validation

Bootstrap

Regression and classification

Introduction
Cross validation
Bootstrap
References

- Consider a sample $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ where $x_i \in \mathcal{X}$ (commonly $\mathbb{R}^p$) and $y_i \in \mathcal{Y}$.
- The aim is to forecast y for a new value x.
- Two cases can be distinguished: regression and classification:
  - Regression case: $\mathcal{Y} = \mathbb{R}$.
  - Classification, for a binary discrimination: $\mathcal{Y} = \{-1, 1\}$.

Prediction

Introduction
Cross validation
Bootstrap
References

- We assume that (x_i, y_i) is a realization of the random variable (X_i, Y_i) with unknown joint probability distribution $P_{X,Y}$ (nonparametric statistical model).
- A prediction of Y is a measurable function $f : \mathcal{X} \to \mathcal{Y}$.

Performance of a model

The performance of a model is evaluated by the prediction risk (or generalization error, or prediction error) which:

- allows model selection,
- provides a measure of the confidence that can be placed in the forecast.

The risk is defined from a loss function.

Loss function

- The function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}^+$ is a loss function if:
  - $\ell(y, y') > 0$,
  - $\ell(y, y) > 0$ for $y \neq y'$.

- Examples of loss functions:
  - Regression:
    $$\ell(y, y') = |y - y'|^q$$
    with $q \in \mathbb{R}^+$.
  - Binary discrimination:
    $$\ell(y, y') = \mathbb{1}_{y \neq y'} = \frac{|y - y'|}{2} = \frac{(y - y')^2}{4} .$$

Risk

Introduction
Cross validation
Bootstrap
References

The risk of a prediction function f is:

$$R(f) = \mathbb{E}\left[\ell\left(Y, f\left(X\right)\right)\right] \ .$$

Minimize the risk

We want to consider prediction functions which minimize this risk.

For example, consider the regression:

- If $\ell(y, y') = |y - y'|$ then:

$$\widehat{f}_{opt}(x) = \mathrm{Med}(Y / X = x) \ .$$

- If $\ell(y, y') = (y - y')^2$ then:

$$\widehat{f}_{opt}(x) = \mathbb{E}(Y / X = x) \ .$$

Empirical risk

Introduction
Cross validation
Bootstrap
References

The empirical risk is:

$$\widehat{R}(f) = \frac{1}{n} \sum_{i=1}^{n} \ell(Y, f(X)) \ .$$

Issue

- The empirical risk under-estimates the risk.
- This can lead to overfitting.
- There are different strategies to obtain an unbiased estimate of this risk:
  - Mathematical adjustment of the training error (which underestimates the test error): AIC, BIC, C_p, ...
  - Resampling methods: cross validation or bootstrap.

Bias-variance trade-off I

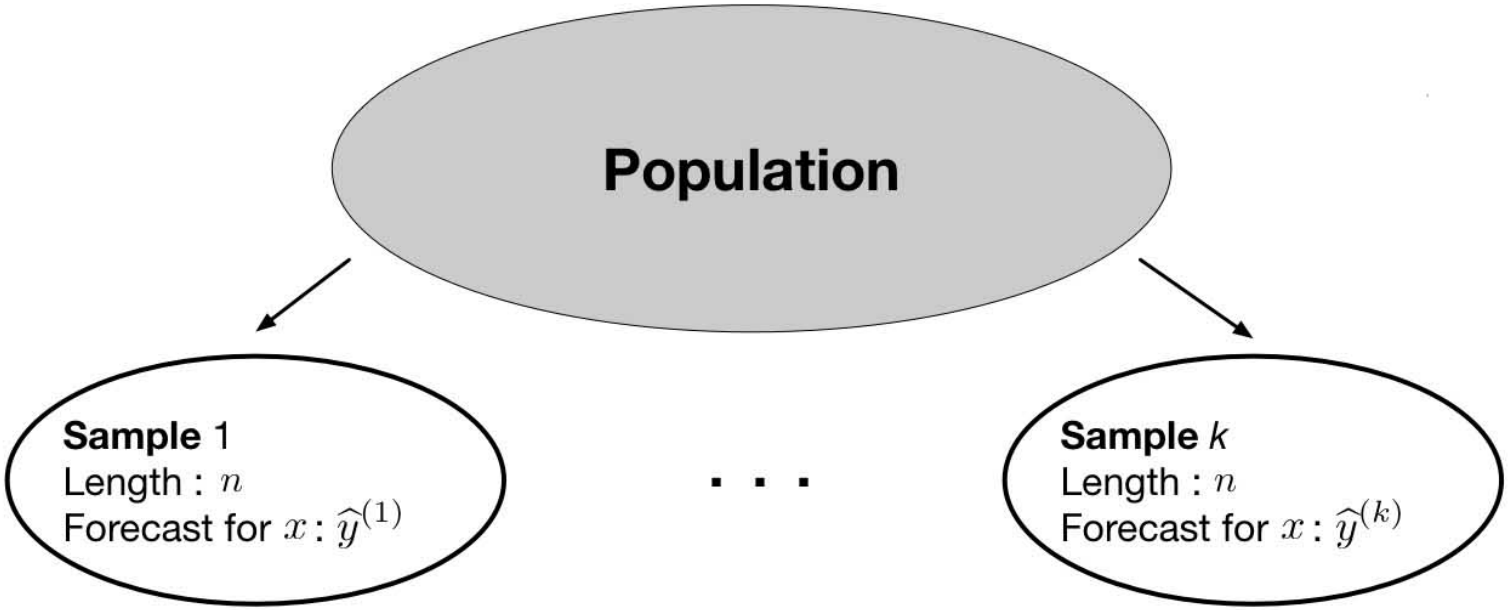
For a regression model:

$$y = f(x) + \varepsilon \ .$$

The bias-variance decomposition of the MSE is:

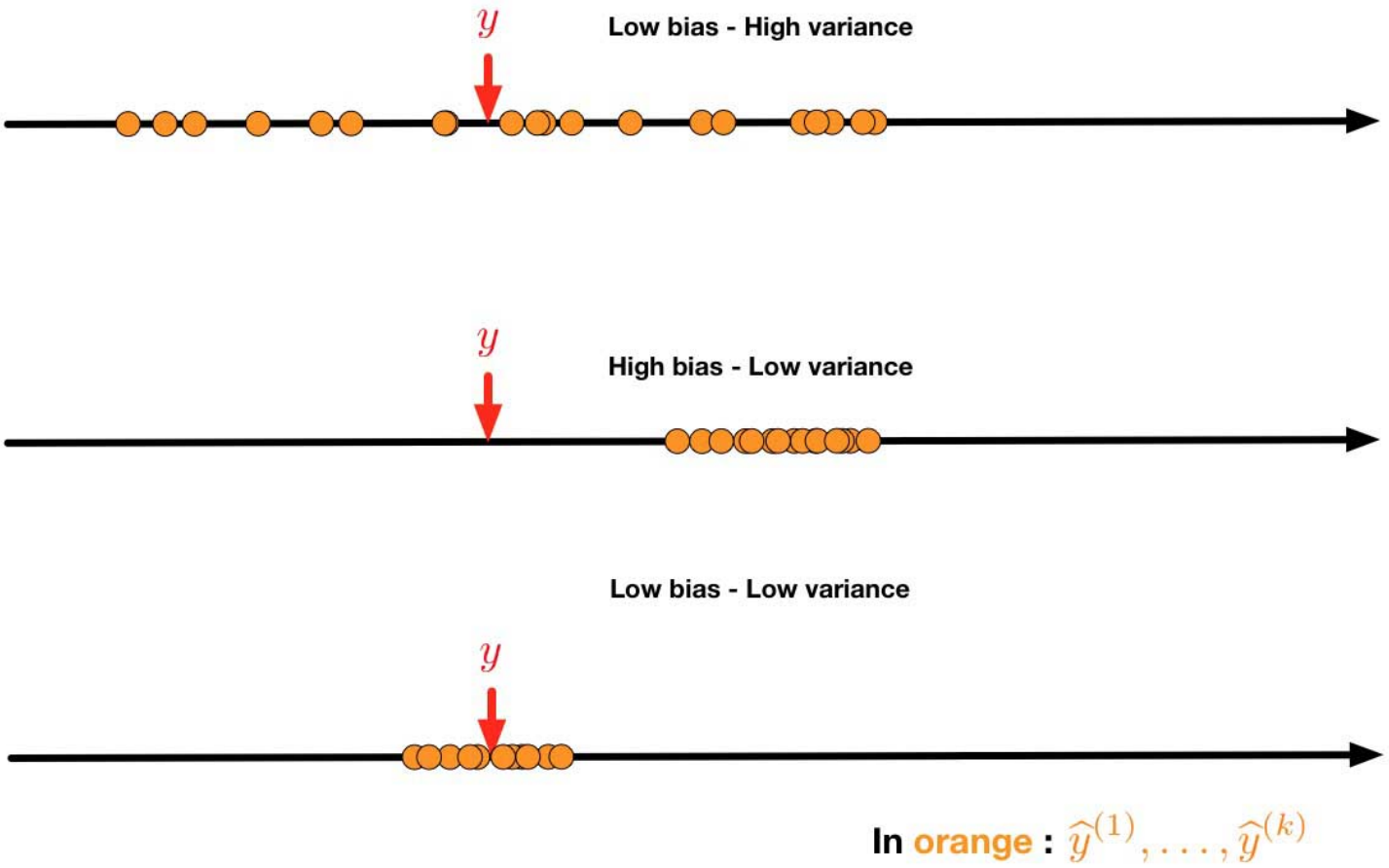
$$\begin{aligned}
&\text{MSE} \\
&:= \mathbb{E}\left[(y - \widehat{y})^2\right] \\
&= \mathbb{E}\left[\left(y - \widehat{f}(x)\right)^2\right] \\
&= \mathrm{Var}\left(f(x) - \widehat{f}(x)\right) + \left[\mathbb{E}\left(f(x)\right) - \mathbb{E}\left(\widehat{f}(x)\right)\right]^2 + \mathrm{Var}(\varepsilon)
\end{aligned}$$

Bias-variance trade-off II



Introduction
Cross validation
Bootstrap
References

Bias-variance trade-off III



Introduction
Cross validation
Bootstrap
References

Table of contents

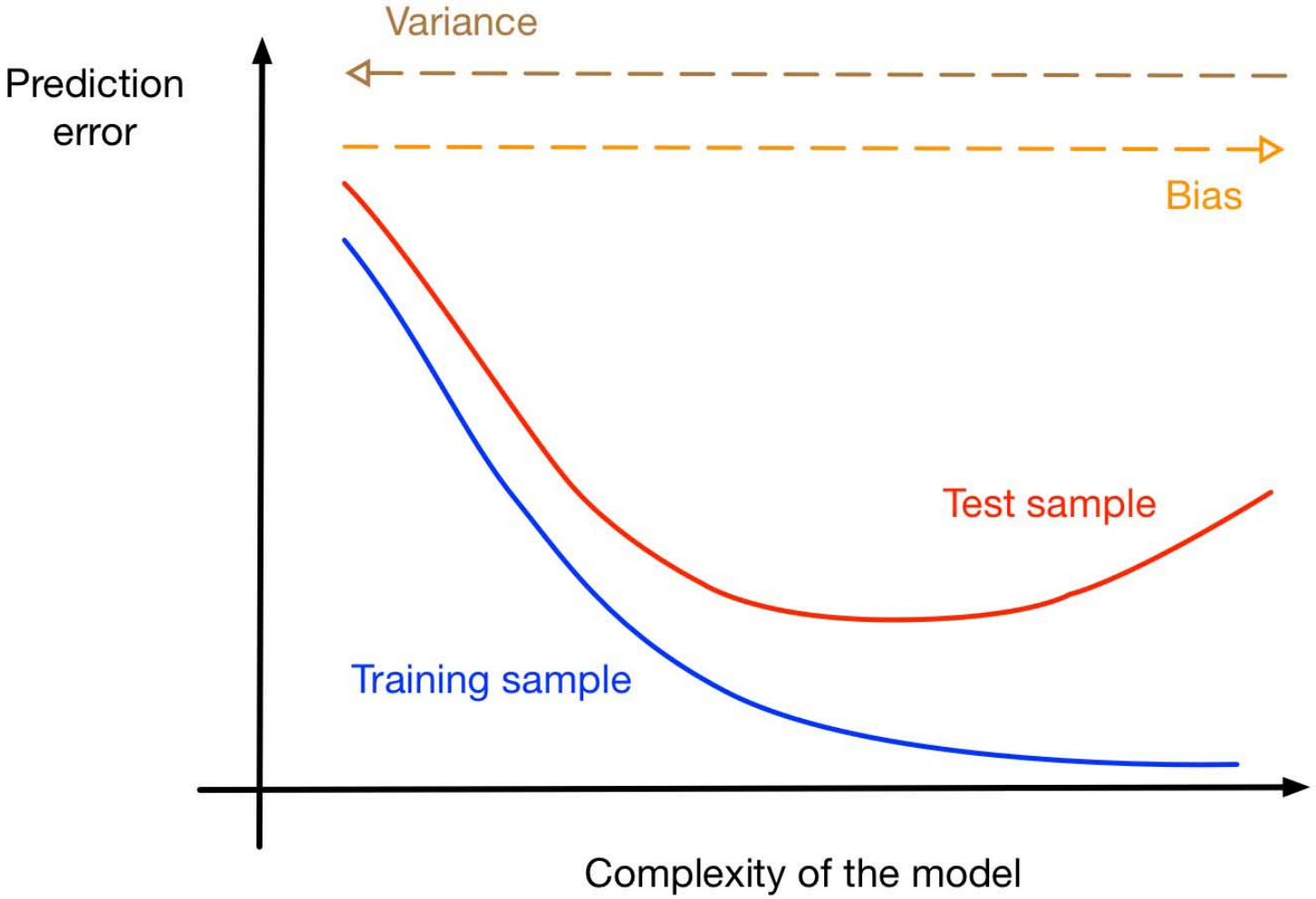
Introduction

Cross validation

Bootstrap

Reminder

The training error rate often underestimate the test error rate.



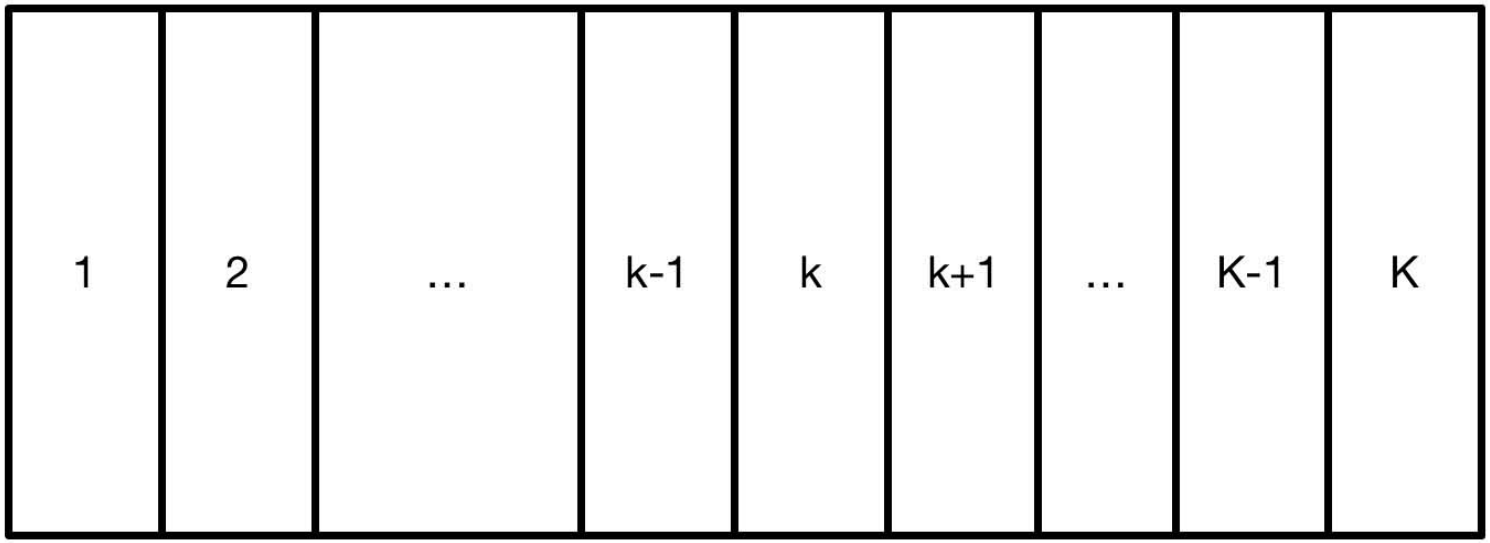
Introduction
Cross validation
Bootstrap
References

K-fold cross-validation I

- Randomly divide the data into K "equal-sized" parts. The part k contains n_k observations, $n_k = \frac{n}{K}$ if n is a multiple of k.
- For $k \in \{1, \ldots, K\}$:
  1. Leave out part k from the training test.
  2. Estimate the prediction function on the training test.
  3. Compute the test error on part k: MSE_k.
- Compute the cross validation criterion:

$$\mathrm{CV} = \sum_{k=1}^{K} \frac{n_k}{n} \mathrm{MSE}_k \ .$$

K-fold cross-validation II



K-fold cross-validation III

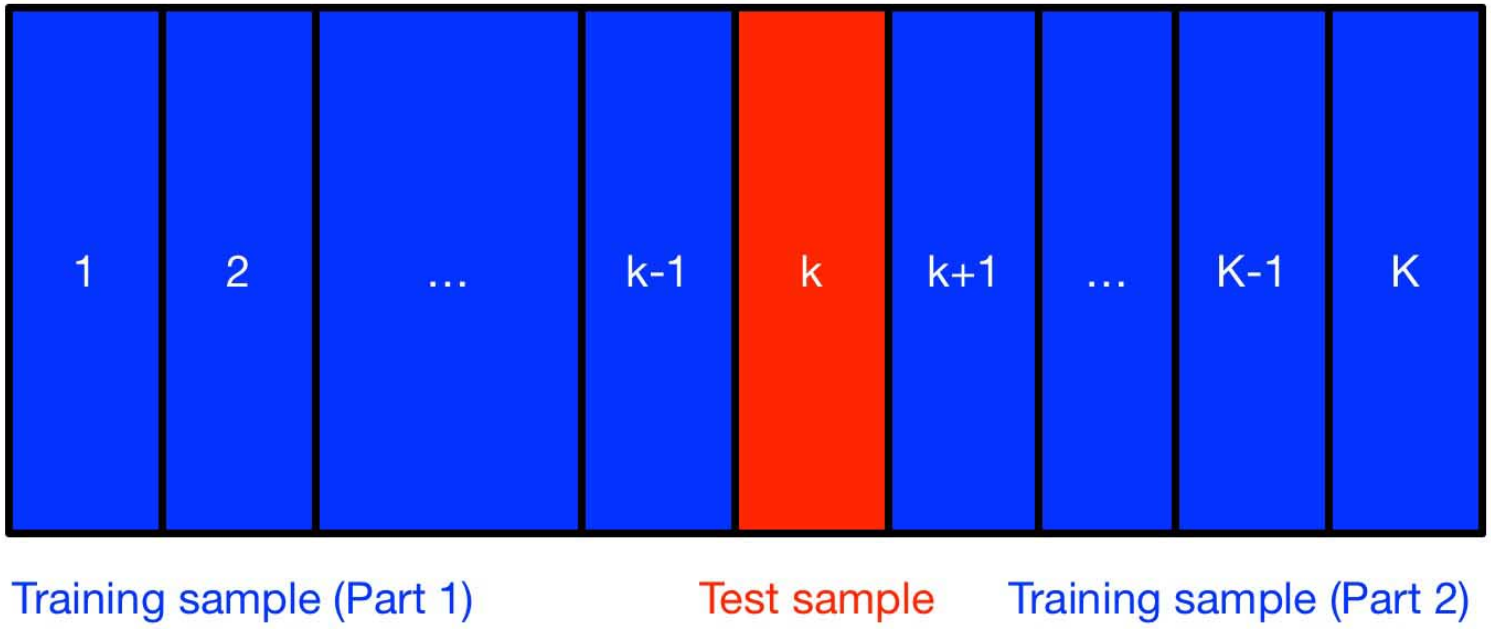


Table of contents

Introduction

Cross validation

Bootstrap

Origin

From **The surprising adventures of Baron Munchausen** from Rudolph Erich Raspe :

"*The Baron had fallen to the bottom of a deep lake. Just when it looked like all was lost, he thought to pick himself up by his own bootstraps.*"

Aim

Quantify the uncertainty of a statistical estimator or a statistical learning method.

Introduction
Cross validation
Bootstrap
References

Principle

- Cross validation: works on independent data sets from the population.
- Bootstrap : works on samples of the same size as the original one, resampling the original data set **with replacement**.
  Some observations may appear more than once, others do not appear in a bootstrap sample.

References

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The elements of statistical learning. Data Mining, inference, and prediction*. Springer Series in Statistics. Springer, 2 edition.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2015). *An introduction to statistical learning with applications in R*. Springer Texts in Statistics. Springer.