Evaluating machine learning models: Introduction

Roman Hornung[1,2], Sebastian Fischer[3], Anne-Laure Boulesteix[1,2]

[1]Institute for Medical Information Processing, Biometry and Epidemiology, LMU Munich, Munich, Germany

[2]Munich Center for Machine Learning (MCML), Munich, Germany

[3]Department of Statistics, LMU Munich, Munich, Germany

Online Webinar (Wiesbaden/Munich)

June, 26th, 2024

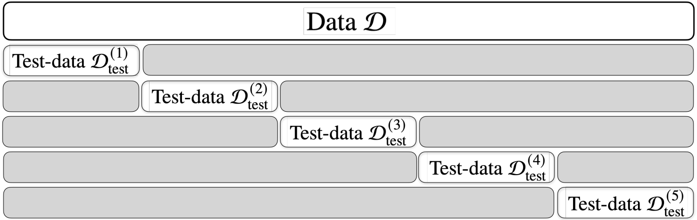
AP1.2: "Evaluation der Prädiktionsgüte in komplexen Situationen"

- PI: Prof. Dr. Anne-Laure Boulesteix
- Staff: Dr. Roman Hornung, H. Schulz-Kümpel, S. Fischer
- Area: Supervised learning (prediction) using ML algorithms
- Focus: Evaluation of prediction performance
- Methods: Resampling-based procedures (e.g., cross-validation)
- Addressed issues:
  - Evaluation of prediction performance in case of violation of the i.i.d. assumption
  - Confidence intervals for performance estimated through resampling

Output: Two reviews/comparison studies

- Hornung, R., Nalenz, N., Schneider, L., Bender, A., Bothmann, L., Bischl, B., Augustin, A., Boulesteix, A.-L., 2023. Evaluating machine learning models in non-standard settings: An overview and new findings. arXiv:2310.15108. Statistical Science (to appear).
- H. Schulz-Kümpel*, S. Fischer*, R. Hornung, A.-L. Boulesteix, Thomas Nagler, Bernd Bischl. Constructing confidence intervals for 'the' generalization error – a comprehensive benchmark study. arXiv:2409.18836. Data-Centric Machine Learning Research (to appear). *contributed equally.

A basic resampling procedure: K-fold cross-validation



- In K-fold cross-validation, the **dataset** is **randomly partitioned** into K **subsets** of (nearly) equal size, termed "folds".
- **Each fold** serves **once** as the **testing set**, while the **remaining folds** collectively form the **training set**.
- More **broadly**, **resampling** involves the **random and repeated splitting** of data into distinct training and testing sets.

Project 1: Complex structures beyond i.i.d.

- Machine learning (**ML**) applications in **official statistics** frequently deal with **complex data structures** most of which violate the standard i.i.d. assumption:
  - clustered data
  - spatial data
  - unequal sampling probabilities
  - concept drift
- The presence of these complex structures **can introduce bias** in generalization error (**GE**) **estimates** derived from **ordinary resampling** methods like cross-validation.
- **Tailored resampling** techniques are **necessary** for each data structure to obtain (largely) unbiased GE estimates.

Project 2: Confidence intervals

- Imagine an accuracy of, say, 90% is estimated using a resampling technique.
- As all estimators, this estimator has a variance.
  How reliable is it?
- To interpret it, we need a confidence interval.
- Naive approach: Consider the K-folds as i.i.d. observations to derive a confidence interval in the "usual way".
- Problem: The results for the K folds are not i.i.d.!
- **More sophisticated techniques are necessary**.