Tutorial Proposal: Applied Machine Learning and Efficient Model Selection with mlr

Bernd Bischl, Michel Lang
Department of Statistics, TU Dortmund University

Brief description of the tutorial R does not define a standardized interface for all its machine learning algorithms. Therefore, for any non-trivial experiments you need to write lengthy, tedious and error-prone wrappers to call the different algorithms and unify their respective output. The mlr package offers a clean, easy-to-use and flexible domain specific language for machine learning experiments in R. It supports classification, regression, clustering and survival analysis and connects to nearly a hundred predictive modeling techniques. The package allows for different hyperparameter optimization and configuration techniques, including iterated F-racing and sequential model based optimization. Variable selection is possible through various filter and wrapper approaches.

Hence, mlr allows data analysts who are neither experts in machine learning nor seasoned R programmers to nevertheless specify and complex machine learning experiments in short, succinct and scalable code. Experienced programmers, on the other hand, get to wield a large, well-designed toolbox, which they can customize and extend to quickly construct their own algorithms.

The course will enable the participants to understand and apply the basic mlr operations for data handling and preprocessing, model building, evaluation and resampling. After these basics are covered we will especially focus on the important aspects of benchmarking, model selection and hyperparameter tuning. As all of these usually require a large amount of computational resources in realistic applications, we will show how to easily parallelize them in common parallel environments. The course will end with a short demonstration on how to access the new OpenML server for open machine learning (`http://www.openml.org`) which provides a large repository of benchmark data sets and enables reproducible experiments and meta analysis.

Project page: `https://www.github.com/berndbischl/mlr/`
Online tutorial: `https://berndbischl.github.io/mlr/tutorial/html/`

Detailed Outline

- Very brief intro to applied machine learning

- Data handling and machine learning tasks

- Classification, regression, clustering and survival modeling with mlr

- Performance evaluation and resampling

- Visual model analysis

- Parallelization and high-performance computing

- Model selection and hyper-parameter tuning

- Interfacing the OpenML server with mlr

Background knowledge required   Basic knowledge of R and machine learning

Potential attendees   Anybody from academia or industry with an interest in modern machine learning