mboost - Componentwise Boosting for Generalised Regression Models

Thomas Kneib & Torsten Hothorn
Department of Statistics
Ludwig-Maximilians-University Munich

In recent years, boosting has emerged into a widely applied technique for fitting various types of generalised regression models. The main reason for its popularity is that it is surprisingly simple in requiring only iterative fitting of some (potentially simple) base-learning procedure such as (penalised) least-squares to working residuals. Moreover, boosting allows to define various types of regression situations by formulating them in terms of a suitable loss function. From a theoretical perspective, boosting then equals a functional gradient descent algorithm for solving the empirical risk minimisation problem and the working residuals are given by the negative gradient of the loss function.

While boosting has mainly been used to fit completely nonparametric black box models in a prediction-oriented framework first, recent research has shown that it can actually be used to estimate structured regression models. Therefore the base-learning procedure is separated into several components and only the best-fitting component is updated in each iteration. For example, when fitting a generalised linear model, each base-learner might correspond to a single covariate and only the effect of the best-fitting covariate is updated in each boosting iteration. Applying a suitable stopping rule to the boosting iterations yields an adaptively regularised model fit that also provides a means of variable selection and model choice.

The package **mboost** provides implementations for the most common types of univariate exponential family responses where the negative log-likelihood provides the loss-function, but also for other types of regression situations such as robust regression based on Huber's loss function. Further extensions, for example to survival modelling are currently being investigated.

A wide range of componentwise base-learning procedures is available based on (penalised) least squares fits, for example

- parametric linear effects as in generalised linear models,
- penalised splines for nonparametric effects and varying coefficient terms,
- penalised tensor product splines for interaction surfaces and spatial effects,
- ridge regression for random intercepts and slopes,
- stumps for piecewise constant functions.

Through its modular formulation, boosting allows to define models consisting of arbitrary combinations of these effects and we will illustrate the versatility of the resulting model class in a spatio-temporal regression model for the analysis of forest health.