A toolkit for cross-validation: The R package cvTools

Andreas Alfons

ORSTAT Research Center, Faculty of Business and Economics, KU Leuven
*Contact author: andreas.alfons@econ.kuleuven.be

Keywords: Cross-validation, Package development, Prediction performance, Regression models

Cross-validation is a widely used technique in applied statistics to estimate the prediction performance of regression models on a given data set. In addition, advanced regression methods often have tuning parameters, and cross-validation is a popular way to select the final model.

The idea of cross-validation is simple and easy to implement: split the data into several blocks, leave out one block for model estimation, and predict the values of the left-out block. Those predictions are then used to compute a certain prediction loss function. Even though the basic procedure is simple, some additional programming effort is necessary for more complex procedures such as repeated (double) cross-validation, or using cross-validation to select the optimal combination of tuning parameters. While many packages for computing regression models already offer functionality for cross-validation, different packages use different interfaces and the returned objects have a different structure. Furthermore, developers often copy and paste the basic code skeleton when implementing cross-validation for different models, which complicates maintaining the code.

The R package cvTools [1] tackles those problems. It is mainly aimed at package developers and offers functions to carry out all iterations of complex cross-validation procedures with one simple command, including performing cross-validation for all possible combinations of tuning parameters of a method. The programming effort of implementing cross-validation for a certain model is thus greatly reduced. A typical function for cross-validation based on package cvTools consists of the following two parts: first the data is extracted from the model, then only one or two more lines of code are necessary to perform cross-validation.

The main advantage of cvTools is that functions building upon the package provide a unified interface to cross-validation and the same object structure for the results, which benefits users. Cross-validation methods for objects returned by popular regression functions are thereby already included in cvTools. Another advantage for users is that several cross-validation objects can be combined into a larger object, which then reports the optimal model. Moreover, different plots are available in cvTools for visual inspection of the cross-validation results.

References