Journal of Statistical Software

MMMMMM YYYY, Volume VV, Issue II. http://www.jstatsoft.org/

Cross-Validated (Nested) Forward Selection: the R Package nestfs

Marco Colombo, Felix Agakov, Paul McKeigue
University of Edinburgh

Abstract

Forward selection is a well-know technique for constructing a predictive model that has the properties of sparsity and interpretability. As for other techniques used to solve a prediction problem, it can be easily mis-used if its parameters are learned outside of a cross-validation setting, thus producing over-optimistic estimates and models that do not generalise well in prediction of withdrawn data.

In this paper we present an implementation of forward selection for the R programming language which adopts cross-validation as a core component of the selection procedure. The **nestfs** package features several selection and termination criteria, and is parallelised over the cross-validation folds.

Keywords: forward selection, cross-validation, R.

1. Literature review

Should describe the history of forward selection: when and why it was introduced, what sort of problems it was originally designed to solve.

Talk about techniques that build on it or are comparable to it (stepwise selection, backward elimination). Mention other types of parametric or non-parametric ways of building a predictive model.

Talk about known advantages (interpretability, sparsity) and disadvantages (speed) compared to lasso-type models.

Talk about types of problems for which forward selection struggles: large number of predictors, correlated predictors. Talk about how these can be solved with filtering, removal of correlated variables.

2. Software review

Talk about what other implementations exist, specifically for R.

3. Introducing nestfs

Talk about the general setup: iterations, inner folds, selection criteria, termination criteria.

Talk about cross-validation and nested forward selection.

Talk about parallelisation options.

4. Practical examples

Using some existing data sets, show how to use the package. Compare and contrasts different options on the same data.

Show what the package saves and reports.

5. Conclusions

Affiliation:

Marco Colombo
Centre for Population Health Sciences
University of Edinburgh
Medical School
Teviot Place
Edinburgh
EH8 9AG
E-mail: m.colombo@ed.ac.uk
URL: http://maths.ed.ac.uk/~mcolombo/

Journal of Statistical Software http://www.jstatsoft.org/
published by the American Statistical Association http://www.amstat.org/
Volume VV, Issue II *Submitted:* yyyy-mm-dd
MMMMMM YYYY *Accepted:* yyyy-mm-dd