An extensive experimental survey of regression methods

Título An extensive experimental survey of regression methods

Autores M. Fernández-Delgado, M.S. Sirsat, E. Cernadas, S. Alawadi, S. Barro, M. Febrero-Bande

Tipo Artículo de revista

Fonte Neural Networks, PERGAMON-ELSEVIER SCIENCE LTD , Vol. 111, No. Marzo, pp. 11-34 , 2019.

Rank Ranked Q1 in Cognitive Neuroscience by SJR

ISSN 0893-6080

DOI 10.1016/j.neunet.2018.12.010

Abstract Regression is a very relevant problem in machine learning, with many different available approaches. The current work presents a comparison of a large collection composed by 77 popular regression models which belong to 19 families: linear and generalized linear models, generalized additive models, least squares, projection methods, LASSO and ridge regression, Bayesian models, Gaussian processes, quantile regression, nearest neighbors, regression trees and rules, random forests, bagging and boosting, neural networks, deep learning and support vector regression. These methods are evaluated using all the regression datasets of the UCI machine learning repository (83 datasets), with some exceptions due to technical reasons. The experimental work identifies several outstanding regression models: the M5 rule-based model with corrections based on nearest neighbors (cubist), the gradient boosted machine (gbm), the boosting ensemble of regression trees (bstTree) and the M5 regression tree. Cubist achieves the best squared correlation (R^2) in 15.7% of datasets being very near to it, with difference below 0.2 for 89.1% of datasets, and the median of these differences over the dataset collection is very low (0.0192), compared e.g. to the classical linear regression (0.150). However, cubist is slow and fails in several large datasets, while other similar regression models as M5 never fail and its difference to the best R^2 is below 0.2 for 92.8% of datasets. Other well-performing regresors are the committee of neural networks (avNNet), extremely randomized regression trees (extraTrees, which achieves the best R^2 in 33.7% of datasets), random forest (rf) and varepsilon-support vector regression (svr), but they are slower and fail in several datasets. The fastest regression model is least angle regression lars, which is 70 and 2,115 times faster than M5 and cubist, respectively. The model which requires least memory is non-negative least squares (nnls), about 2 GB, similarly to cubist, while M5 requires about 8 GB. For 97.6% of datasets there is a regression model among the 10 bests which is very near (difference below 0.1) to the best R^2, which increases to 100% allowing differences of 0.2. Therefore, provided that our dataset and model collection are representative enough, the main conclusion of this study is that, for a new regression problem, some model in our top-10 should achieve R^2 near to the best attainable for that problem.

Palabras chave Regression, UCI machine learning repository, cubist, M5, gradient boosted machine, extremely randomized regression tree, support vector regression penalized linear regression.

LIGAZÓNS

Versión da editorial

DESCARGAS

Referencia BibTex

Descargar preprint

DATOS ADICIONAIS

Datos e software adicionais