CRAN Task View: Machine Learning & Statistical Learning

Maintainer: Torsten Hothorn
Contact: Torsten.Hothorn at R-project.org
Version: 2014-12-18

Several add-on packages implement ideas and methods developed at the borderline between computer science and statistics - this field of research is usually referred to as machine learning. The packages can be roughly structured into the following topics:

- *Neural Networks* : Single-hidden-layer neural network are implemented in package nnet (shipped with base R). Package RSNNS offers an interface to the Stuttgart Neural Network Simulator (SNNS). An interface to the FCNN library allows user-extensible artificial neural networks in package FCNN4R.
- *Recursive Partitioning* : Tree-structured models for regression, classification and survival analysis, following the ideas in the CART book, are implemented in rpart (shipped with base R) and tree. Package rpart is recommended for computing CART-like trees. A rich toolbox of partitioning algorithms is available in Weka , package RWeka provides an interface to this implementation, including the J4.8-variant of C4.5 and M5. The Cubist package fits rule-based models (similar to trees) with linear regression models in the terminal leaves, instance-based corrections and boosting. The C50 package can fit C5.0 classification trees, rule-based models, and boosted versions of these.
  Two recursive partitioning algorithms with unbiased variable selection and statistical stopping criterion are implemented in package party. Function `ctree()` is based on non-parametrical conditional inference procedures for testing independence between response and each input variable whereas `mob()` can be used to partition parametric models. Extensible tools for visualizing binary trees and node distributions of the response are available in package party as well.
  Tree-structured varying coefficient models are implemented in package vcrpart.
  For problems with binary input variables the package LogicReg implements logic regression. Graphical tools for the visualization of trees are available in package maptree.
  Trees for modelling longitudinal data by means of random effects is offered by package REEMtree. Partitioning of mixture models is performed by RPMM.
  Computational infrastructure for representing trees and unified methods for predition and visualization is implemented in partykit. This infrastructure is used by package evtree to implement evolutionary learning of globally optimal trees. Oblique trees are available in package oblique.tree.
- *Random Forests* : The reference implementation of the random forest algorithm for regression and classification is available in package randomForest. Package ipred has bagging for regression, classification and survival analysis as well as bundling, a combination of multiple models via ensemble learning. In addition, a random forest variant for response variables measured at arbitrary scales based on conditional inference trees is implemented in package party. randomForestSRC implements a unified treatment of Breiman's random forests for survival, regression and classification problems. Quantile regression forests quantregForest allow to regress quantiles of a numeric response on exploratory variables via a random forest approach. For binary data, LogicForest is a forest of logic regression trees (package LogicReg. The varSelRF and Boruta packages focus on variable selection by means for random forest algorithms. For large data sets, package bigrf computes random forests in parallel and uses large memory objects to store the data.
- *Regularized and Shrinkage Methods* : Regression models with some constraint on the parameter estimates can be fitted with the lasso2 and lars packages. Lasso with simultaneous updates for groups of parameters (groupwise lasso) is available in package grplasso; the grpreg package implements a number of other group penalization models, such as group MCP and group SCAD. The L1 regularization path for generalized linear models and Cox models can be obtained from functions available in package glmpath, the entire lasso or elastic-net regularization path (also in elasticnet) for linear regression, logistic and multinomial regression models can be obtained from package glmnet. The penalized package provides an alternative implementation of lasso (L1) and ridge (L2) penalized regression models (both GLM and Cox models). Package RXshrink can be used to identify and display TRACEs for a specified shrinkage path and to determine the appropriate extent of shrinkage. Semiparametric additive hazards models under lasso penalties are offered by package ahaz. A generalisation of the Lasso shrinkage technique for linear regression is called relaxed lasso and is available in package relaxo. Fisher's LDA projection with an optional LASSO penalty to produce sparse solutions is implemented in package penalizedLDA. The shrunken centroids classifier and utilities for gene expression analyses are implemented in package pamr. An implementation of multivariate adaptive regression splines is available in package earth. Variable selection through clone selection in SVMs in penalized models (SCAD or L1 penalties) is implemented in package penalizedSVM. Various forms of penalized discriminant analysis are implemented in packages hda, rda, and sda. Package LiblineaR offers an interface to the LIBLINEAR library. The ncvreg package fits linear and logistic regression models under the the SCAD and MCP regression penalties using a coordinate descent algorithm. High-throughput ridge regression (i.e., penalization with many predictor variables) and heteroskedastic effects models are the focus of the bigRR package. An implementation of bundle methods for regularized risk minimization is available form package bmrm.
- *Boosting* : Various forms of gradient boosting are implemented in package gbm (tree-based functional gradient descent boosting). The Hinge-loss is optimized by the boosting implementation in package bst. Package GAMBoost can be used to fit generalized additive models by a boosting algorithm. An extensible boosting framework for generalized linear, additive and nonparametric models is available in package mboost. Likelihood-based boosting for Cox models is implemented in CoxBoost and for mixed models in GMMBoost. GAMLSS models can be fitted using boosting by gamboostLSS.
- *Support Vector Machines and Kernel Methods* : The function `svm()` from e1071 offers an interface to the LIBSVM library and package kernlab implements a flexible framework for kernel learning (including SVMs, RVMs and other kernel learning algorithms). An interface to the SVMlight implementation (only for one-against-all classification) is provided in package klaR. The relevant dimension in kernel feature spaces can be estimated using rdetools which also offers procedures for model selection and prediction.
- *Bayesian Methods* : Bayesian Additive Regression Trees (BART), where the final model is defined in terms of the sum over many weak learners (not unlike ensemble methods), are implemented in package BayesTree. Bayesian nonstationary, semiparametric nonlinear regression and design by treed Gaussian processes including Bayesian CART and treed linear models are made available by package tgp.
- *Optimization using Genetic Algorithms* : Packages rgp and rgenoud offer optimization routines based on genetic algorithms. The package Rmalschains implements memetic algorithms with local search chains, which are a special type of evolutionary algorithms, combining a steady state genetic algorithm with local search for real-valued parameter optimization.
- *Association Rules* : Package arules provides both data structures for efficient handling of sparse binary data as well as interfaces to implementations of Apriori and Eclat for mining frequent itemsets, maximal frequent itemsets, closed frequent itemsets and association rules.
- *Fuzzy Rule-based Systems* : Package frbs implements a host of standard methods for learning fuzzy rule-based systems from data for regression and classification. Package RoughSets provides comprehensive implementations of the rough set theory (RST) and the fuzzy rough set theory (FRST) in a

single package.

- *Model selection and validation* : Package e1071 has function `tune()` for hyper parameter tuning and function `errorest()` (ipred) can be used for error rate estimation. The cost parameter C for support vector machines can be chosen utilizing the functionality of package svmpath. Functions for ROC analysis and other visualisation techniques for comparing candidate classifiers are available from package ROCR. Packages hdi and stabs implement stability selection for a range of models, hdi also offers other inference procedures in high-dimensional models.
- *Meta packages* : Package caret provides miscellaneous functions for building predictive models, including parameter tuning and variable importance measures. The package can be used with various parallel implementations (e.g. MPI, NWS etc). In a similar spirit, package mlr offers a high-level interface to various statistical and machine learning packages.
- *Elements of Statistical Learning* : Data sets, functions and examples from the book The Elements of Statistical Learning: Data Mining, Inference, and Prediction by Trevor Hastie, Robert Tibshirani and Jerome Friedman have been packaged and are available as ElemStatLearn.
- *GUI* rattle is a graphical user interface for data mining in R.

CORElearn implements a rather broad class of machine learning algorithms, such as nearest neighbors, trees, random forests, and several feature selection methods. Similar, package rminer interfaces several learning algorithms implemented in other packages and computes several performance measures.

CRAN packages:

- ahaz
- arules
- BayesTree
- bigrf
- bigRR
- bmrm
- Boruta
- bst
- C50
- caret
- CORElearn
- CoxBoost
- Cubist
- e1071 (core)
- earth
- elasticnet
- ElemStatLearn
- evtree
- FCNN4R
- frbs
- GAMBoost
- gamboostLSS
- gbm (core)
- glmnet
- glmpath
- GMMBoost
- grplasso
- grpreg
- hda
- hdi
- ipred
- kernlab (core)
- klaR
- lars
- lasso2
- LiblineaR
- LogicForest
- LogicReg
- maptree
- mboost (core)
- mlr
- ncvreg
- nnet (core)
- oblique.tree
- pamr
- party
- partykit
- penalized
- penalizedLDA
- penalizedSVM
- quantregForest

- randomForest (core)
- randomForestSRC
- rattle
- rda
- rdetools
- REEMtree
- relaxo
- rgenoud
- rgp
- Rmalschains
- rminer
- ROCR
- RoughSets
- rpart (core)
- RPMM
- RSNNS
- RWeka
- RXshrink
- sda
- stabs
- svmpath
- tgp
- tree
- varSelRF
- vcrpart

Related links:

- MLOSS: Machine Learning Open Source Software
- Boosting Research Site