mlr Tutorial

Julia Schiffner Jakob Richter Florian Pfisterer Tobias Kühn Bernd Bischl Zachary M. Jones Mason Gallo Janek Thomas Michel Lang Philipp Probst Dominik Kirchhoff Lars Kotthoff

Contents

1	Machine Learning in R: mlr Tutorial			5
	1.1	Quick	start	6
2	Basics			6
	2.1	Learni	ing Tasks	6
		2.1.1	Task types and creation	7
		2.1.2	Further settings	11
		2.1.3	Accessing a learning task	12
		2.1.4	Modifying a learning task	14
		2.1.5	Example tasks and convenience functions	16
	2.2	Learne	ers	16
		2.2.1	Constructing a learner	16
		2.2.2	Accessing a learner	18
		2.2.3	Modifying a learner	20
		2.2.4	Listing learners	21
	2.3 Training a Learner		ng a Learner	22
		2.3.1	Accessing learner models	23
		2.3.2	Further options and comments	26
2.4 Predicting Outcomes for New Data			cting Outcomes for New Data	27
		2.4.1	Accessing the prediction	29

CONTENTS CONTENTS

	2.4.2	Adjusting the threshold	33	
	2.4.3	Visualizing the prediction	35	
2.5	Evaluating Learner Performance			
	2.5.1	Available performance measures	38	
	2.5.2	Listing measures	39	
	2.5.3	Calculate performance measures	40	
	2.5.4	Access a performance measure	41	
	2.5.5	Binary classification	42	
2.6	Resampling			
	2.6.1	Stratified resampling	47	
	2.6.2	Accessing individual learner models	48	
	2.6.3	Resample descriptions and resample instances	49	
	2.6.4	Aggregating performance values	52	
	2.6.5	Convenience functions	54	
2.7	Tuning Hyperparameters			
	2.7.1	Basics	55	
	2.7.2	Further comments	64	
2.8	Benchi	mark Experiments	65	
	2.8.1	Conducting benchmark experiments	65	
	2.8.2	Accessing benchmark results	66	
	2.8.3	Merging benchmark results	70	
	2.8.4	Benchmark analysis and visualization	72	
	2.8.5	Further comments	86	
2.9	Parallelization			
	2.9.1	Parallelization levels	87	
	2.9.2	Custom learners and parallelization	88	
	2.9.3	The end \hdots	88	
2.10	Visual	ization	88	
	2.10.1	Generation and plotting functions	88	
	2.10.2	Available generation and plotting functions	95	

CONTENTS CONTENTS

3	Advanced			96
	3.1	Config	guring mlr	96
		3.1.1	Example: Reducing the output on the console	97
		3.1.2	Accessing and resetting the configuration	98
		3.1.3	Example: Turning off parameter checking $\ldots \ldots$.	99
		3.1.4	Example: Handling errors in a learning method	100
	3.2	Wrap	per	101
		3.2.1	Example: Bagging wrapper	102
	3.3	Data	Preprocessing	105
		3.3.1	Fusing learners with preprocessing	106
		3.3.2	Preprocessing with make PreprocWrapperCaret $\ \ldots \ \ldots$	107
		3.3.3	Writing a custom preprocessing wrapper	113
	3.4	Imput	ation of Missing Values	119
		3.4.1	Imputation and reimputation	120
		3.4.2	Fusing a learner with imputation	124
	3.5	Gener	ic Bagging	126
		3.5.1	Changing the type of prediction	127
	3.6	Advar	nced Tuning	129
		3.6.1	Iterated F-Racing for mixed spaces and dependencies $$	129
		3.6.2	Tuning across whole model spaces with ModelMultiplexer	130
		3.6.3	Multi-criteria evaluation and optimization	132
	3.7	Featur	re Selection	133
		3.7.1	Filter methods	134
		3.7.2	Wrapper methods	141
	3.8	Neste	d Resampling	146
		3.8.1	Tuning	148
		3.8.2	Feature selection	151
		3.8.3	Benchmark experiments	156
	3.9	Cost-S	Sensitive Classification	163
		3.9.1	Class-dependent misclassification costs	164
		3.9.2	Example-dependent misclassification costs	177

CONTENTS CONTENTS

3.10	Imbala	nced Classification Problems
3.10		Sampling-based approaches
		(Simple) over- and undersampling
		Cost-based approaches
3.11		Analysis and Performance Curves
	3.11.1	Performance plots with plotROCCurves 187
		Performance plots with asROCRPrediction 194
	3.11.3	Viper charts
3.12	Multila	abel Classification
	3.12.1	Creating a task
	3.12.2	Constructing a learner
	3.12.3	Train
	3.12.4	Predict
	3.12.5	Performance
	3.12.6	Resampling
	3.12.7	Binary performance
3.13	Learni	ng Curve Analysis
	3.13.1	Plotting the learning curve
3.14	Explor	ing Learner Predictions
	3.14.1	Generating partial dependences
	3.14.2	Functional ANOVA
	3.14.3	Plotting partial dependences
3.15	Classif	ier Calibration
3.16	Evalua	ting Hyperparameter Tuning
	3.16.1	Generating hyperparameter tuning data 231
	3.16.2	Visualizing the effect of a single hyperparameter 234
	3.16.3	Visualizing the effect of 2 hyperparameters

4 Extend				245		
	4.1	Integr	ating Another Learner	245		
		4.1.1	Classes, constructors, and naming schemes	246		
		4.1.2	Classification	247		
		4.1.3	Regression	250		
		4.1.4	Survival analysis	251		
		4.1.5	Clustering	252		
		4.1.6	Multilabel classification	253		
		4.1.7	Creating a new feature importance method $\dots \dots$	254		
		4.1.8	Registering your learner	255		
	4.2	Integr	ating Another Measure	255		
		4.2.1	Performance measures and aggregation schemes	256		
		4.2.2	Constructing a performance measure	258		
		4.2.3	Constructing a measure for ordinary misclassification costs	258		
		4.2.4	Creating an aggregation scheme	259		
	4.3	.3 Creating an Imputation Method		262		
		4.3.1	Example: Imputation using the mean	262		
		4.3.2	Writing your own imputation method	262		
	4.4	Integrating Another Filter Method		264		
		4.4.1	Filter objects	264		
		442	Writing a new filter method	265		

1 Machine Learning in R: mlr Tutorial

This document provides an in-depth introduction to the Machine Learning in R: mlr framework for machine learning experiments in R.

We focus on the comprehension of the basic functions and applications. More detailed technical information can be found in the manual pages which are regularly updated and reflect the documentation of the current package version on CRAN.

An offline version of this tutorial is available for download

• here for the current mlr release on CRAN and

• here for the mlr devel version on GitHub.

The tutorial explains the basic analysis of a data set step by step. Please refer to sections of the menu above: Basics, Advanced, Extend and Appendix.

During the tutorial we present various simple examples from classification, regression, cluster and survival analysis to illustrate the main features of the package.

Enjoy reading!

1.1 Quick start

A simple stratified cross-validation of linear discriminant analysis with mlr.

```
library(mlr)
data(iris)
## Define the task
task = makeClassifTask(id = "tutorial", data = iris, target =
    "Species")
## Define the learner
lrn = makeLearner("classif.lda")
## Define the resampling strategy
rdesc = makeResampleDesc(method = "CV", stratify = TRUE)
## Do the resampling
r = resample(learner = lrn, task = task, resampling = rdesc,
    show.info = FALSE)
## Get the mean misclassification error
r$aggr
#> mmce.test.mean
             0.02
```

2 Basics

2.1 Learning Tasks

Learning tasks encapsulate the data set and further relevant information about a machine learning problem, for example the name of the target variable for supervised problems.

2.1.1 Task types and creation

The tasks are organized in a hierarchy, with the generic Task at the top. The following tasks can be instantiated and all inherit from the virtual superclass Task:

- RegrTask for regression problems,
- ClassifTask for binary and multi-class classification problems (costsensitive classification with class-dependent costs can be handled as well),
- SurvTask for survival analysis,
- Cluster Task for cluster analysis,
- MultilabelTask for multilabel classification problems,
- CostSensTask for general cost-sensitive classification (with example-specific costs).

To create a task, just call makeTaskType>, e.g., makeClassifTask. All tasks require an identifier (argument id) and a data.frame (argument data). If no ID is provided it is automatically generated using the variable name of the data. The ID will be later used to name results, for example of benchmark experiments, and to annotate plots. Depending on the nature of the learning problem, additional arguments may be required and are discussed in the following sections.

2.1.1.1 Regression

For supervised learning like regression (as well as classification and survival analysis) we, in addition to data, have to specify the name of the target variable.

```
data(BostonHousing, package = "mlbench")
regr.task = makeRegrTask(id = "bh", data = BostonHousing, target =
    "medv")
regr.task
#> Supervised task: bh
#> Type: regr
#> Target: medv
#> Observations: 506
#> Features:
#> numerics factors ordered
#>
        12
            1
#> Missings: FALSE
#> Has weights: FALSE
#> Has blocking: FALSE
```

As you can see, the Task records the type of the learning problem and basic information about the data set, e.g., the types of the features (numeric vectors,

factors or ordered factors), the number of observations, or whether missing values are present.

Creating tasks for classification and survival analysis follows the same scheme, the data type of the target variables included in **data** is simply different. For each of these learning problems some specifics are described below.

2.1.1.2 Classification

For classification the target column has to be a factor.

In the following example we define a classification task for the BreastCancer data set and exclude the variable Id from all further model fitting and evaluation.

```
data(BreastCancer, package = "mlbench")
df = BreastCancer
df$Id = NULL
classif.task = makeClassifTask(id = "BreastCancer", data = df,
    target = "Class")
classif.task
#> Supervised task: BreastCancer
#> Type: classif
#> Target: Class
#> Observations: 699
#> Features:
#> numerics factors ordered
#>
     0
#> Missings: TRUE
#> Has weights: FALSE
#> Has blocking: FALSE
#> Classes: 2
     benign malignant
       458
#> Positive class: beniqn
```

In binary classification the two classes are usually referred to as *positive* and *negative* class with the positive class being the category of greater interest. This is relevant for many performance measures like the *true positive rate* or ROC curves. Moreover, mlr, where possible, permits to set options (like the decision threshold or class weights) and returns and plots results (like class posterior probabilities) for the positive class only.

makeClassifTask by default selects the first factor level of the target variable as the positive class, in the above example benign. Class malignant can be manually selected as follows:

```
classif.task = makeClassifTask(id = "BreastCancer", data = df,
    target = "Class", positive = "malignant")
```

2.1.1.3 Survival analysis

Survival tasks use two target columns. For left and right censored problems these consist of the survival time and a binary event indicator. For interval censored data the two target columns must be specified in the "interval2" format (see Surv).

```
data(lung, package = "survival")
lung$status = (lung$status == 2) # convert to logical
surv.task = makeSurvTask(data = lung, target = c("time", "status"))
surv.task
#> Supervised task: lung
#> Type: surv
#> Target: time, status
#> Observations: 228
#> Features:
#> numerics factors ordered
#> 8 0 0
#> Missings: TRUE
#> Has weights: FALSE
#> Has blocking: FALSE
```

The type of censoring can be specified via the argument censoring, which defaults to "rcens" for right censored data.

2.1.1.4 Multilabel classification

In multilabel classification each object can belong to more than one category at the same time.

The data are expected to contain as many target columns as there are class labels. The target columns should be logical vectors that indicate which class labels are present. The names of the target columns are taken as class labels and need to be passed to the target argument of makeMultilabelTask.

In the following example we get the data of the yeast data set, extract the label names, and pass them to the target argument in makeMultilabelTask.

```
#> Observations: 2417
#> Features:
#> numerics factors ordered
       103
            0
#> Missings: FALSE
#> Has weights: FALSE
#> Has blocking: FALSE
#> Classes: 14
   label1 label2 label3 label4 label5 label6 label7 label8
   label9
                                                            480
      762
             1038
                     983
                             862
                                     722
                                            597
                                                    428
       178
#> label10 label11 label12 label13 label14
      253
            289
                  1816
                            1799
```

See also the tutorial page on multilabel classification.

2.1.1.5 Cluster analysis

As cluster analysis is unsupervised, the only mandatory argument to construct a cluster analysis task is the data. Below we create a learning task from the data set mtcars.

```
data(mtcars, package = "datasets")
cluster.task = makeClusterTask(data = mtcars)
cluster.task
#> Unsupervised task: mtcars
#> Type: cluster
#> Observations: 32
#> Features:
#> numerics factors ordered
#> 11 0 0
#> Missings: FALSE
#> Has weights: FALSE
#> Has blocking: FALSE
```

2.1.1.6 Cost-sensitive classification

The standard objective in classification is to obtain a high prediction accuracy, i.e., to minimize the number of errors. All types of misclassification errors are thereby deemed equally severe. However, in many applications different kinds of errors cause different costs.

In case of *class-dependent costs*, that solely depend on the actual and predicted class labels, it is sufficient to create an ordinary ClassifTask.

In order to handle example-specific costs it is necessary to generate a CostSen-sTask. In this scenario, each example (x, y) is associated with an individual cost vector of length K with K denoting the number of classes. The k-th component indicates the cost of assigning x to class k. Naturally, it is assumed that the cost of the intended class label y is minimal.

As the cost vector contains all relevant information about the intended class y, only the feature values x and a cost matrix, which contains the cost vectors for all examples in the data set, are required to create the CostSensTask.

In the following example we use the iris data and an artificial cost matrix (which is generated as proposed by Beygelzimer et al., 2005):

```
df = iris
cost = matrix(runif(150 * 3, 0, 2000), 150) * (1 -
   diag(3))[df$Species,]
df$Species = NULL
costsens.task = makeCostSensTask(data = df, cost = cost)
costsens.task
#> Supervised task: df
#> Type: costsens
#> Observations: 150
#> Features:
#> numerics factors ordered
   4 0
#> Missings: FALSE
#> Has blocking: FALSE
#> Classes: 3
#> y1, y2, y3
```

For more details see the page about cost-sensitive classification.

2.1.2 Further settings

The Task help page also lists several other arguments to describe further details of the learning problem.

For example, we could include a blocking factor in the task. This would indicate that some observations "belong together" and should not be separated when splitting the data into training and test sets for resampling.

Another option is to assign **weights** to observations. These can simply indicate observation frequencies or result from the sampling scheme used to collect the data.

Note that you should use this option only if the weights really belong to the task. If you plan to train some learning algorithms with different weights on the same Task, mlr offers several other ways to set observation or class weights

(for supervised classification). See for example the tutorial page about training or function makeWeightedClassesWrapper.

2.1.3 Accessing a learning task

We provide many operators to access the elements stored in a Task. The most important ones are listed in the documentation of Task and getTaskData.

To access the task description that contains basic information about the task you can use:

```
getTaskDescription(classif.task)
#> $id
#> [1] "BreastCancer"
#>
#> $type
#> [1] "classif"
#>
#> $target
#> [1] "Class"
#> $size
#> [1] 699
#>
#> $n.feat
#> numerics factors ordered
    0
            4
#>
#> $has.missings
#> [1] TRUE
#>
#> $has.weights
#> [1] FALSE
#>
#> $has.blocking
#> [1] FALSE
#>
#> $class.levels
#> [1] "beniqn"
                  "malignant"
#>
#> $positive
#> [1] "malignant"
#>
#> $negative
#> [1] "benign"
#>
```

```
#> attr(,"class")
#> [1] "TaskDescClassif" "TaskDescSupervised" "TaskDesc"
```

Note that task descriptions have slightly different elements for different types of Tasks. Frequently required elements can also be accessed directly.

```
### Get the ID
getTaskId(classif.task)
#> [1] "BreastCancer"
### Get the type of task
getTaskType(classif.task)
#> [1] "classif"
### Get the names of the target columns
getTaskTargetNames(classif.task)
#> [1] "Class"
### Get the number of observations
getTaskSize(classif.task)
#> [1] 699
### Get the number of input variables
getTaskNFeats(classif.task)
#> [1] 9
### Get the class levels in classif.task
getTaskClassLevels(classif.task)
#> [1] "benign" "malignant"
```

Moreover, mlr provides several functions to extract data from a Task.

```
### Accessing the data set in classif.task
str(getTaskData(classif.task))
#> 'data.frame':
                  699 obs. of 10 variables:
#> $ Cl.thickness : Ord.factor w/ 10 levels "1"<"2"<"3"<"4"<..: 5
   5 3 6 4 8 1 2 2 4 ...
#> $ Cell.size
                 : Ord.factor w/ 10 levels "1"<"2"<"3"<"4"<..: 1
   4 1 8 1 10 1 1 1 2 ...
#> $ Cell.shape
                  : Ord.factor w/ 10 levels "1"<"2"<"3"<"4"<..: 1
   4 1 8 1 10 1 2 1 1 ...
#> $ Marg.adhesion : Ord.factor w/ 10 levels "1"<"2"<"3"<"4"<..: 1
   5 1 1 3 8 1 1 1 1 ...
#> $ Epith.c.size : Ord.factor w/ 10 levels "1"<"2"<"3"<"4"<..: 2
   723272222...
#> $ Bare.nuclei
                  : Factor w/ 10 levels "1", "2", "3", "4", ...: 1 10
2 4 1 10 10 1 1 1 ...
```

```
#> $ Bl.cromatin : Factor w/ 10 levels "1", "2", "3", "4", ...: 3 3 3
   3393312...
#> $ Normal.nucleoli: Factor w/ 10 levels "1", "2", "3", "4", ...: 1 2 1
    7171111...
                    : Factor w/ 9 levels "1", "2", "3", "4", . . : 1 1 1
#> $ Mitoses
   1 1 1 1 1 5 1 ...
#> $ Class
                     : Factor w/ 2 levels "benign", "malignant": 1 1
    1 1 1 2 1 1 1 1 ...
### Get the names of the input variables in cluster.task
getTaskFeatureNames(cluster.task)
   [1] "mpq" "cyl" "disp" "hp"
                                   "drat" "wt"
                                                 "qsec" "vs"
                                                                "am"
      "gear"
#> [11] "carb"
### Get the values of the target variables in surv.task
head(getTaskTargets(surv.task))
#> time status
#> 1 306
            TRUE
#> 2 455
            TRUE
#> 3 1010 FALSE
#> 4 210
            TRUE
#> 5 883
            TRUE
#> 6 1022 FALSE
### Get the cost matrix in costsens.task
head(getTaskCosts(costsens.task))
       y1
                 y2
#> [1,] 0 1589.5664 674.44434
#> [2,] 0 1173.4364 828.40682
#> [3,] 0 942.7611 1095.33713
#> [4,] 0 1049.5562 477.82496
#> [5,] 0 1121.8899
                      90.85237
#> [6,] 0 1819.9830 841.06686
```

Note that getTaskData offers many options for converting the data set into a convenient format. This especially comes in handy when you integrate a new learner from another ${\bf R}$ package into mlr. In this regard function getTaskFormula is also useful.

2.1.4 Modifying a learning task

mlr provides several functions to alter an existing Task, which is often more convenient than creating a new Task from scratch. Here are some examples.

```
### Select observations and/or features
```

```
cluster.task = subsetTask(cluster.task, subset = 4:17)
### It may happen, especially after selecting observations, that
   features are constant.
### These should be removed.
removeConstantFeatures(cluster.task)
#> Removing 1 columns: am
#> Unsupervised task: mtcars
#> Type: cluster
#> Observations: 14
#> Features:
#> numerics factors ordered
#> 10 0
#> Missings: FALSE
#> Has weights: FALSE
#> Has blocking: FALSE
### Remove selected features
dropFeatures(surv.task, c("meal.cal", "wt.loss"))
#> Supervised task: lung
#> Type: surv
#> Target: time, status
#> Observations: 228
#> Features:
#> numerics factors ordered
   6 0
#> Missings: TRUE
#> Has weights: FALSE
#> Has blocking: FALSE
### Standardize numerical features
task = normalizeFeatures(cluster.task, method = "range")
summary(getTaskData(task))
#>
                      cyl
       mpg
                                      disp
                                                      hp
#> Min. :0.0000 Min. :0.0000 Min. :0.0000 Min. :0.0000
#> 1st Qu.:0.3161 1st Qu.:0.5000 1st Qu.:0.1242
                                                1st Qu.:0.2801
#> Median :0.5107 Median :1.0000 Median :0.4076
                                                Median :0.6311
#> Mean :0.4872 Mean :0.7143 Mean :0.4430
                                                Mean :0.5308
#> 3rd Qu.:0.6196 3rd Qu.:1.0000 3rd Qu.:0.6618
                                                 3rd Qu.:0.7473
#> Max. :1.0000 Max. :1.0000 Max. :1.0000
                                                Max. :1.0000
#>
       drat
                       wt
                                     gsec
                                                      vs
                                Min. :0.0000
#> Min. :0.0000 Min. :0.0000
                                                Min. :0.0000
#> 1st Qu.:0.2672 1st Qu.:0.1275 1st Qu.:0.2302
                                                1st Qu.:0.0000
Median :0.0000
#> Mean :0.4544 Mean :0.3268
                                 Mean :0.3752
                                                Mean :0.4286
#> 3rd Qu.:0.7026 3rd Qu.:0.3727 3rd Qu.:0.4908 3rd Qu.:1.0000
```

```
#>
           :1.0000
                     Max.
                             :1.0000
                                       Max.
                                               :1.0000
                                                         Max.
                                                                :1.0000
#>
                       gear
                                         carb
          a.m
#>
   Min.
           :0.5
                  Min.
                         :0.0000
                                    Min.
                                           :0.0000
#>
   1st Qu.:0.5
                  1st Qu.:0.0000
                                    1st Qu.:0.3333
   Median :0.5
                                    Median :0.6667
                  Median :0.0000
           :0.5
   Mean
                          :0.2857
                                           :0.6429
                  Mean
                                    Mean
   3rd Qu.:0.5
                  3rd Qu.:0.7500
                                    3rd Qu.:1.0000
                                          :1.0000
   Max.
         :0.5
                  Max. :1.0000
                                    Max.
```

For more functions and more detailed explanations have a look at the data preprocessing page.

2.1.5 Example tasks and convenience functions

For your convenience mlr provides pre-defined Tasks for each type of learning problem. These are also used throughout this tutorial in order to get shorter and more readable code. A list of all Tasks can be found in the Appendix.

Moreover, mlr's function convertMLBenchObjToTask can generate Tasks from the data sets and data generating functions in package mlbench.

2.2 Learners

The following classes provide a unified interface to all popular machine learning methods in **R**: (cost-sensitive) classification, regression, survival analysis, and clustering. Many are already integrated in mlr, others are not, but the package is specifically designed to make extensions simple.

Section integrated learners shows the already implemented machine learning methods and their properties. If your favorite method is missing, either open an issue or take a look at how to integrate a learning method yourself. This basic introduction demonstrates how to use already implemented learners.

2.2.1 Constructing a learner

A learner in mlr is generated by calling makeLearner. In the constructor you need to specify which learning method you want to use. Moreover, you can:

- Set hyperparameters.
- Control the output for later prediction, e.g., for classification whether you want a factor of predicted class labels or probabilities.
- Set an ID to name the object (some methods will later use this ID to name results or annotate plots).

The first argument specifies which algorithm to use. The naming convention is classif.<R_method_name> for classification methods, regr.<R_method_name> for regression methods, surv.<R_method_name> for survival analysis, cluster.<R_method_name> for clustering methods, and multilabel.<R_method_name> for multilabel classification.

Hyperparameter values can be specified either via the ... argument or as a list via par.vals.

Occasionally, factor features may cause problems when fewer levels are present in the test data set than in the training data. By setting fix.factors.prediction = TRUE these are avoided by adding a factor level for missing data in the test data set.

Let's have a look at two of the learners created above.

```
#> Class: surv.coxph
#> Properties: numerics, factors, weights, rcens
#> Predict-Type: response
#> Hyperparameters:
```

All generated learners are objects of class Learner. This class contains the properties of the method, e.g., which types of features it can handle, what kind of output is possible during prediction, and whether multi-class problems, observations weights or missing values are supported.

As you might have noticed, there is currently no special learner class for cost-sensitive classification. For ordinary misclassification costs you can use standard classification methods. For example-dependent costs there are several ways to generate cost-sensitive learners from ordinary regression and classification learners. This is explained in greater detail in the section about cost-sensitive classification.

2.2.2 Accessing a learner

The Learner object is a list and the following elements contain information regarding the hyperparameters and the type of prediction.

```
### Get the configured hyperparameter settings that deviate from the
   defaults
cluster.lrn$par.vals
#> $centers
#> [1] 5
### Get the set of hyperparameters
classif.lrn$par.set
#>
                     Type len Def Constr Req Tunable Trafo
#> ntree
                   integer
                               500 1 to Inf -
                                                   TRUE
#> mtry
                  integer
                                - 1 to Inf
                                                   TRUE
                  logical
                           - TRUE
                                                   TRUE
#> replace
#> classwt numericvector <NA>
                               - O to Inf -
                                                   TRUE
#> cutoff
                                  - 0 to 1
            numericvector <NA>
                                                   TRUE
#> strata
                                                   TRUE
            untyped
#> sampsize integervector <NA>
                                - 1 to Inf
                                                   TRUE
#> nodesize integer
                                  1 1 to Inf
                                                   TRUE
                                  - 1 to Inf
#> maxnodes
                  integer
                                                   TRUE
#> importance
                 logical
                            - FALSE
                                                   TRUE
#> localImp
                  logical
                            - FALSE
                                                   TRUE
                             - FALSE
                                                  FALSE
#> proximity
                  logical
#> oob.prox
                   logical
                                                  FALSE
                             - TRUE
                                                  FALSE
#> norm.votes
                   logical
                             - FALSE
#> do.trace
                  logical
                                                  FALSE
```

```
#> keep.forest logical - TRUE - - FALSE -
#> keep.inbag logical - FALSE - - FALSE -
### Get the type of prediction
regr.lrn$predict.type
#> [1] "response"
```

Slot **\$par.set** is an object of class ParamSet. It contains, among others, the type of hyperparameters (e.g., numeric, logical), potential default values and the range of allowed values.

Moreover, mlr provides function getHyperPars to access the current hyperparameter setting of a Learner and getParamSet to get a description of all possible settings. These are particularly useful in case of wrapped Learners, for example if a learner is fused with a feature selection strategy, and both, the learner as well the feature selection method, have hyperparameters. For details see the section on wrapped learners.

```
### Get current hyperparameter settings
getHyperPars(cluster.lrn)
#> $centers
#> [1] 5
### Get a description of all possible hyperparameter settings
getParamSet(classif.lrn)
#>
                       Type len Def Constr Req Tunable Trafo
                                  500 1 to Inf
#> ntree
                    integer
                                                      TRUE
                                  - 1 to Inf
#> mtry
                    integer
                                                      TRUE
                             - TRUE
#> replace
                    logical
                                                      TRUE
#> classwt numericvector <NA>
                                   - O to Inf
                                                      TRUE
#> cutoff
              numericvector <NA>
                                        0 to 1
                                                      TRUE
#> strata
                                                      TRUE
              untyped
#> sampsize
              integervector <NA>
                                     - 1 to Inf
                                                      TRUE
#> nodesize
                   integer
                                    1 1 to Inf
                                                      TRUE
#> maxnodes
                    integer
                                    - 1 to Inf
                                                      TRUE
#> importance
                    logical
                               - FALSE
                                                      TRUE
                               - FALSE
#> localImp
                   logical
                                                      TRUE
                               - FALSE
#> proximity
                    logical
                                                     FALSE
#> oob.prox
                    logical
                                                     FALSE
#> norm.votes
                    logical
                               - TRUE
                                                     FALSE
#> do.trace
                    logical
                               - FALSE
                                                     FALSE
                                                     FALSE
#> keep.forest
                    logical
                               - TRUE
#> keep.inbag
                    logical
                               - FALSE
                                                     FALSE
```

We can also use getParamSet to get a quick overview about the available hyperparameters and defaults of a learning method without explicitly constructing it (by calling makeLearner).

```
getParamSet("classif.randomForest")
                       Type len Def Constr Req Tunable Trafo
                  integer - 500 1 to Inf -
integer - - 1 to Inf -
#> ntree
                                                      TRUE
#> sampsize integervector <NA> - 1 to Inf - TRUE
#> nodesize integer - 1 1 to Inf - TRUE #> maxnodes integer - - 1 to Inf - TRUE #> importance logical - FALSE - - TRUE #> localImp logical - FALSE - - TRUE
                logical - FALSE
logical - -
logical - TRUE
                                           - - FALSE
#> proximity
                                             - Y FALSE
#> oob.prox
                                            - - FALSE
#> norm.votes
#> do.trace
                   logical - FALSE
                                            - - FALSE
                                           - - FALSE
#> keep.forest
                   logical - TRUE
                                         - - FALSE
#> keep.inbag logical - FALSE
```

2.2.3 Modifying a learner

There are also some functions that enable you to change certain aspects of a Learner without needing to create a new Learner from scratch. Here are some examples.

```
### Change the ID
surv.lrn = setLearnerId(surv.lrn, "CoxModel")
surv.lrn
#> Learner CoxModel from package survival
#> Type: surv
#> Name: Cox Proportional Hazard Model; Short name: coxph
#> Class: surv.coxph
#> Properties: numerics, factors, weights, rcens
#> Predict-Type: response
#> Hyperparameters:
### Change the prediction type, predict a factor with class labels
    instead of probabilities
classif.lrn = setPredictType(classif.lrn, "response")
### Change hyperparameter values
cluster.lrn = setHyperPars(cluster.lrn, centers = 4)
### Go back to default hyperparameter values
```

2.2.4 Listing learners

A list of all learners integrated in mlr and their respective properties is shown in the Appendix.

If you would like a list of available learners, maybe only with certain properties or suitable for a certain learning Task use function listLearners.

```
### List everything in mlr
lrns = listLearners()
head(lrns[c("class", "package")])
#>
                 class package
#> 1
           classif.ada
                                ada
#> 2
        classif.avNNet
                               nnet
#> 3 classif.bartMachine bartMachine
           classif.bdk
                        kohonen
#> 5
       classif.binomial
                              stats
#> 6 classif.blackboost mboost,party
### List classifiers that can output probabilities
lrns = listLearners("classif", properties = "prob")
head(lrns[c("class", "package")])
#>
                 class
#> 1
           classif.ada
                               ada
#> 2
         classif.avNNet
#> 3 classif.bartMachine bartMachine
#> 4
          classif.bdk kohonen
#> 5
       classif.binomial
                              stats
#> 6 classif.blackboost mboost,party
### List classifiers that can be applied to iris (i.e., multiclass)
    and output probabilities
lrns = listLearners(iris.task, properties = "prob")
head(lrns[c("class", "package")])
               class
                        package
#> 1
     classif.avNNet
                            nnet
#> 2
        classif.bdk
                         kohonen
#> 3 classif.boosting adabag,rpart
#> 4
       classif.C50
                             C50
#> 5 classif.cforest
                           party
#> 6 classif.ctree
                          party
```

```
### The calls above return character vectors, but you can also
    create learner objects
head(listLearners("cluster", create = TRUE), 2)
#> [[1]]
#> Learner cluster.cmeans from package e1071, clue
#> Type: cluster
#> Name: Fuzzy C-Means Clustering; Short name: cmeans
#> Class: cluster.cmeans
#> Properties: numerics, prob
#> Predict-Type: response
#> Hyperparameters: centers=2
#>
#> [[2]]
#> Learner cluster.Cobweb from package RWeka
#> Type: cluster
#> Name: Cobweb Clustering Algorithm; Short name: cobweb
#> Class: cluster.Cobweb
#> Properties: numerics
#> Predict-Type: response
#> Hyperparameters:
```

2.3 Training a Learner

Training a learner means fitting a model to a given data set. In mlr this can be done by calling function train on a Learner and a suitable Task.

We start with a classification example and perform a linear discriminant analysis on the iris data set.

```
### Generate the task
task = makeClassifTask(data = iris, target = "Species")

### Generate the learner
lrn = makeLearner("classif.lda")

### Train the learner
mod = train(lrn, task)
mod
#> Model for learner.id=classif.lda; learner.class=classif.lda
#> Trained on: task.id = iris; obs = 150; features = 4
#> Hyperparameters:
```

In the above example creating the Learner explicitly is not absolutely necessary. As a general rule, you have to generate the Learner yourself if you want to

change any defaults, e.g., setting hyperparameter values or altering the predict type. Otherwise, train and many other functions also accept the class name of the learner and call makeLearner internally with default settings.

```
mod = train("classif.lda", task)
mod
#> Model for learner.id=classif.lda; learner.class=classif.lda
#> Trained on: task.id = iris; obs = 150; features = 4
#> Hyperparameters:
```

Training a learner works the same way for every type of learning problem. Below is a survival analysis example where a Cox proportional hazards model is fitted to the lung data set. Note that we use the corresponding lung.task provided by mlr. All available Tasks are listed in the Appendix.

```
mod = train("surv.coxph", lung.task)
mod
#> Model for learner.id=surv.coxph; learner.class=surv.coxph
#> Trained on: task.id = lung-example; obs = 167; features = 8
#> Hyperparameters:
```

2.3.1 Accessing learner models

Function train returns an object of class WrappedModel, which encapsulates the fitted model, i.e., the output of the underlying ${\bf R}$ learning method. Additionally, it contains some information about the Learner, the Task, the features and observations used for training, and the training time. A WrappedModel can subsequently be used to make a prediction for new observations.

The fitted model in slot \$learner.model of the WrappedModel object can be accessed using function getLearnerModel.

In the following example we cluster the Ruspini data set (which has four groups and two features) by K-means with K=4 and extract the output of the underlying kmeans function.

```
data(ruspini, package = "cluster")
plot(y ~ x, ruspini)
```



```
### Generate the task
ruspini.task = makeClusterTask(data = ruspini)

### Generate the learner
lrn = makeLearner("cluster.kmeans", centers = 4)

### Train the learner
mod = train(lrn, ruspini.task)
mod
#> Model for learner.id=cluster.kmeans; learner.class=cluster.kmeans
#> Trained on: task.id = ruspini; obs = 75; features = 2
#> Hyperparameters: centers=4

### Peak into mod
names(mod)
#> [1] "learner" "learner.model" "task.desc" "subset"
```

```
#> [5] "features" "factor.levels" "time"
mod$learner
#> Learner cluster.kmeans from package stats, clue
#> Type: cluster
#> Name: K-Means; Short name: kmeans
#> Class: cluster.kmeans
#> Properties: numerics,prob
#> Predict-Type: response
#> Hyperparameters: centers=4
mod$features
#> [1] "x" "y"
mod$time
#> [1] 0.001
### Extract the fitted model
getLearnerModel(mod)
#> K-means clustering with 4 clusters of sizes 23, 17, 15, 20
#>
#> Cluster means:
#>
         \boldsymbol{x}
#> 1 43.91304 146.0435
#> 2 98.17647 114.8824
#> 3 68.93333 19.4000
#> 4 20.15000 64.9500
#>
#> Clustering vector:
#> 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22
   23 24 25
1 1 1
#> 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47
   48 49 50
2 2 2
#> 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72
  73 74 75
3 3 3
#> Within cluster sum of squares by cluster:
#> [1] 3176.783 4558.235 1456.533 3689.500
#> (between_SS / total_SS = 94.7 %)
```

```
#> Available components:
#>
#>
[1] "cluster"     "centers"     "totss"     "withinss"
#> [5] "tot.withinss" "betweenss"     "size"     "iter"
#> [9] "ifault"
```

2.3.2 Further options and comments

By default, the whole data set in the Task is used for training. The **subset** argument of train takes a logical or integer vector that indicates which observations to use, for example if you want to split your data into a training and a test set or if you want to fit separate models to different subgroups in the data.

Below we fit a linear regression model to the BostonHousing data set (bh.task) and randomly select 1/3 of the data set for training.

Note, for later, that all standard resampling strategies are supported. Therefore you usually do not have to subset the data yourself.

Moreover, if the learner supports this, you can specify observation weights that reflect the relevance of observations in the training process. Weights can be useful in many regards, for example to express the reliability of the training observations, reduce the influence of outliers or, if the data were collected over a longer time period, increase the influence of recent data. In supervised classification weights can be used to incorporate misclassification costs or account for class imbalance.

For example in the BreastCancer data set class benign is almost twice as frequent as class malignant. In order to grant both classes equal importance in training the classifier we can weight the examples according to the inverse class frequencies in the data set as shown in the following **R** code.

```
### Calculate the observation weights
```

```
target = getTaskTargets(bc.task)
tab = as.numeric(table(target))
w = 1/tab[target]

train("classif.rpart", task = bc.task, weights = w)
#> Model for learner.id=classif.rpart; learner.class=classif.rpart
#> Trained on: task.id = BreastCancer-example; obs = 683; features =
9
#> Hyperparameters: xval=0
```

Note, for later, that mlr offers much more functionality to deal with imbalanced classification problems.

As another side remark for more advanced readers: By varying the weights in the calls to train, you could also implement your own variant of a general boosting type algorithm on arbitrary mlr base learners.

As you may recall, it is also possible to set observation weights when creating the Task. As a general rule, you should specify them in make*Task if the weights really "belong" to the task and always should be used. Otherwise, pass them to train. The weights in train take precedence over the weights in Task.

2.4 Predicting Outcomes for New Data

Predicting the target values for new observations is implemented the same way as most of the other predict methods in **R**. In general, all you need to do is call predict on the object returned by train and pass the data you want predictions for.

There are two ways to pass the data:

- Either pass the Task via the task argument or
- pass a data frame via the newdata argument.

The first way is preferable if you want predictions for data already included in a Task.

Just as train, the predict function has a **subset** argument, so you can set aside different portions of the data in Task for training and prediction (more advanced methods for splitting the data in train and test set are described in the section on resampling).

In the following example we fit a gradient boosting machine to every second observation of the BostonHousing data set and make predictions on the remaining data in bh.task.

```
n = getTaskSize(bh.task)
train.set = seq(1, n, by = 2)
test.set = seq(2, n, by = 2)
lrn = makeLearner("regr.gbm", n.trees = 100)
mod = train(lrn, bh.task, subset = train.set)
task.pred = predict(mod, task = bh.task, subset = test.set)
task.pred
#> Prediction: 253 observations
#> predict.type: response
#> threshold:
#> time: 0.00
#> id truth response
#> 2 2 21.6 22.28539
#> 6 6 28.7 22.40896
#> 8 8 27.1 22.12750
#> 10 10 18.9 22.12750
#> 12 12 18.9 22.12750
#> ... (253 rows, 3 cols)
```

The second way is useful if you want to predict data not included in the Task.

Here we cluster the iris data set without the target variable. All observations with an odd index are included in the Task and used for training. Predictions are made for the remaining observations.

```
n = nrow(iris)
iris.train = iris[seq(1, n, by = 2), -5]
iris.test = iris[seq(2, n, by = 2), -5]
task = makeClusterTask(data = iris.train)
mod = train("cluster.kmeans", task)
newdata.pred = predict(mod, newdata = iris.test)
newdata.pred
#> Prediction: 75 observations
#> predict.type: response
#> threshold:
#> time: 0.00
#> response
#> 2
         2
            2
#> 4
#> 6
            2
#> 8
           2
#> 10
           2
#> 12
            2
#> ... (75 rows, 1 cols)
```

Note that for supervised learning you do not have to remove the target columns from the data. These columns are automatically removed prior to calling the underlying predict method of the learner.

2.4.1 Accessing the prediction

Function predict returns a named list of class Prediction. Its most important element is \$data which is a data frame that contains columns with the true values of the target variable (in case of supervised learning problems) and the predictions. Use as.data.frame for direct access.

In the following the predictions on the BostonHousing and the iris data sets are shown. As you may recall, the predictions in the first case were made from a Task and in the second case from a data frame.

```
### Result of predict with data passed via task argument
head(as.data.frame(task.pred))
#>
      id truth response
      2 21.6 22.28539
#> 2
      4 33.4 23.33968
      6 28.7 22.40896
#> 8 8 27.1 22.12750
#> 10 10 18.9 22.12750
#> 12 12 18.9 22.12750
### Result of predict with data passed via newdata argument
head(as.data.frame(newdata.pred))
      response
#> 2
             2
#> 4
             2
#> 6
             2
#> 8
#> 10
             2
#> 12
```

As you can see when predicting from a Task, the resulting data frame contains an additional column, called id, which tells us which element in the original data set the prediction corresponds to.

A direct way to access the true and predicted values of the target variable(s) is provided by functions getPredictionTruth and getPredictionResponse.

```
head(getPredictionTruth(task.pred))
#> [1] 21.6 33.4 28.7 27.1 18.9 18.9

head(getPredictionResponse(task.pred))
#> [1] 22.28539 23.33968 22.40896 22.12750 22.12750
```

2.4.1.1 Extract Probabilities

The predicted probabilities can be extracted from the Prediction using the function getPredictionProbabilities. (Function getProbabilities has been deprecated in favor of getPredictionProbabilities in mlr version 2.5.) Here is another cluster analysis example. We use fuzzy c-means clustering on the mtcars data set.

For classification problems there are some more things worth mentioning. By default, class labels are predicted.

```
### Linear discriminant analysis on the iris data set
mod = train("classif.lda", task = iris.task)
pred = predict(mod, task = iris.task)
pred
#> Prediction: 150 observations
#> predict.type: response
#> threshold:
#> time: 0.00
#> id truth response
#> 1 1 setosa setosa
#> 2 2 setosa
               setosa
#> 3 3 setosa setosa
#> 4 4 setosa setosa
#> 5 5 setosa setosa
#> 6 6 setosa
#> ... (150 rows, 3 cols)
```

In order to get predicted posterior probabilities we have to create a Learner with the appropriate predict.type.

```
lrn = makeLearner("classif.rpart", predict.type = "prob")
mod = train(lrn, iris.task)

pred = predict(mod, newdata = iris)
```

```
head(as.data.frame(pred))
      truth\ prob.setosa\ prob.versicolor\ prob.virginica\ response
#> 1 setosa
                     1
                                       0
                                                           setosa
#> 2 setosa
                      1
                                       0
                                                       0
                                                           setosa
#> 3 setosa
                                       0
                                                       0
                      1
                                                           setosa
#> 4 setosa
                      1
                                       0
                                                       0
                                                           setosa
#> 5 setosa
                       1
                                       0
                                                       0
                                                           setosa
#> 6 setosa
                                       0
                                                           setosa
```

In addition to the probabilities, class labels are predicted by choosing the class with the maximum probability and breaking ties at random.

As mentioned above, the predicted posterior probabilities can be accessed via the getPredictionProbabilities function.

```
head(getPredictionProbabilities(pred))
     setosa versicolor virginica
#> 1
         1
                     0
#> 2
                     0
                                0
          1
#> 3
          1
                     0
                                0
#> 4
                     0
                                0
          1
#> 5
          1
                     0
                                0
#> 6
                      0
                                0
```

2.4.1.2 Confusion matrix

A confusion matrix can be obtained by calling calculateConfusionMatrix. The columns represent predicted and the rows true class labels.

```
calculateConfusionMatrix(pred)
              predicted
#>
#> true
               setosa versicolor virginica -err.-
                  50
                                        0
                                                0
#>
    setosa
                              0
#>
    versicolor
                    0
                              49
                                         1
                                                1
#>
    virginica
                    0
                               5
                                        45
                                                5
                                                6
```

You can see the number of correctly classified observations on the diagonal of the matrix. Misclassified observations are on the off-diagonal. The total number of errors for single (true and predicted) classes is shown in the <code>-err.-</code> row and column, respectively.

To get relative frequencies additional to the absolute numbers we can set relative = TRUE.

```
conf.matrix = calculateConfusionMatrix(pred, relative = TRUE)
conf.matrix
#> Relative confusion matrix (normalized by row/column):
```

```
#>
             predicted
#> true
                       versicolor virginica -err.-
              setosa
#>
    setosa
              1.00/1.00 0.00/0.00 0.00/0.00 0.00
#>
    versicolor 0.00/0.00 0.98/0.91 0.02/0.02 0.02
    virginica 0.00/0.00 0.10/0.09 0.90/0.98 0.10
#>
#>
    -err.-
                   0.00
                           0.09
                                     0.02 0.08
#>
#>
#> Absolute confusion matrix:
#>
           predicted
            setosa versicolor virginica -err.-
#> true
              50 0 0
#>
    setosa
    versicolor
                 0
                            49
                                     1
                                             1
#>
                   0
                             5
                                      45
                                             5
    virginica
    -err.-
                   0
                                       1
                                             6
```

It is possible to normalize by either row or column, therefore every element of the above relative confusion matrix contains two values. The first is the relative frequency grouped by row (the true label) and the second value grouped by column (the predicted label).

If you want to access the relative values directly you can do this through the **\$relative.row** and **\$relative.col** members of the returned object **conf.matrix**. For more details see the ConfusionMatrix documentation page.

Finally, we can also add the absolute number of observations for each predicted and true class label to the matrix (both absolute and relative) by setting sums = TRUE.

```
calculateConfusionMatrix(pred, relative = TRUE, sums = TRUE)
#> Relative confusion matrix (normalized by row/column):
#>
             predicted
#> true
               setosa
                        versicolor virginica -err.-
#>
              1.00/1.00 0.00/0.00 0.00/0.00 0.00
                                                      50
    setosa
    versicolor 0.00/0.00 0.98/0.91 0.02/0.02 0.02
#>
                                                      54
#>
    virginica 0.00/0.00 0.10/0.09 0.90/0.98 0.10
                                                      46
#>
    -err.-
                   0.00 0.09 0.02 0.08
                                                      <NA>
                                                      150
#>
               50
                    50
                              50
                                            <NA>
    -n
#>
#>
#> Absolute confusion matrix:
```

```
#>
             setosa versicolor virginica -err.- -n-
#> setosa
                 50
                            0
                                       0
                                              0 50
                                              1 50
#> versicolor
                  0
                            49
                                       1
                                      45
#> virginica
                  0
                            5
                                              5 50
                             5
                  0
#> -err.-
                                       1
                                              6 NA
                                      46
                 50
                            54
#> -n.-
                                             NA 150
```

2.4.2 Adjusting the threshold

We can set the threshold value that is used to map the predicted posterior probabilities to class labels. Note that for this purpose we need to create a Learner that predicts probabilities. For binary classification, the threshold determines when the *positive* class is predicted. The default is 0.5. Now, we set the threshold for the positive class to 0.9 (that is, an example is assigned to the positive class if its posterior probability exceeds 0.9). Which of the two classes is the positive one can be seen by accessing the Task. To illustrate binary classification, we use the Sonar data set from the mlbench package.

```
lrn = makeLearner("classif.rpart", predict.type = "prob")
mod = train(lrn, task = sonar.task)
### Label of the positive class
getTaskDescription(sonar.task)$positive
#> [1] "M"
### Default threshold
pred1 = predict(mod, sonar.task)
pred1$threshold
#> M
       R.
#> 0.5 0.5
### Set the threshold value for the positive class
pred2 = setThreshold(pred1, 0.9)
pred2$threshold
#> M
#> 0.9 0.1
pred2
#> Prediction: 208 observations
#> predict.type: prob
#> threshold: M=0.90, R=0.10
#> time: 0.01
                prob.M
                        prob.R response
    id truth
#> 1 1 R 0.1060606 0.8939394
                                        R
#> 2 2 R 0.7333333 0.2666667
```

```
R
#> 4 4
          R 0.1060606 0.8939394
                                     R
#> 5 5
          R 0.9250000 0.0750000
                                     Μ
#> 6 6
          R 0.0000000 1.0000000
                                     R
#> ... (208 rows, 5 cols)
### We can also set the effect in the confusion matrix
calculateConfusionMatrix(pred1)
         predicted
          M R -err.-
#> true
#>
   Μ
          95 16
          10 87
#>
   R
                   10
   -err.- 10 16
                   26
calculateConfusionMatrix(pred2)
#>
         predicted
#> true
          M R -err.-
   Μ
          84 27
                   27
#>
    R
           6 91
                    6
#>
                   33
    -err.- 6 27
```

Note that in the binary case getPredictionProbabilities by default extracts the posterior probabilities of the positive class only.

It works similarly for multiclass classification. The threshold has to be given by a named vector specifying the values by which each probability will be divided. The class with the maximum resulting value is then selected.

```
lrn = makeLearner("classif.rpart", predict.type = "prob")
mod = train(lrn, iris.task)
pred = predict(mod, newdata = iris)
pred$threshold
#> setosa versicolor virginica
#> 0.3333333 0.3333333 0.33333333
```

```
table(as.data.frame(pred)$response)
#>
#>
       setosa versicolor virginica
                      54
#>
                                 46
pred = setThreshold(pred, c(setosa = 0.01, versicolor = 50,
    virginica = 1))
pred$threshold
#>
       setosa versicolor virginica
#>
         0.01
                   50.00
table(as.data.frame(pred)$response)
#>
#>
       setosa versicolor virginica
```

If you are interested in tuning the threshold (vector) have a look at the section about performance curves and threshold tuning.

2.4.3 Visualizing the prediction

The function plotLearnerPrediction allows to visualize predictions, e.g., for teaching purposes or exploring models. It trains the chosen learning method for 1 or 2 selected features and then displays the predictions with ggplot.

For classification, we get a scatter plot of 2 features (by default the first 2 in the data set). The type of symbol shows the true class labels of the data points. Symbols with white border indicate misclassified observations. The posterior probabilities (if the learner under consideration supports this) are represented by the background color where higher saturation means larger probabilities.

The plot title displays the ID of the Learner (in the following example CART), its parameters, its training performance and its cross-validation performance. mmce stands for *mean misclassification error*, i.e., the error rate. See the sections on performance and resampling for further explanations.

```
lrn = makeLearner("classif.rpart", id = "CART")
plotLearnerPrediction(lrn, task = iris.task)
```



For *clustering* we also get a scatter plot of two selected features. The color of the points indicates the predicted cluster.



For regression, there are two types of plots. The 1D plot shows the target values in relation to a single feature, the regression curve and, if the chosen learner

supports this, the estimated standard error.

plotLearnerPrediction("regr.lm", features = "lstat", task = bh.task)



The 2D variant, as in the classification case, generates a scatter plot of 2 features. The fill color of the dots illustrates the value of the target variable "medv", the background colors show the estimated mean. The plot does not represent the estimated standard error.

```
plotLearnerPrediction("regr.lm", features = c("lstat", "rm"), task =
    bh.task)
```



2.5 Evaluating Learner Performance

The quality of the predictions of a model in mlr can be assessed with respect to a number of different performance measures. In order to calculate the performance measures, call performance on the object returned by predict and specify the desired performance measures.

2.5.1 Available performance measures

mlr provides a large number of performance measures for all types of learning problems. Typical performance measures for *classification* are the mean misclassification error (mmce), accuracy (acc) or measures based on ROC analysis. For *regression* the mean of squared errors (mse) or mean of absolute errors (mae) are usually considered. For *clustering* tasks, measures such as the Dunn index (dunn) are provided, while for *survival* predictions, the Concordance Index (cindex) is supported, and for *cost-sensitive* predictions the misclassification penalty (mcp) and others. It is also possible to access the time to train the learner (timetrain), the time to compute the prediction (timepredict) and their sum (timeboth) as performance measures.

To see which performance measures are implemented, have a look at the table of performance measures and the measures documentation page.

If you want to implement an additional measure or include a measure with non-standard misclassification costs, see the section on creating custom measures.

2.5.2 Listing measures

The properties and requirements of the individual measures are shown in the table of performance measures.

If you would like a list of available measures with certain properties or suitable for a certain learning Task use the function listMeasures.

```
### Performance measures for classification with multiple classes
listMeasures("classif", properties = "classif.multi")
                          "multiclass.brier" "multiclass.aunp"
#> [1] "kappa"
  [4] "multiclass.aunu" "qsr"
                                             "ber"
#> [7] "logloss"
                          "wkappa"
                                             "timeboth"
#> [10] "timepredict"
                          "acc"
                                             "lsr"
                         "multiclass.au1p" "multiclass.au1u"
#> [13] "featperc"
#> [16] "ssr"
                          "timetrain"
                                             "mmce"
### Performance measure suitable for the iris classification task
listMeasures(iris.task)
                          "multiclass.brier" "multiclass.aunp"
#> [1] "kappa"
#> [4] "multiclass.aunu" "qsr"
                                             "ber"
#> [7] "logloss"
                          "wkappa"
                                             "timeboth"
#> [10] "timepredict"
                          "acc"
                                             "lsr"
#> [13] "featperc"
                          "multiclass.au1p" "multiclass.au1u"
#> [16] "ssr"
                          "timetrain"
                                             "mmce"
```

For convenience there exists a default measure for each type of learning problem, which is calculated if nothing else is specified. As defaults we chose the most commonly used measures for the respective types, e.g., the mean squared error (mse) for regression and the misclassification rate (mmce) for classification. The help page of function getDefaultMeasure lists all defaults for all types of learning problems. The function itself returns the default measure for a given task type, Task or Learner.

```
### Get default measure for iris.task
getDefaultMeasure(iris.task)
#> Name: Mean misclassification error
#> Performance measure: mmce
#> Properties: classif, classif.multi, req.pred, req.truth
#> Minimize: TRUE
#> Best: 0; Worst: 1
#> Aggregated by: test.mean
#> Note: Defined as: mean(response != truth)

### Get the default measure for linear regression
getDefaultMeasure(makeLearner("regr.lm"))
#> Name: Mean of squared errors
#> Performance measure: mse
```

```
#> Properties: regr,req.pred,req.truth
#> Minimize: TRUE
#> Best: 0; Worst: Inf
#> Aggregated by: test.mean
#> Note: Defined as: mean((response - truth) 2)
```

2.5.3 Calculate performance measures

In the following example we fit a gradient boosting machine on a subset of the BostonHousing data set and calculate the default measure mean squared error (mse) on the remaining observations.

```
n = getTaskSize(bh.task)
lrn = makeLearner("regr.gbm", n.trees = 1000)
mod = train(lrn, task = bh.task, subset = seq(1, n, 2))
pred = predict(mod, task = bh.task, subset = seq(2, n, 2))

performance(pred)
#> mse
#> 42.68414
```

The following code computes the median of squared errors (medse) instead.

```
performance(pred, measures = medse)
#> medse
#> 9.134965
```

Of course, we can also calculate multiple performance measures at once by simply passing a list of measures which can also include your own measure.

Calculate the mean squared error, median squared error and mean absolute error (mae).

```
performance(pred, measures = list(mse, medse, mae))
#> mse  medse  mae
#> 42.684141  9.134965  4.536750
```

For the other types of learning problems and measures, calculating the performance basically works in the same way.

${\bf 2.5.3.1} \quad {\bf Requirements\ of\ performance\ measures}$

Note that in order to calculate some performance measures it is required that you pass the Task or the fitted model in addition to the Prediction.

For example in order to assess the time needed for training (timetrain), the fitted model has to be passed.

```
performance(pred, measures = timetrain, model = mod)
#> timetrain
#> 0.076
```

For many performance measures in cluster analysis the Task is required.

```
lrn = makeLearner("cluster.kmeans", centers = 3)
mod = train(lrn, mtcars.task)
pred = predict(mod, task = mtcars.task)

### Calculate the Dunn index
performance(pred, measures = dunn, task = mtcars.task)
#> dunn
#> 0.1462919
```

Moreover, some measures require a certain type of prediction. For example in binary classification in order to calculate the AUC (auc) – the area under the ROC (receiver operating characteristic) curve – we have to make sure that posterior probabilities are predicted. For more information on ROC analysis, see the section on ROC analysis.

```
lrn = makeLearner("classif.rpart", predict.type = "prob")
mod = train(lrn, task = sonar.task)
pred = predict(mod, task = sonar.task)

performance(pred, measures = auc)
#> auc
#> 0.9224018
```

Also bear in mind that many of the performance measures that are available for classification, e.g., the false positive rate (fpr), are only suitable for binary problems.

2.5.4 Access a performance measure

Performance measures in mlr are objects of class Measure. If you are interested in the properties or requirements of a single measure you can access it directly. See the help page of Measure for information on the individual slots.

```
#> $ extra.args: list()
#> $ best : num 0
#> $ worst : num 1
#> $ name : chr "Mean misclassification error"
#> $ note : chr "Defined as: mean(response != truth)"
#> $ aggr :List of 4
#> ..$ id : chr "test.mean"
#> ..$ name : chr "Test mean"
#> ..$ fun :function (task, perf.test, perf.train, measure, group, pred)
#> ..$ properties: chr "req.test"
#> .. attr(*, "class")= chr "Aggregation"
#> - attr(*, "class")= chr "Measure"
```

2.5.5 Binary classification

For binary classification specialized techniques exist to analyze the performance.

2.5.5.1 Plot performance versus threshold

As you may recall (see the previous section on making predictions) in binary classification we can adjust the threshold used to map probabilities to class labels. Helpful in this regard is are the functions generate ThreshVsPerfData and plotThreshVsPerf, which generate and plot, respectively, the learner performance versus the threshold.

For more performance plots and automatic threshold tuning see here.

In the following example we consider the Sonar data set and plot the false positive rate (fpr), the false negative rate (fnr) as well as the misclassification rate (mmce) for all possible threshold values.

```
lrn = makeLearner("classif.lda", predict.type = "prob")
n = getTaskSize(sonar.task)
mod = train(lrn, task = sonar.task, subset = seq(1, n, by = 2))
pred = predict(mod, task = sonar.task, subset = seq(2, n, by = 2))

### Performance for the default threshold 0.5
performance(pred, measures = list(fpr, fnr, mmce))
#> fpr fnr mmce
#> 0.2500000 0.3035714 0.2788462
### Plot false negative and positive rates as well as the error rate versus the threshold
d = generateThreshVsPerfData(pred, measures = list(fpr, fnr, mmce))
plotThreshVsPerf(d)
```



There is an experimental ggvis plotting function plotThreshVsPerfGGVIS which performs similarly to plotThreshVsPerf but instead of creating facetted subplots to visualize multiple learners and/or multiple measures, one of them is mapped to an interactive sidebar which selects what to display.

plotThreshVsPerfGGVIS(d)

2.5.5.2 ROC measures

For binary classification a large number of specialized measures exist, which can be nicely formatted into one matrix, see for example the receiver operating characteristic page on wikipedia.

We can generate a similar table with the calculateROCMeasures function.

```
r = calculateROCMeasures(pred)
r
#>
       predicted
                  R
#>
  true M
#>
      M 0.7
                  0.3
                             tpr: 0.7 fnr: 0.3
      R 0.25
                  0.75
#>
                            fpr: 0.25 tnr: 0.75
        ppv: 0.76 for: 0.32 lrp: 2.79 acc: 0.72
#>
        fdr: 0.24 npv: 0.68 lrm: 0.4 dor: 6.88
#>
#>
#>
#> Abbreviations:
#> tpr - True positive rate (Sensitivity, Recall)
#> fpr - False positive rate (Fall-out)
#> fnr - False negative rate (Miss rate)
#> tnr - True negative rate (Specificity)
#> ppv - Positive predictive value (Precision)
#> for - False omission rate
```

```
#> lrp - Positive likelihood ratio (LR+)
#> fdr - False discovery rate
#> npv - Negative predictive value
#> acc - Accuracy
#> lrm - Negative likelihood ratio (LR-)
#> dor - Diagnostic odds ratio
```

The top left 2×2 matrix is the confusion matrix, which shows the relative frequency of correctly and incorrectly classified observations. Below and to the right a large number of performance measures that can be inferred from the confusion matrix are added. By default some additional info about the measures is printed. You can turn this off using the abbreviations argument of the print method: print(r, abbreviations = FALSE).

2.6 Resampling

In order to assess the performance of a learning algorithm, resampling strategies are usually used. The entire data set is split into (multiple) training and test sets. You train a learner on each training set, predict on the corresponding test set (sometimes on the training set as well) and calculate some performance measure. Then the individual performance values are aggregated, typically by calculating the mean. There exist various different resampling strategies, for example cross-validation and bootstrap, to mention just two popular approaches.



If you want to read up further details, the paper Resampling Strategies for Model Assessment and Selection by Simon is proabably not a bad choice. Bernd has also published a paper Resampling methods for meta-model validation with

recommendations for evolutionary computation which contains detailed descriptions and lots of statistical background information on resampling methods.

In mlr the resampling strategy can be chosen via the function makeResampleDesc. The supported resampling strategies are:

- Cross-validation ("CV"),
- Leave-one-out cross-validation ("LOO""),
- Repeated cross-validation ("RepCV"),
- Out-of-bag bootstrap and other variants ("Bootstrap"),
- Subsampling, also called Monte-Carlo cross-validation ("Subsample"),
- Holdout (training/test) ("Holdout").

The resample function evaluates the performance of a Learner using the specified resampling strategy for a given machine learning Task.

In the following example the performance of the Cox proportional hazards model on the lung data set is calculated using 3-fold cross-validation. Generally, in K-fold cross-validation the data set D is partitioned into K subsets of (approximately) equal size. In the i-th step of the K iterations, the i-th subset is used for testing, while the union of the remaining parts forms the training set. The default performance measure in survival analysis is the concordance index (cindex).

```
### Specify the resampling strategy (3-fold cross-validation)
rdesc = makeResampleDesc("CV", iters = 3)
### Calculate the performance
r = resample("surv.coxph", lung.task, rdesc)
#> [Resample] cross-validation iter: 1
#> [Resample] cross-validation iter: 2
#> [Resample] cross-validation iter: 3
#> [Resample] Result: cindex.test.mean=0.627
r
#> Resample Result
#> Task: lung-example
#> Learner: surv.coxph
#> Aggr perf: cindex.test.mean=0.627
#> Runtime: 0.0426955
### peak a little bit into r
names(r)
   [1] "learner.id"
                         "task.id"
                                          "measures.train"
    "measures.test"
                         "pred"
  [5] "aggr"
                                          "models"
                                                            "err.msqs"
                         "runtime"
#> [9] "extract"
r$aggr
#> cindex.test.mean
```

```
0.6271182
r$measures.test
#>
     iter
             cindex
#> 1
        1 0.5783027
#> 2
        2 0.6324074
#> 3
        3 0.6706444
r$measures.train
     iter cindex
#> 1
        1
              NA
#> 2
        2
              NA
        3
#> 3
              NA
```

r\$measures.test gives the value of the performance measure on the 3 individual test data sets. r\$aggr shows the aggregated performance value. Its name, "cindex.test.mean", indicates the performance measure, cindex, and the method used to aggregate the 3 individual performances. test.mean is the default method and, as the name implies, takes the mean over the performances on the 3 test data sets. No predictions on the training data sets were made and thus r\$measures.train contains missing values.

If predictions for the training set are required, too, set **predict** = "train"or **predict** = "both" in makeResampleDesc. This is necessary for some bootstrap methods (b632 and b632+) and we will see some examples later on.

r\$pred is an object of class ResamplePrediction. Just as a Prediction object (see the section on making predictions) r\$pred has an element called "data" which is a data.frame that contains the predictions and in case of a supervised learning problem the true values of the target variable.

```
head(r$pred$data)
#>
     id truth.time truth.event
                                 response iter set
               455
#> 1 1
                          TRUE -0.4951788
                                             1 test
#> 2 2
               210
                          TRUE 0.9573824
                                             1 test
#> 3 4
               310
                          TRUE 0.8069059
                                             1 test
#> 4 10
               613
                          TRUE
                                0.1918188
                                             1 test
                61
#> 5 12
                          TRUE 0.6638736
                                             1 test
#> 6 14
                81
                          TRUE -0.1873917
                                             1 test
```

The columns iter and set indicate the resampling iteration and if an individual prediction was made on the test or the training data set.

In the above example the performance measure is the concordance index (cindex). Of course, it is possible to compute multiple performance measures at once by passing a list of measures (see also the previous section on evaluating learner performance).

In the following we estimate the Dunn index (dunn), the Davies-Bouldin cluster separation measure (db), and the time for training the learner (timetrain) by

subsampling with 5 iterations. In each iteration the data set D is randomly partitioned into a training and a test set according to a given percentage, e.g., 2/3 training and 1/3 test set. If there is just one iteration, the strategy is commonly called *holdout* or test sample estimation.

```
### cluster iris feature data
task = makeClusterTask(data = iris[,-5])
### Subsampling with 5 iterations and default split 2/3
rdesc = makeResampleDesc("Subsample", iters = 5)
### Subsampling with 5 iterations and 4/5 training data
rdesc = makeResampleDesc("Subsample", iters = 5, split = 4/5)
### Calculate the three performance measures
r = resample("cluster.kmeans", task, rdesc, measures = list(dunn,
    db, timetrain))
#> [Resample] subsampling iter: 1
#> [Resample] subsampling iter: 2
#> [Resample] subsampling iter: 3
#> [Resample] subsampling iter: 4
#> [Resample] subsampling iter: 5
#> [Resample] Result:
    dunn.test.mean=0.274,db.test.mean=0.51,timetrain.test.mean=0.001
r$aggr
#>
        dunn.test.mean
                              db.test.mean timetrain.test.mean
#>
             0.2738893
                                 0.5103655 0.0010000
```

2.6.1 Stratified resampling

For classification, it is usually desirable to have the same proportion of the classes in all of the partitions of the original data set. Stratified resampling ensures this. This is particularly useful in case of imbalanced classes and small data sets. Otherwise it may happen, for example, that observations of less frequent classes are missing in some of the training sets which can decrease the performance of the learner, or lead to model crashes In order to conduct stratified resampling, set stratify = TRUE when calling makeResampleDesc.

```
### 3-fold cross-validation
rdesc = makeResampleDesc("CV", iters = 3, stratify = TRUE)

r = resample("classif.lda", iris.task, rdesc)
#> [Resample] cross-validation iter: 1
#> [Resample] cross-validation iter: 2
#> [Resample] cross-validation iter: 3
#> [Resample] Result: mmce.test.mean=0.02
```

Stratification is also available for survival tasks. Here the stratification balances the censoring rate.

Sometimes it is required to also stratify on the input data, e.g. to ensure that all subgroups are represented in all training and test sets. To stratify on the input columns, specify factor columns of your task data via stratify.cols

```
rdesc = makeResampleDesc("CV", iters = 3, stratify.cols = "chas")
r = resample("regr.rpart", bh.task, rdesc)
#> [Resample] cross-validation iter: 1
#> [Resample] cross-validation iter: 2
#> [Resample] cross-validation iter: 3
#> [Resample] Result: mse.test.mean=23.2
```

2.6.2 Accessing individual learner models

In each resampling iteration a Learner is fitted on the respective training set. By default, the resulting WrappedModels are not returned by resample. If you want to keep them, set models = TRUE when calling resample.

```
### 3-fold cross-validation
rdesc = makeResampleDesc("CV", iters = 3)
r = resample("classif.lda", iris.task, rdesc, models = TRUE)
#> [Resample] cross-validation iter: 1
#> [Resample] cross-validation iter: 2
#> [Resample] cross-validation iter: 3
#> [Resample] Result: mmce.test.mean=0.02
r$models
#> [[1]]
#> Model for learner.id=classif.lda; learner.class=classif.lda
#> Trained on: task.id = iris-example; obs = 100; features = 4
#> Hyperparameters:
#>
#> [[2]]
#> Model for learner.id=classif.lda; learner.class=classif.lda
#> Trained on: task.id = iris-example; obs = 100; features = 4
#> Hyperparameters:
#>
#> [[3]]
#> Model for learner.id=classif.lda; learner.class=classif.lda
#> Trained on: task.id = iris-example; obs = 100; features = 4
#> Hyperparameters:
```

Keeping only certain information instead of entire models, for example the variable importance in a regression tree, can be achieved using the extract argu-

ment. The function passed to **extract** is applied to each model fitted on one of the 3 training sets.

```
### 3-fold cross-validation
rdesc = makeResampleDesc("CV", iters = 3)
### Extract the variable importance in a regression tree
r = resample("regr.rpart", bh.task, rdesc,
    extract = function(x) x$learner.model$variable.importance)
#> [Resample] cross-validation iter: 1
#> [Resample] cross-validation iter: 2
#> [Resample] cross-validation iter: 3
#> [Resample] Result: mse.test.mean=30.3
r$extract
#> [[1]]
           rm
                   lstat
                               crim
                                         indus
                                                      age
                                                             ptratio
#> 15228.2872 10742.2277
                          3893.2744
                                     3651.6232
                                                 2601.5262
                                                           2551.8492
#>
          dis
                                rad
                                           tax
                    nox
                                                        zn
   2498.2748 2419.5269 1014.2609
                                      743.3742
                                                 308.8209
#>
#> [[2]]
#>
         lstat
                       nox
                                   age
                                             indus
                                                           crim
#> 15725.19021 9323.20270 8474.23077
                                                    8251.74446
                                        8358.67000
    7332.59637
#>
            zn
                       dis
                                   tax
                                               rad
                                                       ptratio
    6151.29577 2741.12074 2055.67537 1216.01398
                                                      634.78381
    71.00088
#>
#> [[31]
#>
           rm
                   lstat
                                age
                                       ptratio
                                                      nox
                                                                  dis
#> 15890.9279 13262.3672 4296.4175
                                     3678.6651
                                                 3668.4944
                                                            3512.2753
#>
         crim
                     tax
                              indus
                                            zn
                                                        b
   3474.5883 2844.9918 1437.7900 1284.4714
                                                            496.2382
                                                 578.6932
```

2.6.3 Resample descriptions and resample instances

As shown above, the function makeResampleDesc is used to specify the resampling strategy.

```
rdesc = makeResampleDesc("CV", iters = 3)
str(rdesc)
#> List of 4
#> $ id : chr "cross-validation"
#> $ iters : int 3
```

```
#> $ predict : chr "test"
#> $ stratify: logi FALSE
#> - attr(*, "class")= chr [1:2] "CVDesc" "ResampleDesc"
```

The result rdescis an object of class ResampleDesc and contains, as the name implies, a description of the resampling strategy. In principle, this is an instruction for drawing training and test sets including the necessary parameters like the number of iterations, the sizes of the training and test sets etc.

Based on this description, the data set is randomly partitioned into multiple training and test sets. For each iteration, we get a set of index vectors indicating the training and test examples. These are stored in a ResampleInstance.

If a ResampleDesc is passed to resample, it is instantiated internally. Naturally, it is also possible to pass a ResampleInstance directly.

A ResampleInstance can be created through the function makeResampleInstance given a ResampleDesc and either the size of the data set at hand or the Task. It basically performs the random drawing of indices to separate the data into training and test sets according to the description.

```
### Create a resample instance based an a task
rin = makeResampleInstance(rdesc, task = iris.task)
rin
#> Resample instance for 150 cases.
#> Resample description: cross-validation with 3 iterations.
#> Predict: test
#> Stratification: FALSE
### Create a resample instance given the size of the data set
rin = makeResampleInstance(rdesc, size = nrow(iris))
str(rin)
#> List of 5
#> $ desc
               :List of 4
                : chr "cross-validation"
#>
     ..$ id
     ..$ iters
                : int 3
    ..$ predict : chr "test"
     .. $ stratify: logi FALSE
#>
     ..- attr(*, "class")= chr [1:2] "CVDesc" "ResampleDesc"
#>
#>
   $ size
                : int 150
#> $ train.inds:List of 3
#>
    ..$ : int [1:100] 36 81 6 82 120 110 118 132 105 61 ...
     ..$ : int [1:100] 6 119 120 110 121 118 99 100 29 127 ...
    ..$ : int [1:100] 36 81 82 119 121 99 132 105 61 115 ...
#> $ test.inds :List of 3
    ..$ : int [1:50] 2 3 4 5 7 9 11 16 22 24 ...
    ..$ : int [1:50] 8 12 17 19 20 23 25 27 32 33 ...
#> ..$ : int [1:50] 1 6 10 13 14 15 18 21 29 31 ...
```

```
$ group : Factor w/ 0 levels:
  - attr(*, "class")= chr "ResampleInstance"
### Access the indices of the training observations in iteration 3
rin$train.inds[[3]]
    [1] 36 81 82 119 121 99 132 105 61 115 17 42 4 71
    79 30
   [18] 113 138 19 150 77 58 92 114 133
                                        8 109 33 145 22 111
   [35]
                             16 43 149
                                          9 46 32 139 87
         7 44
               3 20 134
                          96
    11 52
   [52] 86 40 141 142 72 54 48 83 64 90 112 148 129 137 116
   143 69
   [69] 84 25 80 37 38
                         75 130 126 135 107 146 26 12 98
   124 60
   [86] 63 117 23 67 73 28 106 76 50 144 59 47 102 56 27
```

While having two separate objects, resample descriptions and instances as well as the resample function seems overly complicated, it has several advantages:

Resample instances allow for paired experiments, that is comparing the
performance of several learners on exactly the same training and test sets.
This is particularly useful if you want to add another method to a comparison experiment you already did.

```
rdesc = makeResampleDesc("CV", iters = 3)
rin = makeResampleInstance(rdesc, task = iris.task)

### Calculate the performance of two learners based on the same
    resample instance
r.lda = resample("classif.lda", iris.task, rin, show.info = FALSE)
r.rpart = resample("classif.rpart", iris.task, rin, show.info =
    FALSE)
r.lda$aggr
#> mmce.test.mean
#> 0.02666667
r.rpart$aggr
#> mmce.test.mean
#> 0.06
```

• It is easy to add other resampling methods later on. You can simply derive from the ResampleInstance class, but you do not have to touch any methods that use the resampling strategy.

As mentioned above, when calling makeResampleInstance the index sets are drawn randomly. Mainly for *holdout* (test sample) estimation you might want full control about the training and tests set and specify them manually. This can be done using the function makeFixedHoldoutInstance.

2.6.4 Aggregating performance values

In resampling we get (for each measure we wish to calculate) one performance value (on the test set, training set, or both) for each iteration. Subsequently, these are aggregated. As mentioned above, mainly the mean over the performance values on the test data sets (test.mean) is calculated.

For example, a 10-fold cross validation computes 10 values for the chosen performance measure. The aggregated value is the mean of these 10 numbers. mlr knows how to handle it because each Measure knows how it is aggregated:

```
### Mean misclassification error
mmce$aggr
#> Aggregation function: test.mean
### Root mean square error
rmse$aggr
#> Aggregation function: test.rmse
```

The aggregation method of a Measure can be changed via the function setAggregation. See the documentation of aggregations for available methods.

2.6.4.1 Example: Different measures and aggregations

test.median computes the median of the performance values on the test sets.

```
#> [Resample] cross-validation iter: 1
#> [Resample] cross-validation iter: 2
#> [Resample] cross-validation iter: 3
#> [Resample] Result: mmce.test.mean=0.293,tpr.test.median=0.735
r$aggr
#> mmce.test.mean tpr.test.median
#> 0.2930987  0.7352941
```

2.6.4.2 Example: Calculating the training error

Here we calculate the mean misclassification error (mmce) on the training and the test data sets. Note that we have to set predict = "both"when calling makeResampleDesc in order to get predictions on both data sets, training and test.

```
mmce.train.mean = setAggregation(mmce, train.mean)
rdesc = makeResampleDesc("CV", iters = 3, predict = "both")
r = resample("classif.rpart", iris.task, rdesc, measures =
   list(mmce, mmce.train.mean))
#> [Resample] cross-validation iter: 1
#> [Resample] cross-validation iter: 2
#> [Resample] cross-validation iter: 3
#> [Resample] Result: mmce.test.mean=0.0467,mmce.train.mean=0.0367
r$measures.train
   iter mmce mmce
#> 1
     1 0.04 0.04
       2 0.03 0.03
#> 2
#> 3
       3 0.04 0.04
r$aggr
#> mmce.test.mean mmce.train.mean
```

2.6.4.3 Example: Bootstrap

In out-of-bag bootstrap estimation B new data sets D_1 to D_B are drawn from the data set D with replacement, each of the same size as D. In the i-th iteration, D_i forms the training set, while the remaining elements from D, i.e., elements not in the training set, form the test set.

The variants b632 and b632+ calculate a convex combination of the training performance and the out-of-bag bootstrap performance and thus require predictions on the training sets and an appropriate aggregation strategy.

```
rdesc = makeResampleDesc("Bootstrap", predict = "both", iters = 10)
b632.mmce = setAggregation(mmce, b632)
b632plus.mmce = setAggregation(mmce, b632plus)
```

```
b632.mmce
#> Name: Mean misclassification error
#> Performance measure: mmce
#> Properties: classif, classif.multi, req. pred, req. truth
#> Minimize: TRUE
#> Best: 0; Worst: 1
#> Aggregated by: b632
#> Note: Defined as: mean(response != truth)
r = resample("classif.rpart", iris.task, rdesc,
    measures = list(mmce, b632.mmce, b632plus.mmce), show.info =
       FALSE)
head(r$measures.train)
   iter
                 mmce
                            mmce
                                        mmce
      1 0.026666667 0.026666667 0.026666667
       2 0.026666667 0.026666667 0.026666667
       3 0.006666667 0.006666667 0.006666667
       4 0.026666667 0.026666667 0.026666667
#> 4
#> 5
       5 0.033333333 0.033333333 0.033333333
       6 0.013333333 0.013333333 0.013333333
#> 6
r$aggr
                      mmce.b632 mmce.b632plus
#> mmce.test.mean
#> 0.07051905 0.05389071 0.05496489
```

2.6.5 Convenience functions

When quickly trying out some learners, it can get tedious to write the ${\bf R}$ code for generating a resample instance, setting the aggregation strategy and so on. For this reason mlr provides some convenience functions for the frequently used resampling strategies, for example holdout, crossval or bootstrapB632. But note that you do not have as much control and flexibility as when using resample with a resample description or instance.

```
holdout("regr.lm", bh.task, measures = list(mse, mae))
crossval("classif.lda", iris.task, iters = 3, measures = list(mmce,
    ber))
```

2.7 Tuning Hyperparameters

Many machine learning algorithms have hyperparameters that need to be set. If selected by the user they can be specified as explained on the tutorial page on Learners – simply pass them to makeLearner. Often suitable parameter values are not obvious and it is preferable to tune the hyperparameters, that is automatically identify values that lead to the best performance.

2.7.1 Basics

In order to tune a machine learning algorithm, you have to specify:

- the search space
- the optimization algorithm (aka tuning method)
- an evaluation method, i.e., a resampling strategy and a performance measure

An example of the search space could be searching values of the C parameter for SVM:

An example of the optimization algorithm could be performing random search on the space:

```
### ex: random search with 100 iterations
ctrl = makeTuneControlRandom(maxit = 100L)
```

An example of an evaluation method could be 3-fold CV using accuracy as the performance measure:

```
rdesc = makeResampleDesc("CV", iters = 3L)
measure = acc
```

The evaluation method is already covered in detail in evaluation of learning methods and resampling.

In this tutorial, we show how to specify the search space and optimization algorithm, how to do the tuning and how to access the tuning result, and how to visualize the hyperparameter tuning effects through several examples.

Throughout this section we consider classification examples. For the other types of learning problems, you can follow the same process analogously.

We use the iris classification task for illustration and tune the hyperparameters of an SVM (function ksvm from the kernlab package) with a radial basis kernel. The following examples tune the cost parameter ${\tt C}$ and the RBF kernel parameter sigma of the ksvm function.

2.7.1.1 Specifying the search space

We first must define a space to search when tuning our learner. For example, maybe we want to tune several specific values of a hyperparameter or perhaps we want to define a space from 10^{-10} to 10^{10} and let the optimization algorithm decide which points to choose.

In order to define a search space, we create a ParamSet object, which describes the parameter space we wish to search. This is done via the function makeParamSet.

For example, we could define a search space with just the values 0.5, 1.0, 1.5, 2.0 for both C and gamma. Notice how we name each parameter as it's defined in the kernlab package:

```
discrete_ps = makeParamSet(
   makeDiscreteParam("C", values = c(0.5, 1.0, 1.5, 2.0)),
   makeDiscreteParam("sigma", values = c(0.5, 1.0, 1.5, 2.0))
)
print(discrete_ps)
#> Type len Def Constr Req Tunable Trafo
#> C discrete - - 0.5,1,1.5,2 - TRUE -
#> sigma discrete - - 0.5,1,1.5,2 - TRUE -
```

We could also define a continuous search space (using makeNumericParam instead of makeDiscreteParam) from 10^{-10} to 10^{10} for both parameters through the use of the trafo argument (trafo is short for transformation). Transformations work like this: All optimizers basically see the parameters on their original scale (from -10 to 10 in this case) and produce values on this scale during the search. Right before they are passed to the learning algorithm, the transformation function is applied.

Notice this time we use makeNumericParam:

Many other parameters can be created, check out the examples in makeParam-Set.

In order to standardize your workflow across several packages, whenever parameters in the underlying \mathbf{R} functions should be passed in a list structure, mlr tries to give you direct access to each parameter and get rid of the list structure!

This is the case with the kpar argument of ksvm which is a list of kernel parameters like sigma. This allows us to interface with learners from different packages in the same way when defining parameters to tune!

2.7.1.2 Specifying the optimization algorithm

Now that we have specified the search space, we need to choose an optimization algorithm for our parameters to pass to the ksvm learner. Optimization algorithms are considered TuneControl objects in mlr.

A grid search is one of the standard – albeit slow – ways to choose an appropriate set of parameters from a given search space.

In the case of discrete_ps above, since we have manually specified the values, grid search will simply be the cross product. We create the grid search object using the defaults, noting that we will have $4 \times 4 = 16$ combinations in the case of discrete_ps:

```
ctrl = makeTuneControlGrid()
```

In the case of num_ps above, since we have only specified the upper and lower bounds for the search space, grid search will create a grid using equally-sized steps. By default, grid search will span the space in 10 equal-sized steps. The number of steps can be changed with the resolution argument. Here we change to 15 equal-sized steps in the space defined within the ParamSet object. For num_ps, this means 15 steps in the form of 10 ^ seq(-10, 10, length.out = 15):

```
ctrl = makeTuneControlGrid(resolution = 15L)
```

Many other types of optimization algorithms are available. Check out TuneControl for some examples.

Since grid search is normally too slow in practice, we'll also examine random search. In the case of discrete_ps, random search will randomly choose from the specified values. The maxit argument controls the amount of iterations.

```
ctrl = makeTuneControlRandom(maxit = 10L)
```

In the case of num_ps, random search will randomly choose points within the space according to the specified bounds. Perhaps in this case we would want to increase the amount of iterations to ensure we adequately cover the space:

```
ctrl = makeTuneControlRandom(maxit = 200L)
```

2.7.1.3 Performing the tuning

Now that we have specified a search space and the optimization algorithm, it's time to perform the tuning. We will need to define a resampling strategy and make note of our performance measure.

We will use 3-fold cross-validation to assess the quality of a specific parameter setting. For this we need to create a resampling description just like in the resampling part of the tutorial.

```
rdesc = makeResampleDesc("CV", iters = 3L)
```

Finally, by combining all the previous pieces, we can tune the SVM parameters by calling tuneParams. We will use discrete_ps with grid search:

```
discrete_ps = makeParamSet(
 makeDiscreteParam("C", values = c(0.5, 1.0, 1.5, 2.0)),
 makeDiscreteParam("sigma", values = c(0.5, 1.0, 1.5, 2.0))
)
ctrl = makeTuneControlGrid()
rdesc = makeResampleDesc("CV", iters = 3L)
res = tuneParams("classif.ksvm", task = iris.task, resampling =
    rdesc,
 par.set = discrete_ps, control = ctrl)
#> [Tune] Started tuning learner classif.ksvm for parameter set:
             Type len Def
                             Constr Req Tunable Trafo
#> C
         discrete - - 0.5,1,1.5,2 -
                                             TRUE
#> sigma discrete - - 0.5,1,1.5,2
                                             TRUE
#> With control class: TuneControlGrid
#> Imputation value: 1
\# [Tune-x] 1: C=0.5; sigma=0.5
#> [Tune-y] 1: mmce.test.mean=0.04; time: 0.0 min; memory: 160Mb
    use, 724Mb max
#> [Tune-x] 2: C=1; sigma=0.5
#> [Tune-y] 2: mmce.test.mean=0.04; time: 0.0 min; memory: 160Mb
    use, 724Mb max
#> [Tune-x] 3: C=1.5; sigma=0.5
#> [Tune-y] 3: mmce.test.mean=0.0467; time: 0.0 min; memory: 160Mb
   use, 724Mb max
#> [Tune-x] 4: C=2; sigma=0.5
#> [Tune-y] 4: mmce.test.mean=0.0467; time: 0.0 min; memory: 160Mb
   use, 724Mb max
#> [Tune-x] 5: C=0.5; sigma=1
#> [Tune-y] 5: mmce.test.mean=0.04; time: 0.0 min; memory: 160Mb
    use, 724Mb max
\# [Tune-x] 6: C=1; sigma=1
#> [Tune-y] 6: mmce.test.mean=0.0467; time: 0.0 min; memory: 160Mb
    use, 724Mb max
#> [Tune-x] 7: C=1.5; sigma=1
#> [Tune-y] 7: mmce.test.mean=0.0467; time: 0.0 min; memory: 160Mb
    use, 724Mb max
#> [Tune-x] 8: C=2; sigma=1
#> [Tune-y] 8: mmce.test.mean=0.0467; time: 0.0 min; memory: 160Mb
    use, 724Mb max
#> [Tune-x] 9: C=0.5; sigma=1.5
#> [Tune-y] 9: mmce.test.mean=0.0333; time: 0.0 min; memory: 160Mb
```

```
use, 724Mb max
#> [Tune-x] 10: C=1; sigma=1.5
#> [Tune-y] 10: mmce.test.mean=0.04; time: 0.0 min; memory: 160Mb
    use, 724Mb max
#> [Tune-x] 11: C=1.5; sigma=1.5
#> [Tune-y] 11: mmce.test.mean=0.04; time: 0.0 min; memory: 160Mb
    use, 724Mb max
#> [Tune-x] 12: C=2; sigma=1.5
#> [Tune-y] 12: mmce.test.mean=0.0467; time: 0.0 min; memory: 160Mb
    use, 724Mb max
#> [Tune-x] 13: C=0.5; sigma=2
#> [Tune-y] 13: mmce.test.mean=0.04; time: 0.0 min; memory: 160Mb
    use, 724Mb max
#> [Tune-x] 14: C=1; siqma=2
#> [Tune-y] 14: mmce.test.mean=0.0333; time: 0.0 min; memory: 160Mb
    use, 724Mb max
#> [Tune-x] 15: C=1.5; sigma=2
#> [Tune-y] 15: mmce.test.mean=0.04; time: 0.0 min; memory: 160Mb
    use, 724Mb max
#> [Tune-x] 16: C=2; sigma=2
#> [Tune-y] 16: mmce.test.mean=0.04; time: 0.0 min; memory: 160Mb
    use, 724Mb max
#> [Tune] Result: C=0.5; sigma=1.5 : mmce.test.mean=0.0333
res
#> Tune result:
#> Op. pars: C=0.5; sigma=1.5
#> mmce.test.mean=0.0333
```

tuneParams simply performs the cross-validation for every element of the cross-product and selects the parameter setting with the best mean performance. As no performance measure was specified, by default the error rate (mmce) is used.

Note that each measure "knows" if it is minimized or maximized during tuning.

```
### error rate
mmce$minimize
#> [1] TRUE

### accuracy
acc$minimize
#> [1] FALSE
```

Of course, you can pass other measures and also a list of measures to tuneParams. In the latter case the first measure is optimized during tuning, the others are simply evaluated. If you are interested in optimizing several measures simultaneously have a look at Advanced Tuning.

In the example below we calculate the accuracy (acc) instead of the error rate. We use function setAggregation, as described on the resampling page, to additionally obtain the standard deviation of the accuracy. We also use random search with 100 iterations on the num_set we defined above and set show.info to FALSE to hide the output for all 100 iterations:

2.7.1.4 Accessing the tuning result

The result object TuneResult allows you to access the best found settings x and their estimated performance y.

```
res$x
#> $C
#> [1] 95.22422
#>
#> $sigma
#> [1] 0.006695534

res$y
#> acc.test.mean acc.test.sd
#> 0.98666667 0.02309401
```

We can generate a Learner with optimal hyperparameter settings as follows:

```
lrn = setHyperPars(makeLearner("classif.ksvm"), par.vals = res$x)
lrn
#> Learner classif.ksvm from package kernlab
#> Type: classif
#> Name: Support Vector Machines; Short name: ksvm
#> Class: classif.ksvm
```

```
#> Properties:
     twoclass, multiclass, numerics, factors, prob, class.weights
#> Predict-Type: response
#> Hyperparameters: fit=FALSE, C=95.2, sigma=0.0067
```

Then you can proceed as usual. Here we refit and predict the learner on the complete iris data set:

```
m = train(lrn, iris.task)
predict(m, task = iris.task)
#> Prediction: 150 observations
#> predict.type: response
#> threshold:
#> time: 0.00
#> id truth response
#> 1 1 setosa setosa
#> 2 2 setosa setosa
#> 3 3 setosa setosa
#> 4 4 setosa setosa
#> 5 5 setosa setosa
#> 6 6 setosa setosa
#> 6 6 setosa setosa
#> ... (150 rows, 3 cols)
```

But what if you wanted to inspect the other points on the search path, not just the optimal?

2.7.1.5 Investigating hyperparameter tuning effects

We can inspect all points evaluated during the search by using generateHyper-ParsEffectData:

```
generateHyperParsEffectData(res)
#> HyperParsEffectData:
#> Hyperparameters: C, sigma
#> Measures: acc.test.mean,acc.test.sd
#> Optimizer: TuneControlRandom
#> Nested CV Used: FALSE
#> Snapshot of data:
            C
                   sigma acc.test.mean acc.test.sd iteration
   exec.time
#> 1 -9.9783231 1.0531818
                           0.2733333 0.02309401
   0.063
#> 2 -0.5292817 3.2214785
                             0.2733333 0.02309401
   0.064
#> 3 -0.3544567 4.1644832
                             0.2733333 0.02309401
  0.064
```

```
#> 4 0.6341910 7.8640461 0.2866667 0.03055050 4
0.064
#> 5 5.7640748 -3.3159251 0.9533333 0.03055050 5
0.066
#> 6 -6.5880397 0.4600323 0.2733333 0.02309401 6
0.065
```

Note that the result of generateHyperParsEffectData contains the parameter values on the original scale. In order to get the transformed parameter values instead, use the trafo argument:

```
generateHyperParsEffectData(res, trafo = TRUE)
#> HyperParsEffectData:
#> Hyperparameters: C,sigma
#> Measures: acc.test.mean,acc.test.sd
#> Optimizer: TuneControlRandom
#> Nested CV Used: FALSE
#> Snapshot of data:
#>
                C
                         sigma acc.test.mean acc.test.sd iteration
    exec. time
#> 1 1.051180e-10 1.130269e+01
                                   0.2733333 0.02309401
                                                                 1
        0.063
#> 2 2.956095e-01 1.665246e+03
                                   0.2733333 0.02309401
        0.064
#> 3 4.421232e-01 1.460438e+04
                                   0.2733333 0.02309401
                                                                 3
        0.064
#> 4 4.307159e+00 7.312168e+07
                                   0.2866667 0.03055050
                                                                 4
        0.064
#> 5 5.808644e+05 4.831421e-04
                                   0.9533333 0.03055050
                                                                 5
        0.066
#> 6 2.582024e-07 2.884246e+00
                                   0.2733333 0.02309401
                                                                 6
        0.065
```

Note that we can also generate performance on the train data along with the validation/test data, as discussed on the resampling tutorial page:

```
#> Snapshot of data:
           C
                   sigma acc.test.mean acc.train.mean iteration
   exec.time
#> 1 9.457202 -4.0536025
                                  0.98
                                                0.97
                                                             1
   0.049
#> 2 9.900523 1.8815923
                                  0.40
                                                1.00
                                                             2
   0.036
#> 3 2.363975 5.3202458
                                  0.26
                                                1.00
                                                             3
   0.035
#> 4 -1.530251 4.7579424
                                  0.26
                                                0.37
   0.035
#> 5 -7.837476 2.4352698
                                  0.26
                                                0.37
                                                             5
   0.035
#> 6 8.782931 -0.4143757
                                                             6
                                  0.92
                                                1.00
   0.034
```

We can also easily visualize the points evaluated by using plotHyperParsEffect. In the example below, we plot the performance over iterations, using the **res** from the previous section but instead with 2 performance measures:

```
res = tuneParams("classif.ksvm", task = iris.task, resampling =
    rdesc, par.set = num_ps,
    control = ctrl, measures = list(acc, mmce), show.info = FALSE)
data = generateHyperParsEffectData(res)
plotHyperParsEffect(data, x = "iteration", y = "acc.test.mean",
    plot.type = "line")
```



Note that by default, we only plot the current global optima. This can be changed with the global.only argument.

For an in-depth exploration of generating hyperparameter tuning effects and plotting the data, check out Hyperparameter Tuning Effects.

2.7.2 Further comments

- Tuning works for all other tasks like regression, survival analysis and so on in a completely similar fashion.
- In longer running tuning experiments it is very annoying if the computation stops due to numerical or other errors. Have a look at on.learner.error in configureMlr as well as the examples given in section Configure mlr of this tutorial. You might also want to inform yourself about impute.val in TuneControl.

• As we continually optimize over the same data during tuning, the estimated performance value might be optimistically biased. A clean approach to ensure unbiased performance estimation is nested resampling, where we embed the whole model selection process into an outer resampling loop.

2.8 Benchmark Experiments

In a benchmark experiment different learning methods are applied to one or several data sets with the aim to compare and rank the algorithms with respect to one or more performance measures.

In mlr a benchmark experiment can be conducted by calling function benchmark on a list of Learners and a list of Tasks. benchmark basically executes resample for each combination of Learner and Task. You can specify an individual resampling strategy for each Task and select one or multiple performance measures to be calculated.

2.8.1 Conducting benchmark experiments

We start with a small example. Two learners, linear discriminant analysis (lda) and a classification tree (rpart), are applied to one classification problem (sonar.task). As resampling strategy we choose "Holdout". The performance is thus calculated on a single randomly sampled test data set.

In the example below we create a resample description (ResampleDesc), which is automatically instantiated by benchmark. The instantiation is done only once per Task, i.e., the same training and test sets are used for all learners. It is also possible to directly pass a ResampleInstance.

If you would like to use a *fixed test data set* instead of a randomly selected one, you can create a suitable ResampleInstance through function makeFixedHold-outInstance.

```
### Two learners to be compared
lrns = list(makeLearner("classif.lda"), makeLearner("classif.rpart"))
### Choose the resampling strategy
rdesc = makeResampleDesc("Holdout")

### Conduct the benchmark experiment
bmr = benchmark(lrns, sonar.task, rdesc)
#> Task: Sonar-example, Learner: classif.lda
#> [Resample] holdout iter: 1
#> [Resample] Result: mmce.test.mean= 0.3
#> Task: Sonar-example, Learner: classif.rpart
#> [Resample] holdout iter: 1
```

```
#> [Resample] Result: mmce.test.mean=0.286

bmr
#> task.id learner.id mmce.test.mean
#> 1 Sonar-example classif.lda 0.3000000
#> 2 Sonar-example classif.rpart 0.2857143
```

In the printed table every row corresponds to one pair of Task and Learner. The entries show the mean misclassification error (mmce), the default performance measure for classification, on the test data set.

The result bmr is an object of class BenchmarkResult. Basically, it contains a list of lists of ResampleResult objects, first ordered by Task and then by Learner.

2.8.2 Accessing benchmark results

mlr provides several accessor functions, named getBMR<WhatToExtract>, that permit to retrieve information for further analyses. This includes for example the performances or predictions of the learning algorithms under consideration.

2.8.2.1 Learner performances

Let's have a look at the benchmark result above. getBMRPerformances returns individual performances in resampling runs, while getBMRAggrPerformances gives the aggregated values.

```
getBMRPerformances(bmr)
#> $ `Sonar-example`
#> $`Sonar-example`$classif.lda
#>
    iter mmce
#> 1
     1 0.3
#>
#> $`Sonar-example`$classif.rpart
   iter
            mmce
#> 1 1 0.2857143
getBMRAggrPerformances(bmr)
#> $ `Sonar-example`
#> $ `Sonar-example `$classif.lda
#> mmce.test.mean
#>
             0.3
#> $`Sonar-example`$classif.rpart
#> mmce.test.mean
#>
   0.2857143
```

Since we used holdout as resampling strategy, individual and aggregated performance values coincide.

Often it is more convenient to work with data.frames. You can easily convert the result structure by setting as.df = TRUE.

2.8.2.2 Predictions

Per default, the BenchmarkResult contains the learner predictions. If you do not want to keep them, e.g., to conserve memory, set keep.pred = FALSE when calling benchmark.

You can access the predictions using function getBMRP redictions. Per default, you get a list of lists of Resample Prediction objects. In most cases you might prefer the data. frame version.

```
getBMRPredictions(bmr)
#> $ `Sonar-example`
#> $`Sonar-example`$classif.lda
#> Resampled Prediction for:
#> Resample description: holdout with 0.67 split rate.
#> Predict: test
#> Stratification: FALSE
#> predict.type: response
#> threshold:
#> time (mean): 0.01
#> id truth response iter set
#> 1 180 M M 1 test
#> 2 100 M
                  R 1 test
#> 3 53 R
                  M 1 test
#> 4 89
          R
                  R
                      1 test
        R
#> 5 92
                   Μ
                       1 test
#> 6 11 R
                   R
                       1 test
#> ... (70 rows, 5 cols)
#>
#> $`Sonar-example`$classif.rpart
```

```
#> Resampled Prediction for:
#> Resample description: holdout with 0.67 split rate.
#> Predict: test
#> Stratification: FALSE
#> predict.type: response
#> threshold:
#> time (mean): 0.01
     id truth response iter set
#> 1 180 M M 1 test
#> 2 100
           Μ
                   Μ
                        1 test
#> 3 53
           R
                   R
                        1 test
#> 4 89 R
                   Μ
                       1 test
#> 5 92
           R
                   Μ
                        1 test
#> 6 11 R
                   R
                        1 test
#> ... (70 rows, 5 cols)
head(getBMRPredictions(bmr, as.df = TRUE))
          task.id learner.id id truth response iter set
#> 1 Sonar-example classif.lda 180
                                 Μ
                                        M 1 test
#> 2 Sonar-example classif.lda 100
                                   Μ
                                            R
                                               1 test
#> 3 Sonar-example classif.lda 53
                                   R
                                            Μ
                                               1 test
#> 4 Sonar-example classif.lda 89
                                    R
                                            R
                                                 1 test
#> 5 Sonar-example classif.lda 92
                                    R
                                            Μ
                                                 1 test
#> 6 Sonar-example classif.lda 11
                                    R
                                                 1 test
```

It is also easily possible to access results for certain learners or tasks via their IDs. For this purpose many "getter" functions have a learner.ids and a task.ids argument.

```
head(getBMRPredictions(bmr, learner.ids = "classif.rpart", as.df =
   TRUE))
          task.id
                    learner.id id truth response iter set
#> 1 Sonar-example classif.rpart 180
                                    Μ
                                             Μ
                                                 1 test
#> 2 Sonar-example classif.rpart 100
                                     Μ
                                              Μ
                                                   1 test
#> 3 Sonar-example classif.rpart 53
                                    R
                                              R
                                                   1 test
#> 4 Sonar-example classif.rpart 89
                                      R
                                              Μ
                                                   1 test
#> 5 Sonar-example classif.rpart 92
                                      R
                                              Μ
                                                   1 test
#> 6 Sonar-example classif.rpart 11 R R 1 test
```

If you don't like the default IDs, you can set the IDs of learners and tasks via the id option of makeLearner and make*Task. Moreover, you can conveniently change the ID of a Learner via function setLearnerId.

2.8.2.3 IDs

The IDs of all Learners, Tasks and Measures in a benchmark experiment can be retrieved as follows:

```
getBMRTaskIds(bmr)
#> [1] "Sonar-example"

getBMRLearnerIds(bmr)
#> [1] "classif.lda" "classif.rpart"

getBMRMeasureIds(bmr)
#> [1] "mmce"
```

2.8.2.4 Learner models

Per default the BenchmarkResult also contains the fitted models for all learners on all tasks. If you do not want to keep them set models = FALSE when calling benchmark. The fitted models can be retrieved by function getBMRModels. It returns a list of lists of WrappedModel objects.

```
getBMRModels(bmr)
#> $ `Sonar-example`
#> $ `Sonar-example `$classif.lda
#> $`Sonar-example`$classif.lda[[1]]
#> Model for learner.id=classif.lda; learner.class=classif.lda
#> Trained on: task.id = Sonar-example; obs = 138; features = 60
#> Hyperparameters:
#>
#>
#> $`Sonar-example`$classif.rpart
#> $`Sonar-example`$classif.rpart[[1]]
#> Model for learner.id=classif.rpart; learner.class=classif.rpart
#> Trained on: task.id = Sonar-example; obs = 138; features = 60
#> Hyperparameters: xval=0
getBMRModels(bmr, learner.ids = "classif.lda")
#> $ `Sonar-example`
#> $`Sonar-example`$classif.lda
#> $`Sonar-example`$classif.lda[[1]]
#> Model for learner.id=classif.lda; learner.class=classif.lda
#> Trained on: task.id = Sonar-example; obs = 138; features = 60
#> Hyperparameters:
```

2.8.2.5 Learners and measures

Moreover, you can extract the employed Learners and Measures.

```
getBMRLearners(bmr)
#> $classif.lda
#> Learner classif.lda from package MASS
```

```
#> Type: classif
#> Name: Linear Discriminant Analysis; Short name: lda
#> Class: classif.lda
#> Properties: twoclass,multiclass,numerics,factors,prob
#> Predict-Type: response
#> Hyperparameters:
#>
#>
#> $classif.rpart
#> Learner classif.rpart from package rpart
#> Type: classif
#> Name: Decision Tree; Short name: rpart
#> Class: classif.rpart
#> Properties:
    twoclass, multiclass, missings, numerics, factors, ordered, prob, weights, featimp
#> Predict-Type: response
#> Hyperparameters: xval=0
getBMRMeasures(bmr)
#> [[1]]
#> Name: Mean misclassification error
#> Performance measure: mmce
#> Properties: classif, classif.multi, req. pred, req. truth
#> Minimize: TRUE
#> Best: 0; Worst: 1
#> Aggregated by: test.mean
#> Note: Defined as: mean(response != truth)
```

2.8.3 Merging benchmark results

Sometimes after completing a benchmark experiment it turns out that you want to extend it by another Learner or another Task. In this case you can perform an additional benchmark experiment and then merge the results to get a single BenchmarkResult object that can be accessed and analyzed as usual.

mlr provides two functions to merge results: mergeBenchmarkResultLearner combines two or more benchmark results for different sets of learners on the same Tasks, while mergeBenchmarkResultTask fuses results obtained with the same Learners on different sets of Tasks.

For example in the benchmark experiment above we applied lda and rpart to the sonar.task. We now perform a second experiment using a random forest and quadratic discriminant analysis (qda) and use mergeBenchmarkResultLearner to combine the results.

```
### First benchmark result
```

```
bmr
#>
           task.id
                     learner.id mmce.test.mean
\#>1 Sonar-example classif.lda 0.3000000
#> 2 Sonar-example classif.rpart
                                     0.2857143
### Benchmark experiment for the additional learners
lrns2 = list(makeLearner("classif.randomForest"),
   makeLearner("classif.qda"))
bmr2 = benchmark(lrns2, sonar.task, rdesc, show.info = FALSE)
bmr2
#>
                            learner.id mmce.test.mean
#> 1 Sonar-example classif.randomForest 0.2000000
#> 2 Sonar-example
                          classif.qda 0.5142857
### Merge the results
mergeBenchmarkResultLearner(bmr, bmr2)
                     learner.id mmce.test.mean classif.lda 0.3000000
          task.id
#> 1 Sonar-example
#> 2 Sonar-example
                       classif.rpart
                                            0.2857143
#> 3 Sonar-example classif.randomForest
                                            0.2000000
#> 4 Sonar-example
                      classif.qda 0.5142857
```

Note that in the above examples in each case a resample description was passed to the benchmark function. For this reason lda and rpart were most likely evaluated on a different training/test set pair than random forest and qda.

Differing training/test set pairs across learners pose an additional source of variation in the results, which can make it harder to detect actual performance differences between learners. Therefore, if you suspect that you will have to extend your benchmark experiment by another Learner later on it's probably easiest to work with ResampleInstances from the start. These can be stored and used for any additional experiments.

Alternatively, if you used a resample description in the first benchmark experiment you could also extract the ResampleInstances from the BenchmarkResult bmr and pass these to all further benchmark calls.

```
rin = getBMRPredictions(bmr)[[1]][[1]]$instance
rin
#> Resample instance for 208 cases.
#> Resample description: holdout with 0.67 split rate.
#> Predict: test
#> Stratification: FALSE

### Benchmark experiment for the additional random forest
bmr3 = benchmark(lrns2, sonar.task, rin, show.info = FALSE)
bmr3
```

```
task.id
                            learner.id mmce.test.mean
#> 1 Sonar-example classif.randomForest
                                            0.2714286
#> 2 Sonar-example
                           classif.qda
                                            0.3857143
### Merge the results
mergeBenchmarkResultLearner(bmr, bmr3)
           task.id
                            learner.id mmce.test.mean
#> 1 Sonar-example
                           classif.lda
                                            0.3000000
#> 2 Sonar-example
                         classif.rpart
                                            0.2857143
#> 3 Sonar-example classif.randomForest
                                            0.2714286
#> 4 Sonar-example
                      classif.qda
                                            0.3857143
```

2.8.4 Benchmark analysis and visualization

mlr offers several ways to analyze the results of a benchmark experiment. This includes visualization, ranking of learning algorithms and hypothesis tests to assess performance differences between learners.

In order to demonstrate the functionality we conduct a slightly larger benchmark experiment with three learning algorithms that are applied to five classification tasks.

2.8.4.1 Example: Comparing Ida, rpart and random Forest

We consider linear discriminant analysis (lda), classification trees (rpart), and random forests (randomForest). Since the default learner IDs are a little long, we choose shorter names in the \mathbf{R} code below.

We use five classification tasks. Three are already provided by mlr, two more data sets are taken from package mlbench and converted to Tasks by function convertMLBenchObjToTask.

For all tasks 10-fold cross-validation is chosen as resampling strategy. This is achieved by passing a single resample description to benchmark, which is then instantiated automatically once for each Task. This way, the same instance is used for all learners applied to a single task.

It is also possible to choose a different resampling strategy for each Task by passing a list of the same length as the number of tasks that can contain both resample descriptions and resample instances.

We use the mean misclassification error mmce as primary performance measure, but also calculate the balanced error rate (ber) and the training time (timetrain).

```
### Create a list of learners
lrns = list(
  makeLearner("classif.lda", id = "lda"),
```

```
makeLearner("classif.rpart", id = "rpart"),
 makeLearner("classif.randomForest", id = "randomForest")
)
### Get additional Tasks from package mlbench
ring.task = convertMLBenchObjToTask("mlbench.ringnorm", n = 600)
wave.task = convertMLBenchObjToTask("mlbench.waveform", n = 600)
tasks = list(iris.task, sonar.task, pid.task, ring.task, wave.task)
rdesc = makeResampleDesc("CV", iters = 10)
meas = list(mmce, ber, timetrain)
bmr = benchmark(lrns, tasks, rdesc, meas, show.info = FALSE)
bmr
#>
                        task.id learner.id mmce.test.mean
   ber.test.mean
                                      lda
#> 1
                   iris-example
                                                0.02000000
   0.02222222
#> 2
                                      rpart
                                                0.08000000
                   iris-example
   0.0755556
#> 3
                   iris-example randomForest
                                                0.05333333
   0.05250000
#> 4
               mlbench.ringnorm
                                        lda
                                                0.35000000
   0.34605671
               mlbench.ringnorm
                                      rpart
                                                0.17333333
   0.17313632
               mlbench.ringnorm randomForest
                                                0.05833333
   0.05806121
               mlbench.waveform
                                                0.19000000
   0.18257244
#> 8
                mlbench.waveform
                                 rpart
                                                0.28833333
   0.28765247
               mlbench.waveform randomForest
                                                0.16500000
   0.16306057
#> 10 PimaIndiansDiabetes-example
                                       lda
                                                0.22778537
   0.27148893
0.25133288
   0.28967870
#> 12 PimaIndiansDiabetes-example randomForest
                                                0.23685919
   0.27543146
#> 13
                  Sonar-example
                                                0.24619048
                                        lda
   0.23986694
                                                0.30785714
#> 14
                  Sonar-example
                                      rpart
   0.31153361
#> 15
                  Sonar-example randomForest
                                                0.17785714
   0.17442696
#> timetrain.test.mean
```

```
#> 1
                    0.0051
#> 2
                    0.0067
#> 3
                    0.0701
#> 4
                    0.0131
#> 5
                    0.0196
#> 6
                    0.4846
#> 7
                    0.0146
#> 8
                    0.0124
#> 9
                    0.4228
#> 10
                    0.0085
#> 11
                    0.0113
                    0.4978
#> 12
#> 13
                    0.0295
#> 14
                    0.0213
#> 15
                    0.3212
```

From the aggregated performance values we can see that for the iris- and PimaIndiansDiabetes-example linear discriminant analysis performs well while for all other tasks the random forest seems superior. Training takes longer for the random forest than for the other learners.

In order to draw any conclusions from the average performances at least their variability has to be taken into account or, preferably, the distribution of performance values across resampling iterations.

The individual performances on the 10 folds for every task, learner, and measure are retrieved below.

```
perf = getBMRPerformances(bmr, as.df = TRUE)
head(perf)
#>
          task.id learner.id iter
                                        mmce
                                                   ber timetrain
#> 1 iris-example
                          lda
                                1 0.0000000 0.0000000
                                                            0.006
#> 2 iris-example
                                 2 0.1333333 0.1666667
                                                            0.006
                          lda
#> 3 iris-example
                                3 0.0000000 0.0000000
                                                            0.006
                          lda
#> 4 iris-example
                          lda
                                 4 0.0000000 0.0000000
                                                            0.005
                                 5 0.0000000 0.0000000
#> 5 iris-example
                          lda
                                                            0.004
#> 6 iris-example
                          lda
                                 6 0.0000000 0.0000000
                                                            0.005
```

A closer look at the result reveals that the random forest outperforms the classification tree in every instance, while linear discriminant analysis performs better than rpart most of the time. Additionally lda sometimes even beats the random forest. With increasing size of such benchmark experiments, those tables become almost unreadable and hard to comprehend.

mlr features some plotting functions to visualize results of benchmark experiments that you might find useful. Moreover, mlr offers statistical hypothesis tests to assess performance differences between learners.

2.8.4.2 Integrated plots

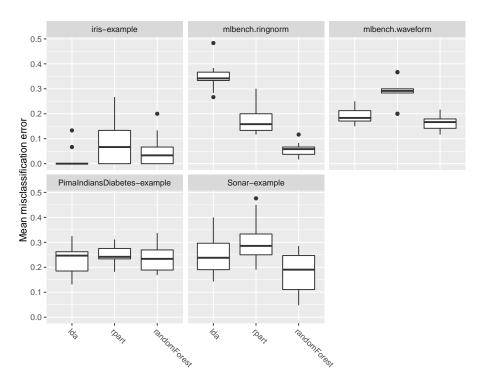
Plots are generated using ggplot2. Further customization, such as renaming plot elements or changing colors, is easily possible.

2.8.4.2.1 Visualizing performances

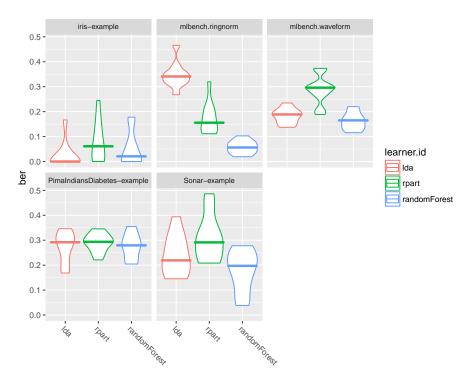
plotBMRBoxplots creates box or violin plots which show the distribution of performance values across resampling iterations for one performance measure and for all learners and tasks (and thus visualize the output of getBMRPerformances).

Below are both variants, box and violin plots. The first plot shows the mmce and the second plot the balanced error rate (ber). Moreover, in the second plot we color the boxes according to the learners to make them better distinguishable.

plotBMRBoxplots(bmr, measure = mmce)



```
plotBMRBoxplots(bmr, measure = ber, style = "violin", pretty.names =
    FALSE) +
    aes(color = learner.id) +
    theme(strip.text.x = element_text(size = 8))
```



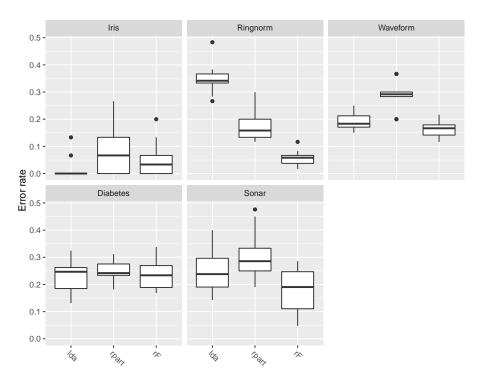
Note that by default the measure names are used as labels for the y-axis.

```
mmce$name
#> [1] "Mean misclassification error"
mmce$id
#> [1] "mmce"
```

If you prefer the shorter ids like mmce and ber set pretty.names = FALSE (as done for the second plot). Of course you can also use the ylab function to choose a completely different label.

Another thing which probably comes up quite often is changing the panel headers (which default to the Task IDs) and the learner names on the x-axis (which default to the Learner IDs). For example looking at the above plots we would like to remove the "example" suffixes and the "mlbench" prefixes from the panel headers. Moreover, compared to the other learner names "randomForest" seems a little long. Currently, the probably simplest solution is to change the factor levels of the plotted data as shown below.

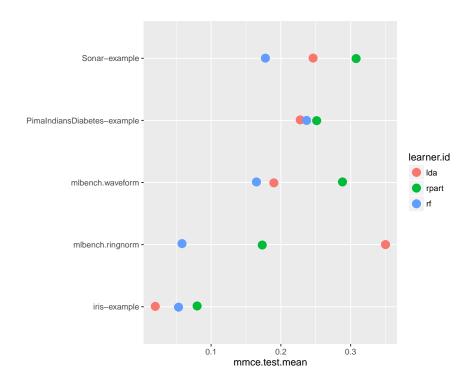
```
#> 3 iris-example
                          lda
                                 3 0.0000000 0.0000000
                                                            0.006
#> 4 iris-example
                          lda
                                 4 0.0000000 0.0000000
                                                            0.005
                                                            0.004
#> 5 iris-example
                          lda
                                 5 0.0000000 0.0000000
#> 6 iris-example
                          lda
                                 6 0.0000000 0.0000000
                                                            0.005
levels(plt$data$task.id) = c("Iris", "Ringnorm", "Waveform",
    "Diabetes", "Sonar")
levels(plt$data$learner.id) = c("lda", "rpart", "rF")
plt + ylab("Error rate")
```



2.8.4.2.2 Visualizing aggregated performances

The aggregated performance values (resulting from getBMRAggrPerformances) can be visualized by function plotBMRSummary. This plot draws one line for each task on which the aggregated values of one performance measure for all learners are displayed. By default, the first measure in the list of Measures passed to benchmark is used, in our example mmce. Moreover, a small vertical jitter is added to prevent overplotting.

```
plotBMRSummary(bmr)
```



2.8.4.2.3 Calculating and visualizing ranks

Additional to the absolute performance, relative performance, i.e., ranking the learners is usually of interest and might provide valuable additional insight.

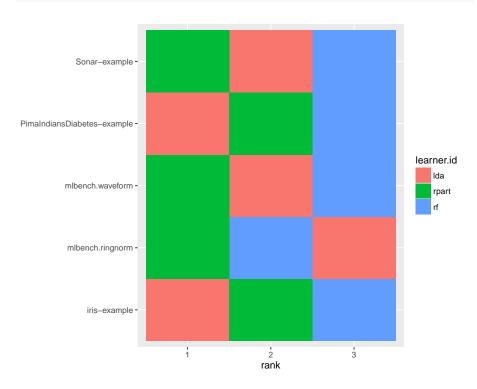
Function convertBMRToRankMatrix calculates ranks based on aggregated learner performances of one measure. We choose the mean misclassification error (mmce). The rank structure can be visualized by plotBMRRanksAsBarChart.

```
m = convertBMRToRankMatrix(bmr, mmce)
m
#>
                 iris-example mlbench.ringnorm mlbench.waveform
#>
  lda
                             1
                                               3
                                                                 2
                             3
                                               2
                                                                 3
#> rpart
                             2
   randomForest
                                               1
                 PimaIndiansDiabetes-example \ Sonar-example
#>
                                             1
#>
   lda
                                             3
                                                            3
#> rpart
                                             2
#> randomForest
                                                            1
```

Methods with best performance, i.e., with lowest mmce, are assigned the lowest rank. Linear discriminant analysis is best for the iris and PimaIndiansDiabetes-examples while the random forest shows best results on the remaining tasks.

plotBMRRanksAsBarChart with option **pos = "tile"** shows a corresponding heat map. The ranks are displayed on the x-axis and the learners are color-coded.





A similar plot can also be obtained via plotBMRSummary. With option trafo = "rank" the ranks are displayed instead of the aggregated performances.

plotBMRSummary(bmr, trafo = "rank", jitter = 0)



Alternatively, you can draw stacked bar charts (the default) or bar charts with juxtaposed bars (pos = "dodge") that are better suited to compare the frequencies of learners within and across ranks.

```
plotBMRRanksAsBarChart(bmr)
plotBMRRanksAsBarChart(bmr, pos = "dodge")
```

2.8.4.3 Comparing learners using hypothesis tests

Many researchers feel the need to display an algorithm's superiority by employing some sort of hypothesis testing. As non-parametric tests seem better suited for such benchmark results the tests provided in mlr are the **Overall Friedmantest** and the **Friedman-Nemenyi post hoc test**.

While the ad hoc Friedman test based on friedman.test from the stats package is testing the hypothesis whether there is a significant difference between the employed learners, the post hoc Friedman-Nemenyi test tests for significant differences between all pairs of learners. Non parametric tests often do have less power then their parametric counterparts but less assumptions about underlying distributions have to be made. This often means many data sets are needed in order to be able to show significant differences at reasonable significance levels.

In our example, we want to compare the three learners on the selected data sets. First we might we want to test the hypothesis whether there is a difference between the learners.

```
friedmanTestBMR(bmr)
#>
#> Friedman rank sum test
#>
#> data: mmce.test.mean and learner.id and task.id
#> Friedman chi-squared = 5.2, df = 2, p-value = 0.07427
```

In order to keep the computation time for this tutorial small, the Learners are only evaluated on five tasks. This also means that we operate on a relatively low significance level $\alpha=0.1$. As we can reject the null hypothesis of the Friedman test at a reasonable significance level we might now want to test where these differences lie exactly.

```
friedmanPostHocTestBMR(bmr, p.value = 0.1)
#>
   Pairwise comparisons using Nemenyi multiple comparison test
#>
#>
                with q approximation for unreplicated blocked data
#>
#> data: mmce.test.mean and learner.id and task.id
#>
#>
                lda
                     rpart
#> rpart
                0.254 -
#> randomForest 0.802 0.069
#>
#> P value adjustment method: none
```

At this level of significance, we can reject the null hypothesis that there exists no performance difference between the decision tree (rpart) and the random Forest.

2.8.4.4 Critical differences diagram

In order to visualize differently performing learners, a critical differences diagram can be plotted, using either the Nemenyi test (test = "nemenyi") or the Bonferroni-Dunn test (test = "bd").

The mean rank of learners is displayed on the x-axis.

• Choosing test = "nemenyi" compares all pairs of Learners to each other, thus the output are groups of not significantly different learners. The diagram connects all groups of learners where the mean ranks do not differ by more than the critical differences. Learners that are not connected by a bar are significantly different, and the learner(s) with the lower mean rank can be considered "better" at the chosen significance level.

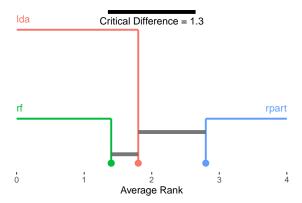
• Choosing test = "bd" performs a pairwise comparison with a baseline. An interval which extends by the given critical difference in both directions is drawn around the Learner chosen as baseline, though only comparisons with the baseline are possible. All learners within the interval are not significantly different, while the baseline can be considered better or worse than a given learner which is outside of the interval.

The critical difference CD is calculated by

$$CD = q_{\alpha} \cdot \sqrt{\frac{k(k+1)}{6N}},$$

where N denotes the number of tasks, k is the number of learners, and q_{α} comes from the studentized range statistic divided by $\sqrt{2}$. For details see Demsar (2006).

Function generateCritDifferencesData does all necessary calculations while function plotCritDifferences draws the plot. See the tutorial page about visualization for details on data generation and plotting functions.



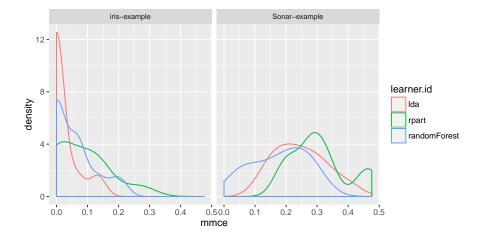
```
### Bonferroni-Dunn test
g = generateCritDifferencesData(bmr, p.value = 0.1, test = "bd",
    baseline = "randomForest")
plotCritDifferences(g) + coord_cartesian(xlim = c(-1,5), ylim =
    c(0,2))
```



2.8.4.5 Custom plots

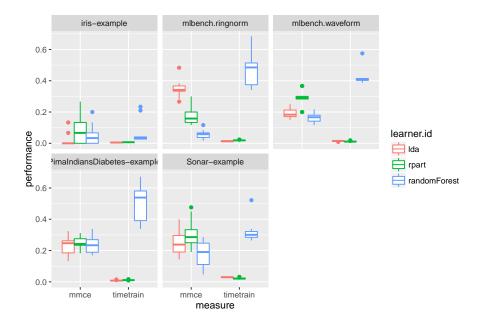
You can easily generate your own visualizations by customizing the ggplot objects returned by the plots above, retrieve the data from the ggplot objects and use them as basis for your own plots, or rely on the data frames returned by getBMRPerformances or getBMRAggrPerformances. Here are some examples.

Instead of boxplots (as in plotBMRBoxplots) we could create density plots to show the performance values resulting from individual resampling iterations.



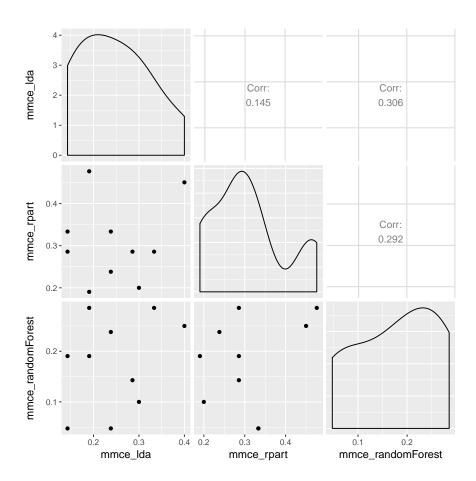
In order to plot multiple performance measures in parallel, perf is reshaped to long format. Below we generate grouped boxplots showing the error rate (mmce) and the training time timetrain.

```
### Compare mmce and timetrain
df = reshape2::melt(perf, id.vars = c("task.id", "learner.id",
    "iter"))
df = df[df$variable != "ber",]
head(df)
#>
          task.id learner.id iter variable
                                                value
#> 1 iris-example
                         lda
                                1
                                      mmce 0.0000000
#> 2 iris-example
                         lda
                                2
                                      mmce 0.1333333
                         lda
#> 3 iris-example
                                3
                                      mmce 0.0000000
#> 4 iris-example
                         lda
                                4
                                      mmce 0.0000000
#> 5 iris-example
                         lda
                                5
                                      mmce 0.0000000
#> 6 iris-example
                         lda
                                6
                                      mmce 0.0000000
qplot(variable, value, data = df, colour = learner.id, geom =
    "boxplot",
  xlab = "measure", ylab = "performance") +
  facet_wrap(~ task.id, nrow = 2)
```



It might also be useful to assess if learner performances in single resampling iterations, i.e., in one fold, are related. This might help to gain further insight, for example by having a closer look at train and test sets from iterations where one learner performs exceptionally well while another one is fairly bad. Moreover, this might be useful for the construction of ensembles of learning algorithms. Below, function ggpairs from package GGally is used to generate a scatterplot matrix of mean misclassification errors (mmce) on the Sonar data set.

```
perf = getBMRPerformances(bmr, task.id = "Sonar-example", as.df =
    TRUE)
df = reshape2::melt(perf, id.vars = c("task.id", "learner.id",
    "iter"))
df = df[df$variable == "mmce",]
df = reshape2::dcast(df, task.id + iter ~ variable + learner.id)
head(df)
#>
           task.id iter mmce_lda mmce_rpart mmce_randomForest
#> 1 Sonar-example
                      1 0.2857143
                                   0.2857143
                                                     0.14285714
                                                     0.23809524
#> 2 Sonar-example
                      2 0.2380952
                                   0.2380952
                                                     0.28571429
#> 3 Sonar-example
                      3 0.3333333
                                   0.2857143
                      4 0.2380952
#> 4 Sonar-example
                                   0.3333333
                                                     0.04761905
#> 5 Sonar-example
                      5 0.1428571 0.2857143
                                                     0.19047619
#> 6 Sonar-example
                      6 0.4000000 0.4500000
                                                     0.25000000
GGally::ggpairs(df, 3:5)
```



2.8.5 Further comments

- Note that for supervised classification mlr offers some more plots that operate on BenchmarkResult objects and allow you to compare the performance of learning algorithms. See for example the tutorial page on ROC curves and functions generateThreshVsPerfData, plotROCCurves, and plotViperCharts as well as the page about classifier calibration and function generateCalibrationData.
- In the examples shown in this section we applied "raw" learning algorithms, but often things are more complicated. At the very least, many learners have hyperparameters that need to be tuned to get sensible results. Reliable performance estimates can be obtained by nested resampling, i.e., by doing the tuning in an inner resampling loop while estimating the performance in an outer loop. Moreover, you might want to combine learners with pre-processing steps like imputation, scaling, outlier removal, dimen-

2.9 Parallelization 2 BASICS

sionality reduction or feature selection and so on. All this can be easily done using mlr's wrapper functionality. The general principle is explained in the section about wrapped learners in the Advanced part of this tutorial. There are also several sections devoted to common pre-processing steps.

• Benchmark experiments can very quickly become computationally demanding. mlr offers some possibilities for parallelization.

2.9 Parallelization

R by default does not make use of parallelization. With the integration of parallelMap into mlr, it becomes easy to activate the parallel computing capabilities already supported by mlr. parallelMap supports all major parallelization backends: local multicore execution using parallel, socket and MPI clusters using snow, makeshift SSH-clusters using BatchJobs and high performance computing clusters (managed by a scheduler like SLURM, Torque/PBS, SGE or LSF) also using BatchJobs.

All you have to do is select a backend by calling one of the parallelStart* functions. The first loop mlr encounters which is marked as parallel executable will be automatically parallelized. It is good practice to call parallelStop at the end of your script.

```
library("parallelMap")
parallelStartSocket(2)
#> Starting parallelization in mode=socket with cpus=2.

rdesc = makeResampleDesc("CV", iters = 3)
r = resample("classif.lda", iris.task, rdesc)
#> Exporting objects to slaves for mode socket: .mlr.slave.options
#> Mapping in parallel: mode = socket; level = mlr.resample; cpus = 2; elements = 3.
#> [Resample] Result: mmce.test.mean=0.02

parallelStop()
#> Stopped parallelization. All cleaned up.
```

On Linux or Mac OS X, you may want to use parallelStartMulticore instead.

2.9.1 Parallelization levels

We offer different parallelization levels for fine grained control over the parallelization. E.g., if you do not want to parallelize the benchmark function because it has only very few iterations but want to parallelize the resampling of each learner instead, you can specifically pass the level "mlr.resample" to the parallelStart* function. Currently the following levels are supported:

```
parallelGetRegisteredLevels()
#> mlr: mlr.benchmark, mlr.resample, mlr.selectFeatures,
    mlr.tuneParams
```

Here is a brief explanation of what these levels do:

- "mlr.resample": Each resampling iteration (a train / test step) is a parallel job.
- "mlr.benchmark": Each experiment "run this learner on this data set" is a parallel job.
- "mlr.tuneParams": Each evaluation in hyperparameter space "resample with these parameter settings" is a parallel job. How many of these can be run independently in parallel, depends on the tuning algorithm. For grid search or random search this is no problem, but for other tuners it depends on how many points are produced in each iteration of the optimization. If a tuner works in a purely sequential fashion, we cannot work magic and the hyperparameter evaluation will also run sequentially. But note that you can still parallelize the underlying resampling.
- "mlr.selectFeatures": Each evaluation in feature space "resample with this feature subset" is a parallel job. The same comments as for "mlr.tuneParams" apply here.

2.9.2 Custom learners and parallelization

If you have implemented a custom learner yourself, locally, you currently need to export this to the slave. So if you see an error after calling, e.g., a parallelized version of resample like this:

```
no applicable method for 'trainLearner' applied to an object of class <my_new_learner>
```

simply add the following line somewhere after calling parallelStart.

2.9.3 The end

For further details, consult the parallelMap tutorial and help.

2.10 Visualization

2.10.1 Generation and plotting functions

mlr's visualization capabilities rely on *generation functions* which generate data for plots, and *plotting functions* which plot this output using either ggplot2 or

ggvis (the latter being currently experimental).

This separation allows users to easily make custom visualizations by taking advantage of the generation functions. The only data transformation that is handled inside plotting functions is reshaping. The reshaped data is also accessible by calling the plotting functions and then extracting the data from the ggplot object.

The functions are named accordingly.

- Names of generation functions start with generate and are followed by a title-case description of their FunctionPurpose, followed by Data, i.e., generateFunctionPurposeData. These functions output objects of class FunctionPurposeData.
- Plotting functions are prefixed by plot followed by their purpose, i.e., plotFunctionPurpose.
- ggvis plotting functions have an additional suffix GGVIS, i.e., plotFunctionPurposeGGVIS.

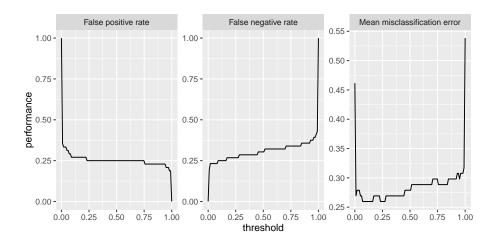
2.10.1.1 Some examples

In the example below we create a plot of classifier performance as function of the decision threshold for the binary classification problem sonar.task. The generation function generateThreshVsPerfData creates an object of class ThreshVsPerfData which contains the data for the plot in slot \$data.

```
lrn = makeLearner("classif.lda", predict.type = "prob")
n = getTaskSize(sonar.task)
mod = train(lrn, task = sonar.task, subset = seq(1, n, by = 2))
pred = predict(mod, task = sonar.task, subset = seq(2, n, by = 2))
d = generateThreshVsPerfData(pred, measures = list(fpr, fnr, mmce))
class(d)
#> [1] "ThreshVsPerfData"
head(d$data)
#>
           fpr
                     fnr
                             mmce threshold
#> 1 1.0000000 0.0000000 0.4615385 0.00000000
#> 2 0.3541667 0.1964286 0.2692308 0.01010101
#> 3 0.3333333 0.2321429 0.2788462 0.02020202
#> 4 0.3333333 0.2321429 0.2788462 0.03030303
#> 5 0.3333333 0.2321429 0.2788462 0.04040404
#> 6 0.3125000 0.2321429 0.2692308 0.05050505
```

For plotting we can use the built-in mlr function plotThreshVsPerf.

```
plotThreshVsPerf(d)
```



Note that by default the Measure names are used to annotate the panels.

```
fpr$name
#> [1] "False positive rate"

fpr$id
#> [1] "fpr"
```

This does not only apply to plotThreshVsPerf, but to other plot functions that show performance measures as well, for example plotLearningCurve. You can use the ids instead of the names by setting pretty.names = FALSE.

2.10.1.2 Customizing plots

As mentioned above it is easily possible to customize the built-in plots or making your own visualizations from scratch based on the generated data.

What will probably come up most often is changing labels and annotations. Generally, this can be done by manipulating the ggplot object, in this example the object returned by plotThreshVsPerf, using the usual ggplot2 functions like ylab or labeller. Moreover, you can change the underlying data, either d\$data (resulting from generateThreshVsPerfData) or the possibly reshaped data contained in the ggplot object (resulting from plotThreshVsPerf), most often by renaming columns or factor levels.

Below are two examples of how to alter the axis and panel labels of the above plot.

Imagine you want to change the order of the panels and also are not satisfied with the panel names, for example you find that "Mean misclassification error" is too long and you prefer "Error rate" instead. Moreover, you want the error rate to be displayed first.

```
plt = plotThreshVsPerf(d, pretty.names = FALSE)
### Reshaped version of the underlying data d
head(plt$data)
      threshold measure performance
#> 1 0.00000000
                          1.0000000
                    fpr
#> 2 0.01010101
                    fpr
                          0.3541667
#> 3 0.02020202
                          0.3333333
                    fpr
#> 4 0.03030303
                          0.3333333
                    fpr
#> 5 0.04040404
                          0.3333333
                    fpr
#> 6 0.05050505
                          0.3125000
                    fpr
levels(plt$data$measure)
#> [1] "fpr" "fnr" "mmce"
### Rename and reorder factor levels
plt$data$measure = factor(plt$data$measure, levels = c("mmce",
    "fpr", "fnr"),
  labels = c("Error rate", "False positive rate", "False negative
      rate"))
plt = plt + xlab("Cutoff") + ylab("Performance")
plt
```

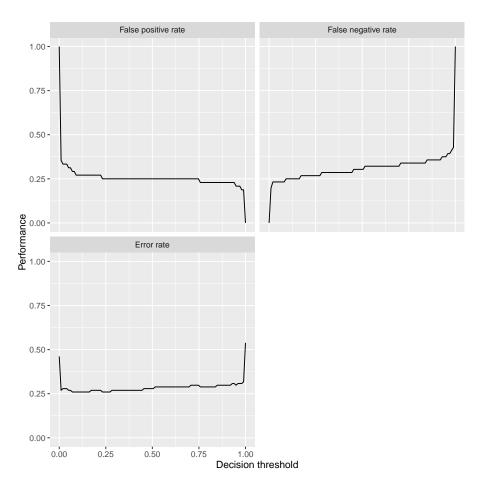


Using the labeller function requires calling facet_wrap (or facet_grid), which can be useful if you want to change how the panels are positioned (number of rows and columns) or influence the axis limits.

```
plt = plotThreshVsPerf(d, pretty.names = FALSE)

measure_names = c(
   fpr = "False positive rate",
```

```
fnr = "False negative rate",
  mmce = "Error rate"
)
### Manipulate the measure names via the labeller function and
### arrange the panels in two columns and choose common axis limits
  for all panels
plt = plt + facet_wrap( ~ measure, labeller = labeller(measure =
      measure_names), ncol = 2)
plt = plt + xlab("Decision threshold") + ylab("Performance")
plt
```



Instead of using the built-in function plotThreshVsPerf we could also manually create the plot based on the output of generateThreshVsPerfData: in this case to plot only one measure.

```
ggplot(d$data, aes(threshold, fpr)) + geom_line()
```



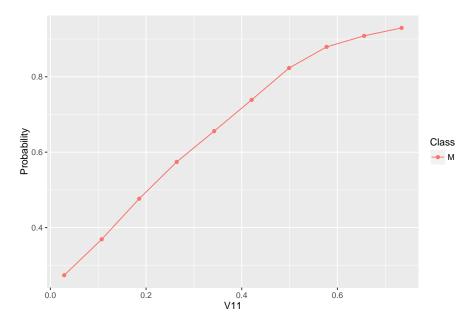
The decoupling of generation and plotting functions is especially practical if you prefer traditional graphics or lattice. Here is a lattice plot which gives a result similar to that of plotThreshVsPerf.

```
lattice::xyplot(fpr + fnr + mmce ~ threshold, data = d$data, type =
    "l", ylab = "performance",
    outer = TRUE, scales = list(relation = "free"),
    strip = strip.custom(factor.levels = sapply(d$measures,
        function(x) x$name),
    par.strip.text = list(cex = 0.8)))
```

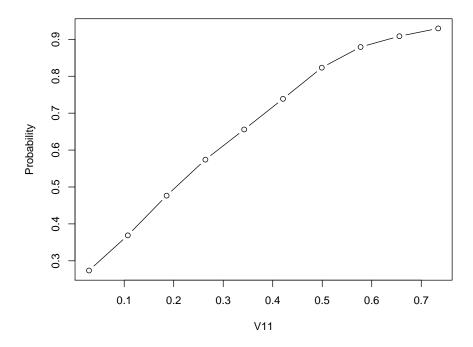


Let's conclude with a brief look on a second example. Here we use plotPartialDependence but extract the data from the ggplot object pltand use it to create a traditional graphics::plot, additional to the ggplot2 plot.

```
sonar = getTaskData(sonar.task)
pd = generatePartialDependenceData(mod, sonar, "V11")
plt = plotPartialDependence(pd)
head(plt$data)
     Class Probability Feature
                                   Value
#> 1
        M 0.2737158
                           V11 0.0289000
#> 2
        Μ
            0.3689970
                           V11 0.1072667
#> 3
        Μ
            0.4765742
                           V11 0.1856333
            0.5741233
                           V11 0.2640000
#> 4
        Μ
        Μ
             0.6557857
                           V11 0.3423667
#> 5
#> 6
             0.7387962
                           V11 0.4207333
plt
```



```
plot(Probability ~ Value, data = plt$data, type = "b", xlab =
    plt$data$Feature[1])
```



2.10.2 Available generation and plotting functions

Below the currently available generation and plotting functions are listed and tutorial pages that provide in depth descriptions of the listed functions are referenced.

Note that some plots, e.g., plotTuneMultiCritResult are not mentioned here since they lack a generation function. Both plotThreshVsPerf and plotROC-Curves operate on the result of generateThreshVsPerfData. Functions plotPartialDependence and plotPartialDependenceGGVIS can be applied to the results of both generatePartialDependenceData and generateFunctionalANOVAData.

The ggvis functions are experimental and are subject to change, though they should work. Most generate interactive shiny applications, that automatically start and run locally.

Performance evaluation and ROC analysis:

- generateThreshVsPerfData
- plotThresVsPerf, plotThreshVsPerfGGVIS
- plotROCCurves

Benchmark experiments:

- generateCritDifferencesData
- plotCritDifferences

Tuning and hyperparameter tuning effects:

- $\bullet \hspace{0.1in} generate Hyper Pars Effect Data$
- plotHyperParsEffect

Feature selection:

- generateFilterValuesData
- plotFilterValues
- plotFilterValuesGGVIS

Learning curves:

- generatePartialDependenceData, generateFunctionalANOVAData
- plotPartialDependence
- $\bullet \quad plot Partial Dependence GGV IS$

Classifier calibration:

- generateCalibrationData
- plotCalibration

3 Advanced

3.1 Configuring mlr

mlr is designed to make usage errors due to typos or invalid parameter values as unlikely as possible. Occasionally, you might want to break those barriers and get full access, for example to reduce the amount of output on the console or to turn off checks. For all available options simply refer to the documentation of configureMlr. In the following we show some common use cases.

Generally, function configure Mlr permits to set options globally for your current ${\bf R}$ session.

It is also possible to set options locally.

- All options referring to the behavior of learners (these are all options except show.info) can be set for an individual learner via the config argument of makeLearner. The local precedes the global configuration.
- Some functions like resample, benchmark, selectFeatures, tuneParams, and tuneParamsMultiCrit have a show.info flag that controls if progress messages are shown. The default value of show.info can be set by configureMlr.

3.1.1 Example: Reducing the output on the console

You are bothered by all the output on the console like in this example?

```
rdesc = makeResampleDesc("Holdout")
r = resample("classif.multinom", iris.task, rdesc)
#> [Resample] holdout iter: 1
#> # weights: 18 (10 variable)
#> initial value 109.861229
#> iter 10 value 12.256619
#> iter 20 value 3.638740
#> iter 30 value 3.228628
#> iter 40 value 2.951100
#> iter 50 value 2.806521
#> iter 60 value 2.739076
#> iter 70 value 2.522206
#> iter 80 value 2.485225
#> iter 90 value 2.381397
#> iter 100 value 2.360602
#> final value 2.360602
#> stopped after 100 iterations
#> [Resample] Result: mmce.test.mean=0.02
```

You can suppress the output for this Learner and this resample call as follows:

```
lrn = makeLearner("classif.multinom", config =
    list(show.learner.output = FALSE))
r = resample(lrn, iris.task, rdesc, show.info = FALSE)
```

(Note that multinom has a **trace** switch that can alternatively be used to turn off the progress messages.)

To globally suppress the output for all subsequent learners and calls to resample, benchmark etc. do the following:

```
configureMlr(show.learner.output = FALSE, show.info = FALSE)
r = resample("classif.multinom", iris.task, rdesc)
```

3.1.2 Accessing and resetting the configuration

Function getMlrOptions returns a list with the current configuration.

```
getMlrOptions()
#> $on.learner.error
#> [1] "stop"
#> $on.learner.warning
#> [1] "warn"
#> $on.par.out.of.bounds
#> [1] "stop"
#>
#> $on.par.without.desc
#> [1] "stop"
#>
#> $show.info
#> [1] FALSE
#>
#> $show.learner.output
#> [1] FALSE
```

To restore the default configuration call configureMlr with an empty argument list.

configureMlr()

```
getMlrOptions()
#> $on.learner.error
#> [1] "stop"
#>
#> $on.learner.warning
#> [1] "warn"
#>
#> $on.par.out.of.bounds
#> [1] "stop"
#> $on.par.without.desc
#> [1] "stop"
#>
#> $show.info
#> [1] TRUE
#>
#> $show.learner.output
#> [1] TRUE
```

3.1.3 Example: Turning off parameter checking

It might happen that you want to set a parameter of a Learner, but the parameter is not registered in the learner's parameter set yet. In this case you might want to contact us or open an issue as well! But until the problem is fixed you can turn off mlr's parameter checking. The parameter setting will then be passed to the underlying function without further ado.

```
### Support Vector Machine with linear kernel and new parameter
    'newParam'
lrn = makeLearner("classif.ksvm", kernel = "vanilladot", newParam =
    3)
#> Error in setHyperPars2.Learner(learner, insert(par.vals, args)):
    classif.ksvm: Setting parameter newParam without available
    description object!
#> Did you mean one of these hyperparameters instead: degree scaled
#> You can switch off this check by using configureMlr!
### Turn off parameter checking completely
configureMlr(on.par.without.desc = "quiet")
lrn = makeLearner("classif.ksvm", kernel = "vanilladot", newParam =
train(lrn, iris.task)
#> Setting default kernel parameters
#> Model for learner.id=classif.ksvm; learner.class=classif.ksvm
#> Trained on: task.id = iris-example; obs = 150; features = 4
#> Hyperparameters: fit=FALSE,kernel=vanilladot,newParam=3
### Option "quiet" also masks typos
lrn = makeLearner("classif.ksvm", kernl = "vanilladot")
train(lrn, iris.task)
#> Model for learner.id=classif.ksvm; learner.class=classif.ksvm
#> Trained on: task.id = iris-example; obs = 150; features = 4
#> Hyperparameters: fit=FALSE, kernl=vanilladot
### Alternatively turn off parameter checking, but still see warnings
configureMlr(on.par.without.desc = "warn")
lrn = makeLearner("classif.ksvm", kernl = "vanilladot", newParam = 3)
#> Warning in setHyperPars2.Learner(learner, insert(par.vals,
    args)): classif.ksvm: Setting parameter kernl without available
    description object!
#> Did you mean one of these hyperparameters instead: kernel nu
#> You can switch off this check by using configureMlr!
```

3.1.4 Example: Handling errors in a learning method

If a learning method throws an error the default behavior of mlr is to generate an exception as well. However, in some situations, for example if you conduct a larger benchmark study with multiple data sets and learners, you usually don't want the whole experiment stopped due to one error. You can prevent this using the on.learner.error option of configureMlr.

```
### This call gives an error caused by the low number of
    observations in class "virginica"
train("classif.qda", task = iris.task, subset = 1:104)
#> Error in qda.default(x, grouping, ...): some group is too small
    for 'qda'
#> Timing stopped at: 0.002 0 0.002
### Get a warning instead of an error
configureMlr(on.learner.error = "warn")
mod = train("classif.qda", task = iris.task, subset = 1:104)
#> Warning in train("classif.qda", task = iris.task, subset =
    1:104): Could not train learner classif.qda: Error in
    qda.default(x, qroupinq, ...):
   some group is too small for 'qda'
mod
#> Model for learner.id=classif.qda; learner.class=classif.qda
#> Trained on: task.id = iris-example; obs = 104; features = 4
#> Hyperparameters:
#> Training failed: Error in qda.default(x, grouping, ...) :
   some group is too small for 'qda'
#>
#> Training failed: Error in qda.default(x, grouping, ...) :
#> some group is too small for 'qda'
```

```
### mod is an object of class FailureModel
isFailureModel(mod)
#> [1] TRUE
### Retrieve the error message
getFailureModelMsg(mod)
\# [1] "Error in qda.default(x, grouping, ...) : \n some group is
    too small for 'qda'\n"
### predict and performance return NA's
pred = predict(mod, iris.task)
pred
#> Prediction: 150 observations
#> predict.type: response
#> threshold:
#> time: NA
   id truth response
#> 1 1 setosa <NA>
#> 2 2 setosa
                  <NA>
#> 3 3 setosa <NA>
#> 4 4 setosa <NA>
#> 5 5 setosa
                  <NA>
#> 6 6 setosa
                  <NA>
#> ... (150 rows, 3 cols)
performance(pred)
#> mmce
  NA
```

If on.learner.error = "warn" a warning is issued instead of an exception and an object of class FailureModel is created. You can extract the error message using function getFailureModelMsg. All further steps like prediction and performance calculation work and return NA's.

3.2 Wrapper

Wrappers can be employed to extend integrated learners with new functionality. The broad scope of operations and methods which are implemented as wrappers underline the flexibility of the wrapping approach:

- Data preprocessing
- Imputation
- Bagging
- Tuning
- Feature selection

- Cost-sensitive classification
- Over- and undersampling for imbalanced classification problems
- Multiclass extension for binary-class learners
- Multilabel classification

All these operations and methods have a few things in common: First, they all wrap around mlr learners and they return a new learner. Therefore learners can be wrapped multiple times. Second, they are implemented using a *train* (pre-model hook) and *predict* (post-model hook) method.

3.2.1 Example: Bagging wrapper

In this section we exemplary describe the bagging wrapper to create a random forest which supports weights. To achieve that we combine several decision trees from the rpart package to create our own custom random forest.

First, we create a weighted toy task.

```
data(iris)
task = makeClassifTask(data = iris, target = "Species", weights =
    as.integer(iris$Species))
```

Next, we use makeBaggingWrapper to create the base learners and the bagged learner. We choose to set equivalents of ntree (100 base learners) and mtry (proportion of randomly selected features).

As we can see in the output, the wrapped learner inherited all properties from the base learner, especially the "weights" attribute is still present. We can use this newly constructed learner like all base learners, i.e. we can use it in train, benchmark, resample, etc.

```
benchmark(tasks = task, learners = list(base.lrn, wrapped.lrn))
#> Task: iris, Learner: classif.rpart
#> [Resample] cross-validation iter: 1
```

```
#> [Resample] cross-validation iter: 2
#> [Resample] cross-validation iter: 3
#> [Resample] cross-validation iter: 4
#> [Resample] cross-validation iter: 5
#> [Resample] cross-validation iter: 6
#> [Resample] cross-validation iter: 7
#> [Resample] cross-validation iter: 8
#> [Resample] cross-validation iter: 9
#> [Resample] cross-validation iter: 10
#> [Resample] Result: mmce.test.mean=0.0667
#> Task: iris, Learner: classif.rpart.bagged
#> [Resample] cross-validation iter: 1
#> [Resample] cross-validation iter: 2
#> [Resample] cross-validation iter: 3
#> [Resample] cross-validation iter: 4
#> [Resample] cross-validation iter: 5
#> [Resample] cross-validation iter: 6
#> [Resample] cross-validation iter: 7
#> [Resample] cross-validation iter: 8
#> [Resample] cross-validation iter: 9
#> [Resample] cross-validation iter: 10
#> [Resample] Result: mmce.test.mean=0.06
    task.id
                      learner.id mmce.test.mean
#> 1
     iris
                   classif.rpart 0.0666667
#> 2 iris classif.rpart.bagged
                                     0.06000000
```

That far we are quite happy with our new learner. But we hope for a better performance by tuning some hyperparameters of both the decision trees and bagging wrapper. Let's have a look at the available hyperparameters of the fused learner:

```
getParamSet(wrapped.lrn)
#>
                 Type len Def Constr Req Tunable Trafo
#> bw.iters
               integer - 10 1 to Inf -
                                          TRUE
             logical - TRUE -
                                          TRUE
#> bw.replace
#> bw.size
              numeric - - 0 to 1 -
                                          TRUE
#> bw.feats
              numeric - 0.667 0 to 1 -
                                          TRUE
               integer - 20 1 to Inf -
#> minsplit
                                          TRUE
#> minbucket
             integer - - 1 to Inf -
                                         TRUE
#> cp
               numeric - 0.01 0 to 1 -
                                         TRUE
#> maxcompete integer
                      - 4 0 to Inf -
                                          TRUE
#> maxsurrogate integer
                           5 0 to Inf
                                          TRUE
#> usesurrogate discrete -
                          2
                                0,1,2 -
                                          TRUE
#> surrogatestyle discrete - 0
                                0,1 -
                                          TRUE
          integer - 30 1 to 30
#> maxdepth
                                          TRUE
               integer - 10 0 to Inf -
#> xval
                                         FALSE
```

```
#> parms untyped - - - TRUE -
```

We choose to tune the parameters minsplit and bw.feats for the mmce using a random search in a 3-fold CV:

```
ctrl = makeTuneControlRandom(maxit = 10)
rdesc = makeResampleDesc("CV", iters = 3)
par.set = makeParamSet(
  makeIntegerParam("minsplit", lower = 1, upper = 10),
  makeNumericParam("bw.feats", lower = 0.25, upper = 1)
)
tuned.lrn = makeTuneWrapper(wrapped.lrn, rdesc, mmce, par.set, ctrl)
print(tuned.lrn)
#> Learner classif.rpart.bagged.tuned from package rpart
#> Type: classif
#> Name: ; Short name:
#> Class: TuneWrapper
#> Properties:
    numerics, factors, ordered, missings, weights, prob, two class, multiclass, featimp
#> Predict-Type: response
#> Hyperparameters: xval=0,bw.iters=100,bw.feats=0.5
```

Calling the train method of the newly constructed learner performs the following steps:

- 1. The tuning wrapper sets parameters for the underlying model in slot \$next.learner and calls its train method.
- 2. Next learner is the bagging wrapper. The passed down argument bw.feats is used in the bagging wrapper training function, the argument minsplit gets passed down to \$next.learner. The base wrapper function calls the base learner bw.iters times and stores the resulting models.
- 3. The bagged models are evaluated using the mean mmce (default aggregation for this performance measure) and new parameters are selected using the tuning method.
- 4. This is repeated until the tuner terminates. Output is a tuned bagged learner.

```
#> [Tune-x] 1: minsplit=5; bw.feats=0.935
#> [Tune-y] 1: mmce.test.mean=0.0467; time: 0.1 min; memory: 161Mb
    use, 724Mb max
#> [Tune-x] 2: minsplit=9; bw.feats=0.675
#> [Tune-y] 2: mmce.test.mean=0.0467; time: 0.1 min; memory: 161Mb
    use, 724Mb max
\# [Tune-x] 3: minsplit=2; bw.feats=0.847
#> [Tune-y] 3: mmce.test.mean=0.0467; time: 0.1 min; memory: 161Mb
    use, 724Mb max
#> [Tune-x] 4: minsplit=4; bw.feats=0.761
#> [Tune-y] 4: mmce.test.mean=0.0467; time: 0.1 min; memory: 161Mb
    use, 724Mb max
\# [Tune-x] 5: minsplit=6; bw.feats=0.338
#> [Tune-y] 5: mmce.test.mean=0.0867; time: 0.1 min; memory: 161Mb
    use, 724Mb max
\# [Tune-x] 6: minsplit=1; bw.feats=0.637
#> [Tune-y] 6: mmce.test.mean=0.0467; time: 0.1 min; memory: 161Mb
    use, 724Mb max
#> [Tune-x] 7: minsplit=1; bw.feats=0.998
#> [Tune-y] 7: mmce.test.mean=0.0467; time: 0.1 min; memory: 161Mb
    use, 724Mb max
#> [Tune-x] 8: minsplit=4; bw.feats=0.698
#> [Tune-y] 8: mmce.test.mean=0.0467; time: 0.1 min; memory: 161Mb
    use, 724Mb max
#> [Tune-x] 9: minsplit=3; bw.feats=0.836
#> [Tune-y] 9: mmce.test.mean=0.0467; time: 0.1 min; memory: 161Mb
    use, 724Mb max
#> [Tune-x] 10: minsplit=10; bw.feats=0.529
#> [Tune-y] 10: mmce.test.mean=0.0533; time: 0.1 min; memory: 161Mb
    use, 724Mb max
#> [Tune] Result: minsplit=1; bw.feats=0.998 : mmce.test.mean=0.0467
print(lrn)
#> Model for learner.id=classif.rpart.bagged.tuned;
    learner.class=TuneWrapper
#> Trained on: task.id = iris; obs = 150; features = 4
#> Hyperparameters: xval=0,bw.iters=100,bw.feats=0.5
```

3.3 Data Preprocessing

Data preprocessing refers to any transformation of the data done before applying a learning algorithm. This comprises for example finding and resolving inconsistencies, imputation of missing values, identifying, removing or replacing outliers, discretizing numerical data or generating numerical dummy variables for categorical data, any kind of transformation like standardization of predictors or Box-Cox, dimensionality reduction and feature extraction and/or selection.

mlr offers several options for data preprocessing. Some of the following simple methods to change a Task (or data.frame) were already mentioned on the page about learning tasks:

- capLargeValues: Convert large/infinite numeric values.
- createDummyFeatures: Generate dummy variables for factor features.
- dropFeatures: Remove selected features.
- joinClassLevels: Only for classification: Merge existing classes to new, larger classes.
- mergeSmallFactorLevels: Merge infrequent levels of factor features.
- normalizeFeatures: Normalize features by different methods, e.g., standardization or scaling to a certain range.
- removeConstantFeatures: Remove constant features.
- subsetTask: Remove observations and/or features from a Task.

Moreover, there are tutorial pages devoted to

- Feature selection and
- Imputation of missing values.

3.3.1 Fusing learners with preprocessing

mlr's wrapper functionality permits to combine learners with preprocessing steps. This means that the preprocessing "belongs" to the learner and is done any time the learner is trained or predictions are made.

This is, on the one hand, very practical. You don't need to change any data or learning Tasks and it's quite easy to combine different learners with different preprocessing steps.

On the other hand this helps to avoid a common mistake in evaluating the performance of a learner with preprocessing: Preprocessing is often seen as completely independent of the later applied learning algorithms. When estimating the performance of the a learner, e.g., by cross-validation all preprocessing is done beforehand on the full data set and only training/predicting the learner is done on the train/test sets. Depending on what exactly is done as preprocessing this can lead to overoptimistic results. For example if imputation by the mean is done on the whole data set before evaluating the learner performance you are using information from the test data during training, which can cause overoptimistic performance results.

To clarify things one should distinguish between data-dependent and data-independent preprocessing steps: Data-dependent steps in some way learn from the data and give different results when applied to different data sets. Data-independent steps always lead to the same results. Clearly, correcting errors in the data or removing data columns like Ids that should not be used

for learning, is data-independent. Imputation of missing values by the mean, as mentioned above, is data-dependent. Imputation by a fixed constant, however, is not.

To get a honest estimate of learner performance combined with preprocessing, all data-dependent preprocessing steps must be included in the resampling. This is automatically done when fusing a learner with preprocessing.

To this end mlr provides two wrappers:

- makePreprocWrapperCaret is an interface to all preprocessing options offered by caret's preProcess function.
- makePreprocWrapper permits to write your own custom preprocessing methods by defining the actions to be taken before training and before prediction.

As mentioned above the specified preprocessing steps then "belong" to the wrapped Learner. In contrast to the preprocessing options listed above like normalizeFeatures

- the Task itself remains unchanged,
- the preprocessing is not done globally, i.e., for the whole data set, but for every pair of training/test data sets in, e.g., resampling,
- any parameters controlling the preprocessing as, e.g., the percentage of outliers to be removed can be tuned together with the base learner parameters.

We start with some examples for makePreprocWrapperCaret.

3.3.2 Preprocessing with makePreprocWrapperCaret

makePreprocWrapperCaret is an interface to caret's preProcess function that provides many different options like imputation of missing values, data transformations as scaling the features to a certain range or Box-Cox and dimensionality reduction via Independent or Principal Component Analysis. For all possible options see the help page of function preProcess.

Note that the usage of makePreprocWrapperCaret is slightly different than that of preProcess.

- makePreprocWrapperCaret takes (almost) the same formal arguments as preProcess, but their names are prefixed by ppc..
- The only exception: makePreprocWrapperCaret does not have a method argument. Instead all preprocessing options that would be passed to pre-Process's method argument are given as individual logical parameters to makePreprocWrapperCaret.

For example the following call to preProcess

```
preProcess(x, method = c("knnImpute", "pca"), pcaComp = 10)
```

with x being a matrix or data.frame would thus translate into

```
makePreprocWrapperCaret(learner, ppc.knnImpute = TRUE, ppc.pca =
    TRUE, ppc.pcaComp = 10)
```

where learner is a mlr Learner or the name of a learner class like "classif.lda".

If you enable multiple preprocessing options (like knn imputation and principal component analysis above) these are executed in a certain order detailed on the help page of function preProcess.

In the following we show an example where principal components analysis (PCA) is used for dimensionality reduction. This should never be applied blindly, but can be beneficial with learners that get problems with high dimensionality or those that can profit from rotating the data.

We consider the sonar.task, which poses a binary classification problem with 208 observations and 60 features.

```
sonar.task
#> Supervised task: Sonar-example
#> Type: classif
#> Target: Class
#> Observations: 208
#> Features:
#> numerics factors ordered
      60
#>
            0
#> Missings: FALSE
#> Has weights: FALSE
#> Has blocking: FALSE
#> Classes: 2
#> M
#> 111 97
#> Positive class: M
```

Below we fuse quadratic discriminant analysis from package MASS with a principal components preprocessing step. The threshold is set to 0.9, i.e., the principal components necessary to explain a cumulative percentage of 90% of the total variance are kept. The data are automatically standardized prior to PCA.

```
#> Type: classif
#> Name: ; Short name:
#> Class: PreprocWrapperCaret
#> Properties: twoclass, multiclass, numerics, factors, prob
#> Predict-Type: response
#> Hyperparameters:
    ppc.BoxCox=FALSE,ppc.YeoJohnson=FALSE,ppc.expoTrans=FALSE,ppc.center=TRUE,ppc.scale=TRUE,ppc
```

The wrapped learner is trained on the sonar.task. By inspecting the underlying qda model, we see that the first 22 principal components have been used for training.

```
mod = train(lrn, sonar.task)
mod
#> Model for learner.id=classif.qda.preproc;
            learner.\,class = PreprocWrapperCaret
#> Trained on: task.id = Sonar-example; obs = 208; features = 60
#> Hyperparameters:
            ppc.BoxCox=FALSE, ppc.YeoJohnson=FALSE, ppc.expoTrans=FALSE, ppc.center=TRUE, ppc.scale=TRUE, ppc.expoTrans=FALSE, ppc.center=TRUE, ppc.scale=TRUE, ppc.expoTrans=FALSE, ppc.center=TRUE, ppc.scale=TRUE, ppc.expoTrans=FALSE, ppc.center=TRUE, ppc.scale=TRUE, ppc.scale=TR
getLearnerModel(mod)
#> Model for learner.id=classif.qda; learner.class=classif.qda
#> Trained on: task.id = Sonar-example; obs = 208; features = 22
#> Hyperparameters:
getLearnerModel(mod, more.unwrap = TRUE)
#> Call:
#> qda(f, data = getTaskData(.task, .subset, recode.target =
             "drop.levels"))
#> Prior probabilities of groups:
#>
                                Μ
#> 0.5336538 0.4663462
#>
#> Group means:
                                   PC1
                                                                     PC2
                                                                                                      PC3
                                                                                                                                          PC4
                                                                                                                                                                              PC5
                                   PC6
#> M 0.5976122 -0.8058235 0.9773518 0.03794232 -0.04568166
            -0.06721702
#> R -0.6838655  0.9221279 -1.1184128 -0.04341853  0.05227489
            0.07691845
                                    PC7
                                                                        PC8
                                                                                                         PC9
                                                                                                                                       PC10
                                                                                                                                                                           PC11
                                      PC12
#> M 0.2278162 -0.01034406 -0.2530606 -0.1793157 -0.04084466
            -0.0004789888
#> R -0.2606969 0.01183702 0.2895848 0.2051963 0.04673977
```

```
0.0005481212
#>
         PC13
                   PC14
                             PC15
                                       PC16
                                                 PC17
         PC18
#> M -0.06138758 -0.1057137 0.02808048 0.05215865 -0.07453265
   0.03869042
#> R 0.07024765 0.1209713 -0.03213333 -0.05968671 0.08528994
   -0.04427460
#>
         PC19
                    PC20
                               PC2.1
                                         PC22.
#> R 0.01364323 -0.006978877 -0.01445851 0.001401586
```

Below the performances of qda with and without PCA preprocessing are compared in a benchmark experiment. Note that we use stratified resampling to prevent errors in qda due to a too small number of observations from either class.

PCA preprocessing in this case turns out to be really beneficial for the performance of Quadratic Discriminant Analysis.

3.3.2.1 Joint tuning of preprocessing options and learner parameters

Let's see if we can optimize this a bit. The threshold value of 0.9 above was chosen arbitrarily and led to 22 out of 60 principal components. But maybe a lower or higher number of principal components should be used. Moreover, qda has several options that control how the class covariance matrices or class probabilities are estimated.

Those preprocessing and learner parameters can be tuned jointly. Before doing this let's first get an overview of all the parameters of the wrapped learner using function getParamSet.

#>	ppc.expoTrans	logical	-	FALSE	-
#>	ppc.center	logical	-	TRUE	-
#>	ppc.scale	logical	_	TRUE	-
#>	- ppc.range	logical	_	FALSE	_
	_				
#>	ppc.knnImpute	logical	-	FALSE	-
#>	ppc.bagImpute	logical	-	FALSE	-
#>	ppc.medianImpute	logical	-	FALSE	-
#>	ppc.pca	logical	_	FALSE	-
	-				
#>	ppc.ica -	logical	_	FALSE	-
#>	<pre>ppc.spatialSign -</pre>	logical	-	FALSE	-
#>	ppc.thresh	numeric	-	0.95	O to Inf
#>	ppc.pcaComp	integer	-	-	1 to Inf
#>	ppc.na.remove	logical	-	TRUE	-
#>	ppc.k	integer	-	5	1 to Inf
#>	- ppc.fudge	numeric	-	0.2	O to Inf
#>	- ppc.numUnique	integer	-	3	1 to Inf
#>	- ppc.n.comp	integer	_	_	1 to Inf
#>	- method	discrete	_	moment	moment,mle,mve,t
	-				
#>	nu Y	numeric	_	5	2 to Inf
#>	predict.method	discrete	- <i>p</i>	olug-in	plug-in, predictive, debiased
#>		Tunable T	rafo		
#>	ppc.BoxCox	TRUE	_		
	ppc.YeoJohnson	TRUE	_		
	ppc.expoTrans	TRUE	_		
	ppc.center	TRUE	_		
	ppc.scale	TRUE	_		

```
#> ppc.range
                      TRUE
#> ppc.knnImpute
                      TRUE
                      TRUE
#> ppc.bagImpute
#> ppc.medianImpute
                      TRUE
                      TRUE
#> ppc.pca
#> ppc.ica
                      TRUE
#> ppc.spatialSign
                      TRUE
#> ppc.thresh
                      TRUE
#> ppc.pcaComp
                     TRUE
#> ppc.na.remove
                      TRUE
#> ppc.k
                      TRUE
#> ppc.fudge
                      TRUE
#> ppc.numUnique
                      TRUE
#> ppc.n.comp
                      TRUE
#> method
                      TRUE
#> nu
                      TRUE
#> predict.method
                      TRUE
```

The parameters prefixed by ppc. belong to preprocessing. method, nu and predict.method are qda parameters.

Instead of tuning the PCA threshold (ppc.thresh) we tune the number of principal components (ppc.pcaComp) directly. Moreover, for qda we try two different ways to estimate the posterior probabilities (parameter predict.method): the usual plug-in estimates and unbiased estimates.

We perform a grid search and set the resolution to 10. This is for demonstration. You might want to use a finer resolution.

```
ps = makeParamSet(
 makeIntegerParam("ppc.pcaComp", lower = 1, upper =
     getTaskNFeats(sonar.task)),
 makeDiscreteParam("predict.method", values = c("plug-in",
     "debiased"))
)
ctrl = makeTuneControlGrid(resolution = 10)
res = tuneParams(lrn, sonar.task, rin, par.set = ps, control = ctrl,
   show.info = FALSE)
res
#> Tune result:
#> Op. pars: ppc.pcaComp=8; predict.method=plug-in
#> mmce.test.mean=0.192
as.data.frame(res$opt.path)[1:3]
     ppc.pcaComp predict.method mmce.test.mean
#> 1
               1
                     plug-in 0.4757074
               8
#> 2
                     plug-in
                                 0.1920635
```

<i>#> 3</i>	14	plug-in	0.2162871
#> 4	21	plug-in	0.2643202
<i>#> 5</i>	27	plug-in	0.2454106
<i>#> 6</i>	34	plug-in	0.2645273
#> 7	40	plug-in	0.2742581
#> 8	47	plug-in	0.3173223
#> 9	<i>53</i>	plug-in	0.3512767
#> 10	60	plug-in	0.3941339
<i>#> 11</i>	1	debiased	0.5336094
<i>#> 12</i>	8	debiased	0.2450656
<i>#> 13</i>	14	debiased	0.2403037
<i>#> 14</i>	21	debiased	0.2546584
<i>#> 15</i>	27	debiased	0.3075224
<i>#> 16</i>	34	debiased	0.3172533
<i>#> 17</i>	40	debiased	0.3125604
#> 18	47	debiased	0.2979986
<i>#> 19</i>	<i>53</i>	debiased	0.3079365
<i>#> 20</i>	60	debiased	0.3654244

There seems to be a preference for a lower number of principal components (<27) for both "plug-in" and "debiased" with "plug-in" achieving slightly lower error rates.

3.3.3 Writing a custom preprocessing wrapper

If the options offered by makePreprocWrapperCaret are not enough, you can write your own preprocessing wrapper using function makePreprocWrapper.

As described in the tutorial section about wrapped learners wrappers are implemented using a *train* and a *predict* method. In case of preprocessing wrappers these methods specify how to transform the data before training and before prediction and are *completely user-defined*.

Below we show how to create a preprocessing wrapper that centers and scales the data before training/predicting. Some learning methods as, e.g., k nearest neighbors, support vector machines or neural networks usually require scaled features. Many, but not all, have a built-in scaling option where the training data set is scaled before model fitting and the test data set is scaled accordingly, that is by using the scaling parameters from the training stage, before making predictions. In the following we show how to add a scaling option to a Learner by coupling it with function scale.

Note that we chose this simple example for demonstration. Centering/scaling the data is also possible with makePreprocWrapperCaret.

3.3.3.1 Specifying the train function

The *train* function has to be a function with the following arguments:

- data is a data.frame with columns for all features and the target variable.
- target is a string and denotes the name of the target variable in data.
- **args** is a list of further arguments and parameters that influence the preprocessing.

It must return a list with elements \$data and \$control, where \$data is the preprocessed data set and \$control stores all information required to preprocess the data before prediction.

The *train* function for the scaling example is given below. It calls scale on the numerical features and returns the scaled training data and the corresponding scaling parameters.

args contains the center and scale arguments of function scale and slot \$control stores the scaling parameters to be used in the prediction stage.

Regarding the latter note that the center and scale arguments of scale can be either a logical value or a numeric vector of length equal to the number of the numeric columns in data, respectively. If a logical value was passed to args we store the column means and standard deviations/ root mean squares in the \$center and \$scale slots of the returned \$control object.

```
trainfun = function(data, target, args = list(center, scale)) {
  ## Identify numerical features
 cns = colnames(data)
 nums = setdiff(cns[sapply(data, is.numeric)], target)
  ## Extract numerical features from the data set and call scale
 x = as.matrix(data[, nums, drop = FALSE])
 x = scale(x, center = args$center, scale = args$scale)
  ## Store the scaling parameters in control
  ## These are needed to preprocess the data before prediction
 control = args
 if (is.logical(control$center) && control$center)
    control$center = attr(x, "scaled:center")
 if (is.logical(control$scale) && control$scale)
    control$scale = attr(x, "scaled:scale")
  ## Recombine the data
 data = data[, setdiff(cns, nums), drop = FALSE]
 data = cbind(data, as.data.frame(x))
 return(list(data = data, control = control))
}
```

3.3.3.2 Specifying the predict function

The *predict* function has the following arguments:

- data is a data.frame containing *only* feature values (as for prediction the target values naturally are not known).
- target is a string indicating the name of the target variable.
- args are the args that were passed to the *train* function.
- control is the object returned by the *train* function.

It returns the preprocessed data.

In our scaling example the *predict* function scales the numerical features using the parameters from the training stage stored in **control**.

```
predictfun = function(data, target, args, control) {
   ## Identify numerical features
   cns = colnames(data)
   nums = cns[sapply(data, is.numeric)]
   ## Extract numerical features from the data set and call scale
   x = as.matrix(data[, nums, drop = FALSE])
   x = scale(x, center = control$center, scale = control$scale)
   ## Recombine the data
   data = data[, setdiff(cns, nums), drop = FALSE]
   data = cbind(data, as.data.frame(x))
   return(data)
}
```

3.3.3.3 Creating the preprocessing wrapper

Below we create a preprocessing wrapper with a regression neural network (which itself does not have a scaling option) as base learner.

The *train* and *predict* functions defined above are passed to makePreprocWrapper via the train and predict arguments. par.vals is a list of parameter values that is relayed to the args argument of the *train* function.

```
lrn = makeLearner("regr.nnet", trace = FALSE, decay = 1e-02)
lrn = makePreprocWrapper(lrn, train = trainfun, predict = predictfun,
    par.vals = list(center = TRUE, scale = TRUE))
lrn
#> Learner regr.nnet.preproc from package nnet
#> Type: regr
#> Name: ; Short name:
#> Class: PreprocWrapper
#> Properties: numerics, factors, weights
#> Predict-Type: response
#> Hyperparameters: size=3, trace=FALSE, decay=0.01
```

Let's compare the cross-validated mean squared error (mse) on the Boston Housing data set with and without scaling.

```
rdesc = makeResampleDesc("CV", iters = 3)
r = resample(lrn, bh.task, resampling = rdesc, show.info = FALSE)
r
#> Resample Result
#> Task: BostonHousing-example
#> Learner: regr.nnet.preproc
#> Aggr perf: mse.test.mean=20.6
#> Runtime: 0.143972
lrn = makeLearner("regr.nnet", trace = FALSE, decay = 1e-02)
r = resample(lrn, bh.task, resampling = rdesc, show.info = FALSE)
r
#> Resample Result
#> Task: BostonHousing-example
#> Learner: regr.nnet
#> Aggr perf: mse.test.mean=55.1
#> Runtime: 0.105009
```

3.3.3.4 Joint tuning of preprocessing and learner parameters

Often it's not clear which preprocessing options work best with a certain learning algorithm. As already shown for the number of principal components in makePreprocWrapperCaret we can tune them easily together with other hyperparameters of the learner.

In our scaling example we can try if nnet works best with both centering and scaling the data or if it's better to omit one of the two operations or do no preprocessing at all. In order to tune **center** and **scale** we have to add appropriate LearnerParams to the parameter set of the wrapped learner.

As mentioned above scale allows for numeric and logical center and scale arguments. As we want to use the latter option we declare center and scale as logical learner parameters.

```
lrn = makeLearner("regr.nnet", trace = FALSE)
lrn = makePreprocWrapper(lrn, train = trainfun, predict = predictfun,
    par.set = makeParamSet(
        makeLogicalLearnerParam("center"),
        makeLogicalLearnerParam("scale")
    ),
    par.vals = list(center = TRUE, scale = TRUE))
lrn
#> Learner regr.nnet.preproc from package nnet
#> Type: regr
#> Name: ; Short name:
```

```
#> Class: PreprocWrapper
#> Properties: numerics, factors, weights
#> Predict-Type: response
#> Hyperparameters: size=3, trace=FALSE, center=TRUE, scale=TRUE
getParamSet(lrn)
           Type len Def Constr Req Tunable Trafo
#>
#> center logical - - - TRUE 
#> scale logical - - TRUE
#> scale logical -
#> size integer -
                     3 O to Inf - TRUE
#> maxit integer - 100 1 to Inf - TRUE
                           - Y TRUE
#> linout logical - FALSE
#> entropy logical - FALSE
                               - Y TRUE
#> softmax logical - FALSE
                               - Y TRUE
                                - Y TRUE
#> censored logical - FALSE
#> skip logical - FALSE -
                                  - TRUE
#> rang numeric - 0.7 -Inf to Inf - TRUE
\# decay numeric - 0 0 to Inf - TRUE
#> Hess logical - FALSE
                                      TRUE
#> trace logical - TRUE
                                  - FALSE
#> MaxNWts integer - 1000 1 to Inf - TRUE
#> abstoll numeric - 0.0001 -Inf to Inf - TRUE
#> reltoll numeric - 1e-08 -Inf to Inf - TRUE
```

Now we do a simple grid search for the decay parameter of nnet and the center and scale parameters.

```
rdesc = makeResampleDesc("Holdout")
ps = makeParamSet(
 makeDiscreteParam("decay", c(0, 0.05, 0.1)),
 makeLogicalParam("center"),
 makeLogicalParam("scale")
)
ctrl = makeTuneControlGrid()
res = tuneParams(lrn, bh.task, rdesc, par.set = ps, control = ctrl,
    show.info = FALSE)
res
#> Tune result:
#> Op. pars: decay=0.05; center=FALSE; scale=TRUE
#> mse.test.mean=14.8
as.data.frame(res$opt.path)
     decay center scale mse.test.mean dob eol error.message
    exec.time
#> 1 0 TRUE TRUE 49.38128 1 NA
```

	0.	043						
				TRUE	20.64761	2	NA	<na></na>
		049						
			TRUE	TRUE	22.42986	3	NA	<na></na>
		.050	E41.0E	an i i	00.05181	,	37.4	477.45
			FALSE	TRUE	96.25474	4	IVA	<na></na>
		026	EALGE	יונותיי	11 01206	_	77.4	ZNT 4.5
		0.05	FALSE	IKUL	14.84300	5	IVA	<na></na>
			FAISE	TRIIF	16.65383	6	πД	<na></na>
		046	TALOL	111011	10.00000	U	IVA	WAY
			TRUE	FALSE	40.51518	7	NA	<na></na>
		049			,			
			TRUE	FALSE	68.00069	8	NA	<na></na>
	0.	046						
#>	9	0.1	TRUE	FALSE	55.42210	9	NA	<na></na>
	0.	050						
#>	10	0	FALSE	FALSE	96.25474	10	NA	<na></na>
		027						
			FALSE	FALSE	56.25758	11	NA	<na></na>
		.049	- 4- <i>c</i> -					
			FALSE	FALSE	42.85529	12	NA	<na></na>
	0.	.047						

3.3.3.5 Preprocessing wrapper functions

If you have written a preprocessing wrapper that you might want to use from time to time it's a good idea to encapsulate it in an own function as shown below. If you think your preprocessing method is something others might want to use as well and should be integrated into mlr just contact us.

```
makePreprocWrapperScale = function(learner, center = TRUE, scale =
   TRUE) {
   trainfum = function(data, target, args = list(center, scale)) {
      cns = colnames(data)
      nums = setdiff(cns[sapply(data, is.numeric)], target)
      x = as.matrix(data[, nums, drop = FALSE])
      x = scale(x, center = args$center, scale = args$scale)
      control = args
   if (is.logical(control$center) && control$center)
      control$center = attr(x, "scaled:center")
   if (is.logical(control$scale) && control$scale)
      control$scale = attr(x, "scaled:scale")
   data = data[, setdiff(cns, nums), drop = FALSE]
   data = cbind(data, as.data.frame(x))
   return(list(data = data, control = control))
```

```
predictfun = function(data, target, args, control) {
    cns = colnames(data)
    nums = cns[sapply(data, is.numeric)]
    x = as.matrix(data[, nums, drop = FALSE])
    x = scale(x, center = control$center, scale = control$scale)
    data = data[, setdiff(cns, nums), drop = FALSE]
    data = cbind(data, as.data.frame(x))
    return(data)
  }
  makePreprocWrapper(
    learner,
    train = trainfun,
    predict = predictfun,
    par.set = makeParamSet(
      makeLogicalLearnerParam("center"),
      makeLogicalLearnerParam("scale")
    par.vals = list(center = center, scale = scale)
}
lrn = makePreprocWrapperScale("classif.lda")
train(lrn, iris.task)
#> Model for learner.id=classif.lda.preproc;
    learner.class=PreprocWrapper
#> Trained on: task.id = iris-example; obs = 150; features = 4
#> Hyperparameters: center=TRUE, scale=TRUE
```

3.4 Imputation of Missing Values

mlr provides several imputation methods which are listed on the help page imputations. These include standard techniques as imputation by a constant value (like a fixed constant, the mean, median or mode) and random numbers (either from the empirical distribution of the feature under consideration or a certain distribution family). Moreover, missing values in one feature can be replaced based on the other features by predictions from any supervised Learner integrated into mlr.

If your favourite option is not implemented in mlr yet, you can easily create your own imputation method.

Also note that some of the learning algorithms included in mlr can deal with missing values in a sensible way, i.e., other than simply deleting observations with missing values. Those Learners have the property "missings" and thus can be identified using listLearners.

```
### Regression learners that can deal with missing values
listLearners("regr", properties = "missings")[c("class", "package")]
#>
                    class
                                  package
#> 1
          regr.blackboost
                             mboost, party
#> 2
            regr.cforest
                                   party
#> 3
              regr.ctree
                                    party
#> 4
              regr.cubist
                                    Cubist
#> 5
                 regr.gbm
                                      gbm
#> 6 regr.randomForestSRC randomForestSRC
#> ... (8 rows, 2 cols)
```

See also the list of integrated learners in the Appendix.

3.4.1 Imputation and reimputation

Imputation can be done by function impute. You can specify an imputation method for each feature individually or for classes of features like numerics or factors. Moreover, you can generate dummy variables that indicate which values are missing, also either for classes of features or for individual features. These allow to identify the patterns and reasons for missing data and permit to treat imputed and observed values differently in a subsequent analysis.

Let's have a look at the airquality data set.

```
data(airquality)
summary(airquality)
#>
       Ozone
                       Solar.R
                                        Wind
                                                         Temp
#>
  Min.
          : 1.00
                   Min. : 7.0
                                          : 1.700
                                                           :56.00
                                   Min.
                                                    Min.
   1st Qu.: 18.00
                   1st Qu.:115.8
#>
                                   1st Qu.: 7.400
                                                    1st Qu.:72.00
#>
  Median : 31.50
                  Median :205.0
                                  Median : 9.700
                                                   Median :79.00
#>
  Mean : 42.13
                   Mean :185.9
                                  Mean : 9.958
                                                   Mean
                                                          :77.88
   3rd Qu.: 63.25
                    3rd Qu.:258.8
#>
                                   3rd Qu.:11.500
                                                    3rd Qu.:85.00
#>
   Max.
          :168.00
                   Max.
                           :334.0
                                   Max.
                                          :20.700
                                                    Max.
                                                           :97.00
   NA's
          :37
                    NA's :7
#>
                       Day
#>
       Month
#>
  Min.
          :5.000
                  Min.
                         : 1.0
                   1st Qu.: 8.0
#>
   1st Qu.:6.000
  Median :7.000
                  Median:16.0
#>
  Mean
         :6.993
                   Mean :15.8
   3rd Qu.:8.000
                   3rd Qu.:23.0
#>
   Max.
          :9.000
                          :31.0
                   Max.
#>
```

There are 37 NA's in variable Ozone (ozone pollution) and 7 NA's in variable Solar.R (solar radiation). For demonstration purposes we insert artificial NA's in column Wind (wind speed) and coerce it into a factor.

```
airq = airquality
ind = sample(nrow(airq), 10)
airq$Wind[ind] = NA
airq$Wind = cut(airq$Wind, c(0,8,16,24))
summary(airq)
#>
        Ozone
                       Solar, R.
                                          Wind
                                                      Temp
#>
   Min. : 1.00
                    Min. : 7.0
                                     (0,8] :51
                                                 Min.
                                                         :56.00
   1st Qu.: 18.00
                    1st Qu.:115.8
                                     (8,16]:86
#>
                                                 1st Qu.:72.00
  Median : 31.50
                    Median :205.0
                                     (16,24]:6
                                                 Median :79.00
  Mean : 42.13
#>
                    Mean
                           :185.9
                                    NA's :10
                                                       :77.88
                                                 Mean
   3rd Qu.: 63.25
                    3rd Qu.:258.8
                                                  3rd Qu.:85.00
          :168.00
                                                 Max. :97.00
#>
   Max.
                    Max.
                           :334.0
   NA's
          :37
                    NA's
                            :7
       Month
                        Day
#>
#>
   Min.
           :5.000
                   Min.
                           : 1.0
#>
   1st Qu.:6.000
                   1st Qu.: 8.0
#>
  Median :7.000
                   Median:16.0
          :6.993
                         :15.8
#>
  Mean
                   Mean
#>
   3rd Qu.:8.000
                   3rd Qu.:23.0
#>
   Max.
          :9.000
                   Max.
                          :31.0
#>
```

If you want to impute NA's in all integer features (these include Ozone and Solar.R) by the mean, in all factor features (Wind) by the mode and additionally generate dummy variables for all integer features, you can do this as follows:

```
imp = impute(airq, classes = list(integer = imputeMean(), factor =
   imputeMode()),
dummy.classes = "integer")
```

impute returns a list where slot \$data contains the imputed data set. Per default, the dummy variables are factors with levels "TRUE" and "FALSE". It is also possible to create numeric zero-one indicator variables.

```
head(imp$data, 10)
         Ozone Solar.R
                            Wind Temp Month Day Ozone.dummy
    Solar.R. dummy
#> 1 41.00000 190.0000
                           (0,8]
                                                       FALSE
    FALSE
#> 2 36.00000 118.0000
                           (0,8]
                                   72
                                           5
                                               2
                                                       FALSE
    FALSE
#> 3 12.00000 149.0000
                          (8, 16]
                                   74
                                           5
                                               3
                                                       FALSE
    FALSE
#> 4 18.00000 313.0000
                          (8, 16]
                                                       FALSE
    FALSE
#> 5 42.12931 185.9315 (8,16]
                                               5
                                                        TRUE
    TRUE
```

```
#> 6 28.00000 185.9315 (8,16]
                                   66
                                              6
                                                      FALSE
   TRUE
#> 7 23.00000 299.0000
                                              7
                         (8, 16]
                                   65
                                                      FALSE
   FALSE
#> 8 19.00000 99.0000
                         (8, 16]
                                   59
                                              8
                                                      FALSE
   FALSE
#> 9 8.00000 19.0000 (16,24]
                                              9
                                                      FALSE
   FALSE
#> 10 42.12931 194.0000
                        (8, 16]
                                             10
                                                       TRUE
   FALSE
```

Slot \$desc is an ImputationDesc object that stores all relevant information about the imputation. For the current example this includes the means and the mode computed on the non-missing data.

```
imp$desc
#> Imputation description
#> Target:
#> Features: 6; Imputed: 6
#> impute.new.levels: TRUE
#> recode.factor.levels: TRUE
#> dummy.type: factor
```

The imputation description shows the name of the target variable (not present), the number of features and the number of imputed features. Note that the latter number refers to the features for which an imputation method was specified (five integers plus one factor) and not to the features actually containing NA's. dummy.type indicates that the dummy variables are factors. For details on impute.new.levels and recode.factor.levels see the help page of function impute.

Let's have a look at another example involving a target variable. A possible learning task associated with the airquality data is to predict the ozone pollution based on the meteorological features. Since we do not want to use columns Day and Month we remove them.

```
airq = subset(airq, select = 1:4)
```

The first 100 observations are used as training data set.

```
airq.train = airq[1:100,]
airq.test = airq[-c(1:100),]
```

In case of a supervised learning problem you need to pass the name of the target variable to impute. This prevents imputation and creation of a dummy variable for the target variable itself and makes sure that the target variable is not used to impute the features.

In contrast to the example above we specify imputation methods for individual features instead of classes of features.

Missing values in Solar.R are imputed by random numbers drawn from the empirical distribution of the non-missing observations.

Function imputeLearner allows to use all supervised learning algorithms integrated into mlr for imputation. The type of the Learner (regr, classif) must correspond to the class of the feature to be imputed. The missing values in Wind are replaced by the predictions of a classification tree (rpart). Per default, all available columns in airq.train except the target variable (Ozone) and the variable to be imputed (Wind) are used as features in the classification tree, here Solar.R and Temp. You can also select manually which columns to use. Note that rpart can deal with missing feature values, therefore the NA's in column Solar.R do not pose a problem.

```
imp = impute(airq.train, target = "Ozone", cols = list(Solar.R =
    imputeHist(),
 Wind = imputeLearner("classif.rpart")), dummy.cols = c("Solar.R",
     "Wind"))
summary(imp$data)
#>
        Ozone
                       Solar.R
                                          Wind
                                                       Temp
                    Min. : 7.00
#>
   Min. : 1.00
                                     (0,8] :34
                                                  Min.
                                                         :56.00
   1st Qu.: 16.00
                                     (8,16]:61
                                                  1st Qu.:69.00
                   1st Qu.: 98.75
#>
  Median: 34.00 Median: 221.50
                                     (16,24]:5
                                                 Median :79.50
   Mean : 41.59
#>
                    Mean :191.54
                                                  Mean
                                                       :76.87
#>
   3rd Qu.: 63.00
                    3rd Qu.:274.25
                                                  3rd Qu.:84.00
  Max. :135.00
                    Max. :334.00
                                                  Max. :93.00
#>
  NA's
         :31
#>
   Solar.R. dummy Wind. dummy
   FALSE:93
                 FALSE: 92
#>
#>
   TRUE : 7
                 TRUE : 8
#>
#>
#>
#>
#>
imp$desc
#> Imputation description
#> Target: Ozone
#> Features: 3; Imputed: 2
#> impute.new.levels: TRUE
#> recode.factor.levels: TRUE
#> dummy.type: factor
```

The ImputationDesc object can be used by function reimpute to impute the test data set the same way as the training data.

```
airq.test.imp = reimpute(airq.test, imp$desc)
head(airq.test.imp)
    Ozone Solar.R
                   Wind Temp Solar.R.dummy Wind.dummy
#> 1
      110
           207 (0,8] 90
                                    FALSE
                                               FALSE
#> 2
           222 (8,16] 92
                                    FALSE
                                               FALSE
      NA
#> 3
       NA
             137 (8,16] 86
                                    FALSE
                                               FALSE
       44
              192 (8,16]
                          86
                                    FALSE
                                               FALSE
       28
              273 (8,16]
                          82
                                    FALSE
                                               FALSE.
#> 5
#> 6
      65
           157 (8,16]
                          80
                                    FALSE
                                               FALSE
```

Especially when evaluating a machine learning method by some resampling technique you might want that impute/reimpute are called automatically each time before training/prediction. This can be achieved by creating an imputation wrapper.

3.4.2 Fusing a learner with imputation

You can couple a Learner with imputation by function makeImputeWrapper which basically has the same formal arguments as impute. Like in the example above we impute Solar.R by random numbers from its empirical distribution, Wind by the predictions of a classification tree and generate dummy variables for both features.

```
lrn = makeImputeWrapper("regr.lm", cols = list(Solar.R =
    imputeHist(),
Wind = imputeLearner("classif.rpart")), dummy.cols = c("Solar.R",
        "Wind"))
lrn
#> Learner regr.lm.imputed from package stats
#> Type: regr
#> Name: ; Short name:
#> Class: ImputeWrapper
#> Properties: numerics, factors, se, weights, missings
#> Predict-Type: response
#> Hyperparameters:
```

Before training the resulting Learner, impute is applied to the training set. Before prediction reimpute is called on the test set and the ImputationDesc object from the training stage.

We again aim to predict the ozone pollution from the meteorological variables. In order to create the Task we need to delete observations with missing values in the target variable.

```
airq = subset(airq, subset = !is.na(airq$Ozone))
task = makeRegrTask(data = airq, target = "Ozone")
```

```
In the following the 3-fold cross-validated mean squared error is calculated.
```

```
rdesc = makeResampleDesc("CV", iters = 3)
r = resample(lrn, task, resampling = rdesc, show.info = FALSE,
   models = TRUE)
r$aggr
#> mse.test.mean
#> 524.3392
lapply(r$models, getLearnerModel, more.unwrap = TRUE)
#> [[1]]
#>
#> Call:
#> stats::lm(formula = f, data = d)
#>
#> Coefficients:
#>
                                                Wind(8,16]
         (Intercept)
                                Solar.R
#>
          -117.0954
                                0.0853
                                                  -27.6763
#>
         Wind(16,24]
                                   Temp Solar.R.dummyTRUE
             -9.0988
#>
                                 2.0505
                                                  -27.4152
#>
      Wind.dummyTRUE
#>
             2.2535
#>
#>
#> [[2]]
#>
#> Call:
#> stats::lm(formula = f, data = d)
#>
#> Coefficients:
#>
         (Intercept)
                                Solar.R
                                                Wind(8,16]
#>
                                0.03936
          -94.84542
                                                 -16.26255
#>
         Wind(16,24]
                                   Temp Solar.R.dummyTRUE
                                                 -11.08578
#>
            -7.00707
                                1.79513
#>
      Wind.dummyTRUE
#>
            -0.68340
#>
#>
#> [[3]]
#>
#> Call:
#> stats::lm(formula = f, data = d)
#>
#> Coefficients:
#>
         (Intercept)
                                Solar.R
                                                Wind(8,16]
#>
           -57.30438
                                0.07426
                                                 -30.70737
        Wind(16,24]
                             {\it Temp Solar.R.dummyTRUE}
```

A second possibility to fuse a learner with imputation is provided by makePre-procWrapperCaret, which is an interface to caret's preProcess function. pre-Process only works for numeric features and offers imputation by k-nearest neighbors, bagged trees, and by the median.

3.5 Generic Bagging

One reason why random forests perform so well is that they are using bagging as a technique to gain more stability. But why do you want to limit yourself to the classifiers already implemented in well known random forests when it is really easy to build your own with mlr?

Just bag an mlr learner already makeBaggingWrapper.

As in a random forest, we need a Learner which is trained on a subset of the data during each iteration of the bagging process. The subsets are chosen according to the parameters given to makeBaggingWrapper:

- bw.iters On how many subsets (samples) do we want to train our Learner?
- bw.replace Sample with replacement (also known as bootstrapping)?
- bw.size Percentage size of the samples. If bw.replace = TRUE, bw.size
 1 is the default. This does not mean that one sample will contain all the observations as observations will occur multiple times in each sample.
- bw.feats Percentage size of randomly selected features for each iteration.

Of course we also need a Learner which we have to pass to makeBaggingWrapper.

```
lrn = makeLearner("classif.rpart")
bag.lrn = makeBaggingWrapper(lrn, bw.iters = 50, bw.replace = TRUE,
    bw.size = 0.8, bw.feats = 3/4)
```

Now we can compare the performance with and without bagging. First let's try it without bagging:

And now with bagging:

```
rdesc = makeResampleDesc("CV", iters = 10)
result = resample(learner = bag.lrn, task = sonar.task, resampling =
    rdesc, show.info = FALSE)
result$aggr
#> mmce.test.mean
#> 0.2069048
```

Training more learners takes more time, but can outperform pure learners on noisy data with many features.

3.5.1 Changing the type of prediction

In case of a *classification* problem the predicted class labels are determined by majority voting over the predictions of the individual models. Additionally, posterior probabilities can be estimated as the relative proportions of the predicted class labels. For this purpose you have to change the predict type of the *bagging learner* as follows.

```
bag.lrn = setPredictType(bag.lrn, predict.type = "prob")
```

Note that it is not relevant if the *base learner* itself can predict probabilities and that for this reason the predict type of the *base learner* always has to be "response".

For regression the mean value across predictions is computed. Moreover, the standard deviation across predictions is estimated if the predict type of the bagging learner is changed to "se". Below, we give a small example for regression.

```
n = getTaskSize(bh.task)
train.inds = seq(1, n, 3)
test.inds = setdiff(1:n, train.inds)
lrn = makeLearner("regr.rpart")
bag.lrn = makeBaggingWrapper(lrn)
bag.lrn = setPredictType(bag.lrn, predict.type = "se")
mod = train(learner = bag.lrn, task = bh.task, subset = train.inds)
```

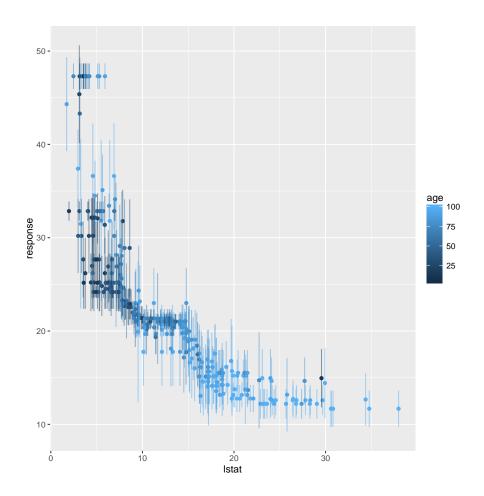
With function getLearnerModel, you can access the models fitted in the individual iterations.

Predict the response and calculate the standard deviation:

In the column labelled se the standard deviation for each prediction is given.

Let's visualise this a bit using ggplot2. Here we plot the percentage of lower status of the population (lstat) against the prediction.

```
library("ggplot2")
library("reshape2")
data = cbind(as.data.frame(pred), getTaskData(bh.task, subset =
    test.inds))
g = ggplot(data, aes(x = lstat, y = response, ymin = response-se,
    ymax = response+se, col = age))
g + geom_point() + geom_linerange(alpha=0.5)
```



3.6 Advanced Tuning

3.6.1 Iterated F-Racing for mixed spaces and dependencies

The package supports a larger number of tuning algorithms, which can all be looked up and selected via TuneControl. One of the cooler algorithms is iterated F-racing from the irace package (technical description here). This not only works for arbitrary parameter types (numeric, integer, discrete, logical), but also for so-called dependent / hierarchical parameters:

```
makeNumericParam("sigma", lower = -12, upper = 12, trafo =
     function(x) 2<sup>x</sup>,
   requires = quote(kernel == "rbfdot")),
 makeIntegerParam("degree", lower = 2L, upper = 5L,
   requires = quote(kernel == "polydot"))
)
ctrl = makeTuneControlIrace(maxExperiments = 200L)
rdesc = makeResampleDesc("Holdout")
res = tuneParams("classif.ksvm", iris.task, rdesc, par.set = ps,
   control = ctrl, show.info = FALSE)
print(head(as.data.frame(res$opt.path)))
            C
                kernel sigma degree mmce.test.mean dob eol
#> 1 -2.8894525
                polydot
                             NA
                                               0.06 1 NA
                                     4
                         NA
#> 2 -2.6793542 vanilladot
                                                  0.04
                                     NA
                                                        1 NA
#> 3 -0.7855061
               rbfdot 11.049783
                                     NA
                                                 0.68
                                                        1 NA
                                     5
#> 4 1.7678978
                polydot NA
                                                 0.14 1 NA
#> 5 9.7729840 vanilladot
                             NA
                                     NA
                                                  0.04 1 NA
               rbfdot -7.198476
                                                  0.36 1 NA
#> 6 -2.7930352
                                     NA
#> error.message exec.time
#> 1
         <NA> 0.044
#> 2
            <NA>
                    0.035
#> 3
            <NA>
                    0.038
#> 4
            <NA>
                    0.033
#> 5
             <NA>
                     0.040
#> 6
             <NA>
                     0.027
```

See how we made the kernel parameters like sigma and degree dependent on the kernel selection parameters? This approach allows you to tune parameters of multiple kernels at once, efficiently concentrating on the ones which work best for your given data set.

3.6.2 Tuning across whole model spaces with ModelMultiplexer

We can now take the following example even one step further. If we use the ModelMultiplexer we can tune over different model classes at once, just as we did with the SVM kernels above.

```
base.learners = list(
  makeLearner("classif.ksvm"),
  makeLearner("classif.randomForest")
)
lrn = makeModelMultiplexer(base.learners)
```

Function makeModelMultiplexerParamSet offers a simple way to construct a parameter set for tuning: The parameter names are prefixed automatically

and the requires element is set, too, to make all parameters subordinate to selected.learner.

```
ps = makeModelMultiplexerParamSet(lrn,
  makeNumericParam("sigma", lower = -12, upper = 12, trafo =
      function(x) 2^x),
 makeIntegerParam("ntree", lower = 1L, upper = 500L)
print(ps)
#>
                                  Type len Def
#> selected.learner
                              discrete
#> classif.ksvm.sigma
                              numeric
#> classif.randomForest.ntree integer
#>
                                                          Constr Req
    Tunable
#> selected.learner
                              {\it classif.ksvm,classif.randomForest}
       TRUE
                                                       -12 to 12
                                                                  Y
#> classif.ksvm.siqma
                                                        1 to 500 Y
#> classif.randomForest.ntree
       TRUE
                              Trafo
#> selected.learner
                                   Y
#> classif.ksvm.sigma
#> classif.randomForest.ntree
rdesc = makeResampleDesc("CV", iters = 2L)
ctrl = makeTuneControlIrace(maxExperiments = 200L)
res = tuneParams(lrn, iris.task, rdesc, par.set = ps, control =
    ctrl, show.info = FALSE)
print(head(as.data.frame(res$opt.path)))
         selected.learner classif.ksvm.sigma
    {\it classif.random} Forest.ntree
#> 1 classif.randomForest
                                          NA
    273
#> 2
            classif.ksvm
                                    10.53605
    NA
#> 3
             classif.ksvm
                                   -11.79057
    NA
             classif.ksvm
                                    10.42478
    NA
#> 5 classif.randomForest
                                          NA
    394
#> 6
             classif.ksvm
                                    11.02356
#> mmce.test.mean dob eol error.message exec.time
```

```
#> 1
          0.04666667
                        1 NA
                                         <NA>
                                                   0.061
#> 2
          0.68000000
                                         <NA>
                                                    0.055
                        1
                           NA
#> 3
          0.52666667
                        1
                           NA
                                         <NA>
                                                    0.053
#> 4
          0.68000000
                        1
                           NA
                                         <NA>
                                                   0.073
#> 5
          0.04666667
                        1 NA
                                         <NA>
                                                    0.087
#> 6
          0.68000000
                        1 NA
                                          \langle NA \rangle
                                                    0.051
```

3.6.3 Multi-criteria evaluation and optimization

During tuning you might want to optimize multiple, potentially conflicting, performance measures simultaneously.

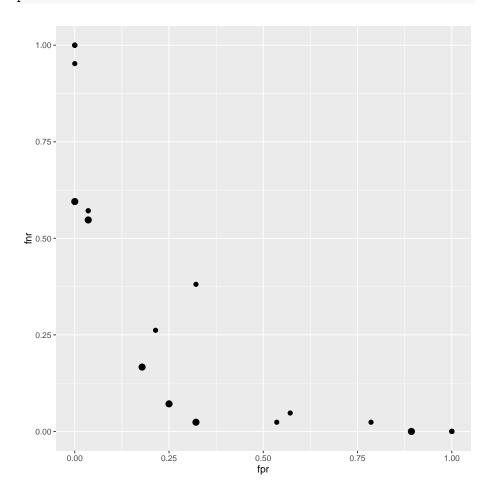
In the following example we aim to minimize both, the false positive and the false negative rates (fpr and fnr). We again tune the hyperparameters of an SVM (function ksvm) with a radial basis kernel and use the sonar classification task for illustration. As search strategy we choose a random search.

For all available multi-criteria tuning algorithms see TuneMultiCritControl.

```
ps = makeParamSet(
 makeNumericParam("C", lower = -12, upper = 12, trafo = function(x)
 makeNumericParam("sigma", lower = -12, upper = 12, trafo =
      function(x) 2^x)
)
ctrl = makeTuneMultiCritControlRandom(maxit = 30L)
rdesc = makeResampleDesc("Holdout")
res = tuneParamsMultiCrit("classif.ksvm", task = sonar.task,
    resampling = rdesc, par.set = ps,
 measures = list(fpr, fnr), control = ctrl, show.info = FALSE)
res
#> Tune multicrit result:
#> Points on front: 6
head(as.data.frame(trafoOptPath(res$opt.path)))
                C
                         sigma fpr.test.mean fnr.test.mean dob eol
#> 1 0.0559057771 2.720696e-04
                               0.03571429
                                                 0.5714286
                                                             1 NA
#> 2 0.3044391309 2.200614e-03
                                  0.32142857
                                                 0.3809524
                                                             2 NA
#> 3 5.1454078985 9.779024e-03
                                0.17857143
                                                 0.1666667
                                                             3 NA
#> 4 6.1938005681 6.443940e+01
                                  0.89285714
                                                 0.0000000
                                                                NA
#> 5 8.1097955432 3.142647e+00
                                  0.00000000
                                                 1.0000000
                                                             5 NA
#> 6 0.0004027434 5.415789e-04
                                  0.03571429
                                                 0.5714286
                                                             6 NA
#>
     error.message exec.time
#> 1
             <NA>
                       0.063
              <NA>
#> 2
                       0.073
#> 3
              <NA>
                       0.070
```

The results can be visualized with function plotTuneMultiCritResult. The plot shows the false positive and false negative rates for all parameter settings evaluated during tuning. Points on the Pareto front are slightly increased.

plotTuneMultiCritResult(res)



3.7 Feature Selection

Often, data sets include a large number of features. The technique of extracting a subset of relevant features is called feature selection. Feature selection can enhance the interpretability of the model, speed up the learning process and

improve the learner performance. There exist different approaches to identify the relevant features. mlr supports filter and wrapper methods.

3.7.1 Filter methods

Filter methods assign an importance value to each feature. Based on these values the features can be ranked and a feature subset can be selected.

3.7.1.1 Calculating the feature importance

Different methods for calculating the feature importance are built into mlr's function generateFilterValuesData (getFilterValues has been deprecated in favor of generateFilterValuesData.). Currently, classification, regression and survival analysis tasks are supported. A table showing all available methods can be found here.

Function generateFilterValuesData requires the Task and a character string specifying the filter method.

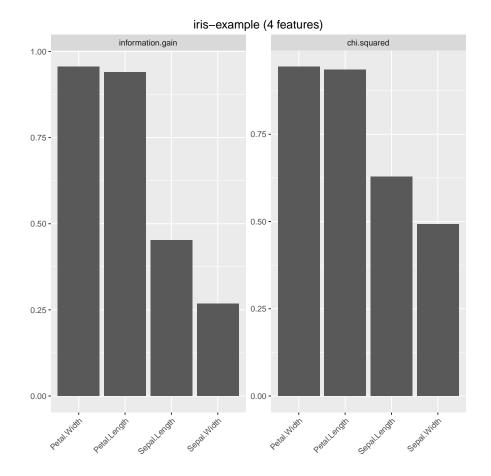
```
fv = generateFilterValuesData(iris.task, method = "information.gain")
fv

#> FilterValues:
#> Task: iris-example
#> name type information.gain
#> 1 Sepal.Length numeric 0.4521286
#> 2 Sepal.Width numeric 0.2672750
#> 3 Petal.Length numeric 0.9402853
#> 4 Petal.Width numeric 0.9554360
```

fv is a FilterValues object and fv\$data contains a data.frame that gives the importance values for all features. Optionally, a vector of filter methods can be passed.

A bar plot of importance values for the individual features can be obtained using function plotFilterValues.

```
plotFilterValues(fv2)
```



By default plotFilterValues will create facetted subplots if multiple filter methods are passed as input to generateFilterValuesData.

There is also an experimental ggvis plotting function, plotFilterValuesGGVIS. This takes the same arguments as plotFilterValues and produces a shiny application that allows the interactive selection of the displayed filter method, the number of features selected, and the sorting method (e.g., ascending or descending).

plotFilterValuesGGVIS(fv2)

According to the "information.gain" measure, Petal.Width and Petal.Length contain the most information about the target variable Species.

3.7.1.2 Selecting a feature subset

With mlr's function filterFeatures you can create a new Task by leaving out features of lower importance.

There are several ways to select a feature subset based on feature importance values:

- Keep a certain absolute number (abs) of features with highest importance.
- Keep a certain *percentage* (perc) of features with highest importance.
- Keep all features whose importance exceeds a certain *threshold value* (threshold).

Function filterFeatures supports these three methods as shown in the following example. Moreover, you can either specify the method for calculating the feature importance or you can use previously computed importance values via argument fval.

```
### Keep the 2 most important features
filtered.task = filterFeatures(iris.task, method =
    "information.gain", abs = 2)
### Keep the 25% most important features
filtered.task = filterFeatures(iris.task, fval = fv, perc = 0.25)
### Keep all features with importance greater than 0.5
filtered.task = filterFeatures(iris.task, fval = fv, threshold = 0.5)
filtered.task
#> Supervised task: iris-example
#> Type: classif
#> Target: Species
#> Observations: 150
#> Features:
#> numerics factors ordered
   2
#> Missings: FALSE
#> Has weights: FALSE
#> Has blocking: FALSE
#> Classes: 3
#>
       setosa versicolor virginica
         50
               50
#> Positive class: NA
```

3.7.1.3 Fuse a learner with a filter method

Often feature selection based on a filter method is part of the data preprocessing and in a subsequent step a learning method is applied to the filtered data. In a proper experimental setup you might want to automate the selection of the features so that it can be part of the validation method of your choice. A Learner can be fused with a filter method by function makeFilterWrapper. The resulting Learner has the additional class attribute FilterWrapper.

In the following example we calculate the 10-fold cross-validated error rate (mmce) of the k nearest neighbor classifier with preceding feature selection on the iris data set. We use "information.gain" as importance measure and select the 2 features with highest importance. In each resampling iteration feature selection is carried out on the corresponding training data set before fitting the learner.

You may want to know which features have been used. Luckily, we have called resample with the argument models = TRUE, which means that r\$models contains a list of models fitted in the individual resampling iterations. In order to access the selected feature subsets we can call getFilteredFeatures on each model.

```
sfeats = sapply(r$models, getFilteredFeatures)
table(sfeats)
#> sfeats
#> Petal.Length Petal.Width
#> 10 10
```

The selection of features seems to be very stable. The features Sepal.Length and Sepal.Width did not make it into a single fold.

3.7.1.4 Tuning the size of the feature subset

In the above examples the number/percentage of features to select or the threshold value have been arbitrarily chosen. If filtering is a preprocessing step before applying a learning method optimal values with regard to the learner performance can be found by tuning.

In the following regression example we consider the BostonHousing data set. We use a linear regression model and determine the optimal percentage value for feature selection such that the 3-fold cross-validated mean squared error of the learner is minimal. As search strategy for tuning a grid search is used.

3.7 Feature Selection 3 ADVANCED

```
res = tuneParams(lrn, task = bh.task, resampling = rdesc, par.set =
 control = makeTuneControlGrid())
#> [Tune] Started tuning learner regr.lm.filtered for parameter set:
               Type len Def
                                                    Constr Req
    Tunable Trafo
#> fw.perc discrete - - 0.2,0.25,0.3,0.35,0.4,0.45,0.5
    TRUE
#> With control class: TuneControlGrid
#> Imputation value: Inf
#> [Tune-x] 1: fw.perc=0.2
#> [Tune-y] 1: mse.test.mean=40.6; time: 0.0 min; memory: 131Mb use,
    643Mb max
#> [Tune-x] 2: fw.perc=0.25
#> [Tune-y] 2: mse.test.mean=40.6; time: 0.0 min; memory: 131Mb use,
    643Mb max
#> [Tune-x] 3: fw.perc=0.3
#> [Tune-y] 3: mse.test.mean=37.1; time: 0.0 min; memory: 131Mb use,
    643Mb max
\# [Tune-x] 4: fw.perc=0.35
#> [Tune-y] 4: mse.test.mean=35.8; time: 0.0 min; memory: 131Mb use,
    643Mb max
#> [Tune-x] 5: fw.perc=0.4
#> [Tune-y] 5: mse.test.mean=35.8; time: 0.0 min; memory: 131Mb use,
    643Mb max
\# [Tune-x] 6: fw.perc=0.45
#> [Tune-y] 6: mse.test.mean=27.4; time: 0.0 min; memory: 131Mb use,
    643Mb max
#> [Tune-x] 7: fw.perc=0.5
#> [Tune-y] 7: mse.test.mean=27.4; time: 0.0 min; memory: 131Mb use,
    643Mb max
#> [Tune] Result: fw.perc=0.5 : mse.test.mean=27.4
res
#> Tune result:
#> Op. pars: fw.perc=0.5
#> mse.test.mean=27.4
```

The performance of all percentage values visited during tuning is:

```
as.data.frame(res$opt.path)
    fw.perc mse.test.mean dob eol error.message exec.time
#>
                40.59578 1 NA
#> 1
       0.2
                                       <NA>
                                                 0.200
#> 2
       0.25
                40.59578 2 NA
                                        <NA>
                                                 0.196
#> 3
       0.3
                37.05592
                         3 NA
                                        <NA>
                                                 0.197
                         4 NA
#> 4
       0.35
                35.83712
                                        <NA>
                                                 0.195
#> 5 0.4 35.83712 5 NA
                                        <NA>
                                                0.194
```

The optimal percentage and the corresponding performance can be accessed as follows:

```
res$x
#> $fw.perc
#> [1] 0.5
res$y
#> mse.test.mean
#> 27.39955
```

After tuning we can generate a new wrapped learner with the optimal percentage value for further use.

```
lrn = makeFilterWrapper(learner = "regr.lm", fw.method =
    "chi.squared", fw.perc = res$x$fw.perc)
mod = train(lrn, bh.task)
mod
#> Model for learner.id=regr.lm.filtered; learner.class=FilterWrapper
#> Trained on: task.id = BostonHousing-example; obs = 506; features
    = 13
#> Hyperparameters: fw.method=chi.squared,fw.perc=0.5
getFilteredFeatures(mod)
#> [1] "crim" "zn" "rm" "dis" "rad" "lstat"
```

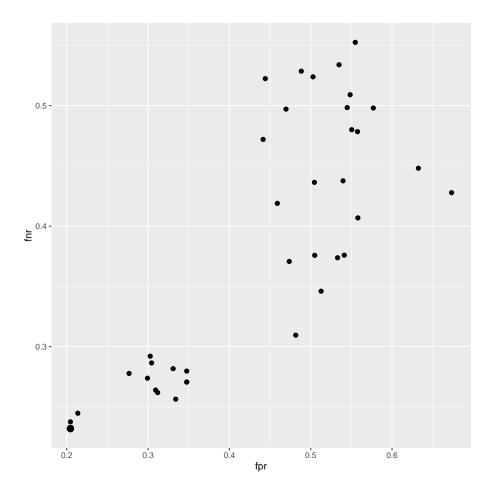
Here is another example using multi-criteria tuning. We consider linear discriminant analysis with precedent feature selection based on the Chi-squared statistic of independence ("chi.squared") on the Sonar data set and tune the threshold value. During tuning both, the false positive and the false negative rate (fpr and fnr), are minimized. As search strategy we choose a random search (see makeTuneMultiCritControlRandom).

7 Feature Selection 3 ADVANCED

```
#> Points on front: 13
head(as.data.frame(res$opt.path))
     fw.threshold fpr.test.mean fnr.test.mean dob eol error.message
    exec.time
        0.4892321
#> 1
                      0.3092818
                                     0.2639033
                                                    NA
                                                                 <NA>
                                                 1
        2.128
#> 2
        0.2481696
                      0.2045499
                                     0.2319697
                                                   NA
                                                                 <NA>
        2.072
#> 3
        0.7691875
                      0.5128000
                                     0.3459740
                                                    NA
                                                                 <NA>
        1.856
#> 4
        0.1470133
                      0.2045499
                                     0.2319697
                                                    NA
                                                                 <NA>
        2.133
#> 5
        0.5958241
                      0.5028216
                                     0.5239538
                                                                 <NA>
        1.907
#> 6
        0.6892421
                      0.6323959
                                     0.4480808
                                                 6 NA
                                                                 <NA>
        1.911
```

The results can be visualized with function plotTuneMultiCritResult. The plot shows the false positive and false negative rates for all parameter values visited during tuning. The size of the points on the Pareto front is slightly increased.

plotTuneMultiCritResult(res)



3.7.2 Wrapper methods

Wrapper methods use the performance of a learning algorithm to assess the usefulness of a feature set. In order to select a feature subset a learner is trained repeatedly on different feature subsets and the subset which leads to the best learner performance is chosen.

In order to use the wrapper approach we have to decide:

- How to assess the performance: This involves choosing a performance measure that serves as feature selection criterion and a resampling strategy.
- Which learning method to use.
- How to search the space of possible feature subsets.

The search strategy is defined by functions following the naming convention

7 Feature Selection 3 ADVANCED

makeFeatSelControl<search_strategy. The following search strategies are available:

- Exhaustive search (makeFeatSelControlExhaustive),
- Genetic algorithm (makeFeatSelControlGA),
- Random search (makeFeatSelControlRandom),
- Deterministic forward or backward search (makeFeatSelControlSequential).

3.7.2.1 Select a feature subset

Feature selection can be conducted with function selectFeatures.

In the following example we perform an exhaustive search on the Wisconsin Prognostic Breast Cancer data set. As learning method we use the Cox proportional hazards model. The performance is assessed by the holdout estimate of the concordance index (cindex).

```
### Specify the search strategy
ctrl = makeFeatSelControlRandom(maxit = 20L)
ctrl

#> FeatSel control: FeatSelControlRandom

#> Same resampling instance: TRUE

#> Imputation value: <worst>

#> Max. features: <not used>

#> Max. iterations: 20

#> Tune threshold: FALSE

#> Further arguments: prob=0.5
```

ctrl is a FeatSelControl object that contains information about the search strategy and potential parameter values.

Feature Selection 3 ADVANCED

sfeats is a FeatSelResult object. The selected features and the corresponding performance can be accessed as follows:

```
sfeats$x
   [1] "mean_radius"
                               "mean_area"
                                                     "mean_smoothness"
   [4] "mean_concavepoints"
                              "mean_symmetry"
                                                     "mean_fractaldim"
#> [7] "SE texture"
                               "SE perimeter"
                                                     "SE smoothness"
#> [10] "SE compactness"
                               "SE_concavity"
    "SE_concavepoints"
#> [13] "worst_area"
                               "worst\_compactness"
    "worst_concavepoints"
#> [16] "tsize"
                               "pnodes"
sfeats$y
#> cindex.test.mean
           0.713799
```

In a second example we fit a simple linear regression model to the BostonHousing data set and use a sequential search to find a feature set that minimizes the mean squared error (mse). method = "sfs" indicates that we want to conduct a sequential forward search where features are added to the model until the performance cannot be improved anymore. See the documentation page make-FeatSelControlSequential for other available sequential search methods. The search is stopped if the improvement is smaller than alpha = 0.02.

```
### Specify the search strategy
ctrl = makeFeatSelControlSequential(method = "sfs", alpha = 0.02)

### Select features
rdesc = makeResampleDesc("CV", iters = 10)
sfeats = selectFeatures(learner = "regr.lm", task = bh.task,
    resampling = rdesc, control = ctrl,
    show.info = FALSE)
sfeats
#> FeatSel result:
#> Features (11): crim, zn, chas, nox, rm, dis, rad, tax, ptratio,
    b, lstat
#> mse.test.mean=23.7
```

Further information about the sequential feature selection process can be obtained by function analyzeFeatSelResult.

```
analyzeFeatSelResult(sfeats)
#> Features : 11
#> Performance : mse.test.mean=23.7
#> crim, zn, chas, nox, rm, dis, rad, tax, ptratio, b, lstat
#>
#> Path to optimum:
```

3 ADVANCED

```
#> - Features:
                 0 Init
                                                    Perf = 84.831
   Diff: NA *
#> - Features:
                  1
                    Add
                            : lstat
                                                    Perf = 38.894
   Diff: 45.936 *
#> - Features:
                 2
                                                    Perf = 31.279
                    Add
   Diff: 7.6156 *
#> - Features:
                 3
                    Add
                            : ptratio
                                                    Perf = 28.108
   Diff: 3.1703 *
#> - Features:
                    Add
                            : dis
                                                    Perf = 27.48
                 4
   Diff: 0.62813 *
                                                    Perf = 26.079
#> - Features:
                 5
                    Add
                            : nox
   Diff: 1.4008 *
#> - Features:
                 6
                                                    Perf = 25.563
                   Add
   Diff: 0.51594 *
#> - Features:
                 7 Add
                            : chas
                                                    Perf = 25.132
   Diff: 0.43097 *
                                                    Perf = 24.792
#> - Features:
                 8 Add
                            : zn
   Diff: 0.34018 *
#> - Features:
                 9 Add
                                                    Perf = 24.599
                            : rad
   Diff: 0.19327 *
                                                    Perf = 24.082
#> - Features:
                10 Add
                            : tax
   Diff: 0.51706 *
#> - Features:
                11 \quad Add
                                                    Perf = 23.732
                            : crim
   Diff: 0.35 *
#> Stopped, because no improving feature was found.
```

3.7.2.2 Fuse a learner with feature selection

A Learner can be fused with a feature selection strategy (i.e., a search strategy, a performance measure and a resampling strategy) by function makeFeatSelWrapper. During training features are selected according to the specified selection scheme. Then, the learner is trained on the selected feature subset.

```
rdesc = makeResampleDesc("CV", iters = 3)
lrn = makeFeatSelWrapper("surv.coxph", resampling = rdesc,
    control = makeFeatSelControlRandom(maxit = 10), show.info = FALSE)
mod = train(lrn, task = wpbc.task)
mod
#> Model for learner.id=surv.coxph.featsel;
    learner.class=FeatSelWrapper
#> Trained on: task.id = wpbc-example; obs = 194; features = 32
#> Hyperparameters:
```

The result of the feature selection can be extracted by function getFeatSelResult.

3.7 Feature Selection 3 ADVANCED

```
sfeats = getFeatSelResult(mod)
sfeats
#> FeatSel result:
#> Features (19): mean_radius, mean_texture, mean_perimeter,
    mean_area, mean_smoothness, mean_compactness,
    mean_concavepoints, mean_fractaldim, SE_compactness,
    SE_concavity, SE_concavepoints, SE_symmetry, worst_texture,
    worst_perimeter, worst_area, worst_concavepoints,
    worst_symmetry, tsize, pnodes
#> cindex.test.mean=0.631
```

The selected features are:

```
sfeats$x
#> [1] "mean_radius"
                                "mean_texture"
                                                         "mean_perimeter"
#> [4] "mean_area"
                                "mean_smoothness"
    "mean_compactness"
#> [7] "mean_concavepoints" "mean_fractaldim"
                                                         "SE_compactness"
#> [10] "SE_concavity" "SE_concavepoints"
#> [13] "worst_texture" "worst_nerimeter"
                                                        "SE_symmetry"
#> [13] "worst_texture"
                                "worst_perimeter"
                                                        "worst area"
#> [16] "worst_concavepoints" "worst_symmetry"
                                                        "tsize"
#> [19] "pnodes"
```

The 5-fold cross-validated performance of the learner specified above can be computed as follows:

```
out.rdesc = makeResampleDesc("CV", iters = 5)

r = resample(learner = lrn, task = wpbc.task, resampling = out.rdesc, models = TRUE, show.info = FALSE)

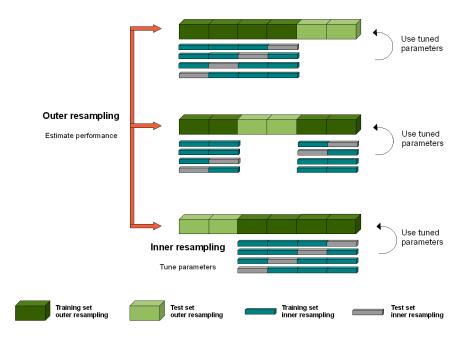
r$aggr
#> cindex.test.mean
#> 0.632357
```

The selected feature sets in the individual resampling iterations can be extracted as follows:

```
#>
#> [[2]]
#> FeatSel result:
#> Features (12): mean_area, mean_compactness, mean_symmetry,
    mean_fractaldim, SE_perimeter, SE_area, SE_concavity,
    SE_symmetry, worst_texture, worst_smoothness, worst_fractaldim,
    tsize
#> cindex.test.mean=0.652
#>
#> [[3]]
#> FeatSel result:
#> Features (14): mean_compactness, mean_symmetry, mean_fractaldim,
    SE_radius, SE_perimeter, SE_smoothness, SE_concavity,
    SE_concavepoints, SE_fractaldim, worst_concavity,
    worst_concavepoints, worst_symmetry, worst_fractaldim, pnodes
#> cindex.test.mean=0.607
#>
#> [[4]]
#> FeatSel result:
#> Features (18): mean_radius, mean_texture, mean_perimeter,
    mean_compactness, mean_concavity, SE_texture, SE_area,
    SE_smoothness, SE_concavity, SE_symmetry, SE_fractaldim,
    worst_radius, worst_compactness, worst_concavepoints,
    worst symmetry, worst fractaldim, tsize, pnodes
#> cindex.test.mean=0.653
#>
#> [[5]]
#> FeatSel result:
#> Features (14): mean_radius, mean_texture, mean_compactness,
    mean_concavepoints, mean_symmetry, SE_texture, SE_compactness,
    SE_symmetry, SE_fractaldim, worst_radius, worst_smoothness,
    worst_compactness, worst_concavity, pnodes
#> cindex.test.mean=0.626
```

3.8 Nested Resampling

In order to obtain honest performance estimates for a learner all parts of the model building like preprocessing and model selection steps should be included in the resampling, i.e., repeated for every pair of training/test data. For steps that themselves require resampling like parameter tuning or feature selection (via the wrapper approach) this results in two nested resampling loops.



The graphic above illustrates nested resampling for parameter tuning with 3-fold cross-validation in the outer and 4-fold cross-validation in the inner loop.

In the outer resampling loop, we have three pairs of training/test sets. On each of these outer training sets parameter tuning is done, thereby executing the inner resampling loop. This way, we get one set of selected hyperparameters for each outer training set. Then the learner is fitted on each outer training set using the corresponding selected hyperparameters and its performance is evaluated on the outer test sets.

In mlr, you can get nested resampling for free without programming any looping by using the wrapper functionality. This works as follows:

- 1. Generate a wrapped Learner via function makeTuneWrapper or makeFeat-SelWrapper. Specify the inner resampling strategy using their resampling argument.
- 2. Call function resample (see also the section about resampling) and pass the outer resampling strategy to its resampling argument.

You can freely combine different inner and outer resampling strategies.

The outer strategy can be a resample description (ResampleDesc) or a resample instance (ResampleInstance). A common setup is prediction and performance evaluation on a fixed outer test set. This can be achieved by using function makeFixedHoldoutInstance to generate the outer ResampleInstance.

The inner resampling strategy should preferably be a ResampleDesc, as the sizes of the outer training sets might differ. Per default, the inner resample description is instantiated once for every outer training set. This way during tuning/feature selection all parameter or feature sets are compared on the same inner training/test sets to reduce variance. You can also turn this off using the same.resampling.instance argument of makeTuneControl* or makeFeatSel-Control*.

Nested resampling is computationally expensive. For this reason in the examples shown below we use relatively small search spaces and a low number of resampling iterations. In practice, you normally have to increase both. As this is computationally intensive you might want to have a look at section parallelization.

3.8.1 Tuning

As you might recall from the tutorial page about tuning, you need to define a search space by function makeParamSet, a search strategy by makeTuneControl*, and a method to evaluate hyperparameter settings (i.e., the inner resampling strategy and a performance measure).

Below is a classification example. We evaluate the performance of a support vector machine (ksvm) with tuned cost parameter C and RBF kernel parameter sigma. We use 3-fold cross-validation in the outer and subsampling with 2 iterations in the inner loop. For tuning a grid search is used to find the hyperparameters with lowest error rate (mmce is the default measure for classification). The wrapped Learner is generated by calling makeTuneWrapper.

Note that in practice the parameter set should be larger. A common recommendation is 2^(-12:12) for both C and sigma.

```
#> Resample Result
#> Task: iris-example
#> Learner: classif.ksvm.tuned
#> Aggr perf: mmce.test.mean=0.0533
#> Runtime: 20.1499
```

You can obtain the error rates on the 3 outer test sets by:

```
r$measures.test

#> iter mmce

#> 1  1  0.02

#> 2  2  0.06

#> 3  3  0.08
```

3.8.1.1 Accessing the tuning result

We have kept the results of the tuning for further evaluations. For example one might want to find out, if the best obtained configurations vary for the different outer splits. As storing entire models may be expensive (but possible by setting models = TRUE) we used the extract option of resample. Function getTuneResult returns, among other things, the optimal hyperparameter values and the optimization path for each iteration of the outer resampling loop. Note that the performance values shown when printing r\$extract are the aggregated performances resulting from inner resampling on the outer training set for the best hyperparameter configurations (not to be confused with r\$measures.test shown above).

```
r$extract
#> [[1]]
#> Tune result:
#> Op. pars: C=2; sigma=0.25
#> mmce.test.mean=0.0147
#>
#> [[2]]
#> Tune result:
#> Op. pars: C=4; sigma=0.25
#> mmce.test.mean= 0
#>
#> [[3]]
#> Tune result:
#> Op. pars: C=4; sigma=0.25
#> mmce.test.mean=0.0735
names(r$extract[[1]])
#> [1] "learner"
                  "control"
                                            "y"
                                                        "threshold"
   "opt.path"
```

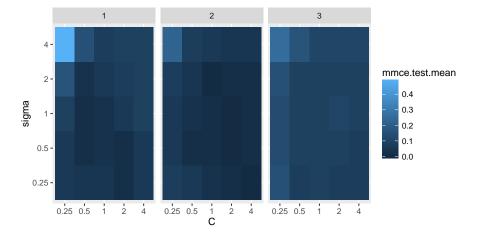
We can compare the optimal parameter settings obtained in the 3 resampling iterations. As you can see, the optimal configuration usually depends on the data. You may be able to identify a *range* of parameter settings that achieve good performance though, e.g., the values for C should be at least 1 and the values for sigma should be between 0 and 1.

With function getNestedTuneResultsOptPathDf you can extract the optimization paths for the 3 outer cross-validation iterations for further inspection and analysis. These are stacked in one data.frame with column iter indicating the resampling iteration.

```
opt.paths = getNestedTuneResultsOptPathDf(r)
head(opt.paths, 10)
          C sigma mmce.test.mean dob eol error.message exec.time iter
#> 1
      0.25
           0.25
                                                               0.044
                       0.05882353
                                     1
                                        NA
                                                     <NA>
                                                                         1
                      0.04411765
#> 2
       0.5
            0.25
                                     2
                                        NA
                                                     <NA>
                                                               0.044
                                                                         1
#> 3
                                        NA
                                                               0.046
          1
             0.25
                       0.04411765
                                     3
                                                     <NA>
                                                                         1
#> 4
          2
             0.25
                       0.01470588
                                        NA
                                                     <NA>
                                                               0.043
                                                                         1
             0.25
                       0.05882353
                                        NA
#> 5
                                     5
                                                     <NA>
                                                               0.044
                                                                         1
#> 6
      0.25
              0.5
                       0.05882353
                                     6
                                        NA
                                                     <NA>
                                                               0.045
                                                                         1
       0.5
#> 7
              0.5
                       0.01470588
                                     7
                                        NA
                                                     <NA>
                                                               0.044
                                                                         1
#> 8
          1
              0.5
                       0.02941176
                                     8
                                        NA
                                                     <NA>
                                                               0.046
                                                                         1
          2
#> 9
              0.5
                                                               0.045
                       0.01470588
                                     9
                                        NA
                                                     <NA>
                                                                         1
                                   10
#> 10
          4
              0.5
                      0.05882353
                                       NA
                                                     <NA>
                                                               0.044
                                                                         1
```

Below we visualize the opt.paths for the 3 outer resampling iterations.

```
g = ggplot(opt.paths, aes(x = C, y = sigma, fill = mmce.test.mean))
g + geom_tile() + facet_wrap(~ iter)
```



Another useful function is getNestedTuneResultsX, which extracts the best found hyperparameter settings for each outer resampling iteration.

```
getNestedTuneResultsX(r)

#> C sigma

#> 1 2 0.25

#> 2 4 0.25

#> 3 4 0.25
```

3.8.2 Feature selection

As you might recall from the section about feature selection, mlr supports the filter and the wrapper approach.

3.8.2.1 Wrapper methods

Wrapper methods use the performance of a learning algorithm to assess the usefulness of a feature set. In order to select a feature subset a learner is trained repeatedly on different feature subsets and the subset which leads to the best learner performance is chosen.

For feature selection in the inner resampling loop, you need to choose a search strategy (function makeFeatSelControl*), a performance measure and the inner resampling strategy. Then use function makeFeatSelWrapper to bind everything together.

Below we use sequential forward selection with linear regression on the Boston-Housing data set (bh.task).

```
### Feature selection in inner resampling loop
inner = makeResampleDesc("CV", iters = 3)
lrn = makeFeatSelWrapper("regr.lm", resampling = inner,
 control = makeFeatSelControlSequential(method = "sfs"), show.info
      = FALSE)
### Outer resampling loop
outer = makeResampleDesc("Subsample", iters = 2)
r = resample(learner = lrn, task = bh.task, resampling = outer,
    extract = getFeatSelResult,
 show.info = FALSE)
#> Resample Result
#> Task: BostonHousing-example
#> Learner: regr.lm.featsel
#> Aggr perf: mse.test.mean=31.7
#> Runtime: 45.3589
r$measures.test
```

3.8.2.1.1 Accessing the selected features

The result of the feature selection can be extracted by function getFeatSelResult. It is also possible to keep whole models by setting models = TRUE when calling resample.

```
r$extract
#> [[1]]
#> FeatSel result:
#> Features (10): crim, zn, indus, nox, rm, dis, rad, tax, ptratio,
    lstat
#> mse.test.mean=20.2
#>
#> [[2]]
#> FeatSel result:
#> Features (9): zn, nox, rm, dis, rad, tax, ptratio, b, lstat
#> mse.test.mean=22.6
### Selected features in the first outer resampling iteration
r$extract[[1]]$x
   [1] "crim"
                  "zn"
                            "indus"
                                      "nox"
                                                "rm"
                                                          "dis"
    "rad"
  [8] "tax"
                  "ptratio" "lstat"
### Resampled performance of the selected feature subset on the
   first inner training set
r$extract[[1]]$y
#> mse.test.mean
#> 20.15939
```

As for tuning, you can extract the optimization paths. The resulting data frames contain, among others, binary columns for all features, indicating if they were included in the linear regression model, and the corresponding performances.

```
opt.paths = lapply(r$extract, function(x) as.data.frame(x$opt.path))
head(opt.paths[[1]])
    crim zn indus chas nox rm age dis rad tax ptratio b lstat
   mse.test.mean
#> 1
     0 0
                     0
                         0 0
                                0
                                    0
                                       0
                                           0
                                                   00
                                                           0
                0
   80.33019
#> 2
      1 0
                0
                     0
                         0 0
                                0
                                  0
                                       0
                                          0
                                                   0 0
                                                           0
   65.95316
```

```
#> 3 0 1
                    0
                       0 0
   69.15417
#> 4
     0 0
                                         0
                                                 0 0
                                                        0
   55.75473
#> 5 0 0
                                                 0 0
                                                        0
   80.48765
#> 6 0 0
                                                 00
   63.06724
    dob eol error.message exec.time
#> 1
      1
                    <NA>
                            0.021
         2
#> 2
     2 2
                    <NA>
                            0.034
#> 3
     2 2
                            0.035
                    <NA>
#> 4
     2 2
                    <NA>
                            0.034
      2 2
#> 5
                    <NA>
                            0.038
#> 6
      2 2
                    <NA>
                            0.033
```

An easy-to-read version of the optimization path for sequential feature selection can be obtained with function analyzeFeatSelResult.

```
analyzeFeatSelResult(r$extract[[1]])
#> Features
#> Performance
                   : mse.test.mean=20.2
#> crim, zn, indus, nox, rm, dis, rad, tax, ptratio, lstat
#>
#> Path to optimum:
#> - Features:
                 0 Init
                                                   Perf = 80.33
   Diff: NA *
#> - Features:
                                                   Perf = 36.451
                 1 \quad Add
                            : lstat
   Diff: 43.879 *
#> - Features:
                 2
                                                   Perf = 27.289
                    Add
                            : rm
   Diff: 9.1623 *
                                                   Perf = 24.004
#> - Features:
                 3
                    Add
                            : ptratio
   Diff: 3.2849 *
                 4
#> - Features:
                    Add
                                                   Perf = 23.513
                            : nox
   Diff: 0.49082 *
#> - Features:
                 5 Add
                            : dis
                                                   Perf = 21.49
   Diff: 2.023 *
#> - Features:
                 6 Add
                            : crim
                                                   Perf = 21.12
   Diff: 0.37008 *
#> - Features:
                 7 Add
                                                   Perf = 20.82
                            : indus
   Diff: 0.29994 *
#> - Features:
                 8 Add
                                                   Perf = 20.609
                            : rad
   Diff: 0.21054 *
#> - Features:
                 9 Add
                                                   Perf = 20.209
                            : tax
   Diff: 0.40059 *
#> - Features: 10 Add
                                                   Perf = 20.159
                            : zn
```

```
Diff: 0.049441 *

#>
#> Stopped, because no improving feature was found.
```

3.8.2.2 Filter methods with tuning

Filter methods assign an importance value to each feature. Based on these values you can select a feature subset by either keeping all features with importance higher than a certain threshold or by keeping a fixed number or percentage of the highest ranking features. Often, neither the theshold nor the number or percentage of features is known in advance and thus tuning is necessary.

In the example below the threshold value (fw.threshold) is tuned in the inner resampling loop. For this purpose the base Learner "regr.lm" is wrapped two times. First, makeFilterWrapper is used to fuse linear regression with a feature filtering preprocessing step. Then a tuning step is added by makeTuneWrapper.

```
### Tuning of the percentage of selected filters in the inner loop
lrn = makeFilterWrapper(learner = "regr.lm", fw.method =
    "chi.squared")
ps = makeParamSet(makeDiscreteParam("fw.threshold", values = seq(0,
    1, 0.2)))
ctrl = makeTuneControlGrid()
inner = makeResampleDesc("CV", iters = 3)
lrn = makeTuneWrapper(lrn, resampling = inner, par.set = ps, control
    = ctrl, show.info = FALSE)
### Outer resampling loop
outer = makeResampleDesc("CV", iters = 3)
r = resample(learner = lrn, task = bh.task, resampling = outer,
   models = TRUE, show.info = FALSE)
#> Resample Result
#> Task: BostonHousing-example
#> Learner: regr.lm.filtered.tuned
#> Aggr perf: mse.test.mean=25.4
#> Runtime: 7.84384
```

3.8.2.2.1 Accessing the selected features and optimal percentage

In the above example we kept the complete models.

Below are some examples that show how to extract information from the models.

```
r$models
#> [[1]]
```

```
#> Model for learner.id=reqr.lm.filtered.tuned;
    learner.class=TuneWrapper
#> Trained on: task.id = BostonHousing-example; obs = 337; features
#> Hyperparameters: fw.method=chi.squared
#>
#> [[2]]
#> Model for learner.id=regr.lm.filtered.tuned;
    learner.class=TuneWrapper
#> Trained on: task.id = BostonHousing-example; obs = 338; features
#> Hyperparameters: fw.method=chi.squared
#>
#> [[3]]
#> Model for learner.id=reqr.lm.filtered.tuned;
    learner.\,class = Tune \textit{W} rapper
#> Trained on: task.id = BostonHousing-example; obs = 337; features
    = 13
#> Hyperparameters: fw.method=chi.squared
```

The result of the feature selection can be extracted by function getFilteredFeatures. Almost always all 13 features are selected.

```
lapply(r$models, function(x)
    getFilteredFeatures(x$learner.model$next.model))
#> [[1]]
#> [1] "crim"
                   "zn"
                             "indus"
                                        "chas"
                                                  "nox"
                                                             "rm"
    "age"
   [8] "dis"
                   "rad"
                             "tax"
                                        "ptratio" "b"
                                                             "lstat"
#>
#>
#> [[2]]
                   "zn"
  [1] "crim"
                             "indus"
                                        "nox"
                                                  "rm"
                                                             "age"
    "dis"
   [8] "rad"
                             "ptratio" "b"
                   "tax"
                                                  "lstat"
#>
#>
#> [[3]]
   [1] "crim"
                   "zn"
                             "indus"
                                        "chas"
                                                   "nox"
                                                             "rm"
    "age"
                   "rad"
  [8] "dis"
                             "tax"
                                        "ptratio" "b"
                                                             "lstat"
```

Below the tune results and optimization paths are accessed.

```
res = lapply(r$models, getTuneResult)
res
#> [[1]]
#> Tune result:
#> Op. pars: fw.threshold=0
```

```
#> mse.test.mean=24.9
#>
#> [[2]]
#> Tune result:
#> Op. pars: fw.threshold=0.4
#> mse.test.mean=27.2
#>
#> [[3]]
#> Tune result:
#> Op. pars: fw.threshold=0
#> mse.test.mean=19.7
opt.paths = lapply(res, function(x) as.data.frame(x$opt.path))
opt.paths[[1]]
    fw.threshold mse.test.mean dob eol error.message exec.time
#> 1
              0
                      24.89160 1 NA
                                                <NA>
                                                         0.214
#> 2
                      25.18817 2 NA
             0.2
                                                <NA>
                                                         0.211
#> 3
                      25.18817 3 NA
                                                <NA>
                                                         0.207
             0.4
#> 4
             0.6
                      32.15930 4 NA
                                                <NA>
                                                         0.202
#> 5
             0.8
                      90.89848 5 NA
                                                <NA>
                                                         0.196
#> 6
                1
                      90.89848 6 NA
                                                 <NA>
                                                         0.187
```

3.8.3 Benchmark experiments

In a benchmark experiment multiple learners are compared on one or several tasks (see also the section about benchmarking). Nested resampling in benchmark experiments is achieved the same way as in resampling:

- First, use makeTuneWrapper or makeFeatSelWrapper to generate wrapped Learners with the inner resampling strategies of your choice.
- Second, call benchmark and specify the outer resampling strategies for all tasks.

The inner resampling strategies should be resample descriptions. You can use different inner resampling strategies for different wrapped learners. For example it might be practical to do fewer subsampling or bootstrap iterations for slower learners.

If you have larger benchmark experiments you might want to have a look at the section about parallelization.

As mentioned in the section about benchmark experiments you can also use different resampling strategies for different learning tasks by passing a list of resampling descriptions or instances to benchmark.

We will see three examples to show different benchmark settings:

- 1. Two data sets + two classification algorithms + tuning
- 2. One data set + two regression algorithms + feature selection
- 3. One data set + two regression algorithms + feature filtering + tuning

3.8.3.1 Example 1: Two tasks, two learners, tuning

Below is a benchmark experiment with two data sets, iris and sonar, and two Learners, ksvm and kknn, that are both tuned.

As inner resampling strategies we use holdout for ksvm and subsampling with 3 iterations for kknn. As outer resampling strategies we take holdout for the iris and bootstrap with 2 iterations for the sonar data (sonar.task). We consider the accuracy (acc), which is used as tuning criterion, and also calculate the balanced error rate (ber).

```
### List of learning tasks
tasks = list(iris.task, sonar.task)
### Tune sum in the inner resampling loop
ps = makeParamSet(
 makeDiscreteParam("C", 2^(-1:1)),
 makeDiscreteParam("sigma", 2^(-1:1)))
ctrl = makeTuneControlGrid()
inner = makeResampleDesc("Holdout")
lrn1 = makeTuneWrapper("classif.ksvm", resampling = inner, par.set =
    ps, control = ctrl,
 show.info = FALSE)
### Tune k-nearest neighbor in inner resampling loop
ps = makeParamSet(makeDiscreteParam("k", 3:5))
ctrl = makeTuneControlGrid()
inner = makeResampleDesc("Subsample", iters = 3)
lrn2 = makeTuneWrapper("classif.kknn", resampling = inner, par.set =
   ps, control = ctrl,
 show.info = FALSE)
#> Loading required package: kknn
#> Attaching package: 'kknn'
#> The following object is masked from 'package:caret':
#>
#>
       contr.dummy
### Learners
lrns = list(lrn1, lrn2)
### Outer resampling loop
```

```
outer = list(makeResampleDesc("Holdout"),
    makeResampleDesc("Bootstrap", iters = 2))
res = benchmark(lrns, tasks, outer, measures = list(acc, ber),
    show.info = FALSE)
res
#> task.id learner.id acc.test.mean ber.test.mean
#> 1 iris-example classif.ksvm.tuned 0.9400000 0.05882353
#> 2 iris-example classif.kknn.tuned 0.9200000 0.08683473
#> 3 Sonar-example classif.ksvm.tuned 0.5289307 0.50000000
#> 4 Sonar-example classif.kknn.tuned 0.8077080 0.19549714
```

The print method for the BenchmarkResult shows the aggregated performances from the outer resampling loop.

As you might recall, mlr offers several accessor function to extract information from the benchmark result. These are listed on the help page of BenchmarkResult and many examples are shown on the tutorial page about benchmark experiments.

The performance values in individual outer resampling runs can be obtained by getBMRPerformances. Note that, since we used different outer resampling strategies for the two tasks, the number of rows per task differ.

```
getBMRPerformances(res, as.df = TRUE)
#> task.id learner.id iter acc ber
#> 1 iris-example classif.ksvm.tuned 1 0.9400000 0.05882353
#> 2 iris-example classif.kknn.tuned 1 0.9200000 0.08683473
#> 3 Sonar-example classif.ksvm.tuned 1 0.5373134 0.50000000
#> 4 Sonar-example classif.ksvm.tuned 2 0.5205479 0.50000000
#> 5 Sonar-example classif.kknn.tuned 1 0.8208955 0.18234767
#> 6 Sonar-example classif.kknn.tuned 2 0.7945205 0.20864662
```

The results from the parameter tuning can be obtained through function ${\it getBM-RTuneResults}.$

```
getBMRTuneResults(res)
#> $`iris-example`$classif.ksvm.tuned
#> $`iris-example`$classif.ksvm.tuned[[1]]
#> Tune result:
#> Op. pars: C=0.5; sigma=0.5
#> mmce.test.mean=0.0588
#>
#>
#>
#>
$`iris-example`$classif.kknn.tuned
#> $`iris-example`$classif.kknn.tuned[[1]]
#> Tune result:
#> Op. pars: k=3
```

```
#> mmce.test.mean=0.049
#>
#>
#>
#> $ `Sonar-example`
#> $ `Sonar-example`$classif.ksvm.tuned
#> $`Sonar-example`$classif.ksvm.tuned[[1]]
#> Tune result:
#> Op. pars: C=1; sigma=2
#> mmce.test.mean=0.343
#> $`Sonar-example`$classif.ksvm.tuned[[2]]
#> Tune result:
#> Op. pars: C=2; sigma=0.5
#> mmce.test.mean= 0.2
#>
#>
#> $ `Sonar-example`$classif.kknn.tuned
#> $ `Sonar-example`$classif.kknn.tuned[[1]]
#> Tune result:
#> Op. pars: k=4
#> mmce.test.mean=0.11
#>
#> $`Sonar-example`$classif.kknn.tuned[[2]]
#> Tune result:
#> Op. pars: k=3
#> mmce.test.mean=0.0667
```

As for several other accessor functions a clearer representation as data.frame can be achieved by setting as.df = TRUE.

```
getBMRTuneResults(res, as.df = TRUE)
#>
          task.id learner.id iter C sigma mmce.test.mean
#> 1 iris-example classif.ksvm.tuned 1 0.5
                                               0.5
                                                      0.05882353
   NA
                                               NA
#> 2 iris-example classif.kknn.tuned
                                       1 NA
                                                      0.04901961
#> 3 Sonar-example classif.ksvm.tuned
                                       1 1.0
                                               2.0
                                                      0.34285714
#> 4 Sonar-example classif.ksvm.tuned
                                       2 2.0
                                               0.5
                                                       0.20000000
#> 5 Sonar-example classif.kknn.tuned
                                                       0.10952381
                                       1 NA
                                               NA
   4
#> 6 Sonar-example classif.kknn.tuned
                                       2 NA
                                               NA
                                                       0.0666667
```

It is also possible to extract the tuning results for individual tasks and learners and, as shown in earlier examples, inspect the optimization path.

```
tune.res = getBMRTuneResults(res, task.ids = "Sonar-example",
    learner.ids = "classif.ksvm.tuned",
    as.df = TRUE)
tune.res
#> task.id learner.id iter C sigma mmce.test.mean
#> 1 Sonar-example classif.ksvm.tuned 1 1 2.0 0.3428571
#> 2 Sonar-example classif.ksvm.tuned 2 2 0.5 0.2000000
getNestedTuneResultsOptPathDf(res$results[["Sonar-example"]][["classif.ksvm.tuned"]])
```

3.8.3.2 Example 2: One task, two learners, feature selection

Let's see how we can do feature selection in a benchmark experiment:

```
### Feature selection in inner resampling loop
ctrl = makeFeatSelControlSequential(method = "sfs")
inner = makeResampleDesc("Subsample", iters = 2)
lrn = makeFeatSelWrapper("regr.lm", resampling = inner, control =
    ctrl, show.info = FALSE)
### Learners
lrns = list(makeLearner("regr.rpart"), lrn)
### Outer resampling loop
outer = makeResampleDesc("Subsample", iters = 2)
res = benchmark(tasks = bh.task, learners = lrns, resampling =
    outer, show.info = FALSE)
res
#>
                                learner.id mse.test.mean
                   task.id
#> 1 BostonHousing-example
                                regr.rpart
                                                25.86232
#> 2 BostonHousing-example regr.lm.featsel
                                                25.07465
```

The selected features can be extracted by function getBMRFeatSelResults.

```
getBMRFeatSelResults(res)
#> $`BostonHousing-example`
#> NULL
#>
#> $`BostonHousing-example`$regr.rpart
#> NULL
#>
#> $`BostonHousing-example`$regr.lm.featsel
#> $`BostonHousing-example`$regr.lm.featsel[[1]]
#> FeatSel result:
#> Features (8): crim, zn, chas, nox, rm, dis, ptratio, lstat
```

```
#> mse.test.mean=26.7
#>
#> $`BostonHousing-example`$regr.lm.featsel[[2]]
#> FeatSel result:
#> Features (10): crim, zn, nox, rm, dis, rad, tax, ptratio, b, lstat
#> mse.test.mean=24.3
```

You can access results for individual learners and tasks and inspect them further.

```
feats = getBMRFeatSelResults(res, learner.id = "regr.lm.featsel")
feats = feats$`BostonHousing-example`$`regr.lm.featsel`
### Selected features in the first outer resampling iteration
feats[[1]]$x
                 "zn"
                                               "rm"
#> [1] "crim"
                           "chas"
                                     "nox"
                                                         "dis"
    "ptratio"
#> [8] "lstat"
### Resampled performance of the selected feature subset on the
    first inner training set
feats[[1]]$y
#> mse.test.mean
   26.72574
```

As for tuning, you can extract the optimization paths. The resulting data.frames contain, among others, binary columns for all features, indicating if they were included in the linear regression model, and the corresponding performances. analyzeFeatSelResult gives a clearer overview.

```
opt.paths = lapply(feats, function(x) as.data.frame(x$opt.path))
head(opt.paths[[1]])
    crim zn indus chas nox rm age dis rad tax ptratio b lstat
   mse.test.mean
#> 1
     0 0
                                       0
                                           0
                                                   0 0
                                                          0
                     0
                         0
                           0
                                   0
   90.16159
#> 2
                     0
                           0
                               0
                                   0
                                       0
                                           0
                                                   0 0
                                                          0
     1 0
   82.85880
#> 3 0 1
                0
                           0
                                   0
                                       0
                                           0
                                                   00
                                                          0
                               0
    79.55202
#> 4 0 0
                1
                     0
                           0
                                0
                                   0
                                       0
                                           0
                                                   0 0
                                                          0
    70.02071
#> 5
     0 0
                0
                           0
                               0
                                   0
                                           0
                                                   0 0
                                                           0
   86.93409
#> 6 0 0
                                                   00
                                                          0
   76.32457
   dob eol error.message exec.time
                     <NA> 0.017
#> 1 1 2
```

```
#> 2 2 2
                    <NA>
                            0.026
#> 3 2 2
                    <NA>
                            0.025
     2 2
#> 4
                    <NA>
                            0.026
#> 5
     2 2
                    <NA>
                            0.028
#> 6 2 2
                            0.025
                    <NA>
analyzeFeatSelResult(feats[[1]])
#> Features
                  : 8
#> Performance
                 : mse.test.mean=26.7
#> crim, zn, chas, nox, rm, dis, ptratio, lstat
#> Path to optimum:
#> - Features: 0 Init
                                               Perf = 90.162
   Diff: NA *
#> - Features: 1 Add
                         : lstat
                                               Perf = 42.646
  Diff: 47.515 *
#> - Features: 2 Add
                         : ptratio
                                               Perf = 34.52
   Diff: 8.1263 *
#> - Features: 3 Add
                                               Perf = 30.454
                         : rm
  Diff: 4.066 *
#> - Features: 4 Add
                                               Perf = 29.405
                          : dis
   Diff: 1.0495 *
#> - Features: 5 Add
                                               Perf = 28.059
                         : nox
  Diff: 1.3454 *
#> - Features: 6 Add
                                               Perf = 27.334
                         : chas
   Diff: 0.72499 *
#> - Features: 7 Add
                                               Perf = 26.901
                          : zn
   Diff: 0.43296 *
              8 Add
#> - Features:
                                               Perf = 26.726
                          : crim
   Diff: 0.17558 *
#> Stopped, because no improving feature was found.
```

3.8.3.3 Example 3: One task, two learners, feature filtering with tuning

Here is a minimal example for feature filtering with tuning of the feature subset size.

```
### Feature filtering with tuning in the inner resampling loop
lrn = makeFilterWrapper(learner = "regr.lm", fw.method =
    "chi.squared")
ps = makeParamSet(makeDiscreteParam("fw.abs", values =
    seq_len(getTaskNFeats(bh.task))))
ctrl = makeTuneControlGrid()
inner = makeResampleDesc("CV", iter = 2)
```

```
lrn = makeTuneWrapper(lrn, resampling = inner, par.set = ps, control
   = ctrl,
 show.info = FALSE)
### Learners
lrns = list(makeLearner("regr.rpart"), lrn)
### Outer resampling loop
outer = makeResampleDesc("Subsample", iter = 3)
res = benchmark(tasks = bh.task, learners = lrns, resampling =
    outer, show.info = FALSE)
res
#>
                   task.id
                                       learner.id mse.test.mean
#> 1 BostonHousing-example
                                       regr.rpart
                                                       22.11687
#> 2 BostonHousing-example regr.lm.filtered.tuned
                                                       23.76666
### Performances on individual outer test data sets
getBMRPerformances(res, as.df = TRUE)
#>
                  task.id
                                       learner.id iter
                                                            mse
#> 1 BostonHousing-example
                                       regr.rpart 1 23.55486
                                                     2 20.03453
#> 2 BostonHousing-example
                                       regr.rpart
#> 3 BostonHousing-example
                                       regr.rpart
                                                     3 22.76121
#> 4 BostonHousing-example regr.lm.filtered.tuned
                                                     1 27.51086
#> 5 BostonHousing-example regr.lm.filtered.tuned
                                                     2 24.87820
#> 6 BostonHousing-example regr.lm.filtered.tuned
                                                     3 18.91091
```

3.9 Cost-Sensitive Classification

In regular classification the aim is to minimize the misclassification rate and thus all types of misclassification errors are deemed equally severe. A more general setting is cost-sensitive classification where the costs caused by different kinds of errors are not assumed to be equal and the objective is to minimize the expected costs.

In case of class-dependent costs the costs depend on the true and predicted class label. The costs c(k,l) for predicting class k if the true label is l are usually organized into a $K \times K$ cost matrix where K is the number of classes. Naturally, it is assumed that the cost of predicting the correct class label y is minimal (that is $c(y,y) \le c(k,y)$ for all $k=1,\ldots,K$).

A further generalization of this scenario are example-dependent misclassification costs where each example (x, y) is coupled with an individual cost vector of length K. Its k-th component expresses the cost of assigning x to class k. A real-world example is fraud detection where the costs do not only depend on the true and predicted status fraud/non-fraud, but also on the amount of money involved in each case. Naturally, the cost of predicting the true class label y is assumed to be minimum. The true class labels are redundant information, as they can be easily inferred from the cost vectors. Moreover, given the cost vector, the expected costs do not depend on the true class label y. The classification problem is therefore completely defined by the feature values x and the corresponding cost vectors.

In the following we show ways to handle cost-sensitive classification problems in mlr. Some of the functionality is currently experimental, and there may be changes in the future.

3.9.1 Class-dependent misclassification costs

There are some classification methods that can accommodate misclassification costs directly. One example is rpart.

Alternatively, we can use cost-insensitive methods and manipulate the predictions or the training data in order to take misclassification costs into account. mlr supports thresholding and rebalancing.

- 1. **Thresholding**: The thresholds used to turn posterior probabilities into class labels are chosen such that the costs are minimized. This requires a Learner that can predict posterior probabilities. During training the costs are not taken into account.
- 2. **Rebalancing**: The idea is to change the proportion of the classes in the training data set in order to account for costs during training, either by weighting or by sampling. Rebalancing does not require that the Learner can predict probabilities.
 - For weighting we need a Learner that supports class weights or observation weights.
 - ii. If the Learner cannot deal with weights the proportion of classes can be changed by *over-* and *undersampling*.

We start with binary classification problems and afterwards deal with multi-class problems.

3.9.1.1 Binary classification problems

The positive and negative classes are labeled 1 and -1, respectively, and we consider the following cost matrix where the rows indicate true classes and the columns predicted classes:

true/pred.
$$+1$$
 -1
 $+1$ $c(+1,+1)$ $c(-1,+1)$
 -1 $c(+1,-1)$ $c(-1,-1)$

Often, the diagonal entries are zero or the cost matrix is rescaled to achieve zeros in the diagonal (see for example O'Brien et al, 2008).

A well-known cost-sensitive classification problem is posed by the German Credit data set (see also the UCI Machine Learning Repository). The corresponding cost matrix (though Elkan (2001) argues that this matrix is economically unreasonable) is given as:

true/pred.	Bad	Good
Bad	0	5
Good	1	0

As in the table above, the rows indicate true and the columns predicted classes.

In case of class-dependent costs it is sufficient to generate an ordinary Classif-Task. A CostSensTask is only needed if the costs are example-dependent. In the ${\bf R}$ code below we create the ClassifTask, remove two constant features from the data set and generate the cost matrix. Per default, Bad is the positive class.

```
data(GermanCredit, package = "caret")
credit.task = makeClassifTask(data = GermanCredit, target = "Class")
credit.task = removeConstantFeatures(credit.task)
#> Removing 2 columns: Purpose. Vacation, Personal. Female. Single
credit.task
#> Supervised task: GermanCredit
#> Type: classif
#> Target: Class
#> Observations: 1000
#> Features:
#> numerics factors ordered
#>
     59 0
#> Missings: FALSE
#> Has weights: FALSE
#> Has blocking: FALSE
#> Classes: 2
#> Bad Good
#> 300 700
#> Positive class: Bad
costs = matrix(c(0, 1, 5, 0), 2)
colnames(costs) = rownames(costs) = getTaskClassLevels(credit.task)
```

```
costs
#> Bad Good
#> Bad 0 5
#> Good 1 0
```

3.9.1.1.1 1. Thresholding

We start by fitting a logistic regression model to the German credit data set and predict posterior probabilities.

```
### Train and predict posterior probabilities
lrn = makeLearner("classif.multinom", predict.type = "prob", trace =
   FALSE)
mod = train(lrn, credit.task)
pred = predict(mod, task = credit.task)
pred
#> Prediction: 1000 observations
#> predict.type: prob
#> threshold: Bad=0.50,Good=0.50
#> time: 0.01
    id truth prob.Bad prob.Good response
#> 1 1 Good 0.03525092 0.9647491
                                    Good
       Bad 0.63222363 0.3677764
                                    Bad
Good
#> 4 4 Good 0.25182703 0.7481730
                                    Good
#> 5 5 Bad 0.75193275 0.2480673
                                     Bad
#> 6 6 Good 0.26230149 0.7376985
                                    Good
#> ... (1000 rows, 5 cols)
```

The default thresholds for both classes are 0.5. But according to the cost matrix we should predict class Good only if we are very sure that Good is indeed the correct label. Therefore we should increase the threshold for class Good and decrease the threshold for class Bad.

i. Theoretical thresholding

The theoretical threshold for the *positive* class can be calculated from the cost matrix as

$$t^* = \frac{c(+1,-1) - c(-1,-1)}{c(+1,-1) - c(+1,+1) + c(-1,+1) - c(-1,-1)}.$$

For more details see Elkan (2001).

Below the theoretical threshold for the German credit example is calculated and used to predict class labels. Since the diagonal of the cost matrix is zero the formula given above simplifies accordingly.

```
### Calculate the theoretical threshold for the positive class
th = costs[2,1]/(costs[2,1] + costs[1,2])
th
#> [1] 0.1666667
```

As you may recall you can change thresholds in mlr either before training by using the **predict.threshold** option of makeLearner or after prediction by calling setThreshold on the Prediction object.

As we already have a prediction we use the setThreshold function. It returns an altered Prediction object with class predictions for the theoretical threshold.

```
### Predict class labels according to the theoretical threshold
pred.th = setThreshold(pred, th)
pred.th
#> Prediction: 1000 observations
#> predict.type: prob
#> threshold: Bad=0.17, Good=0.83
#> time: 0.01
             prob.Bad prob.Good response
   id truth
#> 1 1 Good 0.03525092 0.9647491
                                   Good
Bad
#> 3  3  Good 0.02807414 0.9719259
                                   Good
#> 4 4 Good 0.25182703 0.7481730
                                  Bad
#> 5 5 Bad 0.75193275 0.2480673
                                    Bad
#> 6 6 Good 0.26230149 0.7376985
                                    Bad
#> ... (1000 rows, 5 cols)
```

In order to calculate the average costs over the entire data set we first need to create a new performance Measure. This can be done through function makeCostMeasure. It is expected that the rows of the cost matrix indicate true and the columns predicted class labels.

Then the average costs can be computed by function performance. Below we

compare the average costs and the error rate (mmce) of the learning algorithm with both default thresholds 0.5 and theoretical thresholds.

These performance values may be overly optimistic as we used the same data set for training and prediction, and resampling strategies should be preferred. In the **R** code below we make use of the **predict.threshold** argument of makeLearner to set the threshold before doing a 3-fold cross-validation on the credit.task. Note that we create a ResampleInstance (rin) that is used throughout the next several code chunks to get comparable performance values.

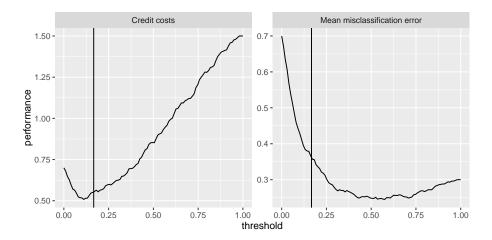
If we are also interested in the cross-validated performance for the default threshold values we can call setThreshold on the resample prediction r\$pred.

Theoretical thresholding is only reliable if the predicted posterior probabilities are correct. If there is bias the thresholds have to be shifted accordingly.

Useful in this regard is function plotThreshVsPerf that you can use to plot the average costs as well as any other performance measure versus possible threshold values for the positive class in [0, 1]. The underlying data is generated by generateThreshVsPerfData.

The following plots show the cross-validated costs and error rate (mmce). The theoretical threshold th calculated above is indicated by the vertical line. As you can see from the left-hand plot the theoretical threshold seems a bit large.

```
d = generateThreshVsPerfData(r, measures = list(credit.costs, mmce))
plotThreshVsPerf(d, mark.th = th)
```



ii. Empirical thresholding

The idea of *empirical thresholding* (see Sheng and Ling, 2006) is to select costoptimal threshold values for a given learning method based on the training data. In contrast to *theoretical thresholding* it suffices if the estimated posterior probabilities are order-correct.

In order to determine optimal threshold values you can use mlr's function tuneThreshold. As tuning the threshold on the complete training data set can lead to overfitting, you should use resampling strategies. Below we perform 3-fold cross-validation and use tuneThreshold to calculate threshold values with lowest average costs over the 3 test data sets.

```
lrn = makeLearner("classif.multinom", predict.type = "prob", trace =
    FALSE)

### 3-fold cross-validation
r = resample(lrn, credit.task, resampling = rin, measures =
    list(credit.costs, mmce), show.info = FALSE)
r
#> Resample Result
#> Task: GermanCredit
#> Learner: classif.multinom
```

```
#> Aggr perf: credit.costs.test.mean=0.852,mmce.test.mean=0.248
#> Runtime: 0.209151

### Tune the threshold based on the predicted probabilities on the 3
    test data sets

tune.res = tuneThreshold(pred = r$pred, measure = credit.costs)

tune.res
#> $th
#> [1] 0.1115426
#>
#> $perf
#> credit.costs
#> 0.507004
```

tuneThreshold returns the optimal threshold value for the positive class and the corresponding performance. As expected the tuned threshold is smaller than the theoretical threshold.

3.9.1.1.2 2. Rebalancing

In order to minimize the average costs, observations from the less costly class should be given higher importance during training. This can be achieved by weighting the classes, provided that the learner under consideration has a 'class weights' or an 'observation weights' argument. To find out which learning methods support either type of weights have a look at the list of integrated learners in the Appendix or use listLearners.

```
### Learners that accept observation weights
listLearners("classif", properties = "weights")[c("class",
    "package")]
#>
                  class
                            package
#> 1
        classif.avNNet
                               nnet
      classif.binomial
                               stats
#> 3 classif.blackboost mboost,party
#> 4
           classif.C50
                                 C50
#> 5
       classif.cforest
                               party
#> 6
         classif.ctree
                               party
#> ... (22 rows, 2 cols)
### Learners that can deal with class weights
listLearners("classif", properties = "class.weights")[c("class",
    "package")]
#>
                         class package
#> 1
                  classif.ksvm
                                kernlab
#> 2 classif.LiblineaRL1L2SVC LiblineaR
#> 3 classif.LiblineaRL1LoqReq LiblineaR
```

```
#> 4 classif.LiblineaRL2L1SVC LiblineaR
#> 5 classif.LiblineaRL2LogReg LiblineaR
#> 6 classif.LiblineaRL2SVC LiblineaR
#> ... (9 rows, 2 cols)
```

Alternatively, over- and undersampling techniques can be used.

i. Weighting

Just as theoretical thresholds, theoretical weights can be calculated from the cost matrix. If t indicates the target threshold and t_0 the original threshold for the positive class the proportion of observations in the positive class has to be multiplied by

$$\frac{1-t}{t}\frac{t_0}{1-t_0}.$$

Alternatively, the proportion of observations in the negative class can be multiplied by the inverse. A proof is given by Elkan (2001).

In most cases, the original threshold is $t_0 = 0.5$ and thus the second factor vanishes. If additionally the target threshold t equals the theoretical threshold t^* the proportion of observations in the positive class has to be multiplied by

$$\frac{1-t^*}{t^*} = \frac{c(-1,+1) - c(+1,+1)}{c(+1,-1) - c(-1,-1)}.$$

For the credit example the theoretical threshold corresponds to a weight of 5 for the positive class.

```
### Weight for positive class corresponding to theoretical treshold w = (1 - th)/th w #> [1] 5
```

A unified and convenient way to assign class weights to a Learner (and tune them) is provided by function makeWeightedClassesWrapper. The class weights are specified using argument wcw.weight. For learners that support observation weights a suitable weight vector is then generated internally during training or resampling. If the learner can deal with class weights, the weights are basically passed on to the appropriate learner parameter. The advantage of using the wrapper in this case is the unified way to specify the class weights.

Below is an example using learner "classif.multinom" (multinom from package nnet) which accepts observation weights. For binary classification problems it is sufficient to specify the weight ${\tt w}$ for the positive class. The negative class then automatically receives weight 1.

```
### Weighted learner
lrn = makeLearner("classif.multinom", trace = FALSE)
lrn = makeWeightedClassesWrapper(lrn, wcw.weight = w)
```

```
lrn
#> Learner weightedclasses.classif.multinom from package nnet
#> Type: classif
#> Name: ; Short name:
#> Class: WeightedClassesWrapper
#> Properties: twoclass, multiclass, numerics, factors, prob
#> Predict-Type: response
#> Hyperparameters: trace=FALSE,wcw.weight=5
r = resample(lrn, credit.task, rin, measures = list(credit.costs,
    mmce), show.info = FALSE)
r
#> Resample Result
#> Task: GermanCredit
#> Learner: weightedclasses.classif.multinom
#> Aggr perf: credit.costs.test.mean=0.526,mmce.test.mean=0.346
#> Runtime: 0.239676
```

For classification methods like "classif.ksvm" (the support vector machine ksvm in package kernlab) that support class weights you can pass them directly.

```
lrn = makeLearner("classif.ksvm", class.weights = c(Bad = w, Good =
     1))
```

Or, more conveniently, you can again use makeWeightedClassesWrapper.

Just like the theoretical threshold, the theoretical weights may not always be suitable, therefore you can tune the weight for the positive class as shown in the following example. Calculating the theoretical weight beforehand may help to narrow down the search interval.

```
lrn = makeLearner("classif.multinom", trace = FALSE)
lrn = makeWeightedClassesWrapper(lrn)
ps = makeParamSet(makeDiscreteParam("wcw.weight", seq(4, 12, 0.5)))
ctrl = makeTuneControlGrid()
tune.res = tuneParams(lrn, credit.task, resampling = rin, par.set = ps,
```

```
measures = list(credit.costs, mmce), control = ctrl, show.info =
      FALSE)
tune.res
#> Tune result:
#> Op. pars: wcw.weight=7.5
#> credit.costs.test.mean=0.501,mmce.test.mean=0.381
as.data.frame(tune.res$opt.path)[1:3]
      wcw.weight credit.costs.test.mean mmce.test.mean
#> 1
                             0.5650291
                                             0.3330127
             4.5
#> 2
                              0.5550251
                                             0.3430167
#> 3
              5
                              0.5260320
                                             0.3460197
#> 4
             5.5
                              0.5130070
                                             0.3530147
#> 5
              6
                              0.5160100
                                             0.3640137
#> 6
             6.5
                              0.5160160
                                             0.3720157
#> 7
             7
                              0.5040250
                                             0.3760167
             7.5
#> 8
                              0.5010040
                                             0.3810038
#> 9
              8
                              0.5100130
                                             0.3900128
#> 10
             8.5
                              0.5100070
                                             0.3940108
#> 11
             9
                              0.5110080
                                             0.4030078
#> 12
             9.5
                              0.5160130
                                             0.4080128
                              0.5260140
#> 13
             10
                                             0.4180138
#> 14
            10.5
                              0.5240060
                                             0.4200098
#> 15
              11
                              0.5319991
                                             0.4280029
#> 16
            11.5
                              0.5289901
                                             0.4330019
#> 17
              12
                              0.5249801
                                             0.4369999
```

ii. Over- and undersampling

If the Learner supports neither observation nor class weights the proportions of the classes in the training data can be changed by over- or undersampling.

In the GermanCredit data set the positive class Bad should receive a theoretical weight of w = (1 - th)/th = 5. This can be achieved by oversampling class Bad with a rate of 5 or by undersampling class Good with a rate of 1/5 (using functions oversample or undersample).

Note that in the above example the learner was trained on the oversampled task credit.task.over. In order to get the training performance on the original task predictions were calculated for credit.task.

We usually prefer resampled performance values, but simply calling resample on the oversampled task does not work since predictions have to be based on the original task. The solution is to create a wrapped Learner via function makeOversampleWrapper. Internally, oversample is called before training, but predictions are done on the original data.

```
lrn = makeLearner("classif.multinom", trace = FALSE)
lrn = makeOversampleWrapper(lrn, osw.rate = w, osw.cl = "Bad")
lrn
#> Learner classif.multinom.oversampled from package mlr,nnet
#> Type: classif
#> Name: ; Short name:
#> Class: OversampleWrapper
#> Properties: numerics, factors, weights, prob, two class, multiclass
#> Predict-Type: response
#> Hyperparameters: trace=FALSE, osw.rate=5, osw.cl=Bad
r = resample(lrn, credit.task, rin, measures = list(credit.costs,
    mmce), show.info = FALSE)
#> Resample Result
#> Task: GermanCredit
#> Learner: classif.multinom.oversampled
#> Aggr perf: credit.costs.test.mean=0.535,mmce.test.mean=0.351
#> Runtime: 0.397176
```

Of course, we can also tune the oversampling rate. For this purpose we again have to create an OversampleWrapper. Optimal values for parameter osw.rate can be obtained using function tuneParams.

3.9.1.2 Multi-class problems

We consider the waveform data set from package mlbench and add an artificial cost matrix:

1	2	3
0	30	80
5	0	4
10	8	0
	5	$ \begin{array}{ccc} 0 & 30 \\ 5 & 0 \end{array} $

We start by creating the Task, the cost matrix and the corresponding performance measure.

```
### Task
df = mlbench::mlbench.waveform(500)
wf.task = makeClassifTask(id = "waveform", data = as.data.frame(df),
    target = "classes")

### Cost matrix
costs = matrix(c(0, 5, 10, 30, 0, 8, 80, 4, 0), 3)
colnames(costs) = rownames(costs) = getTaskClassLevels(wf.task)

### Performance measure
wf.costs = makeCostMeasure(id = "wf.costs", name = "Waveform costs",
    costs = costs,
    best = 0, worst = 10)
```

In the multi-class case, both, thresholding and rebalancing correspond to cost matrices of a certain structure where c(k,l)=c(l) for $k,\ l=1,\ldots,K,\ k\neq l.$ This condition means that the cost of misclassifying an observation is independent of the predicted class label (see Domingos, 1999). Given a cost matrix of this type, theoretical thresholds and weights can be derived in a similar manner as in the binary case. Obviously, the cost matrix given above does not have this special structure.

3.9.1.2.1 1. Thresholding

Given a vector of positive threshold values as long as the number of classes K, the predicted probabilities for all classes are adjusted by dividing them by the corresponding threshold value. Then the class with the highest adjusted probability is predicted. This way, as in the binary case, classes with a low threshold are preferred to classes with a larger threshold.

Again this can be done by function setThreshold as shown in the following example (or alternatively by the predict.threshold option of makeLearner). Note that the threshold vector needs to have names that correspond to the class labels.

```
lrn = makeLearner("classif.rpart", predict.type = "prob")
rin = makeResampleInstance("CV", iters = 3, task = wf.task)
```

```
r = resample(lrn, wf.task, rin, measures = list(wf.costs, mmce),
    show.info = FALSE)
#> Resample Result
#> Task: waveform
#> Learner: classif.rpart
#> Aggr perf: wf.costs.test.mean=7.02,mmce.test.mean=0.262
#> Runtime: 0.0682302
### Calculate thresholds as 1/(average costs of true classes)
th = 2/rowSums(costs)
names(th) = getTaskClassLevels(wf.task)
th
#>
                       2
           1
#> 0.01818182 0.22222222 0.11111111
pred.th = setThreshold(r$pred, threshold = th)
performance(pred.th, measures = list(wf.costs, mmce))
                  mmce
#> wf.costs
#> 5.0372268 0.3502393
```

The threshold vector th in the above example is chosen according to the average costs of the true classes 55, 4.5 and 9. More exactly, th corresponds to an artificial cost matrix of the structure mentioned above with off-diagonal elements c(2,1) = c(3,1) = 55, c(1,2) = c(3,2) = 4.5 and c(1,3) = c(2,3) = 9. This threshold vector may be not optimal but leads to smaller total costs on the data set than the default.

ii. Empirical thresholding

As in the binary case it is possible to tune the threshold vector using function tuneThreshold. Since the scaling of the threshold vector does not change the predicted class labels tuneThreshold returns threshold values that lie in [0,1] and sum to unity.

For comparison we show the standardized version of the theoretically motivated threshold vector chosen above.

```
th/sum(th)
```

```
#> 1 2 3
#> 0.05172414 0.63218391 0.31609195
```

3.9.1.2.2 2. Rebalancing

i. Weighting

In the multi-class case you have to pass a vector of weights as long as the number of classes K to function makeWeightedClassesWrapper. The weight vector can be tuned using function tuneParams.

3.9.2 Example-dependent misclassification costs

In case of example-dependent costs we have to create a special Task via function makeCostSensTask. For this purpose the feature values x and an $n \times K$ cost matrix that contains the cost vectors for all n examples in the data set are required.

We use the iris data and generate an artificial cost matrix (see Beygelzimer et al., 2005).

```
df = iris
cost = matrix(runif(150 * 3, 0, 2000), 150) * (1 -
          diag(3))[df$Species,] + runif(150, 0, 10)
colnames(cost) = levels(iris$Species)
rownames(cost) = rownames(iris)
df$Species = NULL

costsens.task = makeCostSensTask(id = "iris", data = df, cost = cost)
costsens.task
#> Supervised task: iris
#> Type: costsens
```

mlr provides several wrappers to turn regular classification or regression methods into Learners that can deal with example-dependent costs.

- makeCostSensClassifWrapper (wraps a classification Learner): This is a
 naive approach where the costs are coerced into class labels by choosing
 the class label with minimum cost for each example. Then a regular
 classification method is used.
- makeCostSensRegrWrapper (wraps a regression Learner): An individual regression model is fitted for the costs of each class. In the prediction step first the costs are predicted for all classes and then the class with the lowest predicted costs is selected.
- makeCostSensWeightedPairsWrapper (wraps a classification Learner): This is also known as cost-sensitive one-vs-one (CS-OVO) and the most sophisticated of the currently supported methods. For each pair of classes, a binary classifier is fitted. For each observation the class label is defined as the element of the pair with minimal costs. During fitting, the observations are weighted with the absolute difference in costs. Prediction is performed by simple voting.

In the following example we use the third method. We create the wrapped Learner and train it on the CostSensTask defined above.

```
#> Trained on: task.id = iris; obs = 150; features = 4
#> Hyperparameters: trace=FALSE
```

The models corresponding to the individual pairs can be accessed by function getLearnerModel.

```
getLearnerModel(mod)
#> [[1]]
#> Model for learner.id=classif.multinom;
    learner.class=classif.multinom
#> Trained on: task.id = feats; obs = 150; features = 4
#> Hyperparameters: trace=FALSE
#>
#> [[2]]
#> Model for learner.id=classif.multinom;
    learner.\,class = classif.\,multinom
#> Trained on: task.id = feats; obs = 150; features = 4
#> Hyperparameters: trace=FALSE
#>
#> [[3]]
#> Model for learner.id=classif.multinom;
    learner.class=classif.multinom
#> Trained on: task.id = feats; obs = 150; features = 4
#> Hyperparameters: trace=FALSE
```

mlr provides some performance measures for example-specific cost-sensitive classification. In the following example we calculate the mean costs of the predicted class labels (meancosts) and the misclassification penalty (mcp). The latter measure is the average difference between the costs caused by the predicted class labels, i.e., meancosts, and the costs resulting from choosing the class with lowest cost for each observation. In order to compute these measures the costs for the test observations are required and therefore the Task has to be passed to performance.

```
pred = predict(mod, task = costsens.task)
pred

#> Prediction: 150 observations

#> predict.type: response

#> threshold:

#> time: 0.04

#> id response

#> 1 1 setosa

#> 2 2 setosa

#> 3 3 setosa

#> 4 4 setosa

#> 5 5 setosa

#> 6 6 setosa
```

3.10 Imbalanced Classification Problems

In case of binary classification strongly imbalanced classes often lead to unsatisfactory results regarding the prediction of new observations, especially for the small class. In this context imbalanced classes simply means that the number of observations of one class (usu. positive or majority class) by far exceeds the number of observations of the other class (usu. negative or minority class). This setting can be observed fairly often in practice and in various disciplines like credit scoring, fraud detection, medical diagnostics or churn management.

Most classification methods work best when the number of observations per class are roughly equal. The problem with *imbalanced classes* is that because of the dominance of the majority class classifiers tend to ignore cases of the minority class as noise and therefore predict the majority class far more often. In order to lay more weight on the cases of the minority class, there are numerous correction methods which tackle the *imbalanced classification problem*. These methods can generally be divided into *cost- and sampling-based approaches*. Below all methods supported by mlr are introduced.

3.10.1 Sampling-based approaches

The basic idea of *sampling methods* is to simply adjust the proportion of the classes in order to increase the weight of the minority class observations within the model.

The *sampling-based approaches* can be divided further into three different categories:

- 1. **Undersampling methods**: Elimination of randomly chosen cases of the majority class to decrease their effect on the classifier. All cases of the minority class are kept.
- 2. Oversampling methods: Generation of additional cases (copies, artificial observations) of the minority class to increase their effect on the classifier. All cases of the majority class are kept.
- 3. **Hybrid methods**: Mixture of under- and oversampling strategies.

All these methods directly access the underlying data and "rearrange" it. In this way the sampling is done as part of the *preprocessing* and can therefore be combined with every appropriate classifier.

mlr currently supports the first two approaches.

3.10.2 (Simple) over- and undersampling

As mentioned above undersampling always refers to the majority class, while oversampling affects the minority class. By the use of undersampling, randomly chosen observations of the majority class are eliminated. Through (simple) oversampling all observations of the minority class are considered at least once when fitting the model. In addition, exact copies of minority class cases are created by random sampling with repetitions.

First, let's take a look at the effect for a classification task. Based on a simulated ClassifTask with imbalanced classes two new tasks (task.over, task.under) are created via mlr functions oversample and undersample, respectively.

```
data.imbal.train = rbind(
  data.frame(x = rnorm(100, mean = 1), class = "A"),
  data.frame(x = rnorm(5000, mean = 2), class = "B")
task = makeClassifTask(data = data.imbal.train, target = "class")
task.over = oversample(task, rate = 8)
task.under = undersample(task, rate = 1/8)
table(getTaskTargets(task))
#>
#>
      \boldsymbol{A}
           B
    100 5000
table(getTaskTargets(task.over))
#>
           В
      \boldsymbol{A}
   800 5000
table(getTaskTargets(task.under))
#>
#>
     Α
#> 100 625
```

Please note that the *undersampling rate* has to be between 0 and 1, where 1 means no undersampling and 0.5 implies a reduction of the majority class size to 50 percent. Correspondingly, the *oversampling rate* must be greater or equal to 1, where 1 means no oversampling and 2 would result in doubling the minority class size.

As a result the performance should improve if the model is applied to new data.

```
lrn = makeLearner("classif.rpart", predict.type = "prob")
mod = train(lrn, task)
mod.over = train(lrn, task.over)
mod.under = train(lrn, task.under)
data.imbal.test = rbind(
  data.frame(x = rnorm(10, mean = 1), class = "A"),
  data.frame(x = rnorm(500, mean = 2), class = "B")
)
performance(predict(mod, newdata = data.imbal.test), measures =
    list(mmce, ber, auc))
         mmce
                     ber
                                auc
#> 0.01960784 0.50000000 0.50000000
performance(predict(mod.over, newdata = data.imbal.test), measures =
    list(mmce, ber, auc))
#>
         mmce
                     ber
                                auc
#> 0.04509804 0.41500000 0.58500000
performance(predict(mod.under, newdata = data.imbal.test), measures
    = list(mmce, ber, auc))
                     ber
         mmce
                                auc
#> 0.05098039 0.41800000 0.70550000
```

In this case the *performance measure* has to be considered very carefully. As the *misclassification rate* (mmce) evaluates the overall accuracy of the predictions, the *balanced error rate* (ber) and *area under the ROC Curve* (auc) might be more suitable here, as the misclassifications within each class are separately taken into account.

3.10.2.1 Over- and undersampling wrappers

Alternatively, mlr also offers the integration of over- and undersampling via a wrapper approach. This way over- and undersampling can be applied to already existing learners to extend their functionality.

The example given above is repeated once again, but this time with extended learners instead of modified tasks (see makeOversampleWrapper and makeUndersampleWrapper). Just like before the *undersampling rate* has to be between 0 and 1, while the *oversampling rate* has a lower boundary of 1.

```
lrn.over = makeOversampleWrapper(lrn, osw.rate = 8)
lrn.under = makeUndersampleWrapper(lrn, usw.rate = 1/8)
mod = train(lrn, task)
mod.over = train(lrn.over, task)
```

```
mod.under = train(lrn.under, task)
performance(predict(mod, newdata = data.imbal.test), measures =
    list(mmce, ber, auc))
#>
         mmce
                     ber
#> 0.01960784 0.50000000 0.50000000
performance(predict(mod.over, newdata = data.imbal.test), measures =
    list(mmce, ber, auc))
         mmce
#> 0.03333333 0.40900000 0.72020000
performance(predict(mod.under, newdata = data.imbal.test), measures
    = list(mmce, ber, auc))
         mmce
                     ber
                                auc
#> 0.04509804 0.41500000 0.71660000
```

3.10.2.2 Extensions to oversampling

Two extensions to (simple) oversampling are available in mlr.

3.10.2.2.1 1. SMOTE (Synthetic Minority Oversampling Technique)

As the duplicating of the minority class observations can lead to overfitting, within *SMOTE* the "new cases" are constructed in a different way. For each new observation, one randomly chosen minority class observation as well as one of its *randomly chosen next neighbours* are interpolated, so that finally a new *artificial observation* of the minority class is created. The smote function in mlr handles numeric as well as factor features, as the gower distance is used for nearest neighbour calculation. The factor level of the new artificial case is sampled from the given levels of the two input observations.

Analogous to oversampling, SMOTE preprocessing is possible via modification of the task.

```
task.smote = smote(task, rate = 8, nn = 5)
table(getTaskTargets(task))
#>
#> A B
#> 100 5000

table(getTaskTargets(task.smote))
#>
#> A B
#> 800 5000
```

Alternatively, a new wrapped learner can be created via makeSMOTEWrapper.

By default the number of nearest neighbours considered within the algorithm is set to 5.

3.10.2.2.2 2. Overbagging

Another extension of oversampling consists in the combination of sampling with the bagging approach. For each iteration of the bagging process, minority class observations are oversampled with a given rate in obw.rate. The majority class cases can either all be taken into account for each iteration (obw.maxcl = "all") or bootstrapped with replacement to increase variability between training data sets during iterations (obw.maxcl = "boot").

The construction of the **Overbagging Wrapper** works similar to makeBagging-Wrapper. First an existing mlr learner has to be passed to makeOverBagging-Wrapper. The number of iterations or fitted models can be set via obw.iters.

```
lrn = makeLearner("classif.rpart", predict.type = "response")
obw.lrn = makeOverBaggingWrapper(lrn, obw.rate = 8, obw.iters = 3)
```

For binary classification the prediction is based on majority voting to create a discrete label. Corresponding probabilities are predicted by considering the proportions of all the predicted labels. Please note that the benefit of the sampling process is highly dependent on the specific learner as shown in the following example.

First, let's take a look at the tree learner with and without overbagging:

Now let's consider a random forest as initial learner:

```
lrn = makeLearner("classif.randomForest")
obw.lrn = makeOverBaggingWrapper(lrn, obw.rate = 8, obw.iters = 3)
lrn = setPredictType(lrn, "prob")
r1 = resample(learner = lrn, task = task, resampling = rdesc,
    show.info = FALSE,
 measures = list(mmce, ber, auc))
r1$aggr
#> mmce.test.mean ber.test.mean auc.test.mean
      0.03509804
                   0.46089748
                                    0.58514212
obw.lrn = setPredictType(obw.lrn, "prob")
r2 = resample(learner = obw.lrn, task = task, resampling = rdesc,
    show.info = FALSE,
 measures = list(mmce, ber, auc))
r2$aggr
#> mmce.test.mean ber.test.mean auc.test.mean
   0.04098039 0.45961754 0.54926842
```

While overbagging slighty improves the performance of the decision tree, the auc decreases in the second example when additional overbagging is applied. As the random forest itself is already a strong learner (and a bagged one as well), a further bagging step isn't very helpful here and usually won't improve the model.

3.10.3 Cost-based approaches

In contrast to sampling, cost-based approaches usually require particular learners, which can deal with different class-dependent costs (Cost-Sensitive Classification).

3.10.3.1 Weighted classes wrapper

Another approach independent of the underlying classifier is to assign the costs as *class weights*, so that each observation receives a weight, depending on the class it belongs to. Similar to the sampling-based approaches, the effect of the minority class observations is thereby increased simply by a higher weight of these instances and vice versa for majority class observations.

In this way every learner which supports weights can be extended through the wrapper approach. If the learner does not have a direct parameter for class weights, but supports observation weights, the weights depending on the class are internally set in the wrapper.

```
lrn = makeLearner("classif.logreg")
wcw.lrn = makeWeightedClassesWrapper(lrn, wcw.weight = 0.01)
```

For binary classification, the single number passed to the classifier corresponds to the weight of the positive / majority class, while the negative / minority class receives a weight of 1. So actually, no real costs are used within this approach, but the cost ratio is taken into account.

If the underlying learner already has a parameter for class weighting (e.g., class.weights in "classif.ksvm"), the wcw.weight is basically passed to the specific class weighting parameter.

```
lrn = makeLearner("classif.ksvm")
wcw.lrn = makeWeightedClassesWrapper(lrn, wcw.weight = 0.01)
```

3.11 ROC Analysis and Performance Curves

For binary scoring classifiers a *threshold* (or *cutoff*) value controls how predicted posterior probabilities are converted into class labels. ROC curves and other performance plots serve to visualize and analyse the relationship between one or two performance measures and the threshold.

This page is mainly devoted to receiver operating characteristic (ROC) curves that plot the true positive rate (sensitivity) on the vertical axis against the false positive rate (1 - specificity, fall-out) on the horizontal axis for all possible threshold values. Creating other performance plots like lift charts or precision/recall graphs works analogously and is shown briefly.

In addition to performance visualization ROC curves are helpful in

- determining an optimal decision threshold for given class prior probabilities and misclassification costs (for alternatives see also the pages about cost-sensitive classification and imbalanced classification problems in this tutorial),
- identifying regions where one classifier outperforms another and building suitable multi-classifier systems,
- obtaining calibrated estimates of the posterior probabilities.

For more information see the tutorials and introductory papers by Fawcett (2004), Fawcett (2006) as well as Flach (ICML 2004).

In many applications as, e.g., diagnostic tests or spam detection, there is uncertainty about the class priors or the misclassification costs at the time of prediction, for example because it's hard to quantify the costs or because costs and class priors vary over time. Under these circumstances the classifier is expected to work well for a whole range of decision thresholds and the area under the ROC curve (AUC) provides a scalar performance measure for comparing and selecting classifiers. mlr provides the AUC for binary classification (auc based on package ROCR) and also several generalizations of the AUC to the multi-class case (e.g., multiclass.au1p, multiclass.au1u based on Ferri et al. (2009)).

mlr offers three ways to plot ROC and other performance curves.

- 1. Function plotROCCurves can, based on the output of generateThreshVsPerf-Data, plot performance curves for any pair of performance measures available in mlr.
- mlr offers an interface to package ROCR through function as ROCRPrediction.
- 3. mlr's function plotViperCharts provides an interface to ViperCharts.

With mlr version 2.8 functions generateROCRCurvesData, plotROCRCurves, and plotROCRCurvesGGVIS were deprecated.

Below are some examples that demonstrate the three possible ways. Note that you can only use learners that are capable of predicting probabilities. Have a look at the learner table in the Appendix or run listLearners("classif", properties = c("twoclass", "prob")) to get a list of all learners that support this.

3.11.1 Performance plots with plotROCCurves

As you might recall generate ThreshVsPerfData calculates one or several performance measures for a sequence of decision thresholds from 0 to 1. It provides S3 methods for objects of class Prediction, ResampleResult and BenchmarkResult (resulting from predict, resample or benchmark). plotROCCurves plots the result of generate ThreshVsPerfData using ggplot2.

3.11.1.1 Example 1: Single predictions

We consider the Sonar data set from package mlbench, which poses a binary classification problem (sonar.task) and apply linear discriminant analysis.

```
n = getTaskSize(sonar.task)
train.set = sample(n, size = round(2/3 * n))
test.set = setdiff(seq_len(n), train.set)
lrn1 = makeLearner("classif.lda", predict.type = "prob")
```

```
mod1 = train(lrn1, sonar.task, subset = train.set)
pred1 = predict(mod1, task = sonar.task, subset = test.set)
```

Since we want to plot ROC curves we calculate the false and true positive rates (fpr and tpr). Additionally, we also compute error rates (mmce).

```
df = generateThreshVsPerfData(pred1, measures = list(fpr, tpr, mmce))
```

generateThreshVsPerfData returns an object of class ThreshVsPerfData which contains the performance values in the \$data element.

Per default, plotROCCurves plots the performance values of the first two measures passed to generateThreshVsPerfData. The first is shown on the x-axis, the second on the y-axis. Moreover, a diagonal line that represents the performance of a random classifier is added. You can remove the diagonal by setting diagonal = FALSE.

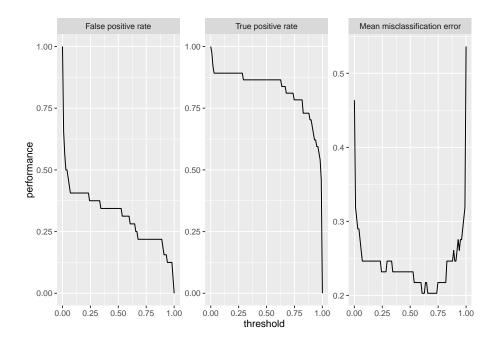
```
plotROCCurves(df)
```

The corresponding area under curve (auc) can be calculated as usual by calling performance.

```
performance(pred1, auc)
#> auc
#> 0.847973
```

plotROCCurves always requires a pair of performance measures that are plotted against each other. If you want to plot individual measures versus the decision threshold you can use function plotThreshVsPerf.

```
plotThreshVsPerf(df)
```



Additional to linear discriminant analysis we try a support vector machine with RBF kernel (ksvm).

```
lrn2 = makeLearner("classif.ksvm", predict.type = "prob")
mod2 = train(lrn2, sonar.task, subset = train.set)
pred2 = predict(mod2, task = sonar.task, subset = test.set)
```

In order to compare the performance of the two learners you might want to display the two corresponding ROC curves in one plot. For this purpose just pass a named list of Predictions to generateThreshVsPerfData.

```
df = generateThreshVsPerfData(list(lda = pred1, ksvm = pred2),
    measures = list(fpr, tpr))
plotROCCurves(df)
```



It's clear from the plot above that ksvm has a slightly higher AUC than lda.

```
performance(pred2, auc)
#> auc
#> 0.9214527
```

Based on the **\$data** member of **df** you can easily generate custom plots. Below the curves for the two learners are superposed.

```
qplot(x = fpr, y = tpr, color = learner, data = df$data, geom =
    "path")
```

It is easily possible to generate other performance plots by passing the appropriate performance measures to generate ThreshVsPerfData and plotROCCurves. Below, we generate a $precision/recall\ graph$ (precision = positive predictive value = ppv, recall = tpr) and a $sensitivity/specificity\ plot$ (sensitivity = tpr, specificity = tnr).

```
df = generateThreshVsPerfData(list(lda = pred1, ksvm = pred2),
    measures = list(ppv, tpr, tnr))

### Precision/recall graph
plotROCCurves(df, measures = list(tpr, ppv), diagonal = FALSE)
#> Warning: Removed 1 rows containing missing values (geom_path).

### Sensitivity/specificity plot
plotROCCurves(df, measures = list(tnr, tpr), diagonal = FALSE)
```

3.11.1.2 Example 2: Benchmark experiment

The analysis in the example above can be improved a little. Instead of writing individual code for training/prediction of each learner, which can become tedious very quickly, we can use function benchmark (see also Benchmark Experiments) and, ideally, the support vector machine should have been tuned.

We again consider the Sonar data set and apply lda as well as ksvm. We first generate a tuning wrapper for ksvm. The cost parameter is tuned on a (for demonstration purposes small) parameter grid. We assume that we are interested in a good performance over the complete threshold range and therefore tune with regard to the auc. The error rate (mmce) for a threshold value of 0.5 is reported as well.

Below the actual benchmark experiment is conducted. As resampling strategy we use 5-fold cross-validation and again calculate the auc as well as the error rate (for a threshold/cutoff value of 0.5).

```
### Benchmark experiment
lrns = list(lrn1, lrn2)
rdesc.outer = makeResampleDesc("CV", iters = 5)

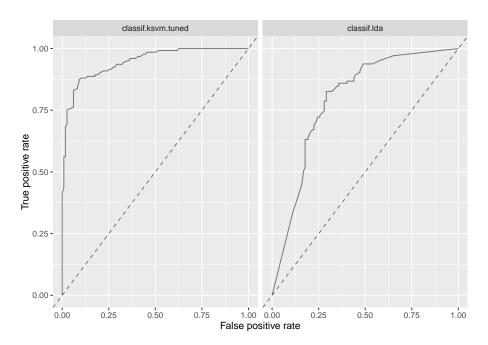
bmr = benchmark(lrns, tasks = sonar.task, resampling = rdesc.outer,
    measures = ms, show.info = FALSE)

bmr

#> task.id learner.id auc.test.mean mmce.test.mean
#> 1 Sonar-example classif.lda 0.7835442 0.2592334
#> 2 Sonar-example classif.ksvm.tuned 0.9454418 0.1390244
```

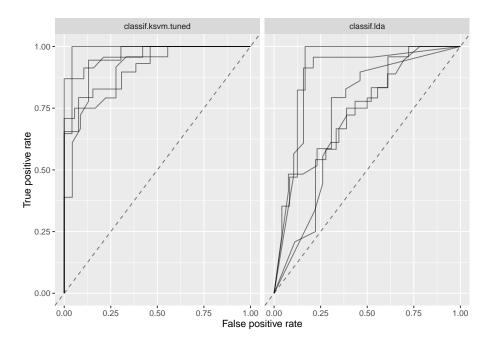
Calling generateThreshVsPerfData and plotROCCurves on the benchmark result produces a plot with ROC curves for all learners in the experiment.

```
df = generateThreshVsPerfData(bmr, measures = list(fpr, tpr, mmce))
plotROCCurves(df)
```



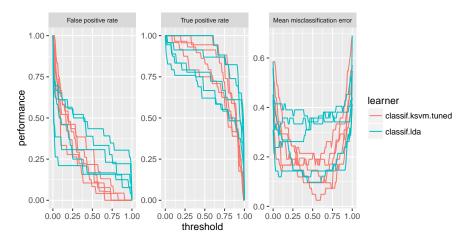
Per default, generateThreshVsPerfData calculates aggregated performances according to the chosen resampling strategy (5-fold cross-validation) and aggregation scheme (test.mean) for each threshold in the sequence. This way we get threshold-averaged ROC curves.

If you want to plot the individual ROC curves for each resample iteration set aggregate = FALSE.



The same applies for plotThreshVsPerf.

```
plotThreshVsPerf(df) +
  theme(strip.text.x = element_text(size = 7))
```



An alternative to averaging is to just merge the 5 test folds and draw a single ROC curve. Merging can be achieved by manually changing the class attribute of the prediction objects from ResamplePrediction to Prediction.

Below, the predictions are extracted from the BenchmarkResult via function getBMRPredictions, the class is changed and the ROC curves are created.

Averaging methods are normally preferred (cp. Fawcett, 2006), as they permit to assess the variability, which is needed to properly compare classifier performance.



Again, you can easily create other standard evaluation plots by passing the appropriate performance measures to generateThreshVsPerfData and plotROC-Curves.

3.11.2 Performance plots with asROCRPrediction

Drawing performance plots with package ROCR works through three basic commands:

- 1. ROCR::prediction: Create a ROCR prediction object.
- 2. ROCR::performance: Calculate one or more performance measures for the given prediction object.
- 3. ROCR::plot: Generate the performance plot.

mlr's function as ROCRPrediction converts an mlr Prediction object to a ROCR prediction object, so you can easily generate performance plots by doing steps 2. and 3. yourself. ROCR's plot method has some nice features which are not (yet) available in plotROCCurves, for example plotting the convex hull of the ROC curves. Some examples are shown below.

3.11.2.1 Example 1: Single predictions (continued)

We go back to out first example where we trained and predicted lda on the sonar classification task.

```
n = getTaskSize(sonar.task)
train.set = sample(n, size = round(2/3 * n))
test.set = setdiff(seq_len(n), train.set)

### Train and predict linear discriminant analysis
lrn1 = makeLearner("classif.lda", predict.type = "prob")
mod1 = train(lrn1, sonar.task, subset = train.set)
pred1 = predict(mod1, task = sonar.task, subset = test.set)
```

Below we use as ROCR Prediction to convert the lda prediction, let ROCR calculate the true and false positive rate and plot the ROC curve.

```
### Convert prediction
ROCRpred1 = asROCRPrediction(pred1)

### Calculate true and false positive rate
ROCRperf1 = ROCR::performance(ROCRpred1, "tpr", "fpr")

### Draw ROC curve
ROCR::plot(ROCRperf1)
```

Below is the same ROC curve, but we make use of some more graphical parameters: The ROC curve is color-coded by the threshold and selected threshold values are printed on the curve. Additionally, the convex hull (black broken line) of the ROC curve is drawn.

```
ch = ROCR::performance(ROCRpred1, "rch")
ROCR::plot(ch, add = TRUE, lty = 2)
```

3.11.2.2 Example 2: Benchmark experiments (continued)

We again consider the benchmark experiment conducted earlier. We first extract the predictions by getBMRPredictions and then convert them via function as-ROCRPrediction.

```
### Extract predictions
preds = getBMRPredictions(bmr)[[1]]

### Convert predictions
ROCRpreds = lapply(preds, asROCRPrediction)

### Calculate true and false positive rate
ROCRperfs = lapply(ROCRpreds, function(x) ROCR::performance(x, "tpr", "fpr"))
```

We draw the vertically averaged ROC curves (solid lines) as well as the ROC curves for the individual resampling iterations (broken lines). Moreover, standard error bars are plotted for selected true positive rates (0.1, 0.2, ..., 0.9). See ROCR's plot function for details.

```
### lda average ROC curve
plot(ROCRperfs[[1]], col = "blue", avg = "vertical", spread.estimate
    = "stderror",
 show.spread.at = seq(0.1, 0.8, 0.1), plotCI.col = "blue",
      plotCI.lwd = 2, lwd = 2)
### lda individual ROC curves
plot(ROCRperfs[[1]], col = "blue", lty = 2, lwd = 0.25, add = TRUE)
### ksvm average ROC curve
plot(ROCRperfs[[2]], col = "red", avg = "vertical", spread.estimate
    = "stderror",
 show.spread.at = seq(0.1, 0.6, 0.1), plotCI.col = "red",
      plotCI.lwd = 2, lwd = 2, add = TRUE)
### ksvm individual ROC curves
plot(ROCRperfs[[2]], col = "red", lty = 2, lwd = 0.25, add = TRUE)
legend("bottomright", legend = getBMRLearnerIds(bmr), lty = 1, lwd =
   2, col = c("blue", "red"))
```

In order to create other evaluation plots like *precision/recall graphs* you just have to change the performance measures when calling ROCR::performance. (Note that you have to use the measures provided by ROCR listed here and not mlr's performance measures.)

If you want to plot a performance measure versus the threshold, specify only one measure when calling ROCR::performance. Below the average accuracy over the 5 cross-validation iterations is plotted against the threshold. Moreover, boxplots for certain threshold values (0.1, 0.2, ..., 0.9) are drawn.

3.11.3 Viper charts

mlr also supports ViperCharts for plotting ROC and other performance curves. Like generateThreshVsPerfData it has S3 methods for objects of class Prediction, ResampleResult and BenchmarkResult. Below plots for the benchmark experiment (Example 2) are generated.

```
z = plotViperCharts(bmr, chart = "rocc", browse = FALSE)
```

You can see the plot created this way here. Note that besides ROC curves you get several other plots like lift charts or cost curves. For details, see plotViper-Charts.

3.12 Multilabel Classification

Multilabel classification is a classification problem where multiple target labels can be assigned to each observation instead of only one like in multiclass classification.

Two different approaches exist for multilabel classification. *Problem transformation methods* try to transform the multilabel classification into binary or multiclass classification problems. *Algorithm adaptation methods* adapt multiclass algorithms so they can be applied directly to the problem.

3.12.1 Creating a task

253

289 1816

The first thing you have to do for multilabel classification in mlr is to get your data in the right format. You need a data frame which consists of the features and a logical vector for each label which indicates if the label is present in the observation or not. After that you can create a Multilabel Task like a normal Classif Task. Instead of one target name you have to specify a vector of targets which correspond to the names of logical variables in the data frame. In the following example we get the yeast data frame from the already existing yeast.task, extract the 14 label names and create the task again.

```
yeast = getTaskData(yeast.task)
labels = colnames(yeast)[1:14]
yeast.task = makeMultilabelTask(id = "multi", data = yeast, target =
    labels)
yeast.task
#> Supervised task: multi
#> Type: multilabel
#> Target:
    label1, label2, label3, label4, label5, label6, label7, label8, label9, label10, label11, label12, label
#> Observations: 2417
#> Features:
#> numerics factors ordered
#>
        103
                   0
#> Missings: FALSE
#> Has weights: FALSE
#> Has blocking: FALSE
#> Classes: 14
   label1 label2 label3 label4
                                    label5 label6 label7 label8
    label9
                                                         428
       762
              1038
                       983
                                862
                                        722
                                                597
                                                                 480
        178
#> label10 label11 label12 label13 label14
```

1799

3.12.2 Constructing a learner

Multilabel classification in mlr can currently be done in two ways:

- Algorithm adaptation methods: Treat the whole problem with a specific algorithm.
- Problem transformation methods: Transform the problem, so that simple binary classification algorithms can be applied.

3.12.2.1 Algorithm adaptation methods

Currently the available algorithm adaptation methods in **R** are the multivariate random forest in the randomForestSRC package and the random ferns multilabel algorithm in the rFerns package. You can create the learner for these algorithms like in multiclass classification problems.

```
lrn.rfsrc = makeLearner("multilabel.randomForestSRC")
lrn.rFerns = makeLearner("multilabel.rFerns")
lrn.rFerns
#> Learner multilabel.rFerns from package rFerns
#> Type: multilabel
#> Name: Random ferns; Short name: rFerns
#> Class: multilabel.rFerns
#> Properties: numerics, factors, ordered
#> Predict-Type: response
#> Hyperparameters:
```

3.12.2.2 Problem transformation methods

For generating a wrapped multilabel learner first create a binary (or multiclass) classification learner with makeLearner. Afterwards apply a function like make-MultilabelBinaryRelevanceWrapper, makeMultilabelClassifierChainsWrapper, makeMultilabelNestedStackingWrapper, makeMultilabelDBRWrapper or makeMultilabelStackingWrapper on the learner to convert it to a learner that uses the respective problem transformation method.

You can also generate a binary relevance learner directly, as you can see in the example.

```
lrn.br = makeLearner("classif.rpart", predict.type = "prob")
lrn.br = makeMultilabelBinaryRelevanceWrapper(lrn.br)
lrn.br

#> Learner multilabel.classif.rpart from package rpart

#> Type: multilabel

#> Name: ; Short name:

#> Class: MultilabelBinaryRelevanceWrapper
```

```
#> Properties:
    numerics, factors, ordered, missings, weights, prob, twoclass, multiclass
#> Predict-Type: prob
#> Hyperparameters: xval=0

lrn.br2 = makeMultilabelBinaryRelevanceWrapper("classif.rpart")
lrn.br2
#> Learner multilabel.classif.rpart from package rpart
#> Type: multilabel
#> Name: ; Short name:
#> Class: MultilabelBinaryRelevanceWrapper
#> Properties:
    numerics, factors, ordered, missings, weights, prob, twoclass, multiclass
#> Predict-Type: response
#> Hyperparameters: xval=0
```

The different methods are shortly described in the following.

3.12.2.2.1 Binary relevance

This problem transformation method converts the multilabel problem to binary classification problems for each label and applies a simple binary classificator on these. In mlr this can be done by converting your binary learner to a wrapped binary relevance multilabel learner.

3.12.2.2.2 Classifier chains

Trains consecutively the labels with the input data. The input data in each step is augmented by the already trained labels (with the real observed values). Therefore an order of the labels has to be specified. At prediction time the labels are predicted in the same order as while training. The required labels in the input data are given by the previous done prediction of the respective label.

3.12.2.2.3 Nested stacking

Same as classifier chains, but the labels in the input data are not the real ones, but estimations of the labels obtained by the already trained learners.

3.12.2.2.4 Dependent binary relevance

Each label is trained with the real observed values of all other labels. In prediction phase for a label the other necessary labels are obtained in a previous step by a base learner like the binary relevance method.

3.12.2.2.5 Stacking

Same as the dependent binary relevance method, but in the training phase the labels used as input for each label are obtained by the binary relevance method.

3.12.3 Train

You can train a model as usual with a multilabel learner and a multilabel task as input. You can also pass **subset** and **weights** arguments if the learner supports this.

3.12.4 Predict

Prediction can be done as usual in mlr with predict and by passing a trained model and either the task to the task argument or some new data to the newdata argument. As always you can specify a subset of the data which should be predicted.

```
pred = predict(mod, task = yeast.task, subset = 1:10)
pred = predict(mod, newdata = yeast[1501:1600,])
names(as.data.frame(pred))
#> [1] "truth.label1"
                            "truth.label2"
                                               "truth.label3"
#> [4] "truth.label4"
                           "truth.label5"
                                               "truth.label6"
#> [7] "truth.label7"
                           "truth.label8"
                                               "truth.label9"
#> [10] "truth.label10"
                           "truth.label11"
                                               "truth.label12"
#> [13] "truth.label13"
                            "truth.label14"
                                               "prob.label1"
                           "prob.label3"
#> [16] "prob.label2"
                                               "prob.label4"
#> [19] "prob.label5"
                           "prob.label6"
                                               "prob.label?"
#> [22] "prob.label8"
                           "prob.label9"
                                               "prob.label10"
```

```
#> [25] "prob.label11"
                           "prob.label12"
                                              "prob.label13"
#> [28] "prob.label14"
                           "response.label1"
                                              "response.label2"
#> [31] "response.label3"
                           "response.label4"
                                              "response.label5"
#> [34] "response.label6"
                           "response.label7" "response.label8"
#> [37] "response.label9" "response.label10" "response.label11"
#> [40] "response.label12" "response.label13" "response.label14"
pred2 = predict(mod2, task = yeast.task)
names(as.data.frame(pred2))
  [1] "id"
                           "truth.label1"
                                              "truth.label2"
   [4] "truth.label3"
                           "truth.label4"
                                              "truth.label5"
   [7] "truth.label6"
                           "truth.label7"
                                              "truth.label8"
#> [10] "truth.label9"
                           "truth.label10"
                                              "truth.label11"
#> [13] "truth.label12"
                           "truth.label13"
                                              "truth.label14"
                           "response.label2"
#> [16] "response.label1"
                                              "response.label3"
                           "response.label5"
                                              "response.label6"
#> [19] "response.label4"
  [22] "response.label7"
                           "response.label8"
                                              "response.label9"
  [25] "response.label10" "response.label11" "response.label12"
  [28] "response.label13" "response.label14"
```

Depending on the chosen **predict.type** of the learner you get true and predicted values and possibly probabilities for each class label. These can be extracted by the usual accessor functions getPredictionTruth, getPredictionResponse and getPredictionProbabilities.

3.12.5 Performance

The performance of your prediction can be assessed via function performance. You can specify via the **measures** argument which **measure(s)** to calculate. The default measure for multilabel classification is the Hamming loss (multilabel.hamloss). All available measures for multilabel classification can be shown by listMeasures and found in the table of performance measures and the measures documentation page.

```
performance(pred)
#> multilabel.hamloss
#>
            0.2257143
performance(pred2, measures = list(multilabel.subset01,
    multilabel.hamloss, multilabel.acc,
  multilabel.f1, timepredict))
#> multilabel.subset01 multilabel.hamloss
                                                 multilabel.acc
#>
             0.8663633
                                 0.2049471
                                                      0.4637509
#>
         multilabel.f1
                               timepredict
#>
             0.5729926
                                  1.3840000
```

3.12.6 Resampling

For evaluating the overall performance of the learning algorithm you can do some resampling. As usual you have to define a resampling strategy, either via makeResampleDesc or makeResampleInstance. After that you can run the resample function. Below the default measure Hamming loss is calculated.

```
rdesc = makeResampleDesc(method = "CV", stratify = FALSE, iters = 3)
r = resample(learner = lrn.br, task = yeast.task, resampling =
    rdesc, show.info = FALSE)
r
#> Resample Result
#> Task: multi
#> Learner: multilabel.classif.rpart
#> Aggr perf: multilabel.hamloss.test.mean=0.225
#> Runtime: 5.46459
r = resample(learner = lrn.rFerns, task = yeast.task, resampling =
    rdesc, show.info = FALSE)
#> Resample Result
#> Task: multi
#> Learner: multilabel.rFerns
#> Aggr perf: multilabel.hamloss.test.mean=0.473
#> Runtime: 0.519855
```

3.12.7 Binary performance

If you want to calculate a binary performance measure like, e.g., the accuracy, the mmce or the auc for each label, you can use function getMultilabelBinaryPerformances. You can apply this function to any multilabel prediction, e.g., also on the resample multilabel prediction. For calculating the auc you need predicted probabilities.

```
#> label2
                     0.64
                                     0.36
                                              0.6547917
#> label3
                     0.68
                                     0.32
                                              0.7118227
#> label4
                     0.69
                                     0.31
                                              0.6764835
#> label5
                     0.73
                                     0.27
                                              0.6676923
#> label6
                                              0.6417739
                     0.70
                                     0.30
#> label7
                                     0.19
                                              0.5968750
                     0.81
#> label8
                     0.73
                                     0.27
                                              0.5164474
                                              0.4688458
#> label9
                     0.89
                                     0.11
#> label10
                     0.86
                                     0.14
                                              0.3996463
                     0.85
                                     0.15
                                              0.5000000
#> label11
#> label12
                     0.76
                                     0.24
                                              0.5330667
#> label13
                     0.75
                                     0.25
                                              0.5938610
#> label14
                                     0.00
                     1.00
                                                      NA
getMultilabelBinaryPerformances(r$pred, measures = list(acc, mmce))
#>
           acc.test.mean mmce.test.mean
#> label1
              0.69383533
                               0.3061647
#> label2
              0.58254034
                               0.4174597
#> label3
              0.70211005
                               0.2978899
#> label4
              0.71369466
                               0.2863053
#> label5
              0.70831609
                               0.2916839
              0.60488209
#> label6
                               0.3951179
#> label7
              0.54447662
                               0.4555234
#> label8
              0.53289201
                                0.4671080
              0.30906082
                               0.6909392
#> label9
                               0.5531651
#> label10
              0.44683492
#> label11
              0.45676458
                               0.5432354
#> label12
              0.52916839
                                0.4708316
#> label13
              0.53702938
                               0.4629706
#> label14
              0.01406703
                                0.9859330
```

3.13 Learning Curve Analysis

To analyse how the increase of observations in the training set improves the performance of a learner the *learning curve* is an appropriate visual tool. The experiment is conducted with an increasing subsample size and the performance is measured. In the plot the x-axis represents the relative subsample size whereas the y-axis represents the performance.

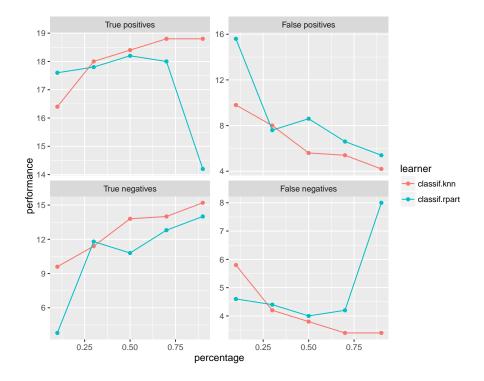
Note that this function internally uses benchmark in combination with make-DownsampleWrapper, so for every run new observations are drawn. Thus the results are noisy. To reduce noise increase the number of resampling iterations. You can define the resampling method in the resampling argument of generate-LearningCurveData. It is also possible to pass a ResampleInstance (which is a

result of makeResampleInstance) to make resampling consistent for all passed learners and each step of increasing the number of observations.

3.13.1 Plotting the learning curve

The mlr function generateLearningCurveData can generate the data for *learning curves* for multiple learners and multiple performance measures at once. With plotLearningCurve the result of generateLearningCurveData can be plotted using ggplot2. plotLearningCurve has an argument facet which can be either "measure" or "learner". By default facet = "measure" and facetted subplots are created for each measure input to generateLearningCurveData. If facet = "measure" learners are mapped to color, and vice versa.

```
r = generateLearningCurveData(
  learners = list("classif.rpart", "classif.knn"),
  task = sonar.task,
  percs = seq(0.1, 1, by = 0.2),
  measures = list(tp, fp, tn, fn),
  resampling = makeResampleDesc(method = "CV", iters = 5),
  show.info = FALSE)
plotLearningCurve(r)
```



What happens in generateLearningCurveData is the following: Each learner will be internally wrapped in a DownsampleWrapper. To measure the performance at the first step of percs, say 0.1, first the data will be split into a training and a test set according to the given resampling strategy. Then a random sample containing 10% of the observations of the training set will be drawn and used to train the learner. The performance will be measured on the complete test set. These steps will be repeated as defined by the given resampling method and for each value of percs.

In the first example a simplified usage of the learners argument was used, so that it's sufficient to give the *name*. It is also possible to create a learner the usual way and even to mix it. Make sure to give different ids in this case.

```
lrns = list(
  makeLearner(cl = "classif.ksvm", id = "ksvm1" , sigma = 0.2, C = 2),
  makeLearner(cl = "classif.ksvm", id = "ksvm2" , sigma = 0.1, C = 1),
  "classif.randomForest"
)
rin = makeResampleDesc(method = "CV", iters = 5)
lc = generateLearningCurveData(learners = lrns, task = sonar.task, percs = seq(0.1, 1, by = 0.1), measures = acc, resampling = rin, show.info = FALSE)
plotLearningCurve(lc)
```



We can display performance on the train set as well as the test set:



There is also an experimental ggvis plotting function, plotLearningCurveGGVIS. Instead of the facet argument to plotLearningCurve there is an argument interactive which plays a similar role. As subplots are not available in ggvis, measures or learners are mapped to an interactive sidebar which allows selection of the displayed measures or learners. The other feature is mapped to color.

plotLearningCurveGGVIS(r, interactive = "measure")

3.14 Exploring Learner Predictions

Learners use features to learn a prediction function and make predictions, but the effect of those features is often not apparent. mlr can estimate the partial dependence of a learned function on a subset of the feature space using generatePartialDependenceData.

Partial dependence plots reduce the potentially high dimensional function estimated by the learner, and display a marginalized version of this function in a lower dimensional space. For example suppose $Y = f(X) + \epsilon$, where $\mathbb{E}[\epsilon|X] = 0$. With (X,Y) pairs drawn independently from this statistical model, a learner may estimate \hat{f} , which, if X is high dimensional, can be uninterpretable. Suppose we want to approximate the relationship between some subset of X. We partition X into two sets, X_s and X_c such that $X = X_s \cup X_c$, where X_s is a subset of X of interest.

The partial dependence of f on X_s is

$$f_{X_s} = \mathbb{E}_{X_c} f(X_s, X_c).$$

 X_c is integrated out. We use the following estimator:

$$\hat{f}_{X_s} = \frac{1}{N} \sum_{i=1}^{N} \hat{f}(X_s, x_{ic}).$$

The individual conditional expectation of an observation can also be estimated using the above algorithm absent the averaging, giving $\hat{f}_{X_s}^{(i)}$. This allows the discovery of features of \hat{f} that may be obscured by an aggregated summary of \hat{f} .

The partial derivative of the partial dependence function, $\frac{\partial f_{X_s}}{\partial X_s}$, and the individual conditional expectation function, $\frac{\partial f_{X_s}^{(i)}}{\partial X_s}$, can also be computed. For regression and survival tasks the partial derivative of a single feature X_s is the gradient of the partial dependence function, and for classification tasks where the learner can output class probabilities the Jacobian. Note that if the learner produces discontinuous partial dependence (e.g., piecewise constant functions such as decision trees, ensembles of decision trees, etc.) the derivative will be 0 (where the function is not changing) or trending towards positive or negative infinity (at the discontinuities where the derivative is undefined). Plotting the partial dependence function of such learners may give the impression that the function is not discontinuous because the prediction grid is not composed of all discontinuous points in the predictor space. This results in a line interpolating that makes the function appear to be piecewise linear (where the derivative would be defined except at the boundaries of each piece).

The partial derivative can be informative regarding the additivity of the learned function in certain features. If $\hat{f}_{X_s}^{(i)}$ is an additive function in a feature X_s , then its partial derivative will not depend on any other features (X_c) that may have been used by the learner. Variation in the estimated partial derivative indicates that there is a region of interaction between X_s and X_c in \hat{f} . Similarly, instead of using the mean to estimate the expected value of the function at different values of X_s , instead computing the variance can highlight regions of interaction between X_s and X_c .

See Goldstein, Kapelner, Bleich, and Pitkin (2014) for more details and their package ICEbox for the original implementation. The algorithm works for any supervised learner with classification, regression, and survival tasks.

3.14.1 Generating partial dependences

Our implementation, following mlr's visualization pattern, consists of the above mentioned function generatePartialDependenceData, as well as two visualiza-

tion functions, plotPartialDependence and plotPartialDependenceGGVIS. The former generates input (objects of class PartialDependenceData) for the latter.

The first step executed by generatePartialDependenceData is to generate a feature grid for every element of the character vector features passed. The data are given by the input argument, which can be a Task or a data.frame. The feature grid can be generated in several ways. A uniformly spaced grid of length gridsize (default 10) from the empirical minimum to the empirical maximum is created by default, but arguments fmin and fmax may be used to override the empirical default (the lengths of fmin and fmax must match the length of features). Alternatively the feature data can be resampled, either by using a bootstrap or by subsampling.

```
lrn.classif = makeLearner("classif.ksvm", predict.type = "prob")
fit.classif = train(lrn.classif, iris.task)
pd = generatePartialDependenceData(fit.classif, iris.task,
    "Petal.Width")
pd
#> PartialDependenceData
#> Task: iris-example
#> Features: Petal.Width
#> Target: setosa, versicolor, virginica
#> Derivative: FALSE
#> Interaction: FALSE
#> Individual: FALSE
      Class Probability Petal. Width
#> 1 setosa 0.4983925 0.1000000
#> 2 setosa 0.4441165
                         0.3666667
#> 3 setosa 0.3808075
                         0.6333333
#> 4 setosa 0.3250243
                         0.9000000
#> 5 setosa
             0.2589014
                         1.1666667
#> 6 setosa
            0.1870692
                         1.4333333
#> ... (30 rows, 3 cols)
```

As noted above, X_s does not have to be unidimensional. If it is not, the **interaction** flag must be set to TRUE. Then the individual feature grids are combined using the Cartesian product, and the estimator above is applied, producing the partial dependence for every combination of unique feature values. If the **interaction** flag is FALSE (the default) then by default X_s is assumed unidimensional, and partial dependencies are generated for each feature separately. The resulting output when **interaction** = FALSE has a column for each feature, and NA where the feature was not used.

```
#> 1 setosa 0.4983925
                        0.1000000
                                           NA
                        0.3666667
#> 2 setosa 0.4441165
                                           NA
            0.3808075
#> 3 setosa
                        0.6333333
                                           NA
#> 4 setosa 0.3250243 0.9000000
                                           NA
#> 5 setosa 0.2589014 1.1666667
                                           NA
#> 6 setosa 0.1870692 1.4333333
                                           NA
tail(pd.lst$data)
         Class Probability Petal. Width Petal. Length
#> 55 virginica 0.2006336 NA
                                         3.622222
#> 56 virginica 0.3114545
                                 NA
                                         4.277778
#> 57 virginica 0.4404613
                                 NA
                                         4.933333
#> 58 virginica 0.6005358
                                  NA
                                         5.588889
#> 59 virginica 0.7099841
                                  NA
                                          6.244444
#> 60 virginica
                0.7242584
                                  NA
                                         6.900000
pd.int = generatePartialDependenceData(fit.classif, iris.task,
    c("Petal.Width", "Petal.Length"), TRUE)
pd.int
#> PartialDependenceData
#> Task: iris-example
#> Features: Petal.Width, Petal.Length
#> Target: setosa, versicolor, virginica
#> Derivative: FALSE
#> Interaction: TRUE
#> Individual: FALSE
     Class Probability Petal.Width Petal.Length
#> 1 setosa  0.6885025  0.1000000
#> 2 setosa 0.6824560 0.3666667
                                            1
#> 3 setosa 0.6459476 0.6333333
                                            1
#> 4 setosa 0.5750861 0.9000000
                                            1
                        1.1666667
#> 5 setosa 0.4745925
                                            1
#> 6 setosa
            0.3749285
                        1.4333333
                                            1
#> ... (300 rows, 4 cols)
```

At each step in the estimation of \hat{f}_{X_s} a set of predictions of length N is generated. By default the mean prediction is used. For classification where $\mathtt{predict.type}$ = "prob" this entails the mean class probabilities. However, other summaries of the predictions may be used. For regression and survival tasks the function used here must either return one number or three, and, if the latter, the numbers must be sorted lowest to highest. For classification tasks the function must return a number for each level of the target feature.

As noted, the fun argument can be a function which returns three numbers (sorted low to high) for a regression task. This allows further exploration of relative feature importance. If a feature is relatively important, the bounds

are necessarily tighter because the feature accounts for more of the variance of the predictions, i.e., it is "used" more by the learner. More directly setting fun = var identifies regions of interaction between X_s and X_c .

```
lrn.regr = makeLearner("regr.ksvm")
fit.regr = train(lrn.regr, bh.task)
pd.regr = generatePartialDependenceData(fit.regr, bh.task, "lstat",
    fun = median)
pd.regr
#> PartialDependenceData
#> Task: BostonHousing-example
#> Features: lstat
#> Target: medv
#> Derivative: FALSE
#> Interaction: FALSE
#> Individual: FALSE
        medv
                 lstat
#> 1 24.69031 1.730000
#> 2 23.72479 5.756667
#> 3 22.34841 9.783333
#> 4 20.78817 13.810000
#> 5 19.76183 17.836667
#> 6 19.33115 21.863333
#> ... (10 rows, 2 cols)
pd.ci = generatePartialDependenceData(fit.regr, bh.task, "lstat",
 fun = function(x) quantile(x, c(.25, .5, .75)))
pd.ci
#> PartialDependenceData
#> Task: BostonHousing-example
#> Features: lstat
#> Target: medv
#> Derivative: FALSE
#> Interaction: FALSE
#> Individual: FALSE
#>
         medv
                lstat
                           lower
#> 1 24.69031 1.730000 21.36068 29.75615
#> 2 23.72479   5.756667   20.80590   28.02338
#> 3 22.34841 9.783333 20.06507 25.22291
#> 4 20.78817 13.810000 18.55592 23.68100
#> 5 19.76183 17.836667 16.52737 22.98520
#> 6 19.33115 21.863333 15.14425 22.12766
#> ... (10 rows, 4 cols)
pd.classif = generatePartialDependenceData(fit.classif, iris.task,
    "Petal.Length", fun = median)
pd.classif
```

```
#> PartialDependenceData
#> Task: iris-example
#> Features: Petal.Length
#> Target: setosa, versicolor, virginica
#> Derivative: FALSE
#> Interaction: FALSE
#> Individual: FALSE
#>
     Class Probability Petal.Length
#> 1 setosa 0.31008788 1.000000
#> 2 setosa 0.24271454
                           1.655556
#> 3 setosa 0.17126036
                           2.311111
                        2.966667
#> 4 setosa 0.09380787
#> 5 setosa 0.04579912
                           3.622222
#> 6 setosa 0.02455344
                           4.277778
#> ... (30 rows, 3 cols)
```

In addition to bounds based on a summary of the distribution of the conditional expectation of each observation, learners which can estimate the variance of their predictions can also be used. The argument bounds is a numeric vector of length two which is added (so the first number should be negative) to the point prediction to produce a confidence interval for the partial dependence. The default is the .025 and .975 quantiles of the Gaussian distribution.

```
fit.se = train(makeLearner("regr.randomForest", predict.type =
   "se"), bh.task)
pd.se = generatePartialDependenceData(fit.se, bh.task, c("lstat",
   "crim"))
head(pd.se$data)
        medv
                 lstat crim
                               lower
                                       upper
#> 1 31.02186 1.730000 NA 27.65357 34.39015
#> 2 25.94429 5.756667 NA 23.43079 28.45779
#> 3 23.52758 9.783333 NA 21.23661 25.81856
#> 4 22.05223 13.810000 NA 20.30446 23.80000
#> 5 20.44293 17.836667 NA 18.72603 22.15982
#> 6 19.80143 21.863333
                        NA 18.04932 21.55353
tail(pd.se$data)
         medv lstat
                        crim
                                lower
                                        upper
#> 15 21.65846 NA 39.54849 19.50827 23.80866
#> 16 21.64409 NA 49.43403 19.49704 23.79114
#> 17 21.63038 NA 59.31957 19.48054 23.78023
#> 18 21.61514 NA 69.20512 19.46092 23.76936
#> 19 21.61969 NA 79.09066 19.46819 23.77119
#> 20 21.61987 NA 88.97620 19.46843 23.77130
```

As previously mentioned if the aggregation function is not used, i.e., it is the

identity, then the conditional expectation of $\hat{f}_{X_s}^{(i)}$ is estimated. If individual = TRUE then generatePartialDependenceData returns n partial dependence estimates made at each point in the prediction grid constructed from the features.

```
pd.ind.regr = generatePartialDependenceData(fit.regr, bh.task,
    "lstat", individual = TRUE)
pd.ind.regr
#> PartialDependenceData
#> Task: BostonHousing-example
#> Features: lstat
#> Target: medv
#> Derivative: FALSE
#> Interaction: FALSE
#> Individual: TRUE
#> Predictions centered: FALSE
#> medv lstat idx
#> 1 25.66995 1.730000
                         1
#> 2 24.71747 5.756667
                         1
#> 3 23.64157 9.783333
#> 4 22.70812 13.810000
#> 5 22.00059 17.836667
#> 6 21.46195 21.863333
#> ... (5060 rows, 3 cols)
```

The resulting output, particularly the element data in the returned object, has an additional column idx which gives the index of the observation to which the row pertains.

For classification tasks this index references both the class and the observation index.

```
pd.ind.classif = generatePartialDependenceData(fit.classif,
   iris.task, "Petal.Length", individual = TRUE)
pd.ind.classif
#> PartialDependenceData
#> Task: iris-example
#> Features: Petal.Length
#> Target: setosa, versicolor, virginica
#> Derivative: FALSE
#> Interaction: FALSE
#> Individual: TRUE
#> Predictions centered: FALSE
     Class Probability Petal.Length
#> 1 setosa 0.9814053
                       1 1.setosa
#> 2 setosa 0.9747355
                                1 2.setosa
#> 3 setosa 0.9815516
                                1 3.setosa
                          1 4.setosa
#> 4 setosa 0.9795761
```

Individual estimates of partial dependence can also be centered by predictions made at all n observations for a particular point in the prediction grid created by the features. This is controlled by the argument center which is a list of the same length as the length of the features argument and contains the values of the features desired.

```
iris = getTaskData(iris.task)
pd.ind.classif = generatePartialDependenceData(fit.classif,
    iris.task, "Petal.Length", individual = TRUE,
    center = list("Petal.Length" = min(iris$Petal.Length)))
```

Partial derivatives can also be computed for individual partial dependence estimates and aggregate partial dependence. This is restricted to a single feature at a time. The derivatives of individual partial dependence estimates can be useful in finding regions of interaction between the feature for which the derivative is estimated and the features excluded.

```
pd.regr.der = generatePartialDependenceData(fit.regr, bh.task,
    "lstat", derivative = TRUE)
head(pd.regr.der$data)
          medv
                    lstat
#> 1 -0.1792626 1.730000
#> 2 -0.3584207 5.756667
#> 3 -0.4557666 9.783333
#> 4 -0.4523905 13.810000
#> 5 -0.3700880 17.836667
#> 6 -0.2471346 21.863333
pd.regr.der.ind = generatePartialDependenceData(fit.regr, bh.task,
    "lstat", derivative = TRUE,
  individual = TRUE)
head(pd.regr.der.ind$data)
          medv
#>
                    lstat idx
#> 1 -0.1931323 1.730000
#> 2 -0.2656911 5.756667
#> 3 -0.2571006 9.783333
#> 4 -0.2033080 13.810000
                            1
#> 5 -0.1511472 17.836667
#> 6 -0.1193129 21.863333
                            1
pd.classif.der = generatePartialDependenceData(fit.classif,
    iris.task, "Petal.Width", derivative = TRUE)
head(pd.classif.der$data)
```

```
Class Probability Petal. Width
#> 1 setosa -0.1479385 0.1000000
#> 2 setosa -0.2422728
                         0.3666667
#> 3 setosa -0.2189893
                        0.6333333
#> 4 setosa -0.2162803 0.9000000
#> 5 setosa -0.2768042
                         1.1666667
#> 6 setosa -0.2394176
                        1.4333333
pd.classif.der.ind = generatePartialDependenceData(fit.classif,
    iris.task, "Petal.Width", derivative = TRUE,
  individual = TRUE)
head(pd.classif.der.ind$data)
     Class Probability Petal. Width
#> 1 setosa 0.02479474
                               0.1 1.setosa
#> 2 setosa 0.01710561
                               0.1 2.setosa
#> 3 setosa 0.01646252
                              0.1 3.setosa
#> 4 setosa 0.01530718
                              0.1 4.setosa
#> 5 setosa 0.02608577
                               0.1 5.setosa
#> 6 setosa 0.03925531
                               0.1 6.setosa
```

3.14.2 Functional ANOVA

Hooker (2004) proposed the decomposition of a learned function \hat{f} as a sum of lower dimensional functions

$$f(\mathbf{x}) = g_0 + \sum_{i=1}^p g_i(X_i) + \sum_{i \neq j} g_{ij}(x_{ij}) + \dots$$

where p is the number of features. generate FunctionalANOVAData estimates the individual g functions using partial dependence. When functions depend only on one feature, they are equivalent to partial dependence, but a g function which depends on more than one feature is the "effect" of only those features: lower dimensional "effects" are removed.

$$\hat{g}_u(X) = \frac{1}{N} \sum_{i=1}^{N} \left(\hat{f}(X) - \sum_{v \subset u} g_v(X) \right)$$

Here u is a subset of $1, \ldots, p$. When |v| = 1 g_v can be directly computed by computing the bivariate partial dependence of \hat{f} on X_u and then subtracting off the univariate partial dependences of the features contained in v.

Although this decomposition is generalizable to classification it is currently only available for regression tasks.

```
lrn.regr = makeLearner("regr.ksvm")
fit.regr = train(lrn.regr, bh.task)
fa = generateFunctionalANOVAData(fit.regr, bh.task, "lstat", depth =
    1, fun = median)
fa
#> FunctionalANOVAData
#> Task: BostonHousing-example
#> Features: lstat
#> Target: medv
#>
#>
   effect
                         lstat
               medv
#> 1 lstat 24.89524 1.730000
#> 2 lstat 23.76025 5.756667
#> 3 lstat 22.35052 9.783333
#> 4 lstat 20.68435 13.810000
#> 5 lstat 19.60062 17.836667
#> 6 lstat 19.01178 21.863333
#> ... (10 rows, 3 cols)
pd.regr = generatePartialDependenceData(fit.regr, bh.task, "lstat",
   fun = median)
pd.regr
#> PartialDependenceData
#> Task: BostonHousing-example
#> Features: lstat
#> Target: medv
#> Derivative: FALSE
#> Interaction: FALSE
#> Individual: FALSE
        medv
                 lstat
#> 1 24.89524 1.730000
#> 2 23.76025 5.756667
#> 3 22.35052 9.783333
#> 4 20.68435 13.810000
#> 5 19.60062 17.836667
#> 6 19.01178 21.863333
#> ... (10 rows, 2 cols)
```

The depth argument is similar to the **interaction** argument in generate Partial Dependence Data but instead of specifying whether all of joint "effect" of all the **features** is computed, it determines whether "effects" of all subsets of the features given the specified depth are computed. So, for example, with p features and depth 1, the univariate partial dependence is returned. If, instead, depth = 2, then all possible bivariate functional ANOVA effects are returned.

This is done by computing the univariate partial dependence for each feature and subtracting it from the bivariate partial dependence for each possible pair.

```
fa.bv = generateFunctionalANOVAData(fit.regr, bh.task, c("crim",
    "lstat", "age"),
 depth = 2)
fa.bv
#> FunctionalANOVAData
#> Task: BostonHousing-example
#> Features: crim, lstat, age
#> Target: medv
#>
#>
#>
        effect
                    medv
                              crim lstat age
#> 1 crim:lstat -22.68734 0.006320 1.73 NA
#> 2 crim:lstat -23.22114 9.891862 1.73 NA
#> 3 crim:lstat -24.77479 19.777404 1.73 NA
#> 4 crim:lstat -26.41395 29.662947 1.73 NA
#> 5 crim:lstat -27.56524 39.548489 1.73 NA
#> 6 crim:lstat -28.27952 49.434031 1.73 NA
#> ... (300 rows, 5 cols)
names(table(fa.bv$data$effect)) ## interaction effects estimated
#> [1] "crim:age" "crim:lstat" "lstat:age"
```

3.14.3 Plotting partial dependences

Results from generatePartialDependenceData and generateFunctionalANOVA-Data can be visualized with plotPartialDependence and plotPartialDependenceGGVIS.

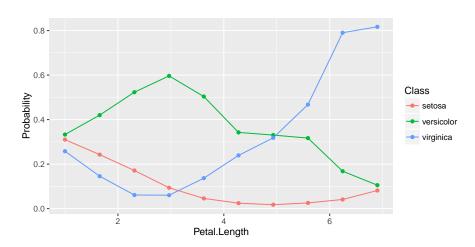
With one feature and a regression task the output is a line plot, with a point for each point in the corresponding feature's grid.

```
plotPartialDependence(pd.regr)
```



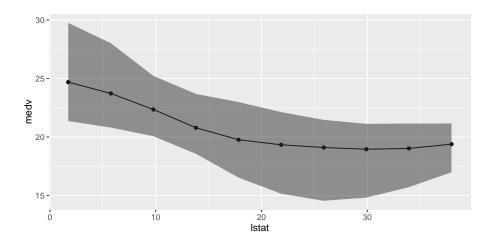
With a classification task, a line is drawn for each class, which gives the estimated partial probability of that class for a particular point in the feature grid.

plotPartialDependence(pd.classif)



For regression tasks, when the fun argument of generatePartialDependenceData is used, the bounds will automatically be displayed using a gray ribbon.

plotPartialDependence(pd.ci)



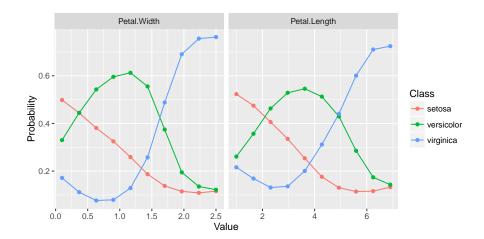
The same goes for plots of partial dependences where the learner has predict.type = "se".

plotPartialDependence(pd.se)



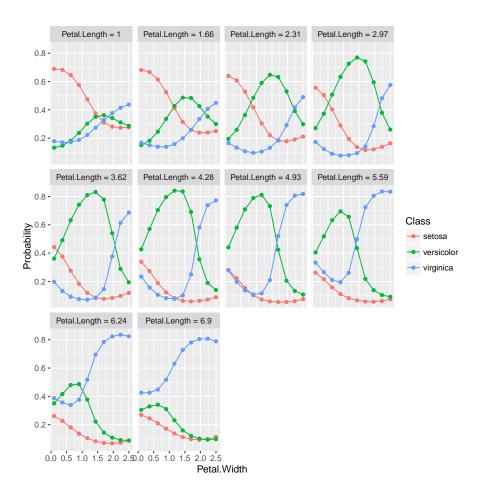
When multiple features are passed to generatePartialDependenceData but interaction = FALSE, facetting is used to display each estimated bivariate relationship.

plotPartialDependence(pd.lst)



When interaction = TRUE in the call to generatePartialDependenceData, one variable must be chosen to be used for facetting, and a subplot for each value in the chosen feature's grid is created, wherein the other feature's partial dependences within the facetting feature's value are shown. Note that this type of plot is limited to two features.

plotPartialDependence(pd.int, facet = "Petal.Length")

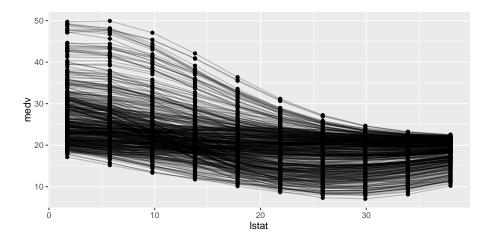


plotPartialDependenceGGVIS can be used similarly, however, since ggvis currently lacks subplotting/facetting capabilities, the argument interact maps one feature to an interactive sidebar where the user can select a value of one feature.

plotPartialDependenceGGVIS(pd.int, interact = "Petal.Length")

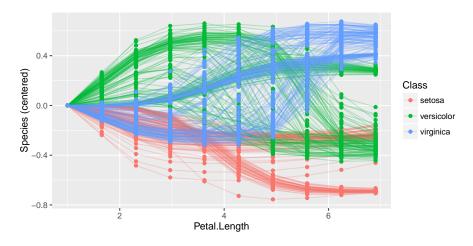
When individual = TRUE each individual conditional expectation curve is plotted.

plotPartialDependence(pd.ind.regr)



When the individual curves are centered by subtracting the individual conditional expectations estimated at a particular value of X_s this results in a fixed intercept which aids in visualizing variation in predictions made by $\hat{f}_{X_s}^{(i)}$.

plotPartialDependence(pd.ind.classif)



Plotting partial derivative functions works the same as partial dependence. Below are estimates of the derivative of the mean aggregated partial dependence function, and the individual partial dependence functions for a regression and a classification task respectively.

plotPartialDependence(pd.regr.der)



This suggests that \hat{f} is not additive in lstat except in the neighborhood of 25. plotPartialDependence(pd.regr.der.ind)

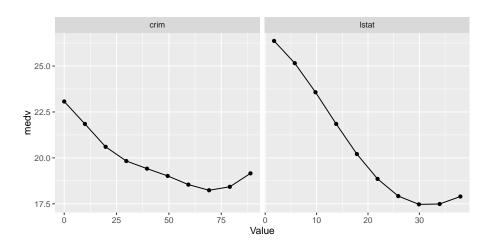


This suggests that Petal.Width interacts with some other feature in the neighborhood of (1.5, 2) for classes "virginica" and "versicolor".

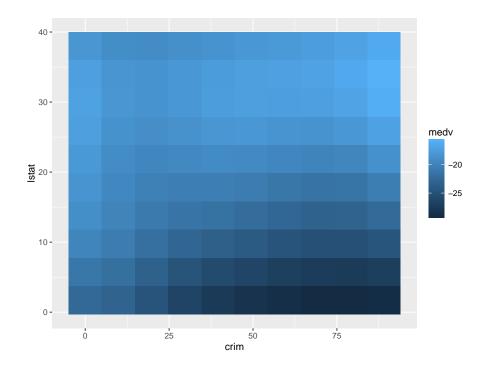
plotPartialDependence(pd.classif.der.ind)



Output from generate Functional ANOVA Data can also be plotted using plot Partial Dependence.



Interactions can often be more easily visualized by using functional ANOVA.



3.15 Classifier Calibration

A classifier is "calibrated" when the predicted probability of a class matches the expected frequency of that class. mlr can visualize this by plotting estimated class probabilities (which are discretized) against the observed frequency of said class in the data using generateCalibrationData and plotCalibration.

generateCalibrationData takes as input Prediction, ResampleResult, BenchmarkResult, or a named list of Prediction or ResampleResult objects on a classification (multiclass or binary) task with learner(s) that are capable of outputting probabilities (i.e., learners must be constructed with predict.type = TRUE). The result is an object of class CalibrationData which has elements proportion, data, and task. proportion gives the proportion of observations labelled with a given class for each predicted probability bin (e.g., for observations which are predicted to have class "A" with probability (0,0.1], what is the proportion of said observations which have class "A"?).

```
lrn = makeLearner("classif.rpart", predict.type = "prob")
mod = train(lrn, task = sonar.task)
pred = predict(mod, task = sonar.task)
cal = generateCalibrationData(pred)
cal$proportion
#> Learner bin Class Proportion
```

```
#> 1 prediction (0.1,0.2]
                            M 0.1060606
#> 2 prediction (0.7,0.8]
                             M 0.7333333
#> 3 prediction [0,0.1]
                            M 0.0000000
#> 4 prediction (0.9,1]
                             M 0.9333333
#> 5 prediction (0.2,0.3]
                            M 0.2727273
#> 6 prediction (0.4,0.5]
                             M 0.4615385
#> 7 prediction (0.8,0.9]
                             M 0.0000000
#> 8 prediction (0.5,0.6]
                            M 0.0000000
```

The manner in which the predicted probabilities are discretized is controlled by two arguments: breaks and groups. By default breaks = "Sturges" which uses the Sturges algorithm in hist. This argument can specify other algorithms available in hist, it can be a numeric vector specifying breakpoints for cut, or a single integer specifying the number of bins to create (which are evenly spaced). Alternatively, groups can be set to a positive integer value (by default groups = NULL) in which case cut2 is used to create bins with an approximately equal number of observations in each bin.

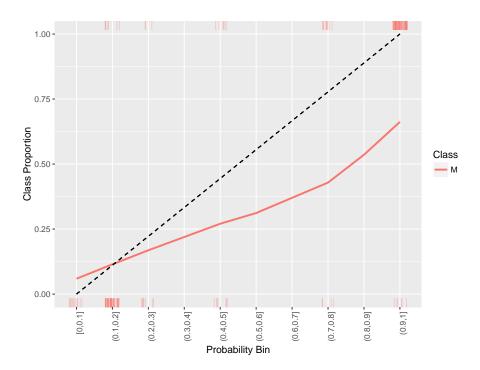
CalibrationData objects can be plotted using plotCalibration. plotCalibration by default plots a reference line which shows perfect calibration and a "rag" plot, which is a rug plot on the top and bottom of the graph, where the top pertains to "positive" cases, where the predicted class matches the observed class, and the bottom pertains to "negative" cases, where the predicted class does not match the observed class. Perfect classifier performance would result in all the positive cases clustering in the top right (i.e., the correct classes are predicted with high probability) and the negative cases clustering in the bottom left.

```
plotCalibration(cal)
```



Because of the discretization of the probabilities, sometimes it is advantageous to smooth the calibration plot. Though <code>smooth = FALSE</code> by default, setting this option to <code>TRUE</code> replaces the estimated proportions with a loess smoother.

cal = generateCalibrationData(pred)
plotCalibration(cal, smooth = TRUE)



All of the above functionality works with multi-class classification as well.

```
lrns = list(
   makeLearner("classif.randomForest", predict.type = "prob"),
   makeLearner("classif.nnet", predict.type = "prob", trace = FALSE)
)
mod = lapply(lrns, train, task = iris.task)
pred = lapply(mod, predict, task = iris.task)
names(pred) = c("randomForest", "nnet")
cal = generateCalibrationData(pred, breaks = c(0, .3, .6, 1))
plotCalibration(cal)
```



3.16 Evaluating Hyperparameter Tuning

As mentioned on the Tuning tutorial page, tuning a machine learning algorithm typically involves:

• the hyperparameter search space:

• the optimization algorithm (aka tuning method):

```
### ex: random search with 100 iterations
ctrl = makeTuneControlRandom(maxit = 100L)
```

• an evaluation method, i.e., a resampling strategy and a performance measure:

```
### ex: 2-fold CV
rdesc = makeResampleDesc("CV", iters = 2L)
```

After tuning, you may want to evaluate the tuning process in order to answer questions such as:

- How does varying the value of a hyperparameter change the performance of the machine learning algorithm?
- What's the relative importance of each hyperparameter?
- How did the optimization algorithm (prematurely) converge?

mlr provides methods to generate and plot the data in order to evaluate the effect of hyperparameter tuning.

3.16.1 Generating hyperparameter tuning data

mlr separates the generation of the data from the plotting of the data in case the user wishes to use the data in a custom way downstream.

The generateHyperParsEffectData method takes the tuning result along with 2 additional arguments: trafo and include.diagnostics. The trafo argument will convert the hyperparameter data to be on the transformed scale in case a transformation was used when creating the parameter (as in the case below). The include.diagnostics argument will tell mlr whether to include the eol and any error messages from the learner.

Below we perform random search on the C parameter for SVM on the famous Pima Indians dataset. We generate the hyperparameter effect data so that the C parameter is on the transformed scale and we do not include diagnostic data:

```
#> Snapshot of data:
#>
           C acc.test.mean mmce.test.mean iteration exec.time
#> 1 0.3770897 0.7695312 0.2304688 1 0.069
                                                 0.068
#> 2 3.4829323
               0.7526042 0.2473958
                                            2
#> 3 2.2050176
                                             3
              0.7630208 0.2369792
                                                 0.067
#> 4 24.9285221
               0.7070312
                             0.2929688
                                             4
                                                  0.075
#> 5 0.2092395
                0.7539062
                             0.2460938
                                             5
                                                  0.067
#> 6 0.1495099
                                             6
                                                  0.069
              0.7395833
                             0.2604167
```

As a reminder from the resampling tutorial, if we wanted to generate data on the training set as well as the validation set, we only need to make a few minor changes:

```
ps = makeParamSet(
 makeNumericParam("C", lower = -5, upper = 5, trafo = function(x)
)
ctrl = makeTuneControlRandom(maxit = 100L)
rdesc = makeResampleDesc("CV", iters = 2L, predict = "both")
res = tuneParams("classif.ksvm", task = pid.task, control = ctrl,
 measures = list(acc, setAggregation(acc, train.mean), mmce,
      setAggregation(mmce,
    train.mean)), resampling = rdesc, par.set = ps, show.info =
        FALSE)
generateHyperParsEffectData(res, trafo = T, include.diagnostics =
    FALSE)
#> HyperParsEffectData:
#> Hyperparameters: C
#> Measures:
    acc.test.mean,acc.train.mean,mmce.test.mean,mmce.train.mean
#> Optimizer: TuneControlRandom
#> Nested CV Used: FALSE
#> Snapshot of data:
              C acc.test.mean acc.train.mean mmce.test.mean
   mmce.train.mean
#> 1 0.03518875
                    0.6510417
                                   0.6510417
                                                  0.3489583
   0.3489583
#> 2 0.17104229
                    0.7356771
                                   0.7721354
                                                  0.2643229
    0.2278646
#> 3 4.35326556
                    0.7304688
                                   0.8828125
                                                  0.2695312
    0.1171875
#> 4 0.33644238
                    0.7486979
                                   0.8138021
                                                  0.2513021
    0.1861979
#> 5 1.28168692
                    0.7500000
                                   0.8476562
                                                  0.2500000
    0.1523438
#> 6 7.36607693
                                   0.8932292
                                                  0.2760417
                    0.7239583
```

```
0.1067708
#> iteration exec.time
     1 0.096
#> 1
            0.093
#> 2
       2
#> 3
       3 0.091
#> 4
        4
            0.092
#> 5
        5
             0.090
#> 6
     6 0.091
```

In the example below, we perform grid search on the C parameter for SVM on the Pima Indians dataset using nested cross validation. We generate the hyperparameter effect data so that the C parameter is on the untransformed scale and we do not include diagnostic data. As you can see below, nested cross validation is supported without any extra work by the user, allowing the user to obtain an unbiased estimator for the performance.

```
ps = makeParamSet(
 makeNumericParam("C", lower = -5, upper = 5, trafo = function(x)
)
ctrl = makeTuneControlGrid()
rdesc = makeResampleDesc("CV", iters = 2L)
lrn = makeTuneWrapper("classif.ksvm", control = ctrl,
 measures = list(acc, mmce), resampling = rdesc, par.set = ps,
    show.info = FALSE)
res = resample(lrn, task = pid.task, resampling = cv2, extract =
   getTuneResult, show.info = FALSE)
generateHyperParsEffectData(res)
#> HyperParsEffectData:
#> Hyperparameters: C
#> Measures: acc.test.mean,mmce.test.mean
#> Optimizer: TuneControlGrid
#> Nested CV Used: TRUE
#> Snapshot of data:
#>
           C acc.test.mean mmce.test.mean iteration exec.time
0.3359375
                                         2 0.048
3 0.049
                                        4 0.049
                                        5 0.048
                                              0.049
#> nested cv run
      1
#> 1
#> 2
             1
#> 3
             1
#> 4
```

After generating the hyperparameter effect data, the next step is to visualize it. mlr has several methods built-in to visualize the data, meant to support the needs of the researcher and the engineer in industry. The next few sections will walk through the visualization support for several use-cases.

3.16.2 Visualizing the effect of a single hyperparameter

In a situation when the user is tuning a single hyperparameter for a learner, the user may wish to plot the performance of the learner against the values of the hyperparameter.

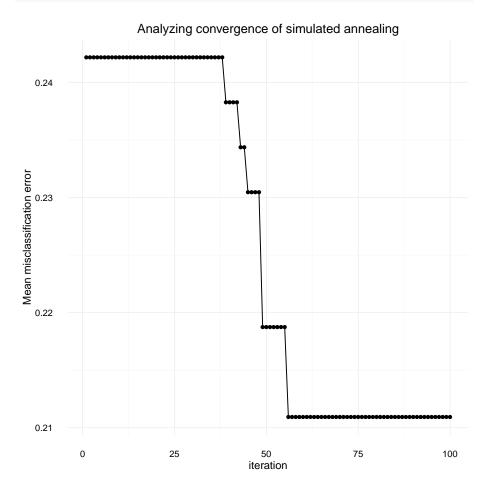
In the example below, we tune the number of clusters against the silhouette score on the Pima dataset. We specify the x-axis with the x argument and the y-axis with the y argument. If the plot.type argument is not specified, mlr will attempt to plot a scatterplot by default. Since plotHyperParsEffect returns a ggplot object, we can easily customize it to our liking!

```
ps = makeParamSet(
 makeDiscreteParam("centers", values = 3:10)
)
ctrl = makeTuneControlGrid()
rdesc = makeResampleDesc("Holdout")
res = tuneParams("cluster.kmeans", task = mtcars.task, control =
 measures = silhouette, resampling = rdesc, par.set = ps, show.info
      = FALSE)
#> This is package 'modeest' written by P. PONCET.
#> For a complete list of functions, use 'library(help = "modeest")'
    or 'help.start()'.
data = generateHyperParsEffectData(res)
plt = plotHyperParsEffect(data, x = "centers", y =
    "silhouette.test.mean")
### add our own touches to the plot
plt + geom_point(colour = "red") +
 ggtitle("Evaluating Number of Cluster Centers on mtcars") +
 scale_x_continuous(breaks = 3:10) +
 theme_bw()
```



In the example below, we tune SVM with the C hyperparameter on the Pima dataset. We will use simulated annealing optimizer, so we are interested in seeing if the optimization algorithm actually improves with iterations. By default, mlr only plots improvements to the global optimum.

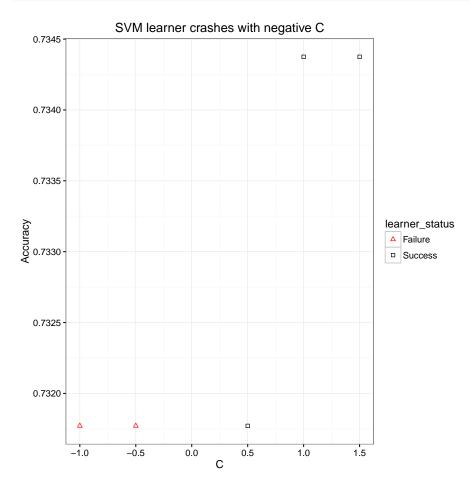
```
plt + ggtitle("Analyzing convergence of simulated annealing") +
   theme_minimal()
```



In the case of a learner crash, mlr will impute the crash with the worst value graphically and indicate the point. In the example below, we give the C parameter negative values, which will result in a learner crash for SVM.

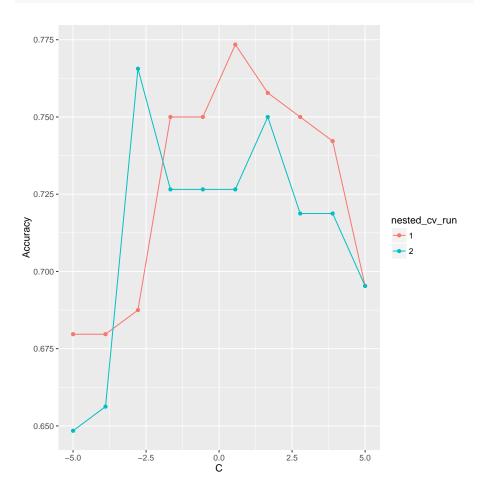
```
ps = makeParamSet(
   makeDiscreteParam("C", values = c(-1, -0.5, 0.5, 1, 1.5))
)
ctrl = makeTuneControlGrid()
rdesc = makeResampleDesc("CV", iters = 2L)
res = tuneParams("classif.ksvm", task = pid.task, control = ctrl,
   measures = list(acc, mmce), resampling = rdesc, par.set = ps,
        show.info = FALSE)
data = generateHyperParsEffectData(res)
plt = plotHyperParsEffect(data, x = "C", y = "acc.test.mean")
```

```
plt + ggtitle("SVM learner crashes with negative C") +
  theme_bw()
```



The example below uses nested cross validation with an outer loop of 2 runs. mlr indicates each run within the visualization.

```
data = generateHyperParsEffectData(res)
plotHyperParsEffect(data, x = "C", y = "acc.test.mean", plot.type =
    "line")
```



3.16.3 Visualizing the effect of 2 hyperparameters

In the case of tuning 2 hyperparameters simultaneously, mlr provides the ability to plot a heatmap and contour plot in addition to a scatterplot or line.

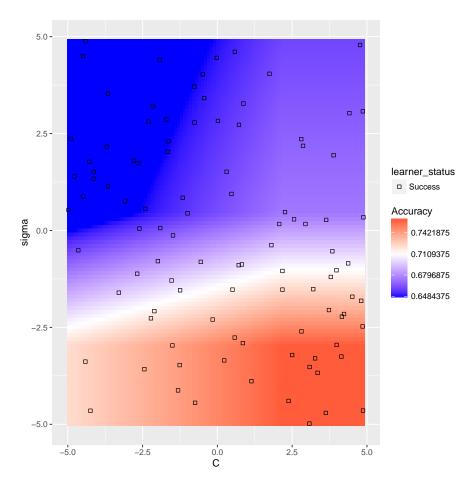
In the example below, we tune the C and sigma parameters for SVM on the Pima dataset. We use interpolation to produce a regular grid for plotting the heatmap. The interpolation argument accepts any regression learner from mlr to perform the interpolation. The z argument will be used to fill the heatmap or color lines, depending on the plot.type used.

```
ps = makeParamSet(
```

```
makeNumericParam("C", lower = -5, upper = 5, trafo = function(x)
 makeNumericParam("sigma", lower = -5, upper = 5, trafo =
     function(x) 2^x))
ctrl = makeTuneControlRandom(maxit = 100L)
rdesc = makeResampleDesc("Holdout")
learn = makeLearner("classif.ksvm", par.vals = list(kernel =
    "rbfdot"))
res = tuneParams(learn, task = pid.task, control = ctrl, measures =
   acc,
 resampling = rdesc, par.set = ps, show.info = FALSE)
data = generateHyperParsEffectData(res)
plt = plotHyperParsEffect(data, x = "C", y = "sigma", z =
    "acc.test.mean",
 plot.type = "heatmap", interpolate = "regr.earth")
min_plt = min(data$data$acc.test.mean, na.rm = TRUE)
max_plt = max(data$data$acc.test.mean, na.rm = TRUE)
med_plt = mean(c(min_plt, max_plt))
plt + scale_fill_gradient2(breaks = seq(min_plt, max_plt, length.out
   = 5),
 low = "blue", mid = "white", high = "red", midpoint = med_plt)
```



We can use the **show.experiments** argument in order to visualize which points were specifically passed to the learner in the original experiment and which points were interpolated by mlr:



We can also visualize how long the optimizer takes to reach an optima for the same example:

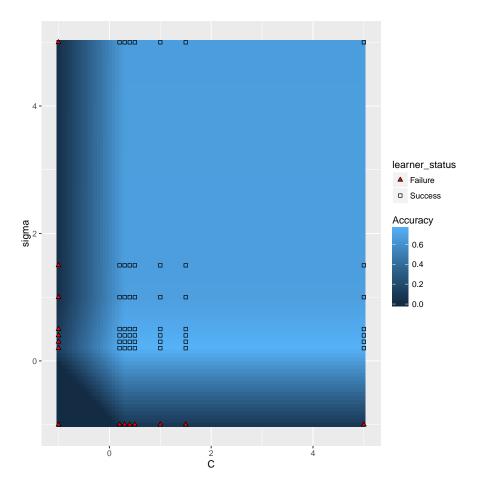
```
plotHyperParsEffect(data, x = "iteration", y = "acc.test.mean",
    plot.type = "line")
```



In the case where we are tuning 2 hyperparameters and we have a learner crash, mlr will indicate the respective points and impute them with the worst value. In the example below, we tune C and sigma, forcing C to be negative for some instances which will crash SVM. We perform interpolation to get a regular grid in order to plot a heatmap. We can see that the interpolation creates axis parallel lines resulting from the learner crashes.

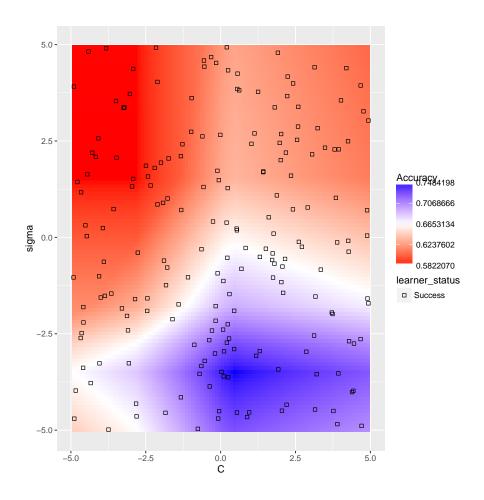
```
ps = makeParamSet(
  makeDiscreteParam("C", values = c(-1, 0.5, 1.5, 1, 0.2, 0.3, 0.4, 5)),
  makeDiscreteParam("sigma", values = c(-1, 0.5, 1.5, 1, 0.2, 0.3, 0.4, 5)))
ctrl = makeTuneControlGrid()
rdesc = makeResampleDesc("Holdout")
learn = makeLearner("classif.ksvm", par.vals = list(kernel = "rbfdot"))
res = tuneParams(learn, task = pid.task, control = ctrl, measures = "restriction of the control of t
```

```
acc,
resampling = rdesc, par.set = ps, show.info = FALSE)
data = generateHyperParsEffectData(res)
plotHyperParsEffect(data, x = "C", y = "sigma", z = "acc.test.mean",
    plot.type = "heatmap", interpolate = "regr.earth")
```



A slightly more complicated example is using nested cross validation while simultaneously tuning 2 hyperparameters. In order to plot a heatmap in this case, mlr will aggregate each of the nested runs by a user-specified function. The default function is **mean**. As expected, we can still take advantage of interpolation.

```
rdesc = makeResampleDesc("Holdout")
learn = makeLearner("classif.ksvm", par.vals = list(kernel =
   "rbfdot"))
lrn = makeTuneWrapper(learn, control = ctrl, measures = list(acc,
   mmce),
 resampling = rdesc, par.set = ps, show.info = FALSE)
res = resample(lrn, task = pid.task, resampling = cv2, extract =
   getTuneResult, show.info = FALSE)
data = generateHyperParsEffectData(res)
plt = plotHyperParsEffect(data, x = "C", y = "sigma", z =
   "acc.test.mean",
 plot.type = "heatmap", interpolate = "regr.earth",
      show.experiments = TRUE,
 nested.agg = mean)
min_plt = min(plt$data$acc.test.mean, na.rm = TRUE)
max_plt = max(plt$data$acc.test.mean, na.rm = TRUE)
med_plt = mean(c(min_plt, max_plt))
plt + scale_fill_gradient2(breaks = seq(min_plt, max_plt, length.out
   = 5),
low = "red", mid = "white", high = "blue", midpoint = med_plt)
```



4 Extend

4.1 Integrating Another Learner

In order to integrate a learning algorithm into mlr some interface code has to be written. Three functions are mandatory for each learner.

- First, define a new learner class with a name, description, capabilities, parameters, and a few other things. (An object of this class can then be generated by makeLearner.)
- Second, you need to provide a function that calls the learner function and builds the model given data (which makes it possible to invoke training by calling mlr's train function).

 Finally, a prediction function that returns predicted values given new data is required (which enables invoking prediction by calling mlr's predict function).

Technically, integrating a learning method means introducing a new S3 class and implementing the corresponding methods for the generic functions makerLerner, trainLearner, and predictLearner. Therefore we start with a quick overview of the involved classes and constructor functions.

4.1.1 Classes, constructors, and naming schemes

As you already know makeLearner generates an object of class Learner.

```
class(makeLearner(cl = "classif.lda"))
#> [1] "classif.lda"
                         "RLearnerClassif" "RLearner"
    "Learner"
class(makeLearner(cl = "regr.lm"))
#> [1] "regr.lm"
                      "RLearnerRegr" "RLearner"
                                                     "Learner"
class(makeLearner(cl = "surv.coxph"))
                     "RLearnerSurv" "RLearner"
#> [1] "surv.coxph"
                                                     "Learner"
class(makeLearner(cl = "cluster.kmeans"))
#> [1] "cluster.kmeans" "RLearnerCluster" "RLearner"
    "Learner"
class(makeLearner(cl = "multilabel.rFerns"))
#> [1] "multilabel.rFerns" "RLearnerMultilabel" "RLearner"
#> [4] "Learner"
```

The first element of each class attribute vector is the name of the learner class passed to the cl argument of makeLearner. Obviously, this adheres to the naming conventions

```
• "classif.<R_method_name>" for classification,
```

- "multilabel.<R_method_name>" for multilabel classification,
- "regr. < R_method_name > " for regression,
- "surv. < R_method_name > " for survival analysis, and
- "cluster.<R_method_name>" for clustering.

Additionally, there exist intermediate classes that reflect the type of learning problem, i.e., all classification learners inherit from RLearnerClassif, all regression learners from RLearnerRegr and so on. Their superclasses are RLearner and finally Learner. For all these (sub)classes there exist constructor functions

makeRLearner, makeRLearnerClassif, makeRLearneRegr etc. that are called internally by makeLearner.

A short side remark: As you might have noticed there does not exist a special learner class for cost-sensitive classification (costsens) with example-specific costs. This type of learning task is currently exclusively handled through wrappers like makeCostSensWeightedPairsWrapper.

In the following we show how to integrate learners for the five types of learning tasks mentioned above. Defining a completely new type of learner that has special properties and does not fit into one of the existing schemes is of course possible, but much more advanced and not covered here.

We use a classification example to explain some general principles (so even if you are interested in integrating a learner for another type of learning task you might want to read the following section). Examples for other types of learning tasks are shown later on.

4.1.2 Classification

We show how the Linear Discriminant Analysis from package MASS has been integrated into the classification learner classif.lda in mlr as an example.

4.1.2.1 Definition of the learner

The minimal information required to define a learner is the mlr name of the learner, its package, the parameter set, and the set of properties of your learner. In addition, you may provide a human-readable name, a short name and a note with information relevant to users of the learner.

First, name your learner. According to the naming conventions above the name starts with classif. and we choose classif.lda.

Second, we need to define the parameters of the learner. These are any options that can be set when running it to change how it learns, how input is interpreted, how and what output is generated, and so on. mlr provides a number of functions to define parameters, a complete list can be found in the documentation of LearnerParam of the ParamHelpers package.

In our example, we have discrete and numeric parameters, so we use makeDiscreteLearnerParam and makeNumericLearnerParam to incorporate the complete description of the parameters. We include all possible values for discrete parameters and lower and upper bounds for numeric parameters. Strictly speaking it is not necessary to provide bounds for all parameters and if this information is not available they can be estimated, but providing accurate and specific information here makes it possible to tune the learner much better (see the section on tuning).

Next, we add information on the properties of the learner (see also the section on learners). Which types of features are supported (numerics, factors)? Are case weights supported? Can the method deal with missing values in the features and deal with NA's in a meaningful way (not na.omit)? Are one-class, two-class, multi-class problems supported? Can the learner predict posterior probabilities?

If the learner supports class weights the name of the relevant learner parameter can be specified via argument class.weights.param.

Below is the complete code for the definition of the LDA learner. It has one discrete parameter, method, and two continuous ones, nu and tol. It supports classification problems with two or more classes and can deal with numeric and factor explanatory variables. It can predict posterior probabilities.

```
makeRLearner.classif.lda = function() {
 makeRLearnerClassif(
    cl = "classif.lda",
    package = "MASS",
    par.set = makeParamSet(
     makeDiscreteLearnerParam(id = "method", default = "moment",
         values = c("moment", "mle", "mve", "t")),
     makeNumericLearnerParam(id = "nu", lower = 2, requires =
          quote(method == "t")),
     makeNumericLearnerParam(id = "tol", default = 1e-4, lower = 0),
      makeDiscreteLearnerParam(id = "predict.method", values =
          c("plug-in", "predictive", "debiased"),
        default = "plug-in", when = "predict"),
      makeLogicalLearnerParam(id = "CV", default = FALSE, tunable =
          FALSE)
   ),
    properties = c("twoclass", "multiclass", "numerics", "factors",
   name = "Linear Discriminant Analysis",
    short.name = "lda",
    note = "Learner param 'predict.method' maps to 'method' in
        predict.lda."
 )
}
```

4.1.2.2 Creating the training function of the learner

Once the learner has been defined, we need to tell mlr how to call it to train a model. The name of the function has to start with trainLearner., followed by the mlr name of the learner as defined above (classif.lda here). The prototype of the function looks as follows.

```
function(.learner, .task, .subset, .weights = NULL, ...) { }
```

This function must fit a model on the data of the task .task with regard to the subset defined in the integer vector .subset and the parameters passed in the ... arguments. Usually, the data should be extracted from the task using getTaskData. This will take care of any subsetting as well. It must return the fitted model. mlr assumes no special data type for the return value – it will be passed to the predict function we are going to define below, so any special code the learner may need can be encapsulated there.

For our example, the definition of the function looks like this. In addition to the data of the task, we also need the formula that describes what to predict. We use the function getTaskFormula to extract this from the task.

```
trainLearner.classif.lda = function(.learner, .task, .subset,
    .weights = NULL, ...) {
    f = getTaskFormula(.task)
    MASS::lda(f, data = getTaskData(.task, .subset), ...)
}
```

4.1.2.3 Creating the prediction method

Finally, the prediction function needs to be defined. The name of this function starts with predictLearner., followed again by the mlr name of the learner. The prototype of the function is as follows.

```
function(.learner, .model, .newdata, ...) { }
```

It must predict for the new observations in the data.frame .newdata with the wrapped model .model, which is returned from the training function. The actual model the learner built is stored in the \$learner.model member and can be accessed simply through .model\$learner.model.

For classification, you have to return a factor of predicted classes if .learner\$predict.type is "response", or a matrix of predicted probabilities if .learner\$predict.type is "prob" and this type of prediction is supported by the learner. In the latter case the matrix must have the same number of columns as there are classes in the task and the columns have to be named by the class names.

The definition for LDA looks like this. It is pretty much just a straight passthrough of the arguments to the predict function and some extraction of prediction data depending on the type of prediction requested.

```
if (.learner$predict.type == "response")
  return(p$class) else return(p$posterior)
}
```

4.1.3 Regression

The main difference for regression is that the type of predictions are different (numeric instead of labels or probabilities) and that not all of the properties are relevant. In particular, whether one-, two-, or multi-class problems and posterior probabilities are supported is not applicable.

Apart from this, everything explained above applies. Below is the definition for the earth learner from the earth package.

```
makeRLearner.regr.earth = function() {
 makeRLearnerRegr(
    cl = "regr.earth",
    package = "earth",
    par.set = makeParamSet(
      makeLogicalLearnerParam(id = "keepxy", default = FALSE,
          tunable = FALSE),
      makeNumericLearnerParam(id = "trace", default = 0, upper = 10,
          tunable = FALSE),
      makeIntegerLearnerParam(id = "degree", default = 1L, lower =
          1L),
      makeNumericLearnerParam(id = "penalty"),
      makeIntegerLearnerParam(id = "nk", lower = 0L),
      makeNumericLearnerParam(id = "thres", default = 0.001),
      makeIntegerLearnerParam(id = "minspan", default = OL),
      makeIntegerLearnerParam(id = "endspan", default = OL),
      makeNumericLearnerParam(id = "newvar.penalty", default = 0),
      makeIntegerLearnerParam(id = "fast.k", default = 20L, lower =
      makeNumericLearnerParam(id = "fast.beta", default = 1),
      makeDiscreteLearnerParam(id = "pmethod", default = "backward",
        values = c("backward", "none", "exhaustive", "forward",
            "seqrep", "cv")),
     makeIntegerLearnerParam(id = "nprune")
    ),
    properties = c("numerics", "factors"),
    name = "Multivariate Adaptive Regression Splines",
    short.name = "earth",
    note = ""
 )
}
```

Again most of the data is passed straight through to/from the train/predict functions of the learner.

4.1.4 Survival analysis

For survival analysis, you have to return so-called linear predictors in order to compute the default measure for this task type, the cindex (for .learner\$predict.type == "response"). For .learner\$predict.type == "prob", there is no substantially meaningful measure (yet). You may either ignore this case or return something like predicted survival curves (cf. example below).

There are three properties that are specific to survival learners: "rcens", "lcens" and "icens", defining the type(s) of censoring a learner can handle – right, left and/or interval censored.

Let's have a look at how the Cox Proportional Hazard Model from package survival has been integrated into the survival learner surv.coxph in mlr as an example:

```
makeIntegerLearnerParam(id = "outer.max", default = 10L, lower
     makeLogicalLearnerParam(id = "model", default = FALSE, tunable
          = FALSE),
     makeLogicalLearnerParam(id = "x", default = FALSE, tunable =
          FALSE),
      makeLogicalLearnerParam(id = "y", default = TRUE, tunable =
          FALSE)
   properties = c("missings", "numerics", "factors", "weights",
        "prob", "rcens"),
   name = "Cox Proportional Hazard Model",
    short.name = "coxph",
    note = ""
}
trainLearner.surv.coxph = function(.learner, .task, .subset,
    .weights = NULL, ...) {
 f = getTaskFormula(.task)
 data = getTaskData(.task, subset = .subset)
 if (is.null(.weights)) {
   mod = survival::coxph(formula = f, data = data, ...)
 } else {
   mod = survival::coxph(formula = f, data = data, weights =
        .weights, ...)
 }
 mod
}
predictLearner.surv.coxph = function(.learner, .model, .newdata,
 if (.learner$predict.type == "response") {
   predict(.model$learner.model, newdata = .newdata, type = "lp",
 }
}
```

4.1.5 Clustering

For clustering, you have to return a numeric vector with the IDs of the clusters that the respective datum has been assigned to. The numbering should start at 1

Below is the definition for the FarthestFirst learner from the RWeka package. Weka starts the IDs of the clusters at 0, so we add 1 to the predicted clusters.

RWeka has a different way of setting learner parameters; we use the special Weka_control function to do this.

```
makeRLearner.cluster.FarthestFirst = function() {
 makeRLearnerCluster(
    cl = "cluster.FarthestFirst",
   package = "RWeka",
   par.set = makeParamSet(
      makeIntegerLearnerParam(id = "N", default = 2L, lower = 1L),
     makeIntegerLearnerParam(id = "S", default = 1L, lower = 1L),
     makeLogicalLearnerParam(id = "output-debug-info", default =
          FALSE, tunable = FALSE)
    ),
    properties = c("numerics"),
   name = "FarthestFirst Clustering Algorithm",
    short.name = "farthestfirst"
 )
}
trainLearner.cluster.FarthestFirst = function(.learner, .task,
    .subset, .weights = NULL, ...) {
 ctrl = RWeka::Weka_control(...)
 RWeka::FarthestFirst(getTaskData(.task, .subset), control = ctrl)
}
predictLearner.cluster.FarthestFirst = function(.learner, .model,
    .newdata, ...) {
 as.integer(predict(.model$learner.model, .newdata, ...)) + 1L
```

4.1.6 Multilabel classification

As stated in the multilabel section, multilabel classification methods can be divided into problem transformation methods and algorithm adaptation methods.

At this moment the only problem transformation method implemented in mlr is the binary relevance method. Integrating more of these methods requires good knowledge of the architecture of the mlr package.

The integration of an algorithm adaptation multilabel classification learner is easier and works very similar to the normal multiclass-classification. In contrast to the multiclass case, not all of the learner properties are relevant. In particular, whether one-, two-, or multi-class problems are supported is not applicable. Furthermore the prediction function output must be a matrix with each prediction of a label in one column and the names of the labels as column names. If .learner\$predict.type is "response" the predictions must be logical. If

.learner\$predict.type is "prob" and this type of prediction is supported by the learner, the matrix must consist of predicted probabilities.

Below is the definition of the rFerns learner from the rFerns package, which does not support probability predictions.

```
makeRLearner.multilabel.rFerns = function() {
 makeRLearnerMultilabel(
    cl = "multilabel.rFerns",
   package = "rFerns",
   par.set = makeParamSet(
     makeIntegerLearnerParam(id = "depth", default = 5L),
     makeIntegerLearnerParam(id = "ferns", default = 1000L)
   properties = c("numerics", "factors", "ordered"),
   name = "Random ferns",
    short.name = "rFerns",
   note = ""
 )
}
trainLearner.multilabel.rFerns = function(.learner, .task, .subset,
    .weights = NULL, ...) {
 d = getTaskData(.task, .subset, target.extra = TRUE)
 rFerns::rFerns(x = d$data, y = as.matrix(d$target), ...)
}
predictLearner.multilabel.rFerns = function(.learner, .model,
    .newdata, ...) {
 as.matrix(predict(.model$learner.model, .newdata, ...))
```

4.1.7 Creating a new feature importance method

Some learners, for example decision trees and random forests, can calculate feature importance values, which can be extracted from a fitted model using function getFeatureImportance.

If your newly integrated learner supports this you need to

- add "featimp" to the learner properties and
- implement a new S3 method for function getFeatureImportanceLearner (which later is called internally by getFeatureImportance).

in order to make this work.

This method takes the Learner .learner, the WrappedModel .model and potential further arguments and calculates or extracts the feature importance. It must return a named vector of importance values.

Below are two simple examples. In case of "classif.rpart" the feature importance values can be easily extracted from the fitted model.

For the random forest from package randomForestSRC function vimp is called.

```
getFeatureImportanceLearner.classif.randomForestSRC =
   function(.learner, .model, ...) {
   mod = getLearnerModel(.model)
   randomForestSRC::vimp(mod, ...)$importance[, "all"]
}
```

4.1.8 Registering your learner

If your interface code to a new learning algorithm exists only locally, i.e., it is not (yet) merged into mlr or does not live in an extra package with a proper namespace you might want to register the new S3 methods to make sure that these are found by, e.g., listLearners. You can do this as follows:

4.2 Integrating Another Measure

In some cases, you might want to evaluate a Prediction or ResamplePrediction with a Measure which is not yet implemented in mlr. This could be either a performance measure which is not listed in the Appendix or a measure that uses a misclassification cost matrix.

4.2.1 Performance measures and aggregation schemes

Performance measures in mlr are objects of class Measure. For example the mse (mean squared error) looks as follows.

```
str(mse)
#> List of 10
                : chr "mse"
#> $ id
#> $ minimize : logi TRUE
#> $ properties: chr [1:3] "regr" "req.pred" "req.truth"
                :function (task, model, pred, feats, extra.args)
#> $ fun
#> $ extra.args: list()
#> $ best : num 0
#> $ worst
               : num Inf
#> $ name
               : chr "Mean of squared errors"
#> $ note : chr "Defined as: mean((response - truth)^2)"
#> $ aggr :List of 4
#> ..$ id : chr "test.mean"
               : chr "test.mean"
                : chr "Test mean"
#>
    ..$ name
#> ..$ fun
                   :function (task, perf.test, perf.train, measure,
   group, pred)
     ..$ properties: chr "req.test"
    ..- attr(*, "class")= chr "Aggregation"
#> - attr(*, "class")= chr "Measure"
mse$fun
#> function (task, model, pred, feats, extra.args)
#>
       measureMSE(pred$data$truth, pred$data$response)
#> }
#> <bytecode: 0x90f7f50>
#> <environment: namespace:mlr>
measureMSE
#> function (truth, response)
#> {
#>
       mean((response - truth)~2)
#> }
#> <bytecode: 0x8cd9ff8>
#> <environment: namespace:mlr>
```

See the Measure documentation page for a detailed description of the object slots.

At the core is slot \$fun which contains the function that calculates the performance value. The actual work is done by function measureMSE. Similar functions, generally adhering to the naming scheme measure followed by the

capitalized measure ID, exist for most performance measures. See the measures help page for a complete list.

Just as Task and Learner objects each Measure has an identifier \$id which is for example used to annotate results and plots. For plots there is also the option to use the longer measure \$name instead. See the tutorial page on Visualization for more information.

Moreover, a Measure includes a number of \$properties that indicate for which types of learning problems it is suitable and what information is required to calculate it. Obviously, most measures need the Prediction object ("req.pred") and, for supervised problems, the true values of the target variable(s) ("req.truth").

For tuning or feature selection each Measure knows its extreme values \$best and \$worst and if it wants to be minimized or maximized (\$minimize).

For resampling slot \$aggr specifies how the overall performance across all resampling iterations is calculated. Typically, this is just a matter of aggregating the performance values obtained on the test sets perf.test or the training sets perf.train by a simple function. The by far most common scheme is test.mean, i.e., the unweighted mean of the performances on the test sets.

```
str(test.mean)
#> List of 4
                : chr "test.mean"
#> $ id
                : chr "Test mean"
#> $ name
  $ fun
                :function (task, perf.test, perf.train, measure,
    group, pred)
#> $ properties: chr "req.test"
   - attr(*, "class")= chr "Aggregation"
test.mean$fun
#> function (task, perf.test, perf.train, measure, group, pred)
#> mean(perf.test)
#> <bytecode: 0x638ab50>
#> <environment: namespace:mlr>
```

All aggregation schemes are objects of class Aggregation with the function in slot **\$fun** doing the actual work. The **\$properties** member indicates if predictions (or performance values) on the training or test data sets are required to calculate the aggregation.

You can change the aggregation scheme of a Measure via function setAggregation. See the tutorial page on resampling for some examples and the aggregations help page for all available aggregation schemes.

You can construct your own Measure and Aggregation objects via functions makeMeasure, makeCostMeasure, makeCustomResampledMeasure and makeAggregation. Some examples are shown in the following.

4.2.2 Constructing a performance measure

Function makeMeasure provides a simple way to construct your own performance measure.

Below this is exemplified by re-implementing the mean misclassification error (mmce). We first write a function that computes the measure on the basis of the true and predicted class labels. Note that this function must have certain formal arguments listed in the documentation of makeMeasure. Then the Measure object is created and we work with it as usual with the performance function.

See the ${f R}$ documentation of make Measure for more details on the various parameters.

```
### Define a function that calculates the misclassification rate
my.mmce.fun = function(task, model, pred, feats, extra.args) {
 tb = table(getPredictionResponse(pred), getPredictionTruth(pred))
 1 - sum(diag(tb)) / sum(tb)
}
### Generate the Measure object
my.mmce = makeMeasure(
 id = "my.mmce", name = "My Mean Misclassification Error",
 properties = c("classif", "classif.multi", "req.pred",
      "req.truth"),
 minimize = TRUE, best = 0, worst = 1,
 fun = my.mmce.fun
### Train a learner and make predictions
mod = train("classif.lda", iris.task)
pred = predict(mod, task = iris.task)
### Calculate the performance using the new measure
performance(pred, measures = my.mmce)
#> my.mmce
#>
      0.02
### Apparently the result coincides with the mlr implementation
performance(pred, measures = mmce)
#> mmce
#> 0.02
```

4.2.3 Constructing a measure for ordinary misclassification costs

For in depth explanations and details see the tutorial page on cost-sensitive classification.

To create a measure that involves ordinary, i.e., class-dependent misclassification costs you can use function makeCostMeasure. You first need to define the cost matrix. The rows indicate true and the columns predicted classes and the rows and columns have to be named by the class labels. The cost matrix can then be wrapped in a Measure object and predictions can be evaluated as usual with the performance function.

See the ${\bf R}$ documentation of function makeCostMeasure for details on the various parameters.

```
### Create the cost matrix
costs = matrix(c(0, 2, 2, 3, 0, 2, 1, 1, 0), ncol = 3)
rownames(costs) = colnames(costs) = getTaskClassLevels(iris.task)

### Encapsulate the cost matrix in a Measure object
my.costs = makeCostMeasure(
   id = "my.costs", name = "My Costs",
   costs = costs,
   minimize = TRUE, best = 0, worst = 3
)

### Train a learner and make a prediction
mod = train("classif.lda", iris.task)
pred = predict(mod, newdata = iris)

### Calculate the average costs
performance(pred, measures = my.costs)

#> my.costs
#> my.costs
#> 0.02666667
```

4.2.4 Creating an aggregation scheme

It is possible to create your own aggregation scheme using function makeAggregation. You need to specify an identifier id, the properties, and write a function that does the actual aggregation. Optionally, you can name your aggregation scheme.

Possible settings for properties are "req.test" and "req.train" if predictions on either the training or test sets are required, and the vector c("req.train", "req.test") if both are needed.

The aggregation function must have a certain signature detailed in the documentation of makeAggregation. Usually, you will only need the performance values on the test sets perf.test or the training sets perf.train. In rare cases, e.g., the Prediction object pred or information stored in the Task object might be required to obtain the aggregated performance. For an example have a look at the definition of function test.join.

4.2.4.1 Example: Evaluating the range of measures

Let's say you are interested in the range of the performance values obtained on individual test sets.

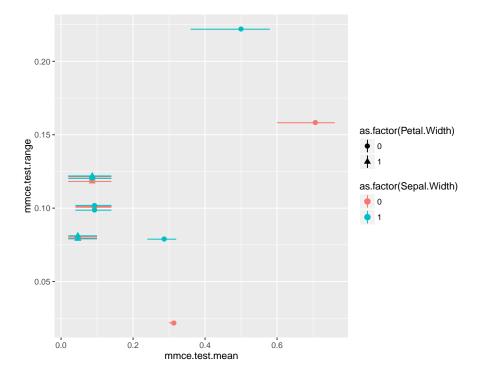
```
my.range.aggr = makeAggregation(id = "test.range", name = "Test
    Range",
    properties = "req.test",
    fun = function (task, perf.test, perf.train, measure, group, pred)
    diff(range(perf.test))
)
```

perf.train and perf.test are both numerical vectors containing the performances on the train and test data sets. In most cases (unless you are using bootstrap as resampling strategy or have set predict = "both" in makeResampleDesc) the perf.train vector is empty.

Now we can run a feature selection based on the first measure in the provided list and see how the other measures turn out.

```
### mmce with default aggregation scheme test.mean
ms1 = mmce
### mmce with new aggregation scheme my.range.aggr
ms2 = setAggregation(ms1, my.range.aggr)
### Minimum and maximum of the mmce over test sets
ms1min = setAggregation(ms1, test.min)
ms1max = setAggregation(ms1, test.max)
### Feature selection
rdesc = makeResampleDesc("CV", iters = 3)
res = selectFeatures("classif.rpart", iris.task, rdesc, measures =
   list(ms1, ms2, ms1min, ms1max),
 control = makeFeatSelControlExhaustive(), show.info = FALSE)
### Optimization path, i.e., performances for the 16 possible
   feature subsets
perf.data = as.data.frame(res$opt.path)
head(perf.data[1:8])
#> Sepal.Length Sepal.Width Petal.Length Petal.Width mmce.test.mean
#> 1
           0
                      0 0.70666667
#> 2
             1
                        0
                                  0
                                             0
                                                  0.31333333
#> 3
            0
                       1
                                  0
                                            0 0.50000000
                       0
#> 4
            0
                                  1
                                            0 0.09333333
                                             1 0.04666667
#> 5
            0
                       0
                                  0
                                   0
#> 6
             1
                       1
                                              0
                                                  0.2866667
#> mmce.test.range mmce.test.min mmce.test.max
```

```
#> 1
                 0.16
                               0.60
                                              0.76
#> 2
                 0.02
                               0.30
                                              0.32
#> 3
                 0.22
                               0.36
                                              0.58
#> 4
                                              0.14
                 0.10
                               0.04
#> 5
                 0.08
                               0.02
                                              0.10
#> 6
                 0.08
                                              0.32
                               0.24
pd = position_jitter(width = 0.005, height = 0)
p = ggplot(aes(x = mmce.test.range, y = mmce.test.mean, ymax =
    mmce.test.max, ymin = mmce.test.min,
  color = as.factor(Sepal.Width), pch = as.factor(Petal.Width)),
      data = perf.data) +
  geom_pointrange(position = pd) +
  coord_flip()
print(p)
```



The plot shows the range versus the mean misclassification error. The value on the y-axis thus corresponds to the length of the error bars. (Note that the points and error bars are jittered in y-direction.)

4.3 Creating an Imputation Method

Function makeImputeMethod permits to create your own imputation method. For this purpose you need to specify a *learn* function that extracts the necessary information and an *impute* function that does the actual imputation. The *learn* and *impute* functions both have at least the following formal arguments:

- data is a data.frame with missing values in some features.
- col indicates the feature to be imputed.
- target indicates the target variable(s) in a supervised learning task.

4.3.1 Example: Imputation using the mean

Let's have a look at function imputeMean.

```
imputeMean = function() {
  makeImputeMethod(learn = function(data, target, col)
      mean(data[[col]], na.rm = TRUE),
  impute = simpleImpute)
}
```

imputeMean calls the unexported mlr function simpleImpute which is defined as follows.

```
simpleImpute = function(data, target, col, const) {
  if (is.na(const))
    stopf("Error imputing column '%s'. Maybe all input data was
        missing?", col)
  x = data[[col]]
  if (is.factor(x) && const %nin% levels(x)) {
    levels(x) = c(levels(x), as.character(const))
  }
  replace(x, is.na(x), const)
}
```

The *learn* function calculates the mean of the non-missing observations in column col. The mean is passed via argument const to the *impute* function that replaces all missing values in feature col.

4.3.2 Writing your own imputation method

Now let's write a new imputation method: A frequently used simple technique for longitudinal data is *last observation carried forward* (LOCF). Missing values are replaced by the most recent observed value.

In the **R** code below the *learn* function determines the last observed value previous to each NA (values) as well as the corresponding number of consecutive NA's (times). The *impute* function generates a vector by replicating the entries in values according to times and replaces the NA's in feature col.

```
imputeLOCF = function() {
  makeImputeMethod(
    learn = function(data, target, col) {
      x = data[[col]]
      ind = is.na(x)
      dind = diff(ind)
      lastValue = which(dind == 1) # position of the last observed
          value previous to NA
      lastNA = which(dind == -1)
                                    # position of the last of
          potentially several consecutive NA's
      values = x[lastValue]
                                    # last observed value previous
          to NA
      times = lastNA - lastValue
                                    # number of consecutive NA's
      return(list(values = values, times = times))
    },
    impute = function(data, target, col, values, times) {
      x = data[[col]]
      replace(x, is.na(x), rep(values, times))
    }
  )
}
```

Note that this function is just for demonstration and is lacking some checks for real-world usage (for example 'What should happen if the first value in x is already missing?'). Below it is used to impute the missing values in features Ozone and Solar.R in the airquality data set.

```
data(airquality)
imp = impute(airquality, cols = list(Ozone = imputeLOCF(), Solar.R =
    imputeLOCF()),
 dummy.cols = c("Ozone", "Solar.R"))
head(imp$data, 10)
#>
      Ozone Solar.R Wind Temp Month Day Ozone.dummy Solar.R.dummy
#> 1
         41
               190 7.4
                           67
                                 5
                                     1
                                              FALSE
                                                            FALSE
#> 2
         36
               118 8.0
                          72
                                  5
                                    2
                                             FALSE
                                                            FALSE
#> 3
        12
               149 12.6
                          74
                                 5
                                    3
                                             FALSE
                                                           FALSE
#> 4
         18
               313 11.5
                          62
                                 5
                                     4
                                              FALSE
                                                            FALSE
         18
               313 14.3
                          56
                                 5
                                              TRUE
                                                            TRUE
#> 5
                                     5
#> 6
        28
               313 14.9
                          66
                                 5
                                     6
                                             FALSE
                                                            TRUE
               299 8.6
#> 7
        23
                          65
                                 5
                                     7
                                             FALSE
                                                           FALSE
#> 8
        19
                 99 13.8
                          59
                                 5
                                     8
                                              FALSE
                                                            FALSE
#> 9
      8
            19 20.1
                          61
                                 5
                                     9
                                             FALSE
                                                           FALSE
```

```
#> 10 8 194 8.6 69 5 10 TRUE FALSE
```

4.4 Integrating Another Filter Method

A lot of feature filter methods are already integrated in mlr and a complete list is given in the Appendix or can be obtained using listFilterMethods. You can easily add another filter, be it a brand new one or a method which is already implemented in another package, via function makeFilter.

4.4.1 Filter objects

In mlr all filter methods are objects of class Filter and are registered in an environment called .FilterRegister (where listFilterMethods looks them up to compile the list of available methods). To get to know their structure let's have a closer look at the "rank.correlation" filter which interfaces function rank.correlation in package FSelector.

```
filters = as.list(mlr:::.FilterRegister)
filters$rank.correlation
#> Filter: 'rank.correlation'
#> Packages: 'Rfast'
#> Supported tasks: regr
#> Supported features: numerics
str(filters$rank.correlation)
#> List of 6
#> $ name
                        : chr "rank.correlation"
#> $ desc
                       : chr "Spearman's correlation between
   feature and target"
                        : chr "Rfast"
#> $ pkg
#> $ supported.tasks : chr "regr"
#> $ supported.features: chr "numerics"
#> $ fun
                       :function (task, nselect, ...)
#> - attr(*, "class")= chr "Filter"
filters$rank.correlation$fun
#> function (task, nselect, ...)
#> {
#>
       d = getTaskData(task, target.extra = TRUE)
#>
       y = Rfast::correls(d$target, d$data, type = "spearman")
#>
       setNames(abs(y[, "correlation"]), getTaskFeatureNames(task))
#> }
#> <bytecode: 0x8941938>
#> <environment: namespace:mlr>
```

The core element is **\$fun** which calculates the feature importance. For the **"rank.correlation"** filter it just extracts the data and formula from the **task** and passes them on to the rank.correlation function.

Additionally, each Filter object has a \$name, which should be short and is for example used to annotate graphics (cp. plotFilterValues), and a slightly more detailed description in slot \$desc. If the filter method is implemented by another package its name is given in the \$pkg member. Moreover, the supported task types and feature types are listed.

4.4.2 Writing a new filter method

You can integrate your own filter method using makeFilter. This function generates a Filter object and also registers it in the .FilterRegister environment.

The arguments of makeFilter correspond to the slot names of the Filter object above. Currently, feature filtering is only supported for supervised learning tasks and possible values for supported.tasks are "regr", "classif" and "surv". supported.features can be "numerics", "factors" and "ordered".

fun must be a function with at least the following formal arguments:

- task is a mlr learning Task.
- nselect corresponds to the argument of generateFilterValuesData of the same name and specifies the number of features for which to calculate importance scores. Some filter methods have the option to stop after a certain number of top-ranked features have been found, in order to save time and ressources when the number of features is high. The majority of filter methods integrated in mlr doesn't support this and thus nselect is ignored in most cases. An exception is the minimum redundancy maximum relevance filter from package mRMRe.
- ... for additional arguments.

fun must return a named vector of feature importance values. By convention the most important features receive the highest scores.

If nselect is actively used fun can either return a vector of nselect scores or a vector as long as the numbers of features in the task that contains NAs for all features whose scores weren't calculated.

For writing fun many of the getter functions for Tasks come in handy, particularly getTaskData, getTaskFormula and getTaskFeatureNames. It's worth having a closer look at getTaskData which provides many options for formatting the data and recoding the target variable.

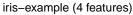
As a short demonstration we write a totally meaningless filter that determines the importance of features according to alphabetical order, i.e., giving highest scores to features with names that come first (decreasing = TRUE) or last (decreasing = FALSE) in the alphabet.

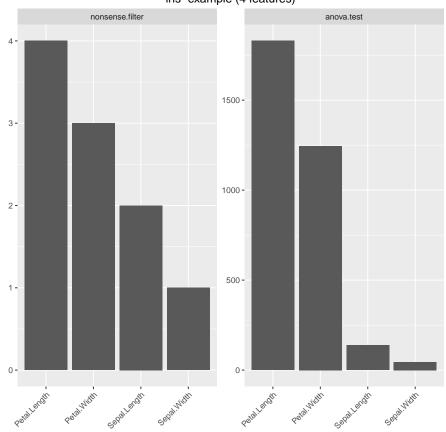
```
makeFilter(
  name = "nonsense.filter",
  desc = "Calculates scores according to alphabetical order of
     features",
  pkg = "",
  supported.tasks = c("classif", "regr", "surv"),
  supported.features = c("numerics", "factors", "ordered"),
  fun = function(task, nselect, decreasing = TRUE, ...) {
    feats = getTaskFeatureNames(task)
    imp = order(feats, decreasing = decreasing)
    names(imp) = feats
    imp
  }
)
#> Filter: 'nonsense.filter'
#> Packages: ''
#> Supported tasks: classif, regr, surv
#> Supported features: numerics, factors, ordered
```

The nonsense.filter is now registered in mlr and shown by listFilterMethods.

```
listFilterMethods()$id
#> [1] anova.test
                                  carscore
#> [3] cforest.importance
                                chi.squared
#> [5] gain.ratio
                                 information.gain
#> [7] kruskal.test
                                 linear.correlation
#> [9] mrmr
                                 nonsense.filter
#> [11] oneR
                                permutation.importance
#> [13] randomForest.importance
                                 randomForestSRC.rfsrc
#> [15] randomForestSRC.var.select rank.correlation
#> [17] relief
                                 rf.importance
#> [19] rf.min.depth
                                symmetrical.uncertainty
#> [21] univariate
                                 univariate.model.score
#> [23] variance
#> 23 Levels: anova.test carscore cforest.importance ... variance
```

You can use it like any other filter method already integrated in mlr (i.e., via the method argument of generateFilterValuesData or the fw.method argument of makeFilterWrapper; see also the page on feature selection).





```
iris.task.filtered = filterFeatures(iris.task, method =
    "nonsense.filter", abs = 2)
iris.task.filtered
#> Supervised task: iris-example
#> Type: classif
#> Target: Species
#> Observations: 150
#> Features:
#> numerics factors ordered
#> 2 0 0
```

```
#> Missings: FALSE
#> Has weights: FALSE
#> Has blocking: FALSE
#> Classes: 3
#> setosa versicolor virginica
#> 50 50 50
#> Positive class: NA

getTaskFeatureNames(iris.task.filtered)
#> [1] "Petal.Length" "Petal.Width"
```

You might also want to have a look at the source code of the filter methods already integrated in mlr for some more complex and meaningful examples.