

Package ‘caret’

February 6, 2012

Version 5.13-20

Date 2012-02-06

Title Classification and Regression Training

Author Max Kuhn. Contributions from Jed Wing, Steve Weston, Andre Williams, Chris Keefer and Allan Engelhardt

Description Misc functions for training and plotting classification and regression models

Maintainer Max Kuhn <Max.Kuhn@pfizer.com>

Depends R (>= 2.10), lattice, reshape, stats, plyr, cluster,foreach

URL <http://caret.r-forge.r-project.org/>

Suggests gbm, pls, mlbench, rpart, ellipse, ipred, klaR, randomForest,gpls, pamr, kernlab, mda, mgcv, nnet, class, MASS, mboost,earth (>= 2.2-3), party (>= 0.9-99992), ada, affy, proxy,e1071, grid, elasticnet, SDDA, caTools, RWeka (>= 0.4-1),superpc, penalized, sparseLDA (>= 0.1-1), spls, sda, glmnet (>= 1.7.1), relaxo, lars, vbmp, nodeHarvest, rrcov, gam, stepPlr,GAMens (>= 1.1.1), rocc, foba, partDSA, hda, fastICA,neuralnet, quantregForest, rda, HD-classif, LogicReg,LogicForest, logicFS, RANN, qrnn, Boruta, Hmisc, Cubist, bst,leaps, obliqueRF, rrllda, evtree, penalizedLDA, rFerns, pROC,kohonen

License GPL-2

Repository CRAN

Date/Publication 2012-02-06 21:06:39

R topics documented:

Alternate Affy Gene Expression Summary Methods.	3
as.table.confusionMatrix	4
avNNNet.default	6
bag.default	8
bagEarth	10

bagFDA	12
BloodBrain	13
BoxCoxTrans.default	14
calibration	16
caretFuncs	18
caretSBF	19
cars	20
classDist	21
confusionMatrix	23
confusionMatrix.train	25
cox2	27
createDataPartition	28
createGrid	29
dhfr	31
diff.resamples	31
dotPlot	33
dotplot.diff.resamples	34
dummyVars	36
featurePlot	38
filterVarImp	40
findCorrelation	41
findLinearCombos	42
format.bagEarth	43
GermanCredit	44
histogram.train	45
icr.formula	46
knn3	48
knnreg	49
lattice.rfe	51
lift	53
maxDissim	55
mdrr	57
modelLookup	58
nearZeroVar	59
normalize.AffyBatch.normalize2Reference	60
normalize2Reference	62
nullModel	63
oil	64
oneSE	65
panel.lift2	67
panel.needle	69
pcaNNet.default	70
plot.train	72
plot.varImp.train	73
plotClassProbs	74
plotObsVsPred	75
plsda	77
postResample	80

pottery	82
prcomp.resamples	82
predict.bagEarth	84
predict.knn3	85
predict.knnreg	86
predict.train	87
predictors	89
preProcess	94
print.confusionMatrix	97
print.train	97
resampleHist	99
resamples	100
resampleSummary	102
rfe	103
rfeControl	106
sbf	109
sbfControl	111
segmentationData	113
sensitivity	114
spatialSign	117
summary.bagEarth	119
teactor	120
train	121
trainControl	128
update.train	130
varImp	132
xyplot.resamples	135

Index**138**

 Alternate Affy Gene Expression Summary Methods.

Generate Expression Values from Probes

Description

Generate an expression from the probes

Usage

```
generateExprVal.method.trimMean(probes, trim = 0.15)
```

Arguments

probes	a matrix of probe intensities with rows representing probes and columns representing samples. Usually <code>pm(probeset)</code> where <code>probeset</code> is a of class <code>ProbeSet</code>
trim	the fraction (0 to 0.5) of observations to be trimmed from each end of the data before the mean is computed.

Value

A list containing entries:

exprs	The expression values.
se.exprs	The standard error estimate.

See Also

[generateExprSet-methods](#)

Examples

```
## Not run:
# first, let affy/expresso know that the method exists
express.summary.stat.methods <- c(express.summary.stat.methods, "trimMean")

example not run, as it would take a while
RawData <- ReadAffy(celfile.path=FilePath)

expresso(RawData,
  bgcorrect.method="rma",
  normalize.method="quantiles",
  normalize.param = list(type= "pmonly"),
  pmcorrect.method="pmonly",
  summary.method="trimMean")

step1 <- bg.correct(RawData, "rma")
step2 <- normalize.AffyBatch.quantiles(step1)
step3 <- computeExprSet(step2, "pmonly", "trimMean")

## End(Not run)
```

as.table.confusionMatrix

Save Confusion Table Results

Description

Conversion functions for class confusionMatrix

Usage

```
## S3 method for class 'confusionMatrix'
as.matrix(x, what = "xtabs", ...)
```

```
## S3 method for class 'confusionMatrix'
as.table(x, ...)
```

Arguments

x an object of class `confusionMatrix`
 what data to convert to matrix. Either "xtabs", "overall" or "classes"
 ... not currently used

Details

For `as.table`, the cross-tabulations are saved. For `as.matrix`, the three object types are saved in matrix format.

Value

A matrix or table

Author(s)

Max Kuhn

See Also

[confusionMatrix](#)

Examples

```
#####
## 2 class example

lvs <- c("normal", "abnormal")
truth <- factor(rep(lvs, times = c(86, 258)),
               levels = rev(lvs))
pred <- factor(
  c(
    rep(lvs, times = c(54, 32)),
    rep(lvs, times = c(27, 231))),
  levels = rev(lvs))

xtab <- table(pred, truth)

results <- confusionMatrix(xtab)
as.table(results)
as.matrix(results)
as.matrix(results, what = "overall")
as.matrix(results, what = "classes")

#####
## 3 class example

library(MASS)

fit <- lda(Species ~ ., data = iris)
model <- predict(fit)$class
```

```

irisTabs <- table(model, iris$Species)

results <- confusionMatrix(irisTabs)
as.table(results)
as.matrix(results)
as.matrix(results, what = "overall")
as.matrix(results, what = "classes")

```

avNNet.default

Neural Networks Using Model Averaging

Description

Aggregate several neural network model

Usage

```

## Default S3 method:
avNNet(x, y, repeats = 5, bag = FALSE, ...)
## S3 method for class 'formula'
avNNet(formula, data, weights, ...,
        repeats = 5, bag = FALSE, subset, na.action, contrasts = NULL)

## S3 method for class 'avNNet'
predict(object, newdata, type = c("raw", "class", "prob"), ...)

```

Arguments

formula	A formula of the form <code>class ~ x1 + x2 + ...</code>
x	matrix or data frame of x values for examples.
y	matrix or data frame of target values for examples.
weights	(case) weights for each example – if missing defaults to 1.
repeats	the number of neural networks with with different random number seeds
bag	a logical for bagging for each repeat
data	Data frame from which variables specified in <code>formula</code> are preferentially to be taken.
subset	An index vector specifying the cases to be used in the training sample. (NOTE: If given, this argument must be named.)
na.action	A function to specify the action to be taken if NAs are found. The default action is for the procedure to fail. An alternative is <code>na.omit</code> , which leads to rejection of cases with missing values on any required variable. (NOTE: If given, this argument must be named.)

contrasts	a list of contrasts to be used for some or all of the factors appearing as variables in the model formula.
object	an object of class avNNet as returned by avNNet.
newdata	matrix or data frame of test examples. A vector is considered to be a row vector comprising a single case.
type	Type of output, either: raw for the raw outputs, code for the predicted class or prob for the class probabilities.
...	arguments passed to nnet

Details

Following Ripley (1996), the same neural network model is fit using different random number seeds. All of the resulting models are used for prediction. For regression, the output from each network are averaged. For classification, the model scores are first averaged, then translated to predicted classes. Bagging can also be used to create the models.

Value

For avNNet, an object of "avNNet" or "avNNet.formula". Items of interest in the output are:

model	a list of the models generated from nnet
repeats	an echo of the model input
names	if any predictors had only one distinct value, this is a character string of the remaining columns. Otherwise a value of NULL

Author(s)

These are heavily based on the nnet code from Brian Ripley.

References

Ripley, B. D. (1996) *Pattern Recognition and Neural Networks*. Cambridge.

See Also

[nnet](#), [preProcess](#)

Examples

```
data(BloodBrain)
modelFit <- avNNet(bbbDescr[, 1:10], logBBB, size = 5, linout = TRUE, trace = FALSE)
modelFit

predict(modelFit, bbbDescr[, 1:10])
```

Description

bag provides a framework for bagging classification or regression models. The user can provide their own functions for model building, prediction and aggregation of predictions (see Details below).

Usage

```
bag(x, ...)

## Default S3 method:
bag(x, y, B = 10, vars = ncol(x), bagControl = bagControl(), ...)

bagControl(fit = NULL,
           predict = NULL,
           aggregate = NULL,
           downSample = FALSE)

ldaBag
plsBag
nbBag
ctreeBag
svmBag
nnetBag

## S3 method for class 'bag'
predict(object, newdata = NULL, ...)
```

Arguments

x	a matrix or data frame of predictors
y	a vector of outcomes
B	the number of bootstrap samples to train over.
bagControl	a list of options.
...	arguments to pass to the model function
fit	a function that has arguments x, y and ... and produces a model object that can later be used for prediction. Example functions are found in ldaBag, plsBag, nbBag, svmBag and nnetBag.
predict	a function that generates predictions for each sub-model. The function should have arguments object and x. The output of the function can be any type of object (see the example below where posterior probabilities are generated. Example functions are found in ldaBag, plsBag, nbBag, svmBag and nnetBag.)

aggregate	a function with arguments <code>x</code> and <code>type</code> . The function that takes the output of the <code>predict</code> function and reduces the bagged predictions to a single prediction per sample. the <code>type</code> argument can be used to switch between predicting classes or class probabilities for classification models. Example functions are found in <code>ldaBag</code> , <code>plsBag</code> , <code>nbBag</code> , <code>svmBag</code> and <code>nnetBag</code> .
downSample	a logical: for classification, should the data set be randomly sampled so that each class has the same number of samples as the smallest class?
vars	an integer. If this argument is not <code>NULL</code> , a random sample of size <code>vars</code> is taken of the predictors in each bagging iteration. If <code>NULL</code> , all predictors are used.
object	an object of class <code>bag</code> .
newdata	a matrix or data frame of samples for prediction. Note that this argument must have a non-null value

Details

The function is basically a framework where users can plug in any model in to assess the effect of bagging. Examples functions can be found in `ldaBag`, `plsBag`, `nbBag`, `svmBag` and `nnetBag`. Each has elements `fit`, `pred` and `aggregate`.

One note: when `vars` is not `NULL`, the sub-setting occurs prior to the `fit` and `predict` functions are called. In this way, the user probably does not need to account for the change in predictors in their functions.

When using `bag` with `train`, classification models should use `type = "prob"` inside of the `predict` function so that `predict.train(object, newdata, type = "prob")` will work.

Value

`bag` produces an object of class `bag` with elements

<code>fits</code>	a list with two sub-objects: the <code>fit</code> object has the actual model fit for that bagged samples and the <code>vars</code> object is either <code>NULL</code> or a vector of integers corresponding to which predictors were sampled for that model
<code>control</code>	a mirror of the arguments passed into <code>bagControl</code>
<code>call</code>	the call
<code>B</code>	the number of bagging iterations
<code>dims</code>	the dimensions of the training set

Author(s)

Max Kuhn

Examples

```
## A simple example of bagging conditional inference regression trees:
data(BloodBrain)

treebag <- bag(bbbDescr, logBBB, B = 10,
              bagControl = bagControl(fit = ctreeBag$fit,
```

```

                                predict = ctreeBag$pred,
                                aggregate = ctreeBag$aggregate))

## An example of pooling posterior probabilities to generate class predictions
data(mdr)

## remove some zero variance predictors and linear dependencies
mdrrDescr <- mdrDescr[, -nearZeroVar(mdrDescr)]
mdrrDescr <- mdrDescr[, -findCorrelation(cor(mdrDescr), .95)]

## Not run:
basicLDA <- train(mdrDescr, mdrClass, "lda")

bagLDA2 <- train(mdrDescr, mdrClass,
                "bag",
                B = 10,
                bagControl(fit = ldaBag$fit,
                           predict = ldaBag$pred,
                           aggregate = ldaBag$aggregate),
                tuneGrid = data.frame(.vars = c((1:10)*10 , ncol(mdrDescr))))

## End(Not run)

```

bagEarth

Bagged Earth

Description

A bagging wrapper for multivariate adaptive regression splines (MARS) via the earth function

Usage

```

## S3 method for class 'formula'
bagEarth(formula, data = NULL, B = 50,
          summary = mean, keepX = TRUE,
          ..., subset, weights, na.action = na.omit)
## Default S3 method:
bagEarth(x, y, weights = NULL, B = 50,
          summary = mean, keepX = TRUE, ...)

```

Arguments

formula	A formula of the form $y \sim x_1 + x_2 + \dots$
x	matrix or data frame of 'x' values for examples.
y	matrix or data frame of numeric values outcomes.

weights	(case) weights for each example - if missing defaults to 1.
data	Data frame from which variables specified in 'formula' are preferentially to be taken.
subset	An index vector specifying the cases to be used in the training sample. (NOTE: If given, this argument must be named.)
na.action	A function to specify the action to be taken if 'NA's are found. The default action is for the procedure to fail. An alternative is na.omit, which leads to rejection of cases with missing values on any required variable. (NOTE: If given, this argument must be named.)
B	the numebr of bootstrap samples
summary	a function with a single argument specifying how the bagged predictions should be summarized
keepX	a logical: should the original training data be kept?
...	arguments passed to the earth function

Details

The function computes a Earth model for each bootstap sample.

Value

A list with elements

fit	a list of B Earth fits
B	the number of bootstrap samples
call	the function call
x	either NULL or the value of x, depending on the value of keepX
oob	a matrix of performance estimates for each bootstrap sample

Author(s)

Max Kuhn (bagEarth.formula is based on Ripley's nnet.formula)

References

J. Friedman, "Multivariate Adaptive Regression Splines" (with discussion) (1991). *Annals of Statistics*, 19/1, 1-141.

See Also

[earth](#), [predict.bagEarth](#)

Examples

```
library(mda)
library(earth)
data(trees)
fit1 <- earth(trees[,-3], trees[,3])
fit2 <- bagEarth(trees[,-3], trees[,3], B = 10)
```

bagFDA

*Bagged FDA***Description**

A bagging wrapper for flexible discriminant analysis (FDA) using multivariate adaptive regression splines (MARS) basis functions

Usage

```
bagFDA(x, ...)
## S3 method for class 'formula'
bagFDA(formula, data = NULL, B = 50, keepX = TRUE,
        ..., subset, weights, na.action = na.omit)
## Default S3 method:
bagFDA(x, y, weights = NULL, B = 50, keepX = TRUE, ...)
```

Arguments

formula	A formula of the form $y \sim x_1 + x_2 + \dots$
x	matrix or data frame of 'x' values for examples.
y	matrix or data frame of numeric values outcomes.
weights	(case) weights for each example - if missing defaults to 1.
data	Data frame from which variables specified in 'formula' are preferentially to be taken.
subset	An index vector specifying the cases to be used in the training sample. (NOTE: If given, this argument must be named.)
na.action	A function to specify the action to be taken if 'NA's are found. The default action is for the procedure to fail. An alternative is na.omit, which leads to rejection of cases with missing values on any required variable. (NOTE: If given, this argument must be named.)
B	the numebr of bootstrap samples
keepX	a logical: should the original training data be kept?
...	arguments passed to the mars function

Details

The function computes a FDA model for each bootstap sample.

Value

A list with elements

fit	a list of B FDA fits
B	the number of bootstrap samples

call	the function call
x	either NULL or the value of x, depending on the value of keepX
oob	a matrix of performance estimates for each bootstrap sample

Author(s)

Max Kuhn (bagFDA . formula is based on Ripley's nnet . formula)

References

J. Friedman, "Multivariate Adaptive Regression Splines" (with discussion) (1991). Annals of Statistics, 19/1, 1-141.

See Also

[fda](#), [predict.bagFDA](#)

Examples

```
library(mlbench)
library(earth)
data(Glass)

set.seed(36)
inTrain <- sample(1:dim(Glass)[1], 150)

trainData <- Glass[ inTrain, ]
testData <- Glass[-inTrain, ]

baggedFit <- bagFDA(Type ~ ., trainData)
confusionMatrix(predict(baggedFit, testData[, -10]),
                 testData[, 10])
```

BloodBrain

Blood Brain Barrier Data

Description

Mente and Lombardo (2005) develop models to predict the log of the ratio of the concentration of a compound in the brain and the concentration in blood. For each compound, they computed three sets of molecular descriptors: MOE 2D, rule-of-five and Charge Polar Surface Area (CPSA). In all, 134 descriptors were calculated. Included in this package are 208 non-proprietary literature compounds. The vector logBBB contains the concentration ratio and the data frame bbbDescr contains the descriptor values.

Usage

```
data(BloodBrain)
```

Value

```
bbbDescr      data frame of chemical descriptors
logBBB        vector of assay results
```

Source

Mente, S.R. and Lombardo, F. (2005). A recursive-partitioning model for blood-brain barrier permeation, *Journal of Computer-Aided Molecular Design*, Vol. 19, pg. 465–481.

```
BoxCoxTrans.default  Box-Cox Transformations
```

Description

This class can be used to estimate a Box-Cox transformation and apply it to existing and future data

Usage

```
BoxCoxTrans(y, ...)

## Default S3 method:
BoxCoxTrans(y, x = rep(1, length(y)),
            fudge = 0.2, numUnique = 3, na.rm = FALSE, ...)

## S3 method for class 'BoxCoxTrans'
predict(object, newdata, ...)
```

Arguments

```
y          a strictly positive numeric vector of data to be transformed
x          an optional dependent variable to be used in a liner model
fudge      a tolerance value: lambda values within +/-fudge will be coerced to 0 and within
           1+/-fudge will be coerced to 1
numUnique  how many unique values should y have to estimate the transformation?
na.rm      a logical value indicating whether NA values should be stripped from y and x
           before the computation proceeds.
...        for BoxCoxTrans: options ot pass to boxcox. plotit should not be passed
           through. For predict.BoxCoxTrans, additional arguments are ignored.
object     an object of class BoxCoxTrans
newdata    a numeric vector of values to transform
```

Details

This function is basically a wrapper for the `boxcox` function in the MASS library. It can be used to estimate the transformation and apply it to new data.

If `any(y <= 0)` or if `length(unique(y)) < numUnique`, lambda is not estimated and no transformation is applied.

Value

`BoxCoxTrans` returns a list of class `BoxCoxTrans` with elements

<code>lambda</code>	estimated transformation value
<code>fudge</code>	value of fudge
<code>n</code>	number of data points used to estimate lambda
<code>summary</code>	the results of <code>summary(y)</code>
<code>ratio</code>	<code>max(y)/min(y)</code>
<code>skewness</code>	sample skewness statistic

`predict.BoxCoxTrans` returns a numeric vector of transformed values

Author(s)

Max Kuhn

References

Box, G. E. P. and Cox, D. R. (1964) An analysis of transformations (with discussion). *Journal of the Royal Statistical Society B*, 26, 211-252.

See Also

[boxcox](#), [preProcess](#)

Examples

```
data(BloodBrain)

ratio <- exp(logBBB)
bc <- BoxCoxTrans(ratio)

bc

predict(bc, ratio[1:5])

ratio[5] <- NA
bc2 <- BoxCoxTrans(ratio, bbbDescr$tpsa, na.rm = TRUE)
bc2
```

 calibration

Probability Calibration Plot

Description

For classification models, this function creates a 'calibration plot' that describes how consistent model probabilities are with observed event rates.

Usage

```
calibration(x, ...)

## S3 method for class 'formula'
calibration(x, data = NULL, class = NULL, cuts = 11, subset = TRUE, lattice.options = NULL, ...)

## S3 method for class 'calibration'
xyplot(x, data, ...)

panel.calibration(...)
```

Arguments

x	a lattice formula (see xyplot for syntax) where the left-hand side of the formula is a factor class variable of the observed outcome and the right-hand side specifies one or model columns corresponding to a numeric ranking variable for a model (e.g. class probabilities). The classification variable should have two levels.
data	For <code>calibration.formula</code> , a data frame (or more precisely, anything that is a valid <code>envir</code> argument in <code>eval</code> , e.g., a list or an environment) containing values for any variables in the formula, as well as groups and subset if applicable. If not found in data, or if data is unspecified, the variables are looked for in the environment of the formula. This argument is not used for <code>xyplot.calibration</code> .
class	a character string for the class of interest
cuts	the number of splits of the data are used to create the plot. By default, it uses as many cuts as there are rows in data
subset	An expression that evaluates to a logical or integer indexing vector. It is evaluated in data. Only the resulting rows of data are used for the plot.
lattice.options	A list that could be supplied to lattice.options
...	options to pass through to xyplot or the panel function (not used in <code>calibration.formula</code>).

Details

`calibration.formula` is used to process the data and `xyplot.calibration` is used to create the plot.

To construct the calibration plot, the following steps are used for each model:

1. The data are split into `cuts - 1` roughly equal groups by their class probabilities
2. the number of samples with true results equal to `class` are determined
3. the event rate is determined for each bin

`xyplot.calibration` produces a plot of the observed event rate by the mid-point of the bins.

This implementation uses the **lattice** function `xyplot`, so plot elements can be changed via panel functions, `trellis.par.set` or other means. `calibration` uses the panel function `panel.calibration` by default, but it can be changed by passing that argument into `xyplot.calibration`.

The following elements are set by default in the plot but can be changed by passing new values into `xyplot.calibration`: `xlab = "Bin Midpoint"`, `ylab = "Observed Event Percentage"`, `type = "o"`, `ylim = extendrange(c(0, 100))`, `xlim = extendrange(c(0, 100))` and `panel = panel.calibration`

Value

`calibration.formula` returns a list with elements:

<code>data</code>	the data used for plotting
<code>cuts</code>	the number of cuts
<code>class</code>	the event class
<code>probNames</code>	the names of the model probabilities

`xyplot.calibration` returns a **lattice** object

Author(s)

Max Kuhn, some **lattice** code and documentation by Deepayan Sarkar

See Also

[xyplot](#), [trellis.par.set](#)

Examples

```
data(mdr)
mdrrDescr <- mdrDescr[, -nearZeroVar(mdrDescr)]

inTrain <- createDataPartition(mdrClass)
trainX <- mdrDescr[inTrain[[1]], 1:10]
trainY <- mdrClass[inTrain[[1]]]
testX <- mdrDescr[-inTrain[[1]], 1:10]
testY <- mdrClass[-inTrain[[1]]]
```

```

library(MASS)

ldaFit <- lda(trainX, trainY)
qdaFit <- qda(trainX, trainY)

testProbs <- data.frame(obs = testY,
                        lda = predict(ldaFit, testX)$posterior[,1],
                        qda = predict(qdaFit, testX)$posterior[,1])

calibration(obs ~ lda + qda, data = testProbs)

calPlotData <- calibration(obs ~ lda + qda, data = testProbs)
calPlotData

xyplot(calPlotData, auto.key = list(columns = 2))

```

caretFuncs

Backwards Feature Selection Helper Functions

Description

Ancillary functions for backwards selection

Usage

```

pickSizeTolerance(x, metric, tol = 1.5, maximize)
pickSizeBest(x, metric, maximize)

```

```

pickVars(y, size)

```

```

caretFuncs
lmFuncs
rffFuncs
treebagFuncs
ldaFuncs
nbFuncs
gamFuncs

```

Arguments

x	a matrix or data frame with the performance metric of interest
metric	a character string with the name of the performance metric that should be used to choose the appropriate number of variables
maximize	a logical; should the metric be maximized?
tol	a scalar to denote the acceptable difference in optimal performance (see Details below)
y	a list of data frames with variables Overall and var
size	an integer for the number of variables to retain

Details

This page describes the functions that are used in backwards selection (aka recursive feature elimination). The functions described here are passed to the algorithm via the `functions` argument of [rfeControl](#).

See [rfeControl](#) for details on how these functions should be defined.

Author(s)

Max Kuhn

See Also

[rfeControl](#), [rfe](#)

caretSBF

Selection By Filtering (SBF) Helper Functions

Description

Ancillary functions for univariate feature selection

Usage

```
anovaScores(x, y)
gamScores(x, y)
```

```
caretSBF
lmSBF
rfSBF
treebagSBF
ldaSBF
nbSBF
```

Arguments

<code>x</code>	a matrix or data frame of numeric predictors
<code>y</code>	a numeric or factor vector of outcomes

Details

This page documents the functions that are used in selection by filtering (SBF). The functions described here are passed to the algorithm via the `functions` argument of [sbfControl](#).

See [sbfControl](#) for details on how these functions should be defined.

`anovaScores` and `gamScores` are two examples of univariate filtering functions. `anovaScores` fits a simple linear model between a single feature and the outcome, then the p-value for the whole model F-test is returned. `gamScores` fits a generalized additive model between a single predictor

and the outcome using a smoothing spline basis function. A p-value is generated using the whole model test from `summary.gam` and is returned.

If a particular model fails for `lm` or `gam`, a p-value of 1 is returned.

Author(s)

Max Kuhn

See Also

`sbfControl`, `sbf`, `summary.gam`

cars

Kelly Blue Book resale data for 2005 model year GM cars

Description

Kuiper (2008) collected data on Kelly Blue Book resale data for 804 GM cars (2005 model year).

Usage

```
data(cars)
```

Value

cars data frame of the suggested retail price (column Price) and various characteristics of each car (columns Mileage, Cylinder, Doors, Cruise, Sound, Leather, Buick, Cadillac, Chevy, Pontiac, Saab, Saturn, convertible, coupe, hatchback, sedan and wagon)

Source

Kuiper, S. (2008). Introduction to Multiple Regression: How Much Is Your Car Worth?, *Journal of Statistics Education*, Vol. 16, www.amstat.org/publications/jse/v16n3/datasets.kuiper.html

classDist	<i>Compute and predict the distances to class centroids</i>
-----------	---

Description

This function computes the class centroids and covariance matrix for a training set for determining Mahalanobis distances of samples to each class centroid.

Usage

```
classDist(x, ...)

## Default S3 method:
classDist(x, y, groups = 5, pca = FALSE, keep = NULL, ...)

## S3 method for class 'classDist'
predict(object, newdata, trans = log, ...)
```

Arguments

<code>x</code>	a matrix or data frame of predictor variables
<code>y</code>	a numeric or factor vector of class labels
<code>groups</code>	an integer for the number of bins for splitting a numeric outcome
<code>pca</code>	a logical: should principal components analysis be applied to the dataset prior to splitting the data by class?
<code>keep</code>	an integer for the number of PCA components that should be used to predict new samples (NULL uses all within a tolerance of $\sqrt{.Machine\$double.eps}$)
<code>object</code>	an object of class <code>classDist</code>
<code>newdata</code>	a matrix or data frame. If <code>vars</code> was previously specified, these columns should be in <code>newdata</code>
<code>trans</code>	an optional function that can be applied to each class distance. <code>trans = NULL</code> will not apply a function
<code>...</code>	optional arguments to pass (not currently used)

Details

For factor outcomes, the data are split into groups for each class and the mean and covariance matrix are calculated. These are then used to compute Mahalanobis distances to the class centers (using `predict.classDist`). The function will check for non-singular matrices.

For numeric outcomes, the data are split into roughly equal sized bins based on groups. Percentiles are used to split the data.

Value

for `classDist`, an object of class `classDist` with elements:

<code>values</code>	a list with elements for each class. Each element contains a mean vector for the class centroid and the inverse of the class covariance matrix
<code>classes</code>	a character vector of class labels
<code>pca</code>	the results of <code>prcomp</code> when <code>pca = TRUE</code>
<code>call</code>	the function call
<code>p</code>	the number of variables
<code>n</code>	a vector of samples sizes per class

For `predict.classDist`, a matrix with columns for each class. The columns names are the names of the class with the prefix `dist..` In the case of numeric `y`, the class labels are the percentiles. For example, of `groups = 9`, the variable names would be `dist.11.11`, `dist.22.22`, etc.

Author(s)

Max Kuhn

References

Forina et al. CAIMAN brothers: A family of powerful classification and class modeling techniques. *Chemometrics and Intelligent Laboratory Systems* (2009) vol. 96 (2) pp. 239-245

See Also

[mahalanobis](#)

Examples

```
trainSet <- sample(1:150, 100)

distData <- classDist(iris[trainSet, 1:4],
                     iris$Species[trainSet])

newDist <- predict(distData,
                  iris[-trainSet, 1:4])

splom(newDist, groups = iris$Species[-trainSet])
```

confusionMatrix *Create a confusion matrix*

Description

Calculates a cross-tabulation of observed and predicted classes with associated statistics.

Usage

```
confusionMatrix(data, ...)

## Default S3 method:
confusionMatrix(data, reference, positive = NULL,
                 dnn = c("Prediction", "Reference"),
                 prevalence = NULL, ...)

## S3 method for class 'table'
confusionMatrix(data, positive = NULL, prevalence = NULL, ...)
```

Arguments

data	a factor of predicted classes (for the default method) or an object of class <code>table</code> .
reference	a factor of classes to be used as the true results
positive	an optional character string for the factor level that corresponds to a "positive" result (if that makes sense for your data). If there are only two factor levels, the first level will be used as the "positive" result.
dnn	a character vector of dimnames for the table
prevalence	a numeric value or matrix for the rate of the "positive" class of the data. When data has two levels, prevalence should be a single numeric value. Otherwise, it should be a vector of numeric values with elements for each class. The vector should have names corresponding to the classes.
...	options to be passed to <code>table</code> . NOTE: do not include <code>dnn</code> here

Details

The functions requires that the factors have exactly the same levels.

For two class problems, the sensitivity, specificity, positive predictive value and negative predictive value is calculated using the `positive` argument. Also, the prevalence of the "event" is computed from the data (unless passed in as an argument), the detection rate (the rate of true events also predicted to be events) and the detection prevalence (the prevalence of predicted events).

Suppose a 2x2 table with notation

	Reference	
Predicted	Event	No Event

Event	A	B
No Event	C	D

The formulas used here are:

$$\text{Sensitivity} = A/(A + C)$$

$$\text{Specificity} = D/(B + D)$$

$$\text{Prevalence} = (A + C)/(A + B + C + D)$$

$$\text{PPV} = (\text{sensitivity} * \text{Prevalence}) / ((\text{sensitivity} * \text{Prevalence}) + ((1 - \text{specificity}) * (1 - \text{Prevalence})))$$

$$\text{NPV} = (\text{specificity} * (1 - \text{Prevalence})) / (((1 - \text{sensitivity}) * \text{Prevalence}) + (\text{specificity} * (1 - \text{Prevalence})))$$

$$\text{DetectionRate} = A/(A + B + C + D)$$

$$\text{DetectionPrevalence} = (A + B)/(A + B + C + D)$$

See the references for discussions of the first five formulas.

For more than two classes, these results are calculated comparing each factor level to the remaining levels (i.e. a "one versus all" approach).

The overall accuracy and unweighted Kappa statistic are calculated. A p-value from McNemar's test is also computed using `mcnemar.test` (which can produce NA values with sparse tables).

The overall accuracy rate is computed along with a 95 percent confidence interval for this rate (using `binom.test`) and a one-sided test to see if the accuracy is better than the "no information rate," which is taken to be the largest class percentage in the data.

Value

a list with elements

table	the results of table on data and reference
positive	the positive result level
overall	a numeric vector with overall accuracy and Kappa statistic values
byClass	the sensitivity, specificity, positive predictive value, negative predictive value, prevalence, dection rate and detection prevalence for each class. For two class systems, this is calculated once using the positive argument

Author(s)

Max Kuhn

References

- Kuhn, M. (2008), "Building predictive models in R using the caret package," *Journal of Statistical Software*, (<http://www.jstatsoft.org/v28/i05/>).
- Altman, D.G., Bland, J.M. (1994) "Diagnostic tests 1: sensitivity and specificity," *British Medical Journal*, vol 308, 1552.
- Altman, D.G., Bland, J.M. (1994) "Diagnostic tests 2: predictive values," *British Medical Journal*, vol 309, 102.

See Also

[as.table.confusionMatrix](#), [as.matrix.confusionMatrix](#), [sensitivity](#), [specificity](#), [posPredValue](#), [negPredValue](#), [print.confusionMatrix](#), [binom.test](#)

Examples

```
#####  
## 2 class example  
  
lvs <- c("normal", "abnormal")  
truth <- factor(rep(lvs, times = c(86, 258)),  
               levels = rev(lvs))  
pred <- factor(  
  c(  
    rep(lvs, times = c(54, 32)),  
    rep(lvs, times = c(27, 231))),  
  levels = rev(lvs))  
  
xtab <- table(pred, truth)  
  
confusionMatrix(xtab)  
confusionMatrix(pred, truth)  
confusionMatrix(xtab, prevalence = 0.25)  
  
#####  
## 3 class example  
  
library(MASS)  
  
fit <- lda(Species ~ ., data = iris)  
model <- predict(fit)$class  
  
irisTabs <- table(model, iris$Species)  
  
confusionMatrix(irisTabs)  
confusionMatrix(model, iris$Species)  
  
newPrior <- c(.05, .8, .15)  
names(newPrior) <- levels(iris$Species)  
  
confusionMatrix(irisTabs, prevalence = newPrior)  
## Need names for prevalence  
## Not run: confusionMatrix(irisTabs, prevalence = c(.05, .8, .15))
```

Description

Using a `train`, `rfe`, `sbf` object, determine a confusion matrix based on the resampling procedure

Usage

```
## S3 method for class 'train'
confusionMatrix(data, norm = "overall",
                dnn = c("Prediction", "Reference"), ...)

## S3 method for class 'rfe'
confusionMatrix(data, norm = "overall",
                dnn = c("Prediction", "Reference"), ...)

## S3 method for class 'sbf'
confusionMatrix(data, norm = "overall",
                dnn = c("Prediction", "Reference"), ...)
```

Arguments

<code>data</code>	an object of class <code>train</code> , <code>rfe</code> , <code>sbf</code> that did not use out-of-bag resampling or leave-one-out cross-validation.
<code>norm</code>	a character string indicating how the table entries should be normalized. Valid values are "none", "overall" or "average".
<code>dnn</code>	a character vector of dimnames for the table
<code>...</code>	not used here

Details

When `train` is used for tuning a model, it tracks the confusion matrix cell entries for the hold-out samples. These can be aggregated and used for diagnostic purposes. For `train`, the matrix is estimated for the final model tuning parameters determined by `train`. For `rfe`, the matrix is associated with the optimal number of variables.

There are several ways to show the table entries. Using `norm = "none"` will show the frequencies of samples on each of the cells (across all resamples). `norm = "overall"` first divides the cell entries by the total number of data points in the table, then averages these percentages. `norm = "average"` takes the raw, aggregate cell counts across resamples and divides by the number of resamples (i.e. to yield an average count for each cell).

Value

a list of class `confusionMatrix.train`, `confusionMatrix.rfe` or `confusionMatrix.sbf` with elements

<code>table</code>	the normalized matrix
<code>nrom</code>	an echo fo the call
<code>text</code>	a character string with details about the resampling procedure (e.g. "Bootstrapped (25 reps) Confusion Matrix")

Author(s)

Max Kuhn

See Also[confusionMatrix](#), [train](#), [rfe](#), [sbf](#), [trainControl](#)**Examples**

```
data(iris)
TrainData <- iris[,1:4]
TrainClasses <- iris[,5]

knnFit <- train(TrainData, TrainClasses,
               method = "knn",
               preProcess = c("center", "scale"),
               tuneLength = 10,
               trControl = trainControl(method = "cv"))
confusionMatrix(knnFit)
confusionMatrix(knnFit, "average")
confusionMatrix(knnFit, "none")
```

`cox2`*COX-2 Activity Data*

Description

From Sutherland, O'Brien, and Weaver (2003): "A set of 467 cyclooxygenase-2 (COX-2) inhibitors has been assembled from the published work of a single research group, with in vitro activities against human recombinant enzyme expressed as IC50 values ranging from 1 nM to >100 uM (53 compounds have indeterminate IC50 values)."

The data are in the Supplemental Data file for the article.

A set of 255 descriptors (MOE2D and QikProp) were generated. To classify the data, we used a cutoff of $2^{2.5}$ to determine activity

Usage`data(cox2)`**Value**

<code>cox2Descr</code>	the descriptors
<code>cox2IC50</code>	the IC50 data used to determine activity
<code>cox2Class</code>	the categorical outcome ("Active" or "Inactive") based on the $2^{2.5}$ cutoff

Source

Sutherland, J. J., O'Brien, L. A. and Weaver, D. F. (2003). Spline-Fitting with a Genetic Algorithm: A Method for Developing Classification Structure-Activity Relationships, *Journal of Chemical Information and Computer Sciences*, Vol. 43, pg. 1906–1915.

createDataPartition *Data Splitting functions*

Description

A series of test/training partitions are created using createDataPartition while createResample creates one or more bootstrap samples. createFolds splits the data into k groups.

Usage

```
createDataPartition(y,
                    times = 1,
                    p = 0.5,
                    list = TRUE,
                    groups = min(5, length(y)))
createResample(y, times = 10, list = TRUE)
createFolds(y, k = 10, list = TRUE, returnTrain = FALSE)
createMultiFolds(y, k = 10, times = 5)
```

Arguments

y	a vector of outcomes
times	the number of partitions to create
p	the percentage of data that goes to training
list	logical - should the results be in a list (TRUE) or a matrix with the number of rows equal to floor(p * length(y)) and times columns.
groups	for numeric y, the number of breaks in the quantiles (see below)
k	an integer for the number of folds.
returnTrain	a logical. When true, the values returned are the sample positions corresponding to the data used during training. This argument only works in conjunction with list = TRUE

Details

For bootstrap samples, simple random sampling is used.

For other data splitting, the random sampling is done within the levels of y when y is a factor in an attempt to balance the class distributions within the splits.

For numeric y, the sample is split into groups sections based on percentiles and sampling is done within these subgroups. For createDataPartition, the number of percentiles is set via the groups

argument. For createFolds and createMultiFolds, the number of groups is set dynamically based on the sample size and k. For smaller samples sizes, these two functions may not do stratified splitting and, at most, will split the data into quartiles.

Also, for createDataPartition, very small class sizes (≤ 3) the classes may not show up in both the training and test data

For multiple k-fold cross-validation, completely independent folds are created. The names of the list objects will denote the fold membership using the pattern "Foldi.Repj" meaning the ith section (of k) of the jth cross-validation set (of times). Note that this function calls createFolds with list = TRUE and returnTrain = TRUE.

Value

A list or matrix of row position integers corresponding to the training data

Author(s)

Max Kuhn

Examples

```
data(oil)
createDataPartition(oilType, 2)

x <- rgamma(50, 3, .5)
inA <- createDataPartition(x, list = FALSE)

plot(density(x[inA]))
rug(x[inA])

points(density(x[-inA]), type = "l", col = 4)
rug(x[-inA], col = 4)

createResample(oilType, 2)

createFolds(oilType, 10)
createFolds(oilType, 5, FALSE)

createFolds(rnorm(21))
```

createGrid

Tuning Parameter Grid

Description

This function creates a data frame that contains a grid of complexity parameters specific methods.

Usage

```
createGrid(method, len = 3, data = NULL)
```

Arguments

method	a string specifying which classification model to use. See train for a full list.
len	an integer specifying the number of points on the grid for each tuning parameter.
data	the training data (only needed in the case where the method is cforest, earth, bagEarth, fda, bagFDA, rpart, svmRadial, pam, lars2, rf or pls). The outcome should be in a column called .outcome.

Details

A grid is created with rows corresponding to complexity parameter combinations. If the model does not use tuning parameters (like a linear model), values of NA are returned. Columns are named the same as the parameter name, but preceded by a period.

For some models (see list above), the data should be passed to the function via the data argument. In these cases, the outcome should be included in a column named .outcome.

Value

A data frame where the rows are combinations of tuning parameters and columns correspond to the parameters. The column names should be the parameter names preceded by a dot (e.g. .mtry)

Author(s)

Max Kuhn

See Also

[train](#)

Examples

```
createGrid("rda", 4)
createGrid("lm")
createGrid("nnet")

## data needed for SVM with RBF:
tmp <- iris
names(tmp)[5] <- ".outcome"
head(tmp)
createGrid("svmRadial", data = tmp, len = 4)
```

 dhfr

Dihydrofolate Reductase Inhibitors Data

Description

Sutherland and Weaver (2004) discuss QSAR models for dihydrofolate reductase (DHFR) inhibition. This data set contains values for 325 compounds. For each compound, 228 molecular descriptors have been calculated. Additionally, each sample is designated as "active" or "inactive".

The data frame `dhfr` contains a column called `Y` with the outcome classification. The remainder of the columns are molecular descriptor values.

Usage

```
data(dhfr)
```

Value

`dhfr` data frame of chemical descriptors and the activity values

Source

Sutherland, J.J. and Weaver, D.F. (2004). Three-dimensional quantitative structure-activity and structure-selectivity relationships of dihydrofolate reductase inhibitors, *Journal of Computer-Aided Molecular Design*, Vol. 18, pg. 309–331.

 diff.resamples

Inferential Assessments About Model Performance

Description

Methods for making inferences about differences between models

Usage

```
## S3 method for class 'resamples'
diff(x, models = x$models, metric = x$metrics,
     test = t.test,
     confLevel = 0.95, adjustment = "bonferroni",
     ...)

## S3 method for class 'diff.resamples'
summary(object, digits = max(3, getOption("digits") - 3), ...)
```

Arguments

x	an object generated by resamples
models	a character string for which models to compare
metric	a character string for which metrics to compare
test	a function to compute differences. The output of this function should have scalar outputs called estimate and p.value
object	a object generated by diff.resamples
adjustment	any p-value adjustment method to pass to p.adjust .
confLevel	confidence level to use for dotplot.diff.resamples . See Details below.
digits	the number of significant differences to display when printing
...	further arguments to pass to test

Details

The ideas and methods here are based on Hothorn et al (2005) and Eugster et al (2008).

For each metric, all pair-wise differences are computed and tested to assess if the difference is equal to zero.

When a Bonferroni correction is used, the confidence level is changed from confLevel to $1 - ((1 - \text{confLevel})/p)$ here p is the number of pair-wise comparisons are being made. For other correction methods, no such change is used.

Value

An object of class "diff.resamples" with elements:

call	the call
difs	a list for each metric being compared. Each list contains a matrix with differences in columns and resamples in rows
statistics	a list of results generated by test
adjustment	the p-value adjustment used
models	a character string for which models were compared.
metrics	a character string of performance metrics that were used

or...

An object of class "summary.diff.resamples" with elements:

call	the call
table	a list of tables that show the differences and p-values

Author(s)

Max Kuhn

References

Hothorn et al. The design and analysis of benchmark experiments. *Journal of Computational and Graphical Statistics* (2005) vol. 14 (3) pp. 675-699

Eugster et al. Exploratory and inferential analysis of benchmark experiments. Ludwigs-Maximilians-Universitat Munchen, Department of Statistics, Tech. Rep (2008) vol. 30

See Also

[resamples](#), [dotplot.diff.resamples](#), [densityplot.diff.resamples](#), [bwplot.diff.resamples](#), [levelplot.diff.resamples](#)

Examples

```
## Not run:
#load(url("http://caret.r-forge.r-project.org/Classification_and_Regression_Training_files/exampleModels.RData

resamps <- resamples(list(CART = rpartFit,
                          CondInfTree = ctreeFit,
                          MARS = earthFit))

difs <- diff(resamps)

difs

summary(difs)

## End(Not run)
```

dotPlot

Create a dotplot of variable importance values

Description

A lattice [dotplot](#) is created from an object of class `varImp.train`.

Usage

```
dotPlot(x, top = min(20, dim(x$importance)[1]), ...)
```

Arguments

<code>x</code>	an object of class <code>varImp.train</code>
<code>top</code>	the number of predictors to plot
<code>...</code>	options passed to dotplot

Value

an object of class `trellis`.

Author(s)

Max Kuhn

See Also[varImp](#), [dotplot](#)**Examples**

```
data(iris)
TrainData <- iris[,1:4]
TrainClasses <- iris[,5]

knnFit <- train(TrainData, TrainClasses, "knn")

knnImp <- varImp(knnFit)

dotPlot(knnImp)
```

`dotplot.diff.resamples`*Lattice Functions for Visualizing Resampling Differences*

Description

Lattice functions for visualizing resampling result differences between models

Usage

```
## S3 method for class 'diff.resamples'
densityplot(x, data, metric = x$metric, ...)

## S3 method for class 'diff.resamples'
bwplot(x, data, metric = x$metric, ...)

## S3 method for class 'diff.resamples'
levelplot(x, data = NULL, metric = x$metric[1], what = "pvalues", ...)

## S3 method for class 'diff.resamples'
dotplot(x, data = NULL, metric = x$metric[1], ...)
```

Arguments

x	an object generated by diff.resamples
data	Not used
what	levelplot only: display either the "pvalues" or "differences"
metric	a character string for which metrics to plot. Note: dotplot and levelplot require exactly two models whereas the other methods can plot more than two.
...	further arguments to pass to either histogram , densityplot , dotplot or levelplot

Details

[densityplot](#) and [bwplot](#) display univariate visualizations of the resampling distributions. [levelplot](#) displays the matrix of pair-wise comparisons. [dotplot](#) shows the differences along with their associated confidence intervals.

Value

a lattice object

Author(s)

Max Kuhn

See Also

[resamples](#), [diff.resamples](#), [bwplot](#), [densityplot](#), [xyplot](#), [splom](#)

Examples

```
## Not run:
#load(url("http://caret.r-forge.r-project.org/Classification_and_Regression_Training_files/exampleModels.RData"))

resamps <- resamples(list(CART = rpartFit,
                          CondInfTree = ctreeFit,
                          MARS = earthFit))

difs <- diff(resamps)

dotplot(difValues)

densityplot(difValues,
            metric = "RMSE",
            auto.key = TRUE,
            pch = "|")

bwplot(difValues,
       metric = "RMSE")

levelplot(difValues, what = "differences")

## End(Not run)
```

dummyVars

*Create A Full Set of Dummy Variables***Description**

dummyVars creates a full set of dummy variables (i.e. less than full rank parameterization)

Usage

```
dummyVars(formula, ...)

## Default S3 method:
dummyVars(formula, data, sep = ".", levelsOnly = FALSE, ...)

## S3 method for class 'dummyVars'
predict(object, newdata, na.action = na.pass, ...)

contr.dummy(n, ...)
```

Arguments

formula	An appropriate R model formula, see References
data	A data frame with the predictors of interest
sep	An optional separator between factor variable names and their levels. Use sep = NULL for no separator (i.e. normal behavior of <code>model.matrix</code> as shown in the Details section)
levelsOnly	A logical; TRUE means to completely remove the variable names from the column names
object	An object of class dummyVars
newdata	A data frame with the required columns
na.action	A function determining what should be done with missing values in newdata. The default is to predict NA.
n	A vector of levels for a factor, or the number of levels.
...	additional arguments to be passed to other methods

Details

Most of the `contrasts` functions in R produce full rank parameterizations of the predictor data. For example, `contr.treatment` creates a reference cell in the data and defines dummy variables for all factor levels except those in the reference cell. For example, if a factor with 5 levels is used in a model formula alone, `contr.treatment` creates columns for the intercept and all the factor levels except the first level of the factor. For the data in the Example section below, this would produce:

	(Intercept)	dayTue	dayWed	dayThu	dayFri	daySat	daySun
1	1	1	0	0	0	0	0
2	1	1	0	0	0	0	0
3	1	1	0	0	0	0	0
4	1	0	0	1	0	0	0
5	1	0	0	1	0	0	0
6	1	0	0	0	0	0	0
7	1	0	1	0	0	0	0
8	1	0	1	0	0	0	0
9	1	0	0	0	0	0	0

In some situations, there may be a need for dummy variables for all of the levels of the factor. For the same example:

	dayMon	dayTue	dayWed	dayThu	dayFri	daySat	daySun
1	0	1	0	0	0	0	0
2	0	1	0	0	0	0	0
3	0	1	0	0	0	0	0
4	0	0	0	1	0	0	0
5	0	0	0	1	0	0	0
6	1	0	0	0	0	0	0
7	0	0	1	0	0	0	0
8	0	0	1	0	0	0	0
9	1	0	0	0	0	0	0

Given a formula and initial data set, the class `dummyVars` gathers all the information needed to produce a full set of dummy variables for any data set. It uses `contr.dummy` as the base function to do this.

Value

The output of `dummyVars` is a list of class `'dummyVars'` with elements

<code>call</code>	the function call
<code>form</code>	the model formula
<code>vars</code>	names of all the variables in the model
<code>facVars</code>	names of all the factor variables in the model
<code>lvls</code>	levels of any factor variables
<code>sep</code>	NULL or a character separator
<code>terms</code>	the <code>terms.formula</code> object
<code>levelsOnly</code>	a logical

The `predict` function produces a data frame.

`contr.dummy` generates a matrix with `n` rows and `n` columns.

Author(s)

The initial code was graciously provided by Gabor Grothendieck on R-Help; refined by Max Kuhn

References

<http://cran.r-project.org/doc/manuals/R-intro.html#Formulae-for-statistical-models>

See Also

[model.matrix](#), [contrasts](#), [formula](#)

Examples

```
when <- data.frame(time = c("afternoon", "night", "afternoon",
                           "morning", "morning", "morning",
                           "morning", "afternoon", "afternoon"),
                  day = c("Mon", "Mon", "Mon",
                          "Wed", "Wed", "Fri",
                          "Sat", "Sat", "Fri"))

levels(when$time) <- c("morning", "afternoon", "night")
levels(when$day) <- c("Mon", "Tue", "Wed", "Thu", "Fri", "Sat", "Sun")

## Default behavior:
model.matrix(~day, when)

mainEffects <- dummyVars(~ day + time, data = when)
mainEffects
predict(mainEffects, when[1:3,])

interactionModel <- dummyVars(~ day + time + day:time,
                              data = when,
                              sep = ".")
predict(interactionModel, when[1:3,])

noNames <- dummyVars(~ day + time + day:time,
                    data = when,
                    levelsOnly = TRUE)
predict(noNames, when)
```

featurePlot

Wrapper for Lattice Plotting of Predictor Variables

Description

A shortcut to produce lattice graphs

Usage

```
featurePlot(x, y,
            plot = if(is.factor(y)) "strip" else "scatter",
            labels = c("Feature", ""),
            ...)
```

Arguments

x	a matrix or data frame of continuous feature/probe/spectra data.
y	a factor indicating class membership.
plot	the type of plot. For classification: box, strip, density, pairs or ellipse. For regression, pairs or scatter
labels	a bad attempt at pre-defined axis labels
...	options passed to lattice calls.

Details

This function “stacks” data to get it into a form compatible with lattice and creates the plots

Value

An object of class “trellis”. The ‘update’ method can be used to update components of the object and the ‘print’ method (usually called by default) will plot it on an appropriate plotting device.

Author(s)

Max Kuhn

Examples

```
x <- matrix(rnorm(50*5), ncol=5)
y <- factor(rep(c("A", "B"), 25))

trellis.par.set(theme = col.whitebg(), warn = FALSE)
featurePlot(x, y, "ellipse")
featurePlot(x, y, "strip", jitter = TRUE)
featurePlot(x, y, "box")
featurePlot(x, y, "pairs")
```

`filterVarImp`*Calculation of filter-based variable importance*

Description

Specific engines for variable importance on a model by model basis.

Usage

```
filterVarImp(x, y, nonpara = FALSE, ...)
```

Arguments

<code>x</code>	A matrix or data frame of predictor data
<code>y</code>	A vector (numeric or factor) of outcomes)
<code>nonpara</code>	should nonparametric methods be used to assess the relationship between the features and response
<code>...</code>	options to pass to either <code>lm</code> or <code>loess</code>

Details

The importance of each predictor is evaluated individually using a “filter” approach.

For classification, ROC curve analysis is conducted on each predictor. For two class problems, a series of cutoffs is applied to the predictor data to predict the class. The sensitivity and specificity are computed for each cutoff and the ROC curve is computed. The trapezoidal rule is used to compute the area under the ROC curve. This area is used as the measure of variable importance. For multi-class outcomes, the problem is decomposed into all pair-wise problems and the area under the curve is calculated for each class pair (i.e class 1 vs. class 2, class 2 vs. class 3 etc.). For a specific class, the maximum area under the curve across the relevant pair-wise AUC’s is used as the variable importance measure.

For regression, the relationship between each predictor and the outcome is evaluated. An argument, `nonpara`, is used to pick the model fitting technique. When `nonpara = FALSE`, a linear model is fit and the absolute value of the t -value for the slope of the predictor is used. Otherwise, a loess smoother is fit between the outcome and the predictor. The R^2 statistic is calculated for this model against the intercept only null model.

Value

A data frame with variable importances. Column names depend on the problem type. For regression, the data frame contains one column: "Overall" for the importance values.

Author(s)

Max Kuhn

Examples

```
data(mdr)
filterVarImp(mdrDescr[, 1:5], mdrClass)

data(BloodBrain)

filterVarImp(bbbDescr[, 1:5], logBBB, nonpara = FALSE)
apply(
  bbbDescr[, 1:5],
  2,
  function(x, y) summary(lm(y~x))$coefficients[2,3],
  y = logBBB)

filterVarImp(bbbDescr[, 1:5], logBBB, nonpara = TRUE)
```

findCorrelation	<i>Determine highly correlated variables</i>
-----------------	--

Description

This function searches through a correlation matrix and returns a vector of integers corresponding to columns to remove to reduce pair-wise correlations.

Usage

```
findCorrelation(x, cutoff = .90, verbose = FALSE)
```

Arguments

x	A correlation matrix
cutoff	A numeric value for the pairwise absolute correlation cutoff
verbose	A boolean for printing the details

Details

The absolute values of pair-wise correlations are considered. If two variables have a high correlation, the function looks at the mean absolute correlation of each variable and removes the variable with the largest mean absolute correlation.

There are several functions in the **subselect** package ([leaps](#), [genetic](#), [anneal](#)) that can also be used to accomplish the same goal.

Value

A vector of indices denoting the columns to remove. If no correlations meet the criteria, `numeric(0)` is returned.

Author(s)

Original R code by Dong Li, modified by Max Kuhn

See Also

[leaps](#), [genetic](#), [anneal](#), [findLinearCombos](#)

Examples

```
corrMatrix <- diag(rep(1, 5))
corrMatrix[2, 3] <- corrMatrix[3, 2] <- .7
corrMatrix[5, 3] <- corrMatrix[3, 5] <- -.7
corrMatrix[4, 1] <- corrMatrix[1, 4] <- -.67

corrDF <- expand.grid(row = 1:5, col = 1:5)
corrDF$correlation <- as.vector(corrMatrix)
levelplot(correlation ~ row+ col, corrDF)

findCorrelation(corrMatrix, cutoff = .65, verbose = TRUE)

findCorrelation(corrMatrix, cutoff = .99, verbose = TRUE)
```

findLinearCombos

Determine linear combinations in a matrix

Description

Enumerate and resolve the linear combinations in a numeric matrix

Usage

```
findLinearCombos(x)
```

Arguments

x a numeric matrix

Details

The QR decomposition is used to determine if the matrix is full rank and then identify the sets of columns that are involved in the dependencies.

To "resolve" them, columns are iteratively removed and the matrix rank is rechecked.

The `trim.matrix` function in the **subselect** package can also be used to accomplish the same goal.

Value

a list with elements:

`linearCombos` If there are linear combinations, this will be a list with elements for each dependency that contains vectors of column numbers.

`remove` a list of column numbers that can be removed to counter the linear combinations

Author(s)

Kirk Mettler and Jed Wing (enumLC) and Max Kuhn (findLinearCombos)

See Also

[trim.matrix](#)

Examples

```
testData1 <- matrix(0, nrow=20, ncol=8)
testData1[,1] <- 1
testData1[,2] <- round(rnorm(20), 1)
testData1[,3] <- round(rnorm(20), 1)
testData1[,4] <- round(rnorm(20), 1)
testData1[,5] <- 0.5 * testData1[,2] - 0.25 * testData1[,3] - 0.25 * testData1[,4]
testData1[1:4,6] <- 1
testData1[5:10,7] <- 1
testData1[11:20,8] <- 1

findLinearCombos(testData1)

testData2 <- matrix(0, nrow=6, ncol=6)
testData2[,1] <- c(1, 1, 1, 1, 1, 1)
testData2[,2] <- c(1, 1, 1, 0, 0, 0)
testData2[,3] <- c(0, 0, 0, 1, 1, 1)
testData2[,4] <- c(1, 0, 0, 1, 0, 0)
testData2[,5] <- c(0, 1, 0, 0, 1, 0)
testData2[,6] <- c(0, 0, 1, 0, 0, 1)

findLinearCombos(testData2)
```

format.bagEarth

Format 'bagEarth' objects

Description

Return a string representing the 'bagEarth' expression.

Usage

```
## S3 method for class 'bagEarth'
format(x, file = "", cat = TRUE, ...)
```

Arguments

<code>x</code>	An <code>bagEarth</code> object. This is the only required argument.
<code>file</code>	A connection, or a character string naming the file to print to. If "" (the default), the output prints to the standard output connection. See <code>link[base]{cat}</code> .
<code>cat</code>	a logical; should the equation be printed?
<code>...</code>	Arguments to <code>format.earth</code> .

Value

A character representation of the bagged earth object.

See Also

[earth](#)

Examples

```
a <- bagEarth(Volume ~ ., data = trees, B= 3)
cat(format(a))

# yields:
# (
# 31.61075
# + 6.587273 * pmax(0, Girth - 14.2)
# - 3.229363 * pmax(0, 14.2 - Girth)
# - 0.3167140 * pmax(0, 79 - Height)
# +
# 22.80225
# + 5.309866 * pmax(0, Girth - 12)
# - 2.378658 * pmax(0, 12 - Girth)
# + 0.793045 * pmax(0, Height - 80)
# - 0.3411915 * pmax(0, 80 - Height)
# +
# 31.39772
# + 6.18193 * pmax(0, Girth - 14.2)
# - 3.660456 * pmax(0, 14.2 - Girth)
# + 0.6489774 * pmax(0, Height - 80)
# )/3
```

Description

Data from Dr. Hans Hofmann of the University of Hamburg.

These data have two classes for the credit worthiness: good or bad. There are predictors related to attributes, such as: checking account status, duration, credit history, purpose of the loan, amount of the loan, savings accounts or bonds, employment duration, Installment rate in percentage of disposable income, personal information, other debtors/guarantors, residence duration, property, age, other installment plans, housing, number of existing credits, job information, Number of people being liable to provide maintenance for, telephone, and foreign worker status.

Many of these predictors are discrete and have been expanded into several 0/1 indicator variables

Usage

```
data(GermanCredit)
```

Source

UCI Machine Learning Repository

histogram.train	<i>Lattice functions for plotting resampling results</i>
-----------------	--

Description

A set of lattice functions are provided to plot the resampled performance estimates (e.g. classification accuracy, RMSE) over tuning parameters (if any).

Usage

```
## S3 method for class 'train'
histogram(x, data = NULL, metric = x$metric, ...)

## S3 method for class 'train'
densityplot(x, data = NULL, metric = x$metric, ...)

## S3 method for class 'train'
xyplot(x, data = NULL, metric = x$metric, ...)

## S3 method for class 'train'
stripplot(x, data = NULL, metric = x$metric, ...)
```

Arguments

x	An object produced by train
data	This argument is not used
metric	A character string specifying the single performance metric that will be plotted
...	arguments to pass to either histogram , densityplot , xyplot or stripplot

Details

By default, only the resampling results for the optimal model are saved in the train object. The function `trainControl` can be used to save all the results (see the example below).

If leave-one-out or out-of-bag resampling was specified, plots cannot be produced (see the method argument of `trainControl`)

For `xyplot` and `stripplot`, the tuning parameter with the most unique values will be plotted on the x-axis. The remaining parameters (if any) will be used as conditioning variables. For `densityplot` and `histogram`, all tuning parameters are used for conditioning.

Using `horizontal = FALSE` in `stripplot` works.

Value

A lattice plot object

Author(s)

Max Kuhn

See Also

`train`, `trainControl`, `histogram`, `densityplot`, `xyplot`, `stripplot`

Examples

icr.formula *Independent Component Regression*

Description

Fit a linear regression model using independent components

Usage

```
## S3 method for class 'formula'  
icr(formula, data, weights, ..., subset, na.action, contrasts = NULL)  
## Default S3 method:  
icr(x, y, ...)  
  
## S3 method for class 'icr'  
predict(object, newdata, ...)
```

Arguments

formula	A formula of the form <code>class ~ x1 + x2 + ...</code>
data	Data frame from which variables specified in <code>formula</code> are preferentially to be taken.
weights	(case) weights for each example – if missing defaults to 1.
subset	An index vector specifying the cases to be used in the training sample. (NOTE: If given, this argument must be named.)
na.action	A function to specify the action to be taken if NAs are found. The default action is for the procedure to fail. An alternative is <code>na.omit</code> , which leads to rejection of cases with missing values on any required variable. (NOTE: If given, this argument must be named.)
contrasts	a list of contrasts to be used for some or all of the factors appearing as variables in the model formula.
...	arguments passed to fastICA
x	matrix or data frame of x values for examples.
y	matrix or data frame of target values for examples.
object	an object of class <code>icr</code> as returned by <code>icr</code> .
newdata	matrix or data frame of test examples.

Details

This produces a model analogous to Principal Components Regression (PCR) but uses Independent Component Analysis (ICA) to produce the scores. The user must specify a value of `n.comp` to pass to [fastICA](#).

The function [preProcess](#) to produce the ICA scores for the original data and for `newdata`.

Value

For `icr`, a list with elements

model	the results of <code>lm</code> after the ICA transformation
ica	pre-processing information
n.comp	number of ICA components
names	column names of the original data

Author(s)

Max Kuhn

See Also

[fastICA](#), [preProcess](#), [lm](#)

Examples

```
data(BloodBrain)

icrFit <- icr(bbbDescr, logBBB, n.comp = 5)

icrFit

predict(icrFit, bbbDescr[1:5,])
```

knn3

k-Nearest Neighbour Classification

Description

`k`-nearest neighbour classification that can return class votes for all classes.

Usage

```
## S3 method for class 'formula'
knn3(formula, data, subset, na.action, k = 5, ...)

## S3 method for class 'matrix'
knn3(x, y, k = 5, ...)

knn3Train(train, test, cl, k=1, l=0, prob = TRUE, use.all=TRUE)
```

Arguments

formula	a formula of the form $lhs \sim rhs$ where lhs is the response variable and rhs a set of predictors.
data	optional data frame containing the variables in the model formula.
subset	optional vector specifying a subset of observations to be used.
na.action	function which indicates what should happen when the data contain NAs.
k	number of neighbours considered.
x	a matrix of training set predictors
y	a factor vector of training set classes
...	additional parameters to pass to <code>knn3Train</code> . However, passing <code>prob = FALSE</code> will be over-ridden.
train	matrix or data frame of training set cases.
test	matrix or data frame of test set cases. A vector will be interpreted as a row vector for a single case.

c1	factor of true classifications of training set
l	minimum vote for definite decision, otherwise doubt. (More precisely, less than k-1 dissenting votes are allowed, even if k is increased by ties.)
prob	If this is true, the proportion of the votes for each class are returned as attribute prob.
use.all	controls handling of ties. If true, all distances equal to the kth largest are included. If false, a random selection of distances equal to the kth is chosen to use exactly k neighbours.

Details

knn3 is essentially the same code as [ipredknn](#) and knn3Train is a copy of [knn](#). The underlying C code from the `class` package has been modified to return the vote percentages for each class (previously the percentage for the winning class was returned).

Value

An object of class `knn3`. See [predict.knn3](#).

Author(s)

[knn](#) by W. N. Venables and B. D. Ripley and [ipredknn](#) by Torsten.Hothorn <Torsten.Hothorn@rzmail.uni-erlangen.de>, modifications by Max Kuhn and Andre Williams

Examples

```
irisFit1 <- knn3(Species ~ ., iris)

irisFit2 <- knn3(as.matrix(iris[, -5]), iris[,5])

data(iris3)
train <- rbind(iris3[1:25,,1], iris3[1:25,,2], iris3[1:25,,3])
test <- rbind(iris3[26:50,,1], iris3[26:50,,2], iris3[26:50,,3])
c1 <- factor(c(rep("s",25), rep("c",25), rep("v",25)))
knn3Train(train, test, c1, k = 5, prob = TRUE)
```

knnreg

k-Nearest Neighbour Regression

Description

k -nearest neighbour classification that can return the average value for the neighbours.

Usage

```
## Default S3 method:
knnreg(x, ...)

## S3 method for class 'formula'
knnreg(formula, data, subset, na.action, k = 5, ...)

## S3 method for class 'matrix'
knnreg(x, y, k = 5, ...)

## S3 method for class 'data.frame'
knnreg(x, y, k = 5, ...)

knnregTrain(train, test, y, k = 5, use.all=TRUE)
```

Arguments

formula	a formula of the form $lhs \sim rhs$ where lhs is the response variable and rhs a set of predictors.
data	optional data frame containing the variables in the model formula.
subset	optional vector specifying a subset of observations to be used.
na.action	function which indicates what should happen when the data contain NAs.
k	number of neighbours considered.
x	a matrix or data frame of training set predictors.
y	a numeric vector of outcomes.
...	additional parameters to pass to <code>knnregTrain</code> .
train	matrix or data frame of training set cases.
test	matrix or data frame of test set cases. A vector will be interpreted as a row vector for a single case.
use.all	controls handling of ties. If true, all distances equal to the kth largest are included. If false, a random selection of distances equal to the kth is chosen to use exactly k neighbours.

Details

`knnreg` is similar to [ipredknn](#) and `knnregTrain` is a modification of [knn](#). The underlying C code from the class package has been modified to return average outcome.

Value

An object of class `knnreg`. See [predict.knnreg](#).

Author(s)

[knn](#) by W. N. Venables and B. D. Ripley and [ipredknn](#) by Torsten.Hothorn <Torsten.Hothorn@rzmail.uni-erlangen.de>, modifications by Max Kuhn and Chris Keefer

Examples

```
data(BloodBrain)

inTrain <- createDataPartition(logBBB, p = .8)[[1]]

trainX <- bbbDescr[inTrain,]
trainY <- logBBB[inTrain]

testX <- bbbDescr[-inTrain,]
testY <- logBBB[-inTrain]

fit <- knnreg(trainX, trainY, k = 3)

plot(testY, predict(fit, testX))
```

lattice.rfe

Lattice functions for plotting resampling results of recursive feature selection

Description

A set of lattice functions are provided to plot the resampled performance estimates (e.g. classification accuracy, RMSE) over different subset sizes.

Usage

```
## S3 method for class 'rfe'
histogram(x, data = NULL, metric = x$metric, ...)

## S3 method for class 'rfe'
densityplot(x, data = NULL, metric = x$metric, ...)

## S3 method for class 'rfe'
xyplot(x, data = NULL, metric = x$metric, ...)

## S3 method for class 'rfe'
stripplot(x, data = NULL, metric = x$metric, ...)
```

Arguments

x	An object produced by rfe
data	This argument is not used
metric	A character string specifying the single performance metric that will be plotted
...	arguments to pass to either histogram , densityplot , xyplot or stripplot

Details

By default, only the resampling results for the optimal model are saved in the `rfe` object. The function `rfeControl` can be used to save all the results using the `returnResamp` argument.

If leave-one-out or out-of-bag resampling was specified, plots cannot be produced (see the method argument of `rfeControl`)

Value

A lattice plot object

Author(s)

Max Kuhn

See Also

[rfe](#), [rfeControl](#), [histogram](#), [densityplot](#), [xyplot](#), [stripplot](#)

Examples

```
## Not run:
library(mlbench)
n <- 100
p <- 40
sigma <- 1
set.seed(1)
sim <- mlbench.friedman1(n, sd = sigma)
x <- cbind(sim$x, matrix(rnorm(n * p), nrow = n))
y <- sim$y
colnames(x) <- paste("var", 1:ncol(x), sep = "")

normalization <- preProcess(x)
x <- predict(normalization, x)
x <- as.data.frame(x)
subsets <- c(10, 15, 20, 25)

ctrl <- rfeControl(
  functions = lmFuncs,
  method = "cv",
  verbose = FALSE,
  returnResamp = "all")

lmProfile <- rfe(x, y,
  sizes = subsets,
  rfeControl = ctrl)

xyplot(lmProfile)
stripplot(lmProfile)

histogram(lmProfile)
densityplot(lmProfile)
```

```
## End(Not run)
```

```
lift
```

```
Lift Plot
```

Description

For classification models, this function creates a 'lift plot' that describes how well a model ranks samples for one class

Usage

```
lift(x, ...)

## S3 method for class 'formula'
lift(x, data = NULL, class = NULL, cuts = NULL,
     subset = TRUE, lattice.options = NULL, ...)

## S3 method for class 'lift'
xyplot(x, data, ...)
```

Arguments

<code>x</code>	a lattice formula (see xyplot for syntax) where the left-hand side of the formula is a factor class variable of the observed outcome and the right-hand side specifies one or model columns corresponding to a numeric ranking variable for a model (e.g. class probabilities). The classification variable should have two levels.
<code>data</code>	For <code>lift.formula</code> , a data frame (or more precisely, anything that is a valid <code>envir</code> argument in <code>eval</code> , e.g., a list or an environment) containing values for any variables in the formula, as well as groups and subset if applicable. If not found in data, or if data is unspecified, the variables are looked for in the environment of the formula. This argument is not used for <code>xyplot.lift</code> .
<code>class</code>	a character string for the class of interest
<code>cuts</code>	the number of splits of the data are used to create the plot. By default, it uses as many cuts as there are rows in data
<code>subset</code>	An expression that evaluates to a logical or integer indexing vector. It is evaluated in data. Only the resulting rows of data are used for the plot.
<code>lattice.options</code>	A list that could be supplied to lattice.options
<code>...</code>	options to pass through to xyplot or the panel function (not used in <code>lift.formula</code>).

Details

`lift.formula` is used to process the data and `xyplot.lift` is used to create the plot.

To construct the lift plot, the following steps are used for each model:

1. The data are ordered by the numeric model prediction used on the right-hand side of the model formula
2. The data are split into `cuts - 1` roughly equal groups
3. the number of samples with true results equal to `class` are determined
4. the lift is calculated as the ratio of the percentage of samples in each split corresponding to `class` over the same percentage in the entire data set

`lift` produces a plot of the cumulative lift values by the percentage of samples evaluated.

This implementation uses the **lattice** function `xyplot`, so plot elements can be changed via panel functions, `trellis.par.set` or other means. `lift` uses the panel function `panel.lift2` by default, but it can be changes using `update.trellis` (see the examples in `panel.lift2`).

The following elements are set by default in the plot but can be changed by passing new values into `xyplot.lift`: `xlab = "% Samples Tested"`, `ylab = "% Samples Found"`, `type = "S"`, `ylim = extendrange(c(0, 100))` and `xlim = extendrange(c(0, 100))`.

Value

`lift.formula` returns a list with elements:

<code>data</code>	the data used for plotting
<code>cuts</code>	the number of cuts
<code>class</code>	the event class
<code>probNames</code>	the names of the model probabilities
<code>pct</code>	the baseline event rate

`xyplot.lift` returns a **lattice** object

Author(s)

Max Kuhn, some **lattice** code and documentation by Deepayan Sarkar

See Also

[xyplot](#), [trellis.par.set](#)

Examples

```
set.seed(1)
simulated <- data.frame(obs = factor(rep(letters[1:2], each = 100)),
                        perfect = sort(runif(200), decreasing = TRUE),
                        random = runif(200))

lift1 <- lift(obs ~ random, data = simulated)
```

```
lift1
xyplot(lift1)

lift2 <- lift(obs ~ random + perfect, data = simulated)
lift2
xyplot(lift2, auto.key = list(columns = 2))
```

maxDissim

Maximum Dissimilarity Sampling

Description

Functions to create a sub-sample by maximizing the dissimilarity between new samples and the existing subset.

Usage

```
maxDissim(a, b, n = 2, obj = minDiss, useNames = FALSE,
          randomFrac = 1, verbose = FALSE, ...)
minDiss(u)
sumDiss(u)
```

Arguments

a	a matrix or data frame of samples to start
b	a matrix or data frame of samples to sample from
n	the size of the sub-sample
obj	an objective function to measure overall dissimilarity
useNames	a logical: should the function return the row names (as opposed to the row index)
randomFrac	a number in (0, 1] that can be used to sub-sample from the remaining candidate values
verbose	a logical; should each step be printed?
...	optional arguments to pass to dist
u	a vector of dissimilarities

Details

Given an initial set of m samples and a larger pool of n samples, this function iteratively adds points to the smaller set by finding with of the n samples is most dissimilar to the initial set. The argument `obj` measures the overall dissimilarity between the initial set and a candidate point. For example, maximizing the minimum or the sum of the m dissimilarities are two common approaches.

This algorithm tends to select points on the edge of the data mainstream and will reliably select outliers. To select more samples towards the interior of the data set, set `randomFrac` to be small (see the examples below).

Value

a vector of integers or row names (depending on useNames) corresponding to the rows of b that comprise the sub-sample.

Author(s)

Max Kuhn <max.kuhn@pfizer.com>

References

Willett, P. (1999), "Dissimilarity-Based Algorithms for Selecting Structurally Diverse Sets of Compounds," *Journal of Computational Biology*, 6, 447-457.

See Also

[dist](#)

Examples

```
example <- function(pct = 1, obj = minDiss, ...)
{
  tmp <- matrix(rnorm(200 * 2), nrow = 200)

  ## start with 15 data points
  start <- sample(1:dim(tmp)[1], 15)
  base <- tmp[start,]
  pool <- tmp[-start,]

  ## select 9 for addition
  newSamp <- maxDissim(
    base, pool,
    n = 9,
    randomFrac = pct, obj = obj, ...)

  allSamp <- c(start, newSamp)

  plot(
    tmp[-newSamp,],
    xlim = extendrange(tmp[,1]), ylim = extendrange(tmp[,2]),
    col = "darkgrey",
    xlab = "variable 1", ylab = "variable 2")
  points(base, pch = 16, cex = .7)

  for(i in seq(along = newSamp))
    points(
      pool[newSamp[i],1],
      pool[newSamp[i],2],
      pch = paste(i), col = "darkred")
}

par(mfrow=c(2,2))
```

```
set.seed(414)
example(1, minDiss)
title("No Random Sampling, Min Score")

set.seed(414)
example(.1, minDiss)
title("10 Pct Random Sampling, Min Score")

set.seed(414)
example(1, sumDiss)
title("No Random Sampling, Sum Score")

set.seed(414)
example(.1, sumDiss)
title("10 Pct Random Sampling, Sum Score")
```

mdrr

Multidrug Resistance Reversal (MDRR) Agent Data

Description

Svetnik et al (2003) describe these data: "Bakken and Jurs studied a set of compounds originally discussed by Klopman et al., who were interested in multidrug resistance reversal (MDRR) agents. The original response variable is a ratio measuring the ability of a compound to reverse a leukemia cell's resistance to adriamycin. However, the problem was treated as a classification problem, and compounds with the ratio >4.2 were considered active, and those with the ratio ≤ 2.0 were considered inactive. Compounds with the ratio between these two cutoffs were called moderate and removed from the data for twoclass classification, leaving a set of 528 compounds (298 actives and 230 inactives). (Various other arrangements of these data were examined by Bakken and Jurs, but we will focus on this particular one.) We did not have access to the original descriptors, but we generated a set of 342 descriptors of three different types that should be similar to the original descriptors, using the DRAGON software."

The data and R code are in the Supplemental Data file for the article.

Usage

```
data(mdrr)
```

Value

mdrrDescr	the descriptors
mdrrClass	the categorical outcome ("Active" or "Inactive")

Source

Svetnik, V., Liaw, A., Tong, C., Culberson, J. C., Sheridan, R. P. Feuston, B. P (2003). Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling, *Journal of Chemical Information and Computer Sciences*, Vol. 43, pg. 1947-1958.

 modelLookup

Descriptions Of Models Available in train()

Description

This function enumerates the parameters and characteristics for models used by [train](#)

Usage

```
modelLookup(model = NULL)
```

Arguments

`model` a character string associated with the method argument of [train](#). If no value is passed, all models are returned

Details

One characteristic listed in the output is whether or not sub-models can be used for prediction. For example, if a PLS model is fit with X components, PLS models with $<X$ components do not need to be fit to obtain predictions. [train](#) exploits this characteristic whenever possible.

These types of tuning parameters are defined as "sequential" parameters since one value can be used to derive a sequences of predictions. Examples of model codes that include sequential tuning parameters are; blackboost, ctree, earth, enet, foba, gamboost, gbm, glmboost, glmnet, lars, lars2, lasso, logitBoost, pam, partDSA, pcr, pls, relaxo, rpart, scrda and superpc.

Value

a data frame with columns

<code>model</code>	a character string for the model code
<code>parameter</code>	the tuning parameter name
<code>label</code>	a tuning parameter label (used in plots)
<code>seq</code>	a logical; can sub-models be used to decrease training time (see the Details section)
<code>forReg</code>	a logical; can the model be used for regression?
<code>forClass</code>	a logical; can the model be used for classification?
<code>probModel</code>	a logical; does the model produce class probabilities?

Author(s)

Max Kuhn

See Also

[train](#)

Examples

```
modelLookup()
modelLookup("gbm")
```

nearZeroVar

Identification of near zero variance predictors

Description

nearZeroVar diagnoses predictors that have one unique value (i.e. are zero variance predictors) or predictors that have both of the following characteristics: they have very few unique values relative to the number of samples and the ratio of the frequency of the most common value to the frequency of the second most common value is large. checkConditionalX looks at the distribution of the columns of x conditioned on the levels of y and identifies columns of x that are sparse within groups of y.

Usage

```
nearZeroVar(x, freqCut = 95/5, uniqueCut = 10, saveMetrics = FALSE)
checkConditionalX(x, y)
checkResamples(index, x, y)
```

Arguments

x	a numeric vector or matrix, or a data frame with all numeric data
freqCut	the cutoff for the ratio of the most common value to the second most common value
uniqueCut	the cutoff for the percentage of distinct values out of the number of total samples
saveMetrics	a logical. If false, the positions of the zero- or near-zero predictors is returned. If true, a data frame with predictor information is returned.
y	a factor vector with at least two levels
index	a list. Each element corresponds to the training set samples in x for a given resample

Details

For example, an example of near zero variance predictor is one that, for 1000 samples, has two distinct values and 999 of them are a single value.

To be flagged, first the frequency of the most prevalent value over the second most frequent value (called the “frequency ratio”) must be above freqCut. Secondly, the “percent of unique values,” the number of unique values divided by the total number of samples (times 100), must also be below uniqueCut.

In the above example, the frequency ratio is 999 and the unique value percentage is 0.0001.

Checking the conditional distribution of x may be needed for some models, such as naive Bayes where the conditional distributions should have at least one data point within a class.

Value

For nearZeroVar: if saveMetrics = FALSE, a vector of integers corresponding to the column positions of the problematic predictors. If saveMetrics = TRUE, a data frame with columns:

freqRatio	the ratio of frequencies for the most common value over the second most common value
percentUnique	the percentage of unique data points out of the total number of data points
zeroVar	a vector of logicals for whether the predictor has only one distinct value
nzv	a vector of logicals for whether the predictor is a near zero variance predictor

For checkResamples or checkConditionalX, a vector of column indicators for predictors with empty conditional distributions in at least one class of y.

Author(s)

Max Kuhn, with speed improvements to nearZeroVar by Allan Engelhardt

Examples

```
nearZeroVar(iris[, -5], saveMetrics = TRUE)

data(BloodBrain)
nearZeroVar(bbbDescr)

set.seed(1)
classes <- factor(rep(letters[1:3], each = 30))
x <- data.frame(x1 = rep(c(0, 1), 45),
               x2 = c(rep(0, 10), rep(1, 80)))

lapply(x, table, y = classes)
checkConditionalX(x, classes)

folds <- createFolds(classes, k = 3, returnTrain = TRUE)
x$x3 <- x$x1
x$x3[folds[[1]]] <- 0

checkResamples(folds, x, classes)
```

normalize.AffyBatch.normalize2Reference

Quantile Normalization to a Reference Distribution

Description

Quantile normalization based upon a reference distribution. This function normalizes a matrix of data (typically Affy probe level intensities).

Usage

```
normalize.AffyBatch.normalize2Reference(  
  abatch,  
  type = c("separate", "pmonly", "mmonly", "together"),  
  ref = NULL)
```

Arguments

abatch	An {AffyBatch}
type	A string specifying how the normalization should be applied. See details for more.
ref	A vector of reference values. See details for more.

Details

This method is based upon the concept of a quantile-quantile plot extended to n dimensions. No special allowances are made for outliers. If you make use of quantile normalization either through rma or *expresso* please cite Bolstad et al, *Bioinformatics* (2003).

The type argument should be one of "separate", "pmonly", "mmonly", "together" which indicates whether to normalize only one probe type (PM,MM) or both together or separately.

The function uses the data supplied in ref to use as the reference distribution. In other words, the PMs in abatch will be normalized to have the same distribution as the data in ref. If ref is NULL, the normalizing takes place using the average quantiles of the PM values in abatch (just as in `normalize.AffyBatch.quantile`).

Value

A normalized AffyBatch.

Author(s)

Max Kuhn, adapted from Ben Bolstad, <bolstad@stat.berkeley.edu>

References

Bolstad, B (2001) *Probe Level Quantile Normalization of High Density Oligonucleotide Array Data*. Unpublished manuscript

Bolstad, B. M., Irizarry R. A., Astrand, M, and Speed, T. P. (2003) *A Comparison of Normalization Methods for High Density Oligonucleotide Array Data Based on Bias and Variance*. *Bioinformatics* 19(2) ,pp 185-193.

See Also

normalize

Examples

```
# first, let affy/expresso know that the method exists
# normalize.AffyBatch.methods <- c(normalize.AffyBatch.methods, "normalize2Reference")

# example not run, as it would take a while
# RawData <- ReadAffy(celfile.path=FilePath)

# Batch1Step1 <- bg.correct(RawData, "rma")
# Batch1Step2 <- normalize.AffyBatch.quantiles(Batch1Step1)
# referencePM <- pm(Batch1Step2)[,1]
# Batch1Step3 <- computeExprSet(Batch1Step2, "pmonly", "medianpolish")

# Batch2Step1 <- bg.correct(RawData2, "rma")
# Batch2Step2 <- normalize.AffyBatch.normalize2Reference(Batch2Step1, ref = referencePM)
# Batch2Step3 <- computeExprSet(Batch2Step2, "pmonly", "medianpolish")
```

normalize2Reference	<i>Quantile Normalize Columns of a Matrix Based on a Reference Distribution</i>
---------------------	---

Description

Normalize the columns of a matrix to have the same quantiles, allowing for missing values. Users do not normally need to call this function directly - use `normalize.AffyBatch.normalize2Reference` instead.

Usage

```
normalize2Reference(data, refData = NULL, ties = TRUE)
```

Arguments

<code>data</code>	numeric matrix. Missing values are allowed.
<code>refData</code>	A vector of reference values.
<code>ties</code>	logical. If TRUE, ties in each column of A are treated in careful way. Tied values will be normalized to the mean of the corresponding pooled quantiles.

Details

This function is intended to normalize single channel or A-value microarray intensities between arrays. Each quantile of each column is set to either: quantiles of the reference distribution (`refData` supplied) or the mean of that quantile across arrays (`refData` is NULL). The intention is to make all the normalized columns have the same empirical distribution. This will be exactly true if there are no missing values and no ties within the columns: the normalized columns are then simply permutations of one another.

If there are ties amongst the intensities for a particular array, then with `ties=FALSE` the ties are broken in an unpredictable order. If `ties=TRUE`, all the tied values for that array will be normalized to the same value, the average of the quantiles for the tied values.

Value

A matrix of the same dimensions as A containing the normalized values.

Author(s)

Max Kuhn, adapted from Gordon Smyth

References

Bolstad, B. M., Irizarry R. A., Astrand, M., and Speed, T. P. (2003), A comparison of normalization methods for high density oligonucleotide array data based on bias and variance. *Bioinformatics* **19**, 185-193.

See Also

`normalize.AffyBatch.normalize2Reference`

nullModel

Fit a simple, non-informative model

Description

Fit a single mean or largest class model

Usage

```
nullModel(x, ...)

## Default S3 method:
nullModel(x = NULL, y, ...)

## S3 method for class 'nullModel'
predict(object, newdata = NULL, type = NULL, ...)
```

Arguments

x	An optional matrix or data frame of predictors. These values are not used in the model fit
y	A numeric vector (for regression) or factor (for classification) of outcomes
...	Optional arguments (not yet used)
object	An object of class nullModel
newdata	A matrix or data frame of predictors (only used to determine the number of predictions to return)
type	Either "raw" (for regression), "class" or "prob" (for classification)

Details

`nullModel` emulates other model building functions, but returns the simplest model possible given a training set: a single mean for numeric outcomes and the most prevalent class for factor outcomes. When class probabilities are requested, the percentage of the training set samples with the most prevalent class is returned.

Value

The output of `nullModel` is a list of class `nullModel` with elements

<code>call</code>	the function call
<code>value</code>	the mean of <code>y</code> or the most prevalent class
<code>levels</code>	when <code>y</code> is a factor, a vector of levels. NULL otherwise
<code>pct</code>	when <code>y</code> is a factor, a data frame with a column for each class (NULL otherwise). The column for the most prevalent class has the proportion of the training samples with that class (the other columns are zero).
<code>n</code>	the number of elements in <code>y</code>

`predict.nullModel` returns either a factor or numeric vector depending on the class of `y`. All predictions are always the same.

Examples

```
outcome <- factor(
  sample(letters[1:2],
        size = 100,
        prob = c(.1, .9),
        replace = TRUE))
useless <- nullModel(y = outcome)
useless
predict(useless, matrix(NA, nrow = 10))
```

 oil

Fatty acid composition of commercial oils

Description

Fatty acid concentrations of commercial oils were measured using gas chromatography. The data is used to predict the type of oil. Note that only the known oils are in the data set. Also, the authors state that there are 95 samples of known oils. However, we count 96 in Table 1 (pgs. 33-35).

Usage

```
data(oil)
```

Value

fattyAcids	data frame of fatty acid compositions: Palmitic, Stearic, Oleic, Linoleic, Linolenic, Eicosanoic and Eicosenoic. When values fell below the lower limit of the assay (denoted as <X in the paper), the limit was used.
oilType	factor of oil types: pumpkin (A), sunflower (B), peanut (C), olive (D), soybean (E), rapeseed (F) and corn (G).

Source

Brodnjak-Voncina et al. (2005). Multivariate data analysis in classification of vegetable oils characterized by the content of fatty acids, *Chemometrics and Intelligent Laboratory Systems*, Vol. 75:31-45.

oneSE *Selecting tuning Parameters*

Description

Various functions for setting tuning parameters

Usage

```
best(x, metric, maximize)
oneSE(x, metric, num, maximize)
tolerance(x, metric, tol = 1.5, maximize)
```

Arguments

x	a data frame of tuning parameters and model results, sorted from least complex models to the most complex
metric	a string that specifies what summary metric will be used to select the optimal model. By default, possible values are "RMSE" and "Rsquared" for regression and "Accuracy" and "Kappa" for classification. If custom performance metrics are used (via the summaryFunction argument in trainControl , the value of metric should match one of the arguments. If it does not, a warning is issued and the first metric given by the summaryFunction is used.
maximize	a logical: should the metric be maximized or minimized?
num	the number of resamples (for oneSE only)
tol	the acceptable percent tolerance (for tolerance only)

Details

These functions can be used by `train` to select the "optimal" model from a series of models. Each requires the user to select a metric that will be used to judge performance. For regression models, values of "RMSE" and "Rsquared" are applicable. Classification models use either "Accuracy" or "Kappa" (for unbalanced class distributions).

By default, `train` uses `best`.

`best` simply chooses the tuning parameter associated with the largest (or lowest for "RMSE") performance.

`oneSE` is a rule in the spirit of the "one standard error" rule of Breiman et al (1984), who suggest that the tuning parameter associated with the best performance may over fit. They suggest that the simplest model within one standard error of the empirically optimal model is the better choice. This assumes that the models can be easily ordered from simplest to most complex (see the Details section below).

`tolerance` takes the simplest model that is within a percent tolerance of the empirically optimal model. For example, if the largest Kappa value is 0.5 and a simpler model within 3 percent is acceptable, we score the other models using $(x - 0.5)/0.5 * 100$. The simplest model whose score is not less than 3 is chosen (in this case, a model with a Kappa value of 0.35 is acceptable).

User-defined functions can also be used. The argument `selectionFunction` in `trainControl` can be used to pass the function directly or to pass the function by name.

Value

an row index

Note

In many cases, it is not very clear how to order the models on simplicity. For simple trees and other models (such as PLS), this is straightforward. However, for others it is not.

For example, many of the boosting models used by `caret` have parameters for the number of boosting iterations and the tree complexity (others may also have a learning rate parameter). In this implementation, we order models on number of iterations, then tree depth. Clearly, this is arguable (please email the author for suggestions though).

For MARS models, they are ordered on the degree of the features, then the number of retained terms.

RBF SVM models are ordered first by the cost parameter, then by the kernel parameter while polynomial models are ordered first on polynomial degree, then cost and scale.

Neural networks are ordered by the number of hidden units and then the amount of weight decay.

k -nearest neighbor models are ordered from most neighbors to least (i.e. smoothest to model jagged decision boundaries).

Elastic net models are ordered first on the L1 penalty, then by the L2 penalty.

Author(s)

Max Kuhn

References

Breiman, Friedman, Olshen, and Stone. (1984) *Classification and Regression Trees*. Wadsworth.

See Also

[train](#), [trainControl](#)

Examples

```
# simulate a PLS regression model
test <- data.frame(
  ncomp = 1:5,
  RMSE = c(3, 1.1, 1.02, 1, 2),
  RMSESD = .4)

best(test, "RMSE", maximize = FALSE)
oneSE(test, "RMSE", maximize = FALSE, num = 10)
tolerance(test, "RMSE", tol = 3, maximize = FALSE)

### usage example

data(BloodBrain)

marsGrid <- data.frame(
  .degree = 1,
  .nprune = (1:10) * 3)

set.seed(1)
marsFit <- train(
  bbbDescr, logBBB,
  "earth",
  tuneGrid = marsGrid,
  trControl = trainControl(
    method = "cv",
    number = 10,
    selectionFunction = "tolerance"))

# around 18 terms should yield the smallest CV RMSE
```

Description

Two panel functions that be used in conjunction with [lift](#).

Usage

```
panel.lift(x, y, ...)
```

```
panel.lift2(x, y, pct = 0, ...)
```

Arguments

x	the percentage of searched to be plotted in the scatterplot
y	the percentage of events found to be plotted in the scatterplot
pct	the baseline percentage of true events in the data
...	options to pass to panel.xyplot

Details

`panel.lift` plots the data with a simple (black) 45 degree reference line.

`panel.lift2` is the default for `lift` and plots the data points with a shaded region encompassing the space between to the random model and perfect model trajectories. The color of the region is determined by the lattice `reference.line` information (see example below).

Author(s)

Max Kuhn

See Also

[lift](#), [panel.xyplot](#), [xyplot](#), [trellis.par.set](#)

Examples

```
set.seed(1)
simulated <- data.frame(obs = factor(rep(letters[1:2], each = 100)),
                        perfect = sort(runif(200), decreasing = TRUE),
                        random = runif(200))

regionInfo <- trellis.par.get("reference.line")
regionInfo$col <- "lightblue"
trellis.par.set("reference.line", regionInfo)

lift2 <- lift(obs ~ random + perfect, data = simulated)
lift2
xyplot(lift2, auto.key = list(columns = 2))

## use a different panel function
xyplot(lift2, panel = panel.lift)
```

`panel.needle`*Needle Plot Lattice Panel*

Description

A variation of `panel.dotplot` that plots horizontal lines from zero to the data point.

Usage

```
panel.needle(x, y, horizontal = TRUE,  
            pch, col, lty, lwd,  
            col.line, levels.fos,  
            groups = NULL,  
            ...)
```

Arguments

<code>x,y</code>	variables to be plotted in the panel. Typically y is the 'factor'
<code>horizontal</code>	logical. If FALSE, the plot is 'transposed' in the sense that the behaviours of x and y are switched. x is now the 'factor'. Interpretation of other arguments change accordingly. See documentation of <code>bwplot</code> for a fuller explanation.
<code>pch, col, lty, lwd, col.line</code>	graphical parameters
<code>levels.fos</code>	locations where reference lines will be drawn
<code>groups</code>	grouping variable (affects graphical parameters)
<code>...</code>	extra parameters, passed to <code>panel.xyplot</code> which is responsible for drawing the foreground points (<code>panel.dotplot</code> only draws the background reference lines).

Details

Creates (possibly grouped) needleplot of x against y or vice versa

Author(s)

Max Kuhn, based on [panel.dotplot](#) by Deepayan Sarkar

See Also

[dotplot](#)

Description

Run PCA on a dataset, then use it in a neural network model

Usage

```
## Default S3 method:
pcaNNet(x, y, thresh = 0.99, ...)
## S3 method for class 'formula'
pcaNNet(formula, data, weights, ...,
         thresh = .99, subset, na.action, contrasts = NULL)

## S3 method for class 'pcaNNet'
predict(object, newdata, type = c("raw", "class"), ...)
```

Arguments

formula	A formula of the form <code>class ~ x1 + x2 + ...</code>
x	matrix or data frame of x values for examples.
y	matrix or data frame of target values for examples.
weights	(case) weights for each example – if missing defaults to 1.
thresh	a threshold for the cumulative proportion of variance to capture from the PCA analysis. For example, to retain enough PCA components to capture 95 percent of variation, set <code>thresh = .95</code>
data	Data frame from which variables specified in <code>formula</code> are preferentially to be taken.
subset	An index vector specifying the cases to be used in the training sample. (NOTE: If given, this argument must be named.)
na.action	A function to specify the action to be taken if NAs are found. The default action is for the procedure to fail. An alternative is <code>na.omit</code> , which leads to rejection of cases with missing values on any required variable. (NOTE: If given, this argument must be named.)
contrasts	a list of contrasts to be used for some or all of the factors appearing as variables in the model formula.
object	an object of class <code>pcaNNet</code> as returned by <code>pcaNNet</code> .
newdata	matrix or data frame of test examples. A vector is considered to be a row vector comprising a single case.
type	Type of output
...	arguments passed to nnet

Details

The function first will run principal component analysis on the data. The cumulative percentage of variance is computed for each principal component. The function uses the `thresh` argument to determine how many components must be retained to capture this amount of variance in the predictors.

The principal components are then used in a neural network model.

When predicting samples, the new data are similarly transformed using the information from the PCA analysis on the training data and then predicted.

Because the variance of each predictor is used in the PCA analysis, the code does a quick check to make sure that each predictor has at least two distinct values. If a predictor has one unique value, it is removed prior to the analysis.

Value

For `pcaNNet`, an object of `"pcaNNet"` or `"pcaNNet.formula"`. Items of interest in the output are:

<code>pc</code>	the output from preProcess
<code>model</code>	the model generated from nnet
<code>names</code>	if any predictors had only one distinct value, this is a character string of the remaining columns. Otherwise a value of <code>NULL</code>

Author(s)

These are heavily based on the `nnet` code from Brian Ripley.

References

Ripley, B. D. (1996) *Pattern Recognition and Neural Networks*. Cambridge.

See Also

[nnet](#), [preProcess](#)

Examples

```
data(BloodBrain)
modelFit <- pcaNNet(bbbDescr[, 1:10], logBBB, size = 5, linout = TRUE, trace = FALSE)
modelFit

predict(modelFit, bbbDescr[, 1:10])
```

plot.train

Plot Method for the train Class

Description

This function takes the output of a `train` object and creates a line or level plot using the `lattice` library.

Usage

```
## S3 method for class 'train'
plot(x,
     plotType = "scatter",
     metric = x$perfNames[1],
     digits = getOption("digits") - 3,
     xTrans = NULL,
     ...)
```

Arguments

<code>x</code>	an object of class <code>train</code> .
<code>metric</code>	What measure of performance to plot. Examples of possible values are "RMSE", "Rsquared", "Accuracy" or "Kappa". Other values can be used depending on what metrics have been calculated.
<code>plotType</code>	a string describing the type of plot ("scatter", "level" or "line")
<code>digits</code>	an integer specifying the number of significant digits used to label the parameter value.
<code>xTrans</code>	a function that will be used to scale the x-axis in scatter plots.
<code>...</code>	specifications to be passed to <code>levelplot</code> , <code>xyplot</code> , <code>stripplot</code> (for line plots). The function automatically sets some arguments (e.g. axis labels) but passing in values here will over-ride the defaults

Details

If there are no tuning parameters, or none were varied, an error is produced.

If the model has one tuning parameter with multiple candidate values, a plot is produced showing the profile of the results over the parameter. Also, a plot can be produced if there are multiple tuning parameters but only one is varied.

If there are two tuning parameters with different values, a plot can be produced where a different line is shown for each value of the other parameter. For three parameters, the same line plot is created within conditioning panels of the other parameter.

Also, with two tuning parameters (with different values), a `levelplot` (i.e. un-clustered heatmap) can be created. For more than two parameters, this plot is created inside conditioning panels.

Author(s)

Max Kuhn

References

Kuhn (2008), “Building Predictive Models in R Using the caret” (<http://www.jstatsoft.org/v28/i05/>)

See Also[train](#), [levelplot](#), [xyplot](#), [stripplot](#)**Examples**

```
pamFit <- train(Species ~ .,
               data = iris,
               "pam",
               tuneLength = 15)
plot(pamFit, scales = list(x = list(rot = 90)))

library(klaR)
rdaFit <- train(Species ~ .,
               data = iris,
               "rda",
               control = trainControl(method = "cv"))
plot(rdaFit, plotType = "line")
plot(rdaFit, plotType = "level")
```

plot.varImp.train *Plotting variable importance measures*

Description

This function produces lattice plots of objects with class "varImp.train". More info will be forthcoming.

Usage

```
## S3 method for class 'varImp.train'
plot(x, top = dim(x$importance)[1], ...)
```

Arguments

x	an object with class varImp.
top	a scalar numeric that specifies the number of variables to be displayed (in order of importance)
...	arguments to pass to the lattice plot function (dotplot and panel.needle)

Details

For models where there is only one importance value, such as regression models, a "Pareto-type" plot is produced where the variables are ranked by their importance and a needle-plot is used to show the top variables.

When there is more than one importance value per predictor, the same plot is produced within conditioning panels for each class. The top predictors are sorted by their average importance.

Value

a lattice plot object

Author(s)

Max Kuhn

plotClassProbs *Plot Predicted Probabilities in Classification Models*

Description

This function takes an object (preferably from the function [extractProb](#)) and creates a lattice plot.

If the call to [extractProb](#) included test data, these data are shown, but if unknowns were also included, these are not plotted

Usage

```
plotClassProbs(object,
                plotType = "histogram",
                useObjects = FALSE,
                ...)
```

Arguments

object	an object (preferably from the function extractProb . There should be columns for each level of the class factor and columns named obs, pred, model (e.g. "rpart", "nnet" etc), dataType (e.g. "Training", "Test" etc) and optionally objects (for giving names to objects with the same model type).
plotType	either "histogram" or "densityplot"
useObjects	a logical; should the object name (if any) be used as a conditioning variable?
...	parameters to pass to histogram or densityplot

Value

A lattice object. Note that the plot has to be printed to be displayed (especially in a loop).

Author(s)

Max Kuhn

Examples

```
data(mdr)
set.seed(90)
inTrain <- createDataPartition(mdrClass, p = .5)[[1]]

trainData <- mdrDescr[inTrain,1:20]
testData <- mdrDescr[-inTrain,1:20]

trainY <- mdrClass[inTrain]
testY <- mdrClass[-inTrain]

ctrl <- trainControl(method = "cv")

nbFit1 <- train(trainData, trainY, "nb",
               trControl = ctrl,
               tuneGrid = data.frame(.usekernel = TRUE, .fL = 0))

nbFit2 <- train(trainData, trainY, "nb",
               trControl = ctrl,
               tuneGrid = data.frame(.usekernel = FALSE, .fL = 0))

models <- list(para = nbFit2,
              nonpara = nbFit1)

predProbs <- extractProb(models,
                        testX = testData,
                        testY = testY)

plotClassProbs(predProbs,
               useObjects = TRUE)
plotClassProbs(predProbs,
               subset = object == "para" & dataType == "Test")
plotClassProbs(predProbs,
               useObjects = TRUE,
               plotType = "densityplot",
               auto.key = list(columns = 2))
```

Description

This function takes an object (preferably from the function `extractPrediction`) and creates a lattice plot. For numeric outcomes, the observed and predicted data are plotted with a 45 degree reference line and a smoothed fit. For factor outcomes, a dotplot plot is produced with the accuracies for the different models.

If the call to `extractPrediction` included test data, these data are shown, but if unknowns were also included, they are not plotted

Usage

```
plotObsVsPred(object, equalRanges = TRUE, ...)
```

Arguments

<code>object</code>	an object (preferably from the function <code>extractPrediction</code> . There should be columns named <code>obs</code> , <code>pred</code> , <code>model</code> (e.g. "rpart", "nnet" etc) and <code>dataType</code> (e.g. "Training", "Test" etc)
<code>equalRanges</code>	a logical; should the x- and y-axis ranges be the same?
<code>...</code>	parameters to pass to <code>xyplot</code> or <code>dotplot</code> , such as <code>auto.key</code>

Value

A lattice object. Note that the plot has to be printed to be displayed (especially in a loop).

Author(s)

Max Kuhn

Examples

```
## Not run:
# regression example
data(BostonHousing)
rpartFit <- train(BostonHousing[1:100, -c(4, 14)],
                 BostonHousing$medv[1:100],
                 "rpart", tuneLength = 9)
plsFit <- train(BostonHousing[1:100, -c(4, 14)],
               BostonHousing$medv[1:100],
               "pls")

predVals <- extractPrediction(list(rpartFit, plsFit),
                              testX = BostonHousing[101:200, -c(4, 14)],
                              testY = BostonHousing$medv[101:200],
                              unkX = BostonHousing[201:300, -c(4, 14)])

plotObsVsPred(predVals)

#classification example
data(Satellite)
```

```
numSamples <- dim(Satellite)[1]
set.seed(716)

varIndex <- 1:numSamples

trainSamples <- sample(varIndex, 150)

varIndex <- (1:numSamples)[-trainSamples]
testSamples <- sample(varIndex, 100)

varIndex <- (1:numSamples)[-c(testSamples, trainSamples)]
unkSamples <- sample(varIndex, 50)

trainX <- Satellite[trainSamples, -37]
trainY <- Satellite[trainSamples, 37]

testX <- Satellite[testSamples, -37]
testY <- Satellite[testSamples, 37]

unkX <- Satellite[unkSamples, -37]

knnFit <- train(trainX, trainY, "knn")
rpartFit <- train(trainX, trainY, "rpart")

predTargets <- extractPrediction(list(knnFit, rpartFit),
                                  testX = testX,
                                  testY = testY,
                                  unkX = unkX)

plotObsVsPred(predTargets)

## End(Not run)
```

plsda

Partial Least Squares and Sparse Partial Least Squares Discriminant Analysis

Description

plsda is used to fit standard PLS models for classification while splsda performs sparse PLS that embeds feature selection and regularization for the same purpose.

Usage

```
plsda(x, ...)
```

Default S3 method:

```
plsda(x, y, ncomp = 2, probMethod = "softmax", prior = NULL, ...)
```

S3 method for class 'plsda'

```

predict(object, newdata = NULL, ncomp = NULL, type = "class", ...)

splsda(x, ...)

## Default S3 method:
splsda(x, y, probMethod = "softmax", prior = NULL, ...)

## S3 method for class 'splsda'
predict(object, newdata = NULL, type = "class", ...)

```

Arguments

x	a matrix or data frame of predictors
y	a factor or indicator matrix for the discrete outcome. If a matrix, the entries must be either 0 or 1 and rows must sum to one
ncomp	the number of components to include in the model. Predictions can be made for models with values less than ncomp.
probMethod	either "softmax" or "Bayes" (see Details)
prior	a vector or prior probabilities for the classes (only used for probMethod = "Bayes")
...	arguments to pass to plsr or spls . For splsda, this is the method for passing tuning parameters specifications (e.g. K, eta or kappa)
object	an object produced by plsda
newdata	a matrix or data frame of predictors
type	either "class", "prob" or "raw" to produce the predicted class, class probabilities or the raw model scores, respectively.

Details

If a factor is supplied, the appropriate indicator matrix is created.

A multivariate PLS model is fit to the indicator matrix using the [plsr](#) or [spls](#) function.

Two prediction methods can be used.

The **softmax function** transforms the model predictions to "probability-like" values (e.g. on [0, 1] and sum to 1). The class with the largest class probability is the predicted class.

Also, **Bayes rule** can be applied to the model predictions to form posterior probabilities. Here, the model predictions for the training set are used along with the training set outcomes to create conditional distributions for each class. When new samples are predicted, the raw model predictions are run through these conditional distributions to produce a posterior probability for each class (along with the prior). This process is repeated ncomp times for every possible PLS model. The [NaiveBayes](#) function is used with usekernel = TRUE for the posterior probability calculations.

Value

For plsda, an object of class "plsda" and "mvr". For splsda, an object of class splsda.

The predict methods produce either a vector, matrix or three-dimensional array, depending on the values of type of ncomp. For example, specifying more than one value of ncomp with type = "class" will produce a three dimensional array but the default specification would produce a factor vector.

See Also

[pls](#), [spls](#)

Examples

```
data(mdr)
set.seed(1)
inTrain <- sample(seq(along = mdrClass), 450)

nzv <- nearZeroVar(mdrDescr)
filteredDescr <- mdrDescr[, -nzv]

training <- filteredDescr[inTrain,]
test <- filteredDescr[-inTrain,]
trainMDRR <- mdrClass[inTrain]
testMDRR <- mdrClass[-inTrain]

preProcValues <- preprocess(training)

trainDescr <- predict(preProcValues, training)
testDescr <- predict(preProcValues, test)

useBayes <- plsda(trainDescr, trainMDRR, ncomp = 5,
                  probMethod = "Bayes")
useSoftmax <- plsda(trainDescr, trainMDRR, ncomp = 5)

confusionMatrix(
  predict(useBayes, testDescr),
  testMDRR)

confusionMatrix(
  predict(useSoftmax, testDescr),
  testMDRR)

histogram(
  ~predict(useBayes, testDescr, type = "prob")[, "Active"]
  | testMDRR, xlab = "Active Prob", xlim = c(-.1, 1.1))
histogram(
  ~predict(useSoftmax, testDescr, type = "prob")[, "Active",]
  | testMDRR, xlab = "Active Prob", xlim = c(-.1, 1.1))

## different sized objects are returned
length(predict(useBayes, testDescr))
dim(predict(useBayes, testDescr, ncomp = 1:3))
dim(predict(useBayes, testDescr, type = "prob"))
```

```

dim(predict(useBayes, testDescr, type = "prob", ncomp = 1:3))

## Using spls:
## (As of 11/09, the spls package now has a similar function with
## the same name. To avoid conflicts, use caret::splsda to
## get this version)

splsFit <- caret::splsda(trainDescr, trainMDRR,
                        K = 5, eta = .9,
                        probMethod = "Bayes")

confusionMatrix(
  caret::predict.splsda(splsFit, testDescr),
  testMDRR)

```

postResample

Calculates performance across resamples

Description

Given two numeric vectors of data, the mean squared error and R-squared are calculated. For two factors, the overall agreement rate and Kappa are determined.

Usage

```

postResample(pred, obs)
defaultSummary(data, lev = NULL, model = NULL)

twoClassSummary(data, lev = NULL, model = NULL)

R2(pred, obs, formula = "corr", na.rm = FALSE)
RMSE(pred, obs, na.rm = FALSE)

```

Arguments

pred	A vector of numeric data (could be a factor)
obs	A vector of numeric data (could be a factor)
data	a data frame or matrix with columns obs and pred for the observed and predicted outcomes. For twoClassSummary, columns should also include predicted probabilities for each class. See the classProbs argument to trainControl
lev	a character vector of factors levels for the response. In regression cases, this would be NULL.
model	a character string for the model name (as taken from the method argument of train).
formula	which R^2 formula should be used? Either "corr" or "traditional". See Kvalseth (1985) for a summary of the different equations.
na.rm	a logical value indicating whether NA values should be stripped before the computation proceeds.

Details

postResample is meant to be used with apply across a matrix. For numeric data the code checks to see if the standard deviation of either vector is zero. If so, the correlation between those samples is assigned a value of zero. NA values are ignored everywhere.

Note that many models have more predictors (or parameters) than data points, so the typical mean squared error denominator (n - p) does not apply. Root mean squared error is calculated using $\sqrt{\text{mean}((\text{pred} - \text{obs})^2)}$. Also, R^2 is calculated wither using as the square of the correlation between the observed and predicted outcomes when form = "corr". when form = "traditional",

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y}_i)^2}$$

For defaultSummary is the default function to compute performance metrics in [train](#). It is a wrapper around postResample.

twoClassSummary computes sensitivity, specificity and the area under the ROC curve. To use this function, the classProbs argument of [trainControl](#) should be TRUE.

Other functions can be used via the summaryFunction argument of [trainControl](#). Custom functions must have the same arguments as defaultSummary.

Value

A vector of performance estimates.

Author(s)

Max Kuhn

References

Kvalseth. Cautionary note about R^2 . American Statistician (1985) vol. 39 (4) pp. 279-285

See Also

[trainControl](#)

Examples

```
predicted <- matrix(rnorm(50), ncol = 5)
observed <- rnorm(10)
apply(predicted, 2, postResample, obs = observed)
```

pottery

Pottery from Pre-Classical Sites in Italy

Description

Measurements of 58 pottery samples.

Usage

```
data(pottery)
```

Value

pottery 11 elemental composition measurements
 potteryClass factor of pottery type: black carbon containing bulks (A) and clayey (B)

Source

R. G. Brereton (2003). *Chemometrics: Data Analysis for the Laboratory and Chemical Plant*, pg. 261.

prcomp.resamples

Principal Components Analysis of Resampling Results

Description

Performs a principal components analysis on an object of class `resamples` and returns the results as an object with classes `prcomp.resamples` and `prcomp`.

Usage

```
## S3 method for class 'resamples'
prcomp(x, metric = x$metrics[1], ...)

## S3 method for class 'resamples'
cluster(x, metric = x$metrics[1], ...)

## S3 method for class 'prcomp.resamples'
plot(x, what = "scree", dims = max(2, ncol(x$rotation)), ...)
```

Arguments

x	For prcomp, an object of class <code>resamples</code> and for <code>plot.prcomp.resamples</code> , an object of class <code>plot.prcomp.resamples</code>
metric	a performance metric that was estimated for every resample
what	the type of plot: "scree" produces a bar chart of standard deviations, "cumulative" produces a bar chart of the cumulative percent of variance, "loadings" produces a scatterplot matrix of the loading values and "components" produces a scatterplot matrix of the PCA components
dims	The number of dimensions to plot when what = "loadings" or what = "components"
...	For <code>prcomp.resamples</code> , options to pass to <code>prcomp</code> , for <code>plot.prcomp.resamples</code> , options to pass to Lattice objects (see Details below) and, for <code>cluster.resamples</code> , options to pass to <code>hclust</code> .

Details

The principal components analysis treats the models as variables and the resamples are realizations of the variables. In this way, we can use PCA to "cluster" the assays and look for similarities. Most of the methods for `prcomp` can be used, although custom print and plot methods are used.

The plot method uses lattice graphics. When what = "scree" or what = "cumulative", `barchart` is used. When what = "loadings" or what = "components", either `xyplot` or `splom` are used (the latter when `dims > 2`). Options can be passed to these methods using ...

When what = "loadings" or what = "components", the plots are put on a common scale so that later components are less likely to be over-interpreted. See Geladi et al (2003) for examples of why this can be important.

For clustering, `hclust` is used to determine clusters of models based on the resampled performance values.

Value

For `prcomp.resamples`, an object with classes `prcomp.resamples` and `prcomp`. This object is the same as the object produced by `prcomp`, but with additional elements:

metric	the value for the metric argument
call	the call

For `plot.prcomp.resamples`, a Lattice object (see Details above)

Author(s)

Max Kuhn

References

Geladi, P.; Manley, M.; and Lestander, T. (2003), "Scatter plotting in multivariate data analysis," *J. Chemometrics*, 17: 503-511

See Also

[resamples](#), [barchart](#), [xyplot](#), [splom](#), [hclust](#)

Examples

```
## Not run:
#load(url("http://caret.r-forge.r-project.org/Classification_and_Regression_Training_files/exampleModels.RData

resamps <- resamples(list(CART = rpartFit,
                          CondInfTree = ctreeFit,
                          MARS = earthFit))
resampPCA <- prcomp(resamps)

resampPCA

plot(resampPCA, "scree")

plot(resampPCA, "components")

plot(resampPCA, "components", dims = 2, auto.key = list(columns = 3))

## End(Not run)
```

predict.bagEarth	<i>Predicted values based on bagged Earth and FDA models</i>
------------------	--

Description

Predicted values based on bagged Earth and FDA models

Usage

```
## S3 method for class 'bagEarth'
predict(object, newdata = NULL, type = "response", ...)
## S3 method for class 'bagFDA'
predict(object, newdata = NULL, type = "class", ...)
```

Arguments

object	Object of class inheriting from bagEarth
newdata	An optional data frame or matrix in which to look for variables with which to predict. If omitted, the fitted values are used (see note below).
type	The type of prediction. For bagged earth regression model, type = "response" will produce a numeric vector of the usual model predictions. earth also allows the user to fit generalized linear models. In this case, type = "response" produces the inverse link results as a vector. In the case of a binomial generalized

linear model, type = "response" produces a vector of probabilities, type = "class" generates a factor vector and type = "prob" produces a 2 column matrix with probabilities for both classes (averaged across the individual models). Similarly, for bagged [fda](#) models, type = "class" generates a factor vector and type = "probs" outputs a matrix of class probabilities.

... not used

Value

a vector of predictions

Note

If the predictions for the original training set are needed, there are two ways to calculate them. First, the original data set can be predicted by each bagged earth model. Secondly, the predictions from each bootstrap sample could be used (but are more likely to overfit). If the original call to `bagEarth` or `bagFDA` had `keepX = TRUE`, the first method is used, otherwise the values are calculated via the second method.

Author(s)

Max Kuhn

See Also

[bagEarth](#)

Examples

```
data(trees)
fit1 <- bagEarth(Volume ~ ., data = trees, keepX = TRUE)
fit2 <- bagEarth(Volume ~ ., data = trees, keepX = FALSE)
hist(predict(fit1) - predict(fit2))
```

predict.knn3

Predictions from k-Nearest Neighbors

Description

Predict the class of a new observation based on k-NN.

Usage

```
## S3 method for class 'knn3'
predict(object, newdata, type=c("prob", "class"), ...)
```

Arguments

object	object of class knn3.
newdata	a data frame of new observations.
type	return either the predicted class or the the proportion of the votes for the winning class.
...	additional arguments.

Details

This function is a method for the generic function [predict](#) for class knn3. For the details see [knn3](#). This is essentially a copy of [predict.ipredknn](#).

Value

Either the predicted class or the the proportion of the votes for each class.

Author(s)

[predict.ipredknn](#) by Torsten.Hothorn <Torsten.Hothorn@rzmail.uni-erlangen.de>

predict.knnreg

Predictions from k-Nearest Neighbors Regression Model

Description

Predict the outcome of a new observation based on k-NN.

Usage

```
## S3 method for class 'knnreg'  
predict(object, newdata, ...)
```

Arguments

object	object of class knnreg.
newdata	a data frame or matrix of new observations.
...	additional arguments.

Details

This function is a method for the generic function [predict](#) for class knnreg. For the details see [knnreg](#). This is essentially a copy of [predict.ipredknn](#).

Value

a numeric vector

Author(s)

Max Kuhn, Chris Keefer, adapted from [knn](#) and [predict.ipredknn](#)

predict.train *Extract predictions and class probabilities from train objects*

Description

These functions can be used for a single train object or to loop through a number of train objects to calculate the training and test data predictions and class probabilities.

Usage

```
## S3 method for class 'list'
predict(object, ...)

## S3 method for class 'train'
predict(object, newdata = NULL, type = "raw", ...)

extractPrediction(models,
                  testX = NULL, testY = NULL,
                  unkX = NULL,
                  unkOnly = !is.null(unkX) & is.null(testX),
                  verbose = FALSE)

extractProb(models,
            testX = NULL, testY = NULL,
            unkX = NULL,
            unkOnly = !is.null(unkX) & is.null(testX),
            verbose = FALSE)
```

Arguments

object	For predict.train, an object of class train . For predict.list, a list of objects of class train .
newdata	an optional set of data to predict on. If NULL, then the original training data are used
type	either "raw" or "prob", for the number/class predictions or class probabilities, respectively. Class probabilities are not available for all classification models
models	a list of objects of the class train. The objects must have been generated with fitBest = FALSE and returnData = TRUE.
testX	an optional set of data to predict
testY	an optional outcome corresponding to the data given in testX
unkX	another optional set of data to predict without known outcomes

unkOnly	a logical to bypass training and test set predictions. This is useful if speed is needed for unknown samples.
verbose	a logical for printing messages
...	additional arguments to be passed to other methods

Details

These functions are wrappers for the specific prediction functions in each modeling package. In each case, the optimal tuning values given in the `tuneValue` slot of the `finalModel` object are used to predict.

To get simple predictions for a new data set, the `predict` function can be used. Limits can be imposed on the range of predictions. See `trainControl` for more information.

To get predictions for a series of models at once, a list of `train` objects can be passes to the `predict` function and a list of model predictions will be returned.

The two extraction functions can be used to get the predictions and observed outcomes at once for the training, test and/or unknown samples at once in a single data frame (instead of a list of just the predictions). These objects can then be passes to `plotObsVsPred` or `plotClassProbs`.

Value

For `predict.train`, a vector of predictions if `type = "raw"` or a data frame of class probabilities for `type = "probs"`. In the latter case, there are columns for each class.

For `predict.list`, a list results. Each element is produced by `predict.train`.

For `extractPrediction`, a data frame with columns:

<code>obs</code>	the observed training and test data
<code>pred</code>	predicted values
<code>model</code>	the type of model used to predict
<code>object</code>	the names of the objects within models. If <code>models</code> is an un-named list, the values of <code>object</code> will be "Object1", "Object2" and so on
<code>dataType</code>	"Training", "Test" or "Unknown" depending on what was specified

For `extractProb`, a data frame. There is a column for each class containing the probabilities. The remaining columns are the same as above (although the `pred` column is the predicted class)

Author(s)

Max Kuhn

References

Kuhn (2008), "Building Predictive Models in R Using the caret" (<http://www.jstatsoft.org/v28/i05/>)

See Also

`plotObsVsPred`, `plotClassProbs`, `trainControl`

Examples

```
## Not run:
library(mlbench)
data(Satellite)
numSamples <- dim(Satellite)[1]
set.seed(716)

varIndex <- 1:numSamples

trainSamples <- sample(varIndex, 150)

varIndex <- (1:numSamples)[-trainSamples]
testSamples <- sample(varIndex, 100)

varIndex <- (1:numSamples)[-c(testSamples, trainSamples)]
unkSamples <- sample(varIndex, 50)

trainX <- Satellite[trainSamples, -37]
trainY <- Satellite[trainSamples, 37]

testX <- Satellite[testSamples, -37]
testY <- Satellite[testSamples, 37]

unkX <- Satellite[unkSamples, -37]

knnFit <- train(trainX, trainY, "knn")
rpartFit <- train(trainX, trainY, "rpart")

predict(knnFit)
predict(knnFit, newdata = testX)
predict(knnFit, type = "prob")

bothModels <- list(
  knn = knnFit,
  tree = rpartFit)

predict(bothModels)

predTargets <- extractPrediction(
  bothModels,
  testX = testX,
  testY = testY,
  unkX = unkX)

## End(Not run)
```

Description

This class uses a model fit to determine which predictors were used in the final model.

Usage

```
predictors(x, ...)  
  
## S3 method for class 'bagEarth'  
predictors(x, ...)  
  
## S3 method for class 'bagFDA'  
predictors(x, ...)  
  
## S3 method for class 'BinaryTree'  
predictors(x, surrogate = TRUE, ...)  
  
## S3 method for class 'blackboost'  
predictors(x, ...)  
  
## S3 method for class 'classbagg'  
predictors(x, surrogate = TRUE, ...)  
  
## S3 method for class 'dsa'  
predictors(x, cuts = NULL, ...)  
  
## S3 method for class 'earth'  
predictors(x, ...)  
  
## S3 method for class 'fda'  
predictors(x, ...)  
  
## S3 method for class 'foba'  
predictors(x, k = NULL, ...)  
  
## S3 method for class 'formula'  
predictors(x, ...)  
  
## S3 method for class 'gam'  
predictors(x, ...)  
  
## S3 method for class 'gamboost'  
predictors(x, ...)  
  
## S3 method for class 'gausspr'  
predictors(x, ...)  
  
## S3 method for class 'gbm'  
predictors(x, ...)
```

```
## S3 method for class 'glm'  
predictors(x, ...)  
  
## S3 method for class 'glmboost'  
predictors(x, ...)  
  
## S3 method for class 'glmnet'  
predictors(x, lambda = NULL, ...)  
  
## S3 method for class 'gpls'  
predictors(x, ...)  
  
## S3 method for class 'knn3'  
predictors(x, ...)  
  
## S3 method for class 'knnreg'  
predictors(x, ...)  
  
## S3 method for class 'ksvm'  
predictors(x, ...)  
  
## S3 method for class 'lars'  
predictors(x, s = NULL, ...)  
  
## S3 method for class 'lda'  
predictors(x, ...)  
  
## S3 method for class 'list'  
predictors(x, ...)  
  
## S3 method for class 'lm'  
predictors(x, ...)  
  
## S3 method for class 'logforest'  
predictors(x, ...)  
  
## S3 method for class 'logicBagg'  
predictors(x, ...)  
  
## S3 method for class 'LogitBoost'  
predictors(x, ...)  
  
## S3 method for class 'logreg'  
predictors(x, ...)  
  
## S3 method for class 'lssvm'  
predictors(x, ...)
```

```
## S3 method for class 'mda'  
predictors(x, ...)  
  
## S3 method for class 'multinom'  
predictors(x, ...)  
  
## S3 method for class 'mvr'  
predictors(x, ...)  
  
## S3 method for class 'NaiveBayes'  
predictors(x, ...)  
  
## S3 method for class 'nnet'  
predictors(x, ...)  
  
## S3 method for class 'pamrtrained'  
predictors(x, newdata = NULL, threshold = NULL, ...)  
  
## S3 method for class 'pcaNNet'  
predictors(x, ...)  
  
## S3 method for class 'penfit'  
predictors(x, ...)  
  
## S3 method for class 'ppr'  
predictors(x, ...)  
  
## S3 method for class 'qda'  
predictors(x, ...)  
  
## S3 method for class 'randomForest'  
predictors(x, ...)  
  
## S3 method for class 'RandomForest'  
predictors(x, ...)  
  
## S3 method for class 'rda'  
predictors(x, ...)  
  
## S3 method for class 'regbagg'  
predictors(x, surrogate = TRUE, ...)  
  
## S3 method for class 'rfe'  
predictors(x, ...)  
  
## S3 method for class 'rpart'  
predictors(x, surrogate = TRUE, ...)
```

```

## S3 method for class 'rvm'
predictors(x, ...)

## S3 method for class 'sbf'
predictors(x, ...)

## S3 method for class 'sda'
predictors(x, ...)

## S3 method for class 'slda'
predictors(x, ...)

## S3 method for class 'smda'
predictors(x, ...)

## S3 method for class 'spl'
predictors(x, ...)

## S3 method for class 'splda'
predictors(x, ...)

## S3 method for class 'superpc'
predictors(x, newdata = NULL, threshold = NULL, n.components = NULL, ...)

## S3 method for class 'terms'
predictors(x, ...)

## S3 method for class 'train'
predictors(x, ...)

## S3 method for class 'trocc'
predictors(x, ...)

## S3 method for class 'Weka_classifier'
predictors(x, ...)

```

Arguments

x	a model object, list or terms
newdata	for pamr.train and superpc.train : the training data
threshold	for pamr.train and superpc.train : the feature selection threshold
n.components	for superpc.train : the number of PCA components used
surrogate	a logical for rpart , ipredbag , ctree and cforest : should variables used as surrogate splits also be returned?
lambda	for glmnet : the L1 regularization value
s	for lars : the path index. See predict.lars

k	for <code>foba</code> : the sparsity level (i.e. the number of selected terms for the model). See <code>predict.foba</code>
cuts	the number of rule sets to use in the model (for <code>partDSA</code> only)
...	not currently used

Details

For `randomForest`, `cforest`, `ctree`, `rpart`, `ipredbag`, `bagging`, `earth`, `fda`, `pamr.train`, `superpc.train`, `bagEarth` and `bagFDA`, an attempt was made to report the predictors that were actually used in the final model.

In cases where the predictors cannot be determined, NA is returned. For example, `nnet` may return missing values from predictors.

Value

a character string of predictors or NA.

preProcess	<i>Pre-Processing of Predictors</i>
------------	-------------------------------------

Description

Pre-processing transformation (centering, scaling etc) can be estimated from the training data and applied to any data set with the same variables.

Usage

```
preProcess(x, ...)

## Default S3 method:
preProcess(x,
  method = c("center", "scale"),
  thresh = 0.95,
  pcaComp = NULL,
  na.remove = TRUE,
  k = 5,
  knnSummary = mean,
  outcome = NULL,
  fudge = .2,
  numUnique = 3,
  verbose = FALSE,
  ...)

## S3 method for class 'preProcess'
predict(object, newdata, ...)
```

Arguments

x	a matrix or data frame. All variables must be numeric.
method	a character vector specifying the type of processing. Possible values are "Box-Cox", "center", "scale", "range", "knnImpute", "bagImpute", "pca", "ica" and "spatialSign" (see Details below)
thresh	a cutoff for the cumulative percent of variance to be retained by PCA
pcaComp	the specific number of PCA components to keep. If specified, this over-rides thresh
na.remove	a logical; should missing values be removed from the calculations?
object	an object of class preProcess
newdata	a matrix or data frame of new data to be pre-processed
k	the number of nearest neighbors from the training set to use for imputation
knnSummary	function to average the neighbor values per column during imputation
outcome	a numeric or factor vector for the training set outcomes. This can be used to help estimate the Box-Cox transformation of the predictor variables (see Details below)
fudge	a tolerance value: Box-Cox transformation lambda values within +/-fudge will be coerced to 0 and within 1+/-fudge will be coerced to 1
numUnique	how many unique values should y have to estimate the Box-Cox transformation?
verbose	a logical: prints a log as the computations proceed
...	additional arguments to pass to fastICA , such as n.comp

Details

The Box-Cox transformation has been "repurposed" here: it is being used to transform the predictor variables. This method was developed for transforming the response variable while another method, the Box-Tidwell transformation, was created to estimate transformations of predictor data. However, the Box-Cox method is simpler, more computationally efficient and is equally effective for estimating power transformations.

The "range" transformation scales the data to be within [0, 1]. If new samples have values larger or smaller than those in the training set, values will be outside of this range.

The operations are applied in this order: Box-Cox transformation, centering, scaling, range, imputation, PCA, ICA then spatial sign. This is a departure from versions of **caret** prior to version 4.76 (where imputation was done first) and is not backwards compatible if bagging was used for imputation.

If PCA is requested but centering and scaling are not, the values will still be centered and scaled. Similarly, when ICA is requested, the data are automatically centered and scaled.

k-nearest neighbor imputation is carried out by finding the k closest samples (Euclidian distance) in the training set. Imputation via bagging fits a bagged tree model for each predictor (as a function of all the others). This method is simple, accurate and accepts missing values, but it has much higher computational cost.

A warning is thrown if both PCA and ICA are requested. ICA, as implemented by the [fastICA](#) package automatically does a PCA decomposition prior to finding the ICA scores.

The function will throw an error if any variables in x has less than two unique values.

Value

preProcess results in a list with elements

call	the function call
dim	the dimensions of x
bc	Box-Cox transformation values, see BoxCoxTrans
mean	a vector of means (if centering was requested)
std	a vector of standard deviations (if scaling or PCA was requested)
rotation	a matrix of eigenvectors if PCA was requested
method	the value of method
thresh	the value of thresh
ranges	a matrix of min and max values for each predictor when method includes "range" (and NULL otherwise)
numComp	the number of principal components required of capture the specified amount of variance
ica	contains values for the W and K matrix of the decomposition

Author(s)

Max Kuhn

References

- Kuhn (2008), "Building Predictive Models in R Using the caret" (<http://www.jstatsoft.org/v28/i05/>)
- Box, G. E. P. and Cox, D. R. (1964) An analysis of transformations (with discussion). *Journal of the Royal Statistical Society B*, 26, 211-252.
- Box, G. E. P. and Tidwell, P. W. (1962) Transformation of the independent variables. *Technometrics* 4, 531-550.

See Also

[BoxCoxTrans](#), [boxcox](#), [prcomp](#), [fastICA](#), [spatialSign](#)

Examples

```
data(BloodBrain)
# one variable has one unique value
## Not run: preProc <- preProcess(bbbDescr[1:100,])

preProc <- preProcess(bbbDescr[1:100,-3])
training <- predict(preProc, bbbDescr[1:100,-3])
test <- predict(preProc, bbbDescr[101:208,-3])
```

print.confusionMatrix *Print method for confusionMatrix*

Description

a print method for confusionMatrix

Usage

```
## S3 method for class 'confusionMatrix'  
print(x, digits = max(3, getOption("digits") - 3),  
      printStats = TRUE, ...)
```

Arguments

x	an object of class confusionMatrix
digits	number of significant digits when printed
printStats	a logical: if TRUE then table statistics are also printed
...	optional arguments to pass to print.table

Value

x is invisibly returned

Author(s)

Max Kuhn

See Also

[confusionMatrix](#)

print.train *Print Method for the train Class*

Description

Print the results of a [train](#) object.

Usage

```
## S3 method for class 'train'
print(x,
      digits = min(3, getOption("digits") - 3),
      printCall = FALSE,
      details = FALSE,
      selectCol = FALSE,
      ...)
```

Arguments

x	an object of class train .
digits	an integer specifying the number of significant digits to print.
printCall	a logical to print the call at the top of the output
details	a logical to show print or summary methods for the final model. In some cases (such as <code>gbm</code> , <code>knn</code> , <code>lvq</code> , naive Bayes and bagged tree models), no information will be printed even if <code>details = TRUE</code>
selectCol	a logical to a column with a star next to the final model
...	options passed to the generic print method

Details

The table of complexity parameters used, their resampled performance and a flag for which rows are optimal.

Value

A data frame with the complexity parameter(s) and performance (invisibly).

Author(s)

Max Kuhn

See Also

[train](#)

Examples

```
data(iris)
TrainData <- iris[,1:4]
TrainClasses <- iris[,5]

library(klaR)
rdaFit <- train(TrainData, TrainClasses, "rda",
               control = trainControl(method = "cv"))
print(rdaFit)
```

resampleHist	<i>Plot the resampling distribution of the model statistics</i>
--------------	---

Description

Create a lattice histogram or densityplot from the resampled outcomes from a `train` object.

Usage

```
resampleHist(object, type = "density", ...)
```

Arguments

<code>object</code>	an object resulting from a call to train
<code>type</code>	a character string. Either "hist" or "density"
<code>...</code>	options to pass to histogram or densityplot

Details

All the metrics from the object are plotted, but only for the final model. For more comprehensive plots functions, see [histogram.train](#), [densityplot.train](#), [xyplot.train](#), [stripplot.train](#).

For the the plot to be made, the `returnResamp` argument in [trainControl](#) should be either "final" or "all".

Value

a object of class `trellis`

Author(s)

Max Kuhn

See Also

[train](#), [histogram](#), [densityplot](#), [histogram.train](#), [densityplot.train](#), [xyplot.train](#), [stripplot.train](#)

Examples

```
data(iris)
TrainData <- iris[,1:4]
TrainClasses <- iris[,5]

knnFit <- train(TrainData, TrainClasses, "knn")

resampleHist(knnFit)
```

Description

These functions provide methods for collection, analyzing and visualizing a set of resampling results from a common data set.

Usage

```
resamples(x, ...)

## Default S3 method:
resamples(x, modelNames = names(x), ...)

## S3 method for class 'resamples'
summary(object, ...)
```

Arguments

x	a list of two or more objects of class <code>train</code> , <code>sbf</code> or <code>rfe</code> with a common set of resampling indices in the control object.
modelNames	an optional set of names to give to the resampling results
object	an object generated by <code>resamples</code>
...	not currently used

Details

The ideas and methods here are based on Hothorn et al (2005) and Eugster et al (2008).

The results from `train` can have more than one performance metric per resample. Each metric in the input object is saved.

`resamples` checks that the resampling results match; that is, the indices in the object `trainObject$control$index` are the same. Also, the argument `trainControl` `returnResamp` should have a value of "final" for each model.

The summary function computes summary statistics across each model/metric combination.

Value

An object with class "resamples" with elements

call	the call
values	a data frame of results where rows correspond to resampled data sets and columns indicate the model and metric
models	a character string of model labels
metrics	a character string of performance metrics
methods	a character string of the <code>train</code> method argument values for each model

Author(s)

Max Kuhn

References

Hothorn et al. The design and analysis of benchmark experiments. *Journal of Computational and Graphical Statistics* (2005) vol. 14 (3) pp. 675-699

Eugster et al. Exploratory and inferential analysis of benchmark experiments. *Ludwigs-Maximilians-Universitat Munchen, Department of Statistics, Tech. Rep* (2008) vol. 30

See Also

[train](#), [trainControl](#), [diff.resamples](#), [xyplot.resamples](#), [densityplot.resamples](#), [bwplot.resamples](#), [splom.resamples](#)

Examples

```
## Not run:
data(BloodBrain)
set.seed(1)
tmp <- createDataPartition(logBBB,
                           p = .8,
                           times = 100)

rpartFit <- train(bbbDescr, logBBB,
                 "rpart",
                 tuneLength = 16,
                 trControl = trainControl(
                   method = "LGOCV", index = tmp))

ctreeFit <- train(bbbDescr, logBBB,
                 "ctree",
                 trControl = trainControl(
                   method = "LGOCV", index = tmp))

earthFit <- train(bbbDescr, logBBB,
                 "earth",
                 tuneLength = 20,
                 trControl = trainControl(
                   method = "LGOCV", index = tmp))

## or load pre-calculated results using:
## load(url("http://caret.r-forge.r-project.org/Classification_and_Regression_Training_files/exampleModels.RData"))

resamps <- resamples(list(CART = rpartFit,
                         CondInfTree = ctreeFit,
                         MARS = earthFit))

resamps
```

```
summary(resamps)
```

```
## End(Not run)
```

```
resampleSummary      Summary of resampled performance estimates
```

Description

This function uses the out-of-bag predictions to calculate overall performance metrics and returns the observed and predicted data.

Usage

```
resampleSummary(obs, resampled, index = NULL, keepData = TRUE)
```

Arguments

obs	A vector (numeric or factor) of the outcome data
resampled	For bootstrapping, this is either a matrix (for numeric outcomes) or a data frame (for factors). For cross-validation, a vector is produced.
index	The list to index of samples in each cross-validation fold (only used for cross-validation).
keepData	A logical for returning the observed and predicted data.

Details

The mean and standard deviation of the values produced by [postResample](#) are calculated.

Value

A list with:

metrics	A vector of values describing the bootstrap distribution.
data	A data frame or NULL. Columns include obs, pred and group (for tracking cross-validation folds or bootstrap samples)

Author(s)

Max Kuhn

See Also

[postResample](#)

Examples

```
resampleSummary(rnorm(10), matrix(rnorm(50), ncol = 5))
```

Description

A simple backwards selection, a.k.a. recursive feature selection (RFE), algorithm

Usage

```
rfe(x, ...)

## Default S3 method:
rfe(x, y,
    sizes = 2^(2:4),
    metric = ifelse(is.factor(y), "Accuracy", "RMSE"),
    maximize = ifelse(metric == "RMSE", FALSE, TRUE),
    rfeControl = rfeControl(),
    ...)

rfeIter(x, y,
        testX, testY,
        sizes,
        rfeControl = rfeControl(),
        ...)

## S3 method for class 'rfe'
predict(object, newdata, ...)
```

Arguments

<code>x</code>	a matrix or data frame of predictors for model training. This object must have unique column names.
<code>y</code>	a vector of training set outcomes (either numeric or factor)
<code>testX</code>	a matrix or data frame of test set predictors. This must have the same column names as <code>x</code>
<code>testY</code>	a vector of test set outcomes
<code>sizes</code>	a numeric vector of integers corresponding to the number of features that should be retained
<code>metric</code>	a string that specifies what summary metric will be used to select the optimal model. By default, possible values are "RMSE" and "Rsquared" for regression and "Accuracy" and "Kappa" for classification. If custom performance metrics are used (via the <code>functions</code> argument in <code>rfeControl</code> , the value of <code>metric</code> should match one of the arguments.
<code>maximize</code>	a logical: should the metric be maximized or minimized?

<code>rfeControl</code>	a list of options, including functions for fitting and prediction. See the package vignette or rfeControl for examples
<code>object</code>	an object of class <code>rfe</code>
<code>newdata</code>	a matrix or data frame of new samples for prediction
<code>...</code>	options to pass to the model fitting function (ignored in <code>predict.rfe</code>)

Details

This function implements backwards selection of predictors based on predictor importance ranking. The predictors are ranked and the less important ones are sequentially eliminated prior to modeling. The goal is to find a subset of predictors that can be used to produce an accurate model. The package vignette for feature selection has detailed descriptions of the algorithms.

`rfe` can be used with "explicit parallelism", where different resamples (e.g. cross-validation group) can be split up and run on multiple machines or processors. By default, `rfe` will use a single processor on the host machine. As of version 4.99 of this package, the framework used for parallel processing uses the **foreach** package. To run the resamples in parallel, the code for `rfe` does not change; prior to the call to `rfe`, a parallel backend is registered with **foreach** (see the examples below).

`rfeIter` is the basic algorithm while `rfe` wraps these operations inside of resampling. To avoid selection bias, it is better to use the function `rfe` than `rfeIter`.

Value

A list with elements

<code>finalVariables</code>	a list of size <code>length(sizes) + 1</code> containing the column names of the "surviving" predictors at each stage of selection. The first element corresponds to all the predictors (i.e. <code>size = ncol(x)</code>)
<code>pred</code>	a data frame with columns for the test set outcome, the predicted outcome and the subset size.

Author(s)

Max Kuhn

See Also

[rfeControl](#)

Examples

```
## Not run:
data(BloodBrain)

x <- scale(bbbDescr[, -nearZeroVar(bbbDescr)])
x <- x[, -findCorrelation(cor(x), .8)]
x <- as.data.frame(x)

set.seed(1)
```

```

lmProfile <- rfe(x, logBBB,
               sizes = c(2:25, 30, 35, 40, 45, 50, 55, 60, 65),
               rfeControl = rfeControl(functions = lmFuncs,
                                       number = 200))

set.seed(1)
lmProfile2 <- rfe(x, logBBB,
                 sizes = c(2:25, 30, 35, 40, 45, 50, 55, 60, 65),
                 rfeControl = rfeControl(functions = lmFuncs,
                                         rerank = TRUE,
                                         number = 200))

xyplot(lmProfile$results$RMSE + lmProfile2$results$RMSE ~
      lmProfile$results$Variables,
      type = c("g", "p", "l"),
      auto.key = TRUE)

rfProfile <- rfe(x, logBBB,
               sizes = c(2, 5, 10, 20),
               rfeControl = rfeControl(functions = rfFuncs))

bagProfile <- rfe(x, logBBB,
                 sizes = c(2, 5, 10, 20),
                 rfeControl = rfeControl(functions = treebagFuncs))

set.seed(1)
svmProfile <- rfe(x, logBBB,
                 sizes = c(2, 5, 10, 20),
                 rfeControl = rfeControl(functions = caretFuncs,
                                         number = 200),
                 ## pass options to train()
                 method = "svmRadial")

## classification with no resampling

data(mdr)
mdrrDescr <- mdrDescr[, -nearZeroVar(mdrDescr)]
mdrrDescr <- mdrDescr[, -findCorrelation(cor(mdrDescr), .8)]

set.seed(1)
inTrain <- createDataPartition(mdrClass, p = .75, list = FALSE)[,1]

train <- mdrDescr[ inTrain, ]
test <- mdrDescr[-inTrain, ]
trainClass <- mdrClass[ inTrain]
testClass <- mdrClass[-inTrain]

set.seed(2)
ldaProfile <- rfe(train, trainClass,
                 sizes = c(1:10, 15, 30),
                 rfeControl = rfeControl(functions = ldaFuncs, method = "cv"))
plot(ldaProfile, type = c("o", "g"))

postResample(predict(ldaProfile, test), testClass)

```

```
## End(Not run)

#####
## Parallel Processing Example via multicore

## Not run:
library(doMC)

## Note: if the underlying model also uses foreach, the
## number of cores specified above will double (along with
## the memory requirements)
registerDoMC(cores = 2)

set.seed(1)
lmProfile <- rfe(x, logBBB,
                 sizes = c(2:25, 30, 35, 40, 45, 50, 55, 60, 65),
                 rfeControl = rfeControl(functions = lmFuncs,
                                         number = 200))

## End(Not run)
```

rfeControl

Controlling the Feature Selection Algorithms

Description

This function generates a control object that can be used to specify the details of the feature selection algorithms used in this package.

Usage

```
rfeControl(functions = NULL,
           rerank = FALSE,
           method = "boot",
           saveDetails = FALSE,
           number = ifelse(method %in% c("cv", "repeatedcv"), 10, 25),
           repeats = ifelse(method %in% c("cv", "repeatedcv"), 1, number),
           verbose = FALSE,
           returnResamp = "all",
           p = .75,
           index = NULL,
           timingSamps = 0)
```

Arguments

functions	a list of functions for model fitting, prediction and variable importance (see Details below)
rerank	a logical: should variable importance be re-calculated each time features are removed?
method	The external resampling method: boot, cv, LOOCV or LGOCV (for repeated training/test splits)
number	Either the number of folds or number of resampling iterations
repeats	For repeated k-fold cross-validation only: the number of complete sets of folds to compute
saveDetails	a logical to save the predictions and variable importances from the selection process
verbose	a logical to print a log for each external resampling iteration
returnResamp	A character string indicating how much of the resampled summary metrics should be saved. Values can be “final”, “all” or “none”
p	For leave-group out cross-validation: the training percentage
index	a list with elements for each external resampling iteration. Each list element is the sample rows used for training at that iteration.
timingSamps	the number of training set samples that will be used to measure the time for predicting samples (zero indicates that the prediction time should not be estimated).

Details

Backwards selection requires function to be specified for some operations.

The `fit` function builds the model based on the current data set. The arguments for the function must be:

- `x` the current training set of predictor data with the appropriate subset of variables
- `y` the current outcome data (either a numeric or factor vector)
- `first` a single logical value for whether the current predictor set has all possible variables
- `last` similar to `first`, but TRUE when the last model is fit with the final subset size and predictors.
- ... optional arguments to pass to the `fit` function in the call to `rfe`

The function should return a model object that can be used to generate predictions.

The `pred` function returns a vector of predictions (numeric or factors) from the current model. The arguments are:

- `object` the model generated by the `fit` function
- `x` the current set of predictor set for the held-back samples

The `rank` function is used to return the predictors in the order of the most important to the least important. Inputs are:

- `object` the model generated by the `fit` function

- `x` the current set of predictor set for the training samples
- `y` the current training outcomes

The function should return a data frame with a column called `var` that has the current variable names. The first row should be the most important predictor etc. Other columns can be included in the output and will be returned in the final `rfe` object.

The `selectSize` function determines the optimal number of predictors based on the resampling output. Inputs for the function are:

- `xa` matrix with columns for the performance metrics and the number of variables, called "Variables"
- `metrica` character string of the performance measure to optimize (e.g. "RMSE", "Rsquared", "Accuracy" or "Kappa")
- `maximizea` single logical for whether the metric should be maximized

This function should return an integer corresponding to the optimal subset size. `caret` comes with two examples functions for this purpose: `pickSizeBest` and `pickSizeTolerance`.

After the optimal subset size is determined, the `selectVar` function will be used to calculate the best rankings for each variable across all the resampling iterations. Inputs for the function are:

- `y` a list of variables importance for each resampling iteration and each subset size (generated by the user-defined rank function). In the example, each each of the cross-validation groups the output of the rank function is saved for each of the subset sizes (including the original subset). If the rankings are not recomputed at each iteration, the values will be the same within each cross-validation iteration.
- `size` the integer returned by the `selectSize` function

This function should return a character string of predictor names (of length `size`) in the order of most important to least important

Examples of these functions are included in the package: `lmFuncs`, `rffFuncs`, `treebagFuncs` and `nbFuncs`.

Model details about these functions, including examples, are in the package vignette for feature selection.

Value

A list

Author(s)

Max Kuhn

See Also

`rfe`, `lmFuncs`, `rffFuncs`, `treebagFuncs`, `nbFuncs`, `pickSizeBest`, `pickSizeTolerance`

sbf

Selection By Filtering (SBF)

Description

Model fitting after applying univariate filters

Usage

```

sbf(x, ...)

## Default S3 method:
sbf(x, y, sbfControl = sbfControl(), ...)

## S3 method for class 'formula'
sbf(form, data, ..., subset, na.action, contrasts = NULL)

## S3 method for class 'sbf'
predict(object, newdata = NULL, ...)

```

Arguments

x	a data frame containing training data where samples are in rows and features are in columns.
y	a numeric or factor vector containing the outcome for each sample.
form	A formula of the form $y \sim x_1 + x_2 + \dots$
data	Data frame from which variables specified in formula are preferentially to be taken.
subset	An index vector specifying the cases to be used in the training sample. (NOTE: If given, this argument must be named.)
na.action	A function to specify the action to be taken if NAs are found. The default action is for the procedure to fail. An alternative is na.omit, which leads to rejection of cases with missing values on any required variable. (NOTE: If given, this argument must be named.)
contrasts	a list of contrasts to be used for some or all of the factors appearing as variables in the model formula.
sbfControl	a list of values that define how this function acts. See sbfControl . (NOTE: If given, this argument must be named.)
object	an object of class sbf
newdata	a matrix or data frame of predictors. The object must have non-null column names
...	for sbf: arguments passed to the classification or regression routine (such as randomForest). For predict.sbf: arguments cannot be passed to the prediction function using predict.sbf as it uses the function originally specified for prediction.

Details

This function can be used to get resampling estimates for models when simple, filter-based feature selection is applied to the training data.

For each iteration of resampling, the predictor variables are univariately filtered prior to modeling. Performance of this approach is estimated using resampling. The same filter and model are then applied to the entire training set and the final model (and final features) are saved.

sbf can be used with "explicit parallelism", where different resamples (e.g. cross-validation group) can be split up and run on multiple machines or processors. By default, sbf will use a single processor on the host machine. As of version 4.99 of this package, the framework used for parallel processing uses the **foreach** package. To run the resamples in parallel, the code for sbf does not change; prior to the call to sbf, a parallel backend is registered with **foreach** (see the examples below).

The modeling and filtering techniques are specified in [sbfControl](#). Example functions are given in [lmSBF](#).

Value

for sbf, an object of class sbf with elements:

pred	if sbfControl\$saveDetails is TRUE, this is a list of predictions for the hold-out samples at each resampling iteration. Otherwise it is NULL
variables	a list of variable names that survived the filter at each resampling iteration
results	a data frame of results aggregated over the resamples
fit	the final model fit with only the filtered variables
optVariables	the names of the variables that survived the filter using the training set
call	the function call
control	the control object
resample	if sbfControl\$returnResamp is "all", a data frame of the resampled performance measures. Otherwise, NULL
metrics	a character vector of names of the performance measures
dots	a list of optional arguments that were passed in

For predict.sbf, a vector of predictions.

Author(s)

Max Kuhn

See Also

[sbfControl](#)

Examples

```

data(BloodBrain)

## Use a GAM as the filter, then fit a random forest model
RFwithGAM <- sbf(bbbDescr, logBBB,
                sbfControl(functions = rfSBF,
                           verbose = FALSE,
                           method = "cv"))

RFwithGAM

predict(RFwithGAM, bbbDescr[1:10,])

## classification example with parallel processing
## Not run:
library(doMC)

## Note: if the underlying model also uses foreach, the
## number of cores specified above will double (along with
## the memory requirements)
registerDoMC(cores = 2)

data(mdrd)
mdrdDescr <- mdrdDescr[,-nearZeroVar(mdrdDescr)]
mdrdDescr <- mdrdDescr[, -findCorrelation(cor(mdrdDescr), .8)]

set.seed(1)
filteredNB <- sbf(mdrdDescr, mdrdClass,
                 sbfControl(functions = nbSBF,
                           verbose = FALSE,
                           method = "repeatedcv",
                           repeats = 5))

confusionMatrix(filteredNB)

## End(Not run)

```

sbfControl

Control Object for Selection By Filtering (SBF)

Description

Controls the execution of models with simple filters for feature selection

Usage

```

sbfControl(functions = NULL,
           method = "boot",
           saveDetails = FALSE,

```

```

number = ifelse(method %in% c("cv", "repeatedcv"), 10, 25),
repeats = ifelse(method %in% c("cv", "repeatedcv"), 1, number),
verbose = FALSE,
returnResamp = "all",
p = 0.75,
index = NULL,
timingSamps = 0)

```

Arguments

functions	a list of functions for model fitting, prediction and variable filtering (see Details below)
method	The external resampling method: boot, cv, LOOCV or LGOVCV (for repeated training/test splits)
number	Either the number of folds or number of resampling iterations
repeats	For repeated k-fold cross-validation only: the number of complete sets of folds to compute
saveDetails	a logical to save the predictions and variable importances from the selection process
verbose	a logical to print a log for each external resampling iteration
returnResamp	A character string indicating how much of the resampled summary metrics should be saved. Values can be "all" or "none"
p	For leave-group out cross-validation: the training percentage
index	a list with elements for each external resampling iteration. Each list element is the sample rows used for training at that iteration.
timingSamps	the number of training set samples that will be used to measure the time for predicting samples (zero indicates that the prediction time should not be estimated).

Details

Simple filter-based feature selection requires function to be specified for some operations.

The `fit` function builds the model based on the current data set. The arguments for the function must be:

- `x` the current training set of predictor data with the appropriate subset of variables (i.e. after filtering)
- `y` the current outcome data (either a numeric or factor vector)
- ... optional arguments to pass to the fit function in the call to `sbf`

The function should return a model object that can be used to generate predictions.

The `pred` function returns a vector of predictions (numeric or factors) from the current model. The arguments are:

- `object` the model generated by the `fit` function
- `x` the current set of predictor set for the held-back samples

The `score` function is used to return a vector of scores with names for each predictor (such as a p-value). Inputs are:

- `x` the predictors for the training samples
- `y` the current training outcomes

The function should return a vector, as previously stated. Examples are give by [anovaScores](#) for classification and [gamScores](#) for regression.

The `filter` function is used to return a logical vector with names for each predictor (TRUE indicates that the predictor should be retained). Inputs are:

- `score` the output of the `score` function
- `x` the predictors for the training samples
- `y` the current training outcomes

The function should return a named logical vector.

Examples of these functions are included in the package: [caretSBF](#), [lmSBF](#), [rfSBF](#), [treebagSBF](#), [ldaSBF](#) and [nbSBF](#).

Model details about these functions, including examples, are in the package vignette for feature selection.

Value

a list that echos the specified arguments

Author(s)

Max Kuhn

See Also

[sbf](#), [caretSBF](#), [lmSBF](#), [rfSBF](#), [treebagSBF](#), [ldaSBF](#) and [nbSBF](#)

segmentationData

Cell Body Segmentation

Description

Hill, LaPan, Li and Haney (2007) develop models to predict which cells in a high content screen were well segmented. The data consists of 119 imaging measurements on 2019. The original analysis used 1009 for training and 1010 as a test set (see the column called Case).

The outcome class is contained in a factor variable called `Class` with levels "PS" for poorly segmented and "WS" for well segmented.

The raw data used in the paper can be found at the Biomedcentral website. Versions of `caret` < 4.98 contained the original data. The version now contained in `segmentationData` is modified. First, several discrete versions of some of the predictors (with the suffix "Status") were removed. Second,

there are several skewed predictors with minimum values of zero (that would benefit from some transformation, such as the log). A constant value of 1 was added to these fields: AvgIntenCh2, FiberAlign2Ch3, FiberAlign2Ch4, SpotFiberCountCh4 and TotalIntenCh2.

A binary version of the original data is at http://caret.r-forge.r-project.org/Classification_and_Regression_Training_files/segmentationOriginal.RData.

Usage

```
data(segmentationData)
```

Value

```
segmentationData
  data frame of cells
```

Source

Hill, LaPan, Li and Haney (2007). Impact of image segmentation on high-content screening data quality for SK-BR-3 cells, *BMC Bioinformatics*, Vol. 8, pg. 340, <http://www.biomedcentral.com/1471-2105/8/340>.

sensitivity

Calculate sensitivity, specificity and predictive values

Description

These functions calculate the sensitivity, specificity or predictive values of a measurement system compared to a reference results (the truth or a gold standard). The measurement and "truth" data must have the same two possible outcomes and one of the outcomes must be thought of as a "positive" results.

The sensitivity is defined as the proportion of positive results out of the number of samples which were actually positive. When there are no positive results, sensitivity is not defined and a value of NA is returned. Similarly, when there are no negative results, specificity is not defined and a value of NA is returned. Similar statements are true for predictive values.

The positive predictive value is defined as the percent of predicted positives that are actually positive while the negative predictive value is defined as the percent of negative positives that are actually negative.

Usage

```
sensitivity(data, ...)
## Default S3 method:
sensitivity(data, reference, positive = levels(reference)[1], ...)
## S3 method for class 'table'
sensitivity(data, positive = rownames(data)[1], ...)
## S3 method for class 'matrix'
sensitivity(data, positive = rownames(data)[1], ...)
```

```

specificity(data, ...)
## Default S3 method:
specificity(data, reference, negative = levels(reference)[-1], ...)
## S3 method for class 'table'
specificity(data, negative = rownames(data)[-1], ...)
## S3 method for class 'matrix'
specificity(data, negative = rownames(data)[-1], ...)

posPredValue(data, ...)
## Default S3 method:
posPredValue(data, reference, positive = levels(reference)[1],
              prevalence = NULL, ...)
## S3 method for class 'table'
posPredValue(data, positive = rownames(data)[1], prevalence = NULL, ...)
## S3 method for class 'matrix'
posPredValue(data, positive = rownames(data)[1], prevalence = NULL, ...)

negPredValue(data, ...)
## Default S3 method:
negPredValue(data, reference, negative = levels(reference)[2],
             prevalence = NULL, ...)
## S3 method for class 'table'
negPredValue(data, negative = rownames(data)[-1], prevalence = NULL, ...)
## S3 method for class 'matrix'
negPredValue(data, negative = rownames(data)[-1], prevalence = NULL, ...)

```

Arguments

<code>data</code>	for the default functions, a factor containing the discrete measurements. For the table or matrix functions, a table or matrix object, respectively.
<code>reference</code>	a factor containing the reference values
<code>positive</code>	a character string that defines the factor level corresponding to the "positive" results
<code>negative</code>	a character string that defines the factor level corresponding to the "negative" results
<code>prevalence</code>	a numeric value for the rate of the "positive" class of the data
<code>...</code>	not currently used

Details

Suppose a 2x2 table with notation

	Reference	
Predicted	Event	No Event
Event	A	B
No Event	C	D

The formulas used here are:

$$\text{Sensitivity} = A/(A + C)$$

$$\text{Specificity} = D/(B + D)$$

$$\text{Prevalence} = (A + C)/(A + B + C + D)$$

$$\text{PPV} = (\text{sensitivity} * \text{Prevalence}) / ((\text{sensitivity} * \text{Prevalence}) + ((1 - \text{specificity}) * (1 - \text{Prevalence})))$$

$$\text{NPV} = (\text{specificity} * (1 - \text{Prevalence})) / (((1 - \text{sensitivity}) * \text{Prevalence}) + (\text{specificity} * (1 - \text{Prevalence})))$$

See the references for discussions of the statistics.

Value

A number between 0 and 1 (or NA).

Author(s)

Max Kuhn

References

Kuhn, M. (2008), "Building predictive models in R using the caret package," *Journal of Statistical Software*, (<http://www.jstatsoft.org/v28/i05/>).

Altman, D.G., Bland, J.M. (1994) "Diagnostic tests 1: sensitivity and specificity," *British Medical Journal*, vol 308, 1552.

Altman, D.G., Bland, J.M. (1994) "Diagnostic tests 2: predictive values," *British Medical Journal*, vol 309, 102.

See Also

[confusionMatrix](#)

Examples

```
#####
## 2 class example

lvs <- c("normal", "abnormal")
truth <- factor(rep(lvs, times = c(86, 258)),
               levels = rev(lvs))
pred <- factor(
  c(
    rep(lvs, times = c(54, 32)),
    rep(lvs, times = c(27, 231))),
  levels = rev(lvs))

xtab <- table(pred, truth)

sensitivity(pred, truth)
sensitivity(xtab)
posPredValue(pred, truth)
```

```

posPredValue(pred, truth, prevalence = 0.25)

specificity(pred, truth)
negPredValue(pred, truth)
negPredValue(xtab)
negPredValue(pred, truth, prevalence = 0.25)

prev <- seq(0.001, .99, length = 20)
npvVals <- ppvVals <- prev * NA
for(i in seq(along = prev))
{
  ppvVals[i] <- posPredValue(pred, truth, prevalence = prev[i])
  npvVals[i] <- negPredValue(pred, truth, prevalence = prev[i])
}

plot(prev, ppvVals,
      ylim = c(0, 1),
      type = "l",
      ylab = "",
      xlab = "Prevalence (i.e. prior)")
points(prev, npvVals, type = "l", col = "red")
abline(h=sensitivity(pred, truth), lty = 2)
abline(h=specificity(pred, truth), lty = 2, col = "red")
legend(.5, .5,
       c("ppv", "npv", "sens", "spec"),
       col = c("black", "red", "black", "red"),
       lty = c(1, 1, 2, 2))

#####
## 3 class example

library(MASS)

fit <- lda(Species ~ ., data = iris)
model <- predict(fit)$class

irisTabs <- table(model, iris$Species)

## When passing factors, an error occurs with more
## than two levels
## Not run: sensitivity(model, iris$Species)

## When passing a table, more than two levels can
## be used
sensitivity(irisTabs, "versicolor")
specificity(irisTabs, c("setosa", "virginica"))

```

Description

Compute the spatial sign (a projection of a data vector to a unit length circle). The spatial sign of a vector w is $w / \text{norm}(w)$.

Usage

```
## Default S3 method:
spatialSign(x)
## S3 method for class 'matrix'
spatialSign(x)
## S3 method for class 'data.frame'
spatialSign(x)
```

Arguments

x an object full of numeric data (which should probably be scaled). Factors are not allowed. This could be a vector, matrix or data frame.

Value

A vector, matrix or data frame with the same dim names of the original data.

Author(s)

Max Kuhn

References

Serneels et al. Spatial sign preprocessing: a simple way to impart moderate robustness to multivariate estimators. *J. Chem. Inf. Model* (2006) vol. 46 (3) pp. 1402-1409

Examples

```
spatialSign(rnorm(5))

spatialSign(matrix(rnorm(12), ncol = 3))

# should fail since the fifth column is a factor
try(spatialSign(iris), silent = TRUE)

spatialSign(iris[,-5])

trellis.par.set(caretTheme())
featurePlot(iris[,-5], iris[,5], "pairs")
featurePlot(spatialSign(scale(iris[,-5])), iris[,5], "pairs")
```

summary.bagEarth	<i>Summarize a bagged earth or FDA fit</i>
------------------	--

Description

The function shows a summary of the results from a bagged earth model

Usage

```
## S3 method for class 'bagEarth'  
summary(object, ...)  
## S3 method for class 'bagFDA'  
summary(object, ...)
```

Arguments

object	an object of class "bagEarth" or "bagFDA"
...	optional arguments (not used)

Details

The out-of-bag statistics are summarized, as well as the distribution of the number of model terms and number of variables used across all of the bootstrap samples.

Value

a list with elements	
modelInfo	a matrix with the number of model terms and variables used
oobStat	a summary of the out-of-bag statistics
bmarsCall	the original call to bagEarth

Author(s)

Max Kuhn

Examples

```
data(trees)  
fit <- bagEarth(trees[,-3], trees[3])  
summary(fit)
```

tecator

Fat, Water and Protein Content of Meat Samples

Description

"These data are recorded on a Tecator Infratec Food and Feed Analyzer working in the wavelength range 850 - 1050 nm by the Near Infrared Transmission (NIT) principle. Each sample contains finely chopped pure meat with different moisture, fat and protein contents.

If results from these data are used in a publication we want you to mention the instrument and company name (Tecator) in the publication. In addition, please send a preprint of your article to Karin Thente, Tecator AB, Box 70, S-263 21 Hoganas, Sweden

The data are available in the public domain with no responsibility from the original data source. The data can be redistributed as long as this permission note is attached."

"For each meat sample the data consists of a 100 channel spectrum of absorbances and the contents of moisture (water), fat and protein. The absorbance is $-\log_{10}$ of the transmittance measured by the spectrometer. The three contents, measured in percent, are determined by analytic chemistry."

Included here are the training, monitoring and test sets.

Usage

```
data(tecator)
```

Value

absorp	absorbance data for 215 samples. The first 129 were originally used as a training set
endpoints	the percentages of water, fat and protein

Examples

```
data(tecator)

splom(~endpoints)

# plot 10 random spectra
set.seed(1)
inSubset <- sample(1:dim(endpoints)[1], 10)

absorpSubset <- absorp[inSubset,]
endpointSubset <- endpoints[inSubset, 3]

newOrder <- order(absorpSubset[,1])
absorpSubset <- absorpSubset[newOrder,]
endpointSubset <- endpointSubset[newOrder]

plotColors <- rainbow(10)
```

```

plot(absorpSubset[1,],
     type = "n",
     ylim = range(absorpSubset),
     xlim = c(0, 105),
     xlab = "Wavelength Index",
     ylab = "Absorption")

for(i in 1:10)
{
  points(absorpSubset[i,], type = "l", col = plotColors[i], lwd = 2)
  text(105, absorpSubset[i,100], endpointSubset[i], col = plotColors[i])
}
title("Predictor Profiles for 10 Random Samples")

```

train

Fit Predictive Models over Different Tuning Parameters

Description

This function sets up a grid of tuning parameters for a number of classification and regression routines, fits each model and calculates a resampling based performance measure.

Usage

```

train(x, ...)

## Default S3 method:
train(x, y,
      method = "rf",
      preProcess = NULL,
      ...,
      weights = NULL,
      metric = ifelse(is.factor(y), "Accuracy", "RMSE"),
      maximize = ifelse(metric == "RMSE", FALSE, TRUE),
      trControl = trainControl(),
      tuneGrid = NULL,
      tuneLength = 3)

## S3 method for class 'formula'
train(form, data, ..., weights, subset, na.action, contrasts = NULL)

```

Arguments

x	a data frame containing training data where samples are in rows and features are in columns.
y	a numeric or factor vector containing the outcome for each sample.
form	A formula of the form $y \sim x_1 + x_2 + \dots$

data	Data frame from which variables specified in formula are preferentially to be taken.
weights	a numeric vector of case weights. This argument will only affect models that allow case weights.
subset	An index vector specifying the cases to be used in the training sample. (NOTE: If given, this argument must be named.)
na.action	A function to specify the action to be taken if NAs are found. The default action is for the procedure to fail. An alternative is na.omit, which leads to rejection of cases with missing values on any required variable. (NOTE: If given, this argument must be named.)
contrasts	a list of contrasts to be used for some or all of the factors appearing as variables in the model formula.
method	a string specifying which classification or regression model to use. Possible values are: ada, avNNet, bag, bagEarth, bagFDA, bdk, blackboost, Boruta, bstTree, bstLs, bstSm, cforest, ctree, ctree2, cubist, earth, enet, fda, foba, gam, gamboost, gamLoess, gamSpline, gaussprPoly, gaussprRadial, gaussprLinear, gbm, gcvEarth, glm, glmboost, glmnet, gpls, hda, hdda, icr, J48, JRip, kernelppls, knn, lars, lasso, lda, leapForward, leapBackward, leapSeq, Linda, lm, lmStepAIC, LMT, logforest, logitBoost, logreg, lssvmPoly, lssvmRadial, lvq, M5, M5Rules, mda, multinom, nb, neuralnet, nnet, nodeHarvest, OneR, pam, parRF, partDSA, pcaNNet, pcr, pda, pda2, penalized, PenalizedLDA, plr, pls, ppr, qda, QdaCov, qrnn, rda, rf, ridge, rlm, rocc, rpart, rpart2, rrllda, rvmLinear, rvmPoly, rvmRadial, sda, sddaLDA, sddaQDA, simpls, sllda, smda, sparseLDA, spls, stepLDA, stepQDA, superpc, svmPoly, svmRadial, svmRadialCost, svmlinear, treebag, vbmpRadial, widekernelppls and xkf. See the Details section below.
...	arguments passed to the classification or regression routine (such as randomForest). Errors will occur if values for tuning parameters are passed here.
preProcess	a string vector that defines an pre-processing of the predictor data. Current possibilities are center, scale, spatialSign, pca, ica, and knnImpute. See preProcess and trainControl on the procedures and how to adjust them.
metric	a string that specifies what summary metric will be used to select the optimal model. By default, possible values are "RMSE" and "Rsquared" for regression and "Accuracy" and "Kappa" for classification. If custom performance metrics are used (via the summaryFunction argument in trainControl , the value of metric should match one of the arguments. If it does not, a warning is issued and the first metric given by the summaryFunction is used. (NOTE: If given, this argument must be named.)
maximize	a logical: should the metric be maximized or minimized?
trControl	a list of values that define how this function acts. See trainControl . (NOTE: If given, this argument must be named.)
tuneGrid	a data frame with possible tuning values. The columns are named the same as the tuning parameters in each method preceded by a period (e.g. .decay, .lambda). Also, a function can be passed to tuneGrid with arguments called len and data. The output of the function should be the same as the output produced by createGrid . (NOTE: If given, this argument must be named.)

tuneLength an integer denoting the number of levels for each tuning parameters that should be generated by `createGrid`. (NOTE: If given, this argument must be named.)

Details

`train` can be used to tune models by picking the complexity parameters that are associated with the optimal resampling statistics. For particular model, a grid of parameters (if any) is created and the model is trained on slightly different data for each candidate combination of tuning parameters. Across each data set, the performance of held-out samples is calculated and the mean and standard deviation is summarized for each combination. The combination with the optimal resampling statistic is chosen as the final model and the entire training set is used to fit a final model.

A variety of models are currently available. The table below enumerates the models and the values of the method argument, as well as the complexity parameters used by `train`.

Model	method	Value	Package	Tuning Parameter(s)
Generalized linear model	glm		stats	none
	glmStepAIC		MASS	none
Generalized additive model	gam		mgcv	select, method
	gamLoess		gam	span, degree
	gamSpline		gam	df
Recursive partitioning	rpart		rpart	cp
	rpart2		rpart	maxdepth
	ctree		party	mincriterion
	ctree2		party	maxdepth
Boosted trees	gbm		gbm	interaction depth, n.trees, shrinkage
	blackboost		mboost	maxdepth, mstop
	ada		ada	maxdepth, iter, nu
	bstTree		bst	maxdepth, mstop, nu
	glmboost		mboost	mstop
Boosted regression models	gamboost		mboost	mstop
	logitBoost		caTools	nIter
	bstLs		bst	mstop, nu
	bstSm		bst	mstop, nu
	Random forests	rf		randomForest
parRF			randomForest, foreach	mtry
cforest			party	mtry
Boruta			Boruta	mtry
Bagging	treebag		ipred	None
	bag		caret	vars
	logicBag		logicFS	ntrees, nleaves
Other Trees	nodeHarvest		nodeHarvest	maxinter, node
	partDSA		partDSA	cut.off.growth, MPD
Logic Regression	logreg		LogicReg	ntrees, treesize
Elastic net (glm)	glmnet		glmnet	alpha, lambda
Neural networks	nnet		nnet	decay, size
	neuralnet		neuralnet	layer1, layer2, layer3
	pcaNNet		caret	decay, size
	avNNet		caret	decay, size, bag

Self-Organizing Maps	xyf bdk	kohonen kohonen	xdim, ydim, topo, xweight xdim, ydim, topo, xweight
Projection pursuit regression	ppr	stats	nterms
Principal component regression	pcr	pls	ncomp
Independent component regression	icr	caret	n.comp
Partial least squares	pls kernelpls simpls widekernelpls	pls, caret pls, caret pls, caret pls, caret	ncomp ncomp ncomp ncomp
Sparse partial least squares	spls	spls, caret	K, eta, kappa
Support vector machines	svmLinear svmRadial svmRadialCost svmPoly	kernlab kernlab kernlab kernlab	C sigma, C C scale, degree, C
Relevance vector machines	rvmLinear rvmRadial rvmPoly	kernlab kernlab kernlab	none sigma scale, degree
Least squares support vector machines	lssvmRadial	kernlab	sigma
Gaussian processes	guassprLinear1 guassprRadial guassprPoly	kernlab kernlab kernlab	none sigma scale, degree
Linear least squares	lm lmStepAIC leapForward leapBackward leapSeq	stats MASS leaps leaps leaps	None None nvmax nvmax nvmax
Robust linear regression	rlm	MASS	None
Multivariate adaptive regression splines	earth gcvEarth	earth earth	degree, nprune degree
Bagged MARS	bagEarth	caret, earth	degree, nprune
Rule Based Regression	M5Rules M5 cubist	RWeka RWeka Cubist	pruned, smoothed pruned, smoothed, rule committees, neighbors
Penalized linear models	penalized ridge enet lars lars2 enet foba	penalized elasticnet elasticnet lars lars elasticnet foba	lambda1, lambda2 lambda lambda, fraction fraction steps fraction lambda, k
Supervised principal components	superpc	superpc	n.components, threshold
Quantile regression forests	qrf	quantregForest	mtry
Quantile regression neural networks	qrnn	qrnn	n.hidden, penalty, bag
Linear discriminant analysis	lda Linda	MASS rrcov	None None
Quadratic discriminant analysis	qda QdaCov	MASS rrcov	None None
Stabilized linear discriminant analysis	sllda	ipred	None

Heteroscedastic discriminant analysis	hda	hda	newdim, lambda, gamma
Stepwise discriminant analysis	stepLDA	klaR	maxvar, direction
	stepQDA	klaR	maxvar, direction
Stepwise diagonal discriminant analysis	sddaLDA	SDDA	None
	sddaQDA	SDDA	None
Shrinkage discriminant analysis	sda	sda	diagonal
Sparse linear discriminant analysis	sparseLDA	sparseLDA	NumVars, lambda
	PenalizedLDA	penalizedLDA	lambda, K
Regularized discriminant analysis	rda	klaR	lambda, gamma
Mixture discriminant analysis	mda	mda	subclasses
Sparse mixture discriminant analysis	smda	sparseLDA	NumVars, R, lambda
Penalized discriminant analysis	pda	mda	lambda
	pda2	mda	df
Stabilised linear discriminant analysis	sllda	ipred	None
High dimensional discriminant analysis	hdda	HDclassif	model, threshold
Flexible discriminant analysis (MARS)	fda	mda, earth	degree, nprune
Robust Regularized Linear Discriminant Analysis	rlda	rlda	lambda, alpha
Bagged FDA	bagFDA	caret, earth	degree, nprune
Logistic/multinomial regression	multinom	nnet	decay
Penalized logistic regression	plr	stepPlr	lambda, cp
Rule-based classification	J48	RWeka	C
	OneR	RWeka	None
	PART	RWeka	threshold, pruned
	JRip	RWeka	NumOpt
Random Ferns	rFerns	rFerns	depth
Logic Forests	logforest	LogicForest	None
Bayesian multinomial probit model	vbmpRadial	vbmp	estimateTheta
k nearest neighbors	knn3	caret	k
Nearest shrunken centroids	pam	pamr	threshold
	scrda	rda	alpha, delta
Naive Bayes	nb	klaR	usekernel, fL
Generalized partial least squares	gpls	gpls	K.prov
Learned vector quantization	lvq	class	size, k
ROC Curves	rocc	rocc	xgenes

By default, the function `createGrid` is used to define the candidate values of the tuning parameters. The user can also specify their own. To do this, a data frame is created with columns for each tuning parameter in the model. The column names must be the same as those listed in the table above with a leading dot. For example, `ncomp` would have the column heading `.ncomp`. This data frame can then be passed to `createGrid`.

In some cases, models may require control arguments. These can be passed via the three dots argument. Note that some models can specify tuning parameters in the control objects. If specified, these values will be superseded by those given in the `createGrid` argument.

The vignette entitled "caret Manual – Model Building" has more details and examples related to this function.

`train` can be used with "explicit parallelism", where different resamples (e.g. cross-validation group) and models can be split up and run on multiple machines or processors. By default, `train`

will use a single processor on the host machine. As of version 4.99 of this package, the framework used for parallel processing uses the **foreach** package. To run the resamples in parallel, the code for `train` does not change; prior to the call to `train`, a parallel backend is registered with **foreach** (see the examples below).

Value

A list is returned of class `train` containing:

<code>method</code>	the chosen model.
<code>modelType</code>	an identifier of the model type.
<code>results</code>	a data frame the training error rate and values of the tuning parameters.
<code>bestTune</code>	a data frame with the final parameters.
<code>call</code>	the (matched) function call with dots expanded
<code>dots</code>	a list containing any ... values passed to the original call
<code>metric</code>	a string that specifies what summary metric will be used to select the optimal model.
<code>control</code>	the list of control parameters.
<code>preProcess</code>	either NULL or an object of class <code>preProcess</code>
<code>finalModel</code>	an fit object using the best parameters
<code>trainingData</code>	a data frame
<code>resample</code>	A data frame with columns for each performance metric. Each row corresponds to each resample. If leave-one-out cross-validation or out-of-bag estimation methods are requested, this will be NULL. The <code>returnResamp</code> argument of <code>trainControl</code> controls how much of the resampled results are saved.
<code>perfNames</code>	a character vector of performance metrics that are produced by the summary function
<code>maximize</code>	a logical recycled from the function arguments.
<code>yLimits</code>	the range of the training set outcomes.
<code>times</code>	a list of execution times: everything is for the entire call to <code>train</code> , <code>final</code> for the final model fit and, optionally, prediction for the time to predict new samples (see <code>trainControl</code>)

Author(s)

Max Kuhn (the guts of `train.formula` were based on Ripley's `nnet.formula`)

References

Kuhn (2008), "Building Predictive Models in R Using the caret" (<http://www.jstatsoft.org/v28/i05/>)

See Also

`trainControl`, `update.train`, `modelLookup`, `createGrid`, `createFolds`

Examples

```
#####
## Classification Example

data(iris)
TrainData <- iris[,1:4]
TrainClasses <- iris[,5]

knnFit1 <- train(TrainData, TrainClasses,
                 method = "knn",
                 preProcess = c("center", "scale"),
                 tuneLength = 10,
                 trControl = trainControl(method = "cv"))

knnFit2 <- train(TrainData, TrainClasses,
                 method = "knn",
                 preProcess = c("center", "scale"),
                 tuneLength = 10,
                 trControl = trainControl(method = "boot"))

library(MASS)
nnetFit <- train(TrainData, TrainClasses,
                 method = "nnet",
                 preProcess = "range",
                 tuneLength = 2,
                 trace = FALSE,
                 maxit = 100)

#####
## Regression Example

library(mlbench)
data(BostonHousing)

lmFit <- train(medv ~ . + rm:lstat,
               data = BostonHousing,
               "lm")

library(rpart)
rpartFit <- train(medv ~ .,
                  data = BostonHousing,
                  "rpart",
                  tuneLength = 9)

#####
## Example with a custom metric

madSummary <- function (data,
                         lev = NULL,
                         model = NULL)
```

```

{
  out <- mad(data$obs - data$pred,
             na.rm = TRUE)
  names(out) <- "MAD"
  out
}

robustControl <- trainControl(summaryFunction = madSummary)
marsGrid <- expand.grid(.degree = 1,
                      .nprune = (1:10) * 2)

earthFit <- train(medv ~ .,
                 data = BostonHousing,
                 "earth",
                 tuneGrid = marsGrid,
                 metric = "MAD",
                 maximize = FALSE,
                 trControl = robustControl)

#####
## Parallel Processing Example via multicore package

## Not run:

library(doMC)
registerDoMC(2)

## NOTE: don't run models form RWeka when using
### multicore. The session will crash.

## The code for train() does not change:
set.seed(1)
usingMC <- train(medv ~ .,
                data = BostonHousing,
                "glmboost")

## or use:
## library(doMPI) or
## library(doSMP) and so on

## End(Not run)

```

trainControl

Control parameters for train

Description

Control the computational nuances of the `train` function

Usage

```
trainControl(method = "boot",
             number = ifelse(method %in% c("cv", "repeatedcv"), 10, 25),
             repeats = ifelse(method %in% c("cv", "repeatedcv"), 1, number),
             verboseIter = FALSE,
             returnData = TRUE,
             returnResamp = "final",
             savePredictions = FALSE,
             p = 0.75,
             classProbs = FALSE,
             summaryFunction = defaultSummary,
             selectionFunction = "best",
             custom = NULL,
             preProcOptions = list(thresh = 0.95, ICComp = 3, k = 5),
             index = NULL,
             timingSamps = 0,
             predictionBounds = rep(FALSE, 2))
```

Arguments

method	The resampling method: boot, boot632, cv, repeatedcv, LOOCV, LGOCV (for repeated training/test splits), or oob (only for random forest, bagged trees, bagged earth, bagged flexible discriminant analysis, or conditional tree forest models)
number	Either the number of folds or number of resampling iterations
repeats	For repeated k-fold cross-validation only: the number of complete sets of folds to compute
verboseIter	A logical for printing a training log.
returnData	A logical for saving the data
returnResamp	A character string indicating how much of the resampled summary metrics should be saved. Values can be "final", "all" or "none"
savePredictions	a logical to save the hold-out predictions for each resample
p	For leave-group out cross-validation: the training percentage
classProbs	a logical; should class probabilities be computed for classification models (along with predicted values) in each resample?
summaryFunction	a function to compute performance metrics across resamples. The arguments to the function should be the same as those in defaultSummary .
custom	an optional list of functions that can be used to fit custom models. See the details below and worked examples in the "train" package vignette. This is an "experimental" version for testing. Please send emails to the maintainer for suggestions or problems.
selectionFunction	the function used to select the optimal tuning parameter. This can be a name of the function or the function itself. See best for details and other options.

preProcOptions	A list of options to pass to <code>preProcess</code> . The type of pre-processing (e.g. center, scaling etc) is passed in via the <code>preProc</code> option in <code>train</code> .
index	a list with elements for each resampling iteration. Each list element is the sample rows used for training at that iteration.
timingSamps	the number of training set samples that will be used to measure the time for predicting samples (zero indicates that the prediction time should not be estimated).
predictionBounds	a logical or numeric vector of length 2 (regression only). If logical, the predictions can be constrained to be within the limit of the training set outcomes. For example, a value of <code>c(TRUE, FALSE)</code> would only constrain the lower end of predictions. If numeric, specific bounds can be used. For example, if <code>c(10, NA)</code> , values below 10 would be predicted as 10 (with no constraint in the upper side).

Details

For custom modeling functions, several functions can be specified using the `custom` argument:

- `parameters` a data frame or function of tuning parameters
- `model` a function that trains the model
- `prediction` a function that predicts new samples (either numbers or character/factor vectors)
- `probability` an optional function for classification models that returns a matrix or data frame of class probabilities (in columns)
- `sort` a function that sorts the tuning parameters by complexity

For more details and worked examples, see the “caret Manual – Model Building” vignette.

Value

An echo of the parameters specified

Author(s)

Max Kuhn

update.train

Update and Re-fit a Model

Description

`update` allows a user to over-ride the tuning parameter selection process by specifying a set of tuning parameters.

Usage

```
## S3 method for class 'train'
update(object, param = NULL, ...)
```

Arguments

object	an object of class <code>train</code>
param	a data frame or named list of all tuning parameters
...	not currently used

Details

To update the model, the training data must be stored in the model object (see the option `returnData` in `trainControl`). Also, all tuning parameters must be specified (with the preceding dot in the name).

All other options are held constant, including the original pre-processing (if any), options passed in using `code...` and so on.

When printing, the verbiage "The tuning parameter was set manually." is used to describe how the tuning parameters were created.

Value

a new `train` object

Author(s)

Max Kuhn

See Also

`train`, `trainControl`

Examples

```
data(iris)
TrainData <- iris[,1:4]
TrainClasses <- iris[,5]

knnFit1 <- train(TrainData, TrainClasses,
  method = "knn",
  preProcess = c("center", "scale"),
  tuneLength = 10,
  trControl = trainControl(method = "cv"))

update(knnFit1, list(.k = 3))
```

varImp	<i>Calculation of variable importance for regression and classification models</i>
--------	--

Description

A generic method for calculating variable importance for objects produced by train and method specific methods

Usage

```
## S3 method for class 'train'  
varImp(object, useModel = TRUE, nonpara = TRUE, scale = TRUE, ...)
```

```
## S3 method for class 'earth'  
varImp(object, value = "gcv", ...)
```

```
## S3 method for class 'fda'  
varImp(object, value = "gcv", ...)
```

```
## S3 method for class 'rpart'  
varImp(object, surrogates = FALSE, competes = TRUE, ...)
```

```
## S3 method for class 'randomForest'  
varImp(object, ...)
```

```
## S3 method for class 'gbm'  
varImp(object, numTrees, ...)
```

```
## S3 method for class 'classbagg'  
varImp(object, ...)
```

```
## S3 method for class 'regbagg'  
varImp(object, ...)
```

```
## S3 method for class 'pamrtrained'  
varImp(object, threshold, data, ...)
```

```
## S3 method for class 'lm'  
varImp(object, ...)
```

```
## S3 method for class 'mvr'  
varImp(object, estimate = NULL, ...)
```

```
## S3 method for class 'bagEarth'  
varImp(object, ...)
```

```

## S3 method for class 'RandomForest'
varImp(object, ...)

## S3 method for class 'rfe'
varImp(object, drop = FALSE, ...)

## S3 method for class 'dsa'
varImp(object, cuts = NULL, ...)

## S3 method for class 'multinom'
varImp(object, ...)

## S3 method for class 'gam'
varImp(object, ...)

## S3 method for class 'cubist'
varImp(object, weights = c(0.5, 0.5), ...)

```

Arguments

object	an object corresponding to a fitted model
useModel	use a model based technique for measuring variable importance? This is only used for some models (lm, pls, rf, rpart, gbm, pam and mars)
nonpara	should nonparametric methods be used to assess the relationship between the features and response (only used with useModel = FALSE and only passed to filterVarImp).
scale	should the importance values be scaled to 0 and 100?
...	parameters to pass to the specific varImp methods
numTrees	the number of iterations (trees) to use in a boosted tree model
threshold	the shrinkage threshold (pamr models only)
data	the training set predictors (pamr models only)
value	the statistic that will be used to calculate importance: either gcv, nsubsets, or rss
surrogates	should surrogate splits contribute to the importance calculation?
competes	should competing splits contribute to the importance calculation?
estimate	which estimate of performance should be used? See mvrVal
drop	a logical: should variables not included in the final set be calculated?
cuts	the number of rule sets to use in the model (for par tDSA only)
weights	a numeric vector of length two that weighs the usage of variabls in the rule conditions and the usage in the linear models (see details below).

Details

For models that do not have corresponding `varImp` methods, see `filerVarImp`.

Otherwise:

Linear Models: the absolute value of the t -statistic for each model parameter is used.

Random Forest: `varImp.randomForest` and `varImp.RandomForest` are wrappers around the importance functions from the **randomForest** and **party** packages, respectively.

Partial Least Squares: the variable importance measure here is based on weighted sums of the absolute regression coefficients. The weights are a function of the reduction of the sums of squares across the number of PLS components and are computed separately for each outcome. Therefore, the contribution of the coefficients are weighted proportionally to the reduction in the sums of squares.

Recursive Partitioning: The reduction in the loss function (e.g. mean squared error) attributed to each variable at each split is tabulated and the sum is returned. Also, since there may be candidate variables that are important but are not used in a split, the top competing variables are also tabulated at each split. This can be turned off using the `maxcompete` argument in `rpart.control`. This method does not currently provide class-specific measures of importance when the response is a factor.

Bagged Trees: The same methodology as a single tree is applied to all bootstrapped trees and the total importance is returned

Boosted Trees: `varImp.gbm` is a wrapper around the function from that package (see the **gbm** package vignette)

Multivariate Adaptive Regression Splines: MARS models include a backwards elimination feature selection routine that looks at reductions in the generalized cross-validation (GCV) estimate of error. The `varImp` function tracks the changes in model statistics, such as the GCV, for each predictor and accumulates the reduction in the statistic when each predictor's feature is added to the model. This total reduction is used as the variable importance measure. If a predictor was never used in any of the MARS basis functions in the final model (after pruning), it has an importance value of zero. Prior to June 2008, the package used an internal function for these calculations. Currently, the `varImp` is a wrapper to the `evimp` function in the `earth` package. There are three statistics that can be used to estimate variable importance in MARS models. Using `varImp(object, value = "gcv")` tracks the reduction in the generalized cross-validation statistic as terms are added. However, there are some cases when terms are retained in the model that result in an increase in GCV. Negative variable importance values for MARS are set to zero. Alternatively, using `varImp(object, value = "rss")` monitors the change in the residual sums of squares (RSS) as terms are added, which will never be negative. Also, the option `varImp(object, value = "nsubsets")`, which counts the number of subsets where the variable is used (in the final, pruned model).

Nearest shrunken centroids: The difference between the class centroids and the overall centroid is used to measure the variable influence (see `pamr.predict`). The larger the difference between the class centroid and the overall center of the data, the larger the separation between the classes. The training set predictions must be supplied when an object of class `pamrtrained` is given to `varImp`.

Cubist: The Cubist output contains variable usage statistics. It gives the percentage of times where each variable was used in a condition and/or a linear model. Note that this output will probably be inconsistent with the rules shown in the output from `summary.cubist`. At each split of the tree, Cubist saves a linear model (after feature selection) that is allowed to have terms for each variable

used in the current split or any split above it. Quinlan (1992) discusses a smoothing algorithm where each model prediction is a linear combination of the parent and child model along the tree. As such, the final prediction is a function of all the linear models from the initial node to the terminal node. The percentages shown in the Cubist output reflects all the models involved in prediction (as opposed to the terminal models shown in the output). The variable importance used here is a linear combination of the usage in the rule conditions and the model.

Value

A data frame with class `c("varImp.train", "data.frame")` for `varImp.train` or a matrix for other models.

Author(s)

Max Kuhn

xyplot.resamples

Lattice Functions for Visualizing Resampling Results

Description

Lattice functions for visualizing resampling results across models

Usage

```
## S3 method for class 'resamples'
xyplot(x, data = NULL, what = "scatter", models = NULL,
       metric = x$metric[1], units = "min", ...)

## S3 method for class 'resamples'
dotplot(x, data = NULL, models = x$models,
        metric = x$metric, conf.level = 0.95, ...)

## S3 method for class 'resamples'
densityplot(x, data = NULL, models = x$models, metric = x$metric, ...)

## S3 method for class 'resamples'
bwplot(x, data = NULL, models = x$models, metric = x$metric, ...)

## S3 method for class 'resamples'
splom(x, data = NULL, variables = "models",
      models = x$models, metric = NULL, panelRange = NULL, ...)

## S3 method for class 'resamples'
parallel(x, data = NULL, models = x$models, metric = x$metric[1], ...)
```

Arguments

x	an object generated by resamples
data	Not used
models	a character string for which models to plot. Note: xyplot requires one or two models whereas the other methods can plot more than two.
metric	a character string for which metrics to use as conditioning variables in the plot. splom requires exactly one metric when variables = "models" and at least two when variables = "metrics".
variables	either "models" or "metrics"; which variable should be treated as the scatter plot variables?
panelRange	a common range for the panels. If NULL, the panel ranges are derived from the values across all the models
what	for xyplot, the type of plot. Valid options are: "scatter" (for a plot of the resampled results between two models), "BlandAltman" (a Bland-Altman, aka MA plot between two models), "tTime" (for the total time to run train versus the metric), "mTime" (for the time to build the final model) or "pTime" (the time to predict samples - see the timingSamps options in trainControl , rfeControl , or sbfControl)
units	either "sec", "min" or "hour"; which what is either "tTime", "mTime" or "pTime", how should the timings be scaled?
conf.level	the confidence level for intervals about the mean (obtained using t.test)
...	further arguments to pass to either histogram , densityplot , xyplot , dotplot or splom

Details

The ideas and methods here are based on Hothorn et al (2005) and Eugster et al (2008).

dotplot plots the average performance value (with two-sided confidence limits) for each model and metric.

densityplot and bwplot display univariate visualizations of the resampling distributions while splom shows the pair-wise relationships.

Value

a lattice object

Author(s)

Max Kuhn

References

Hothorn et al. The design and analysis of benchmark experiments. *Journal of Computational and Graphical Statistics* (2005) vol. 14 (3) pp. 675-699

Eugster et al. Exploratory and inferential analysis of benchmark experiments. *Ludwigs-Maximilians-Universitat Munchen, Department of Statistics, Tech. Rep* (2008) vol. 30

See Also

[resamples](#), [dotplot](#), [bwplot](#), [densityplot](#), [xyplot](#), [splom](#)

Examples

```
## Not run:
#load(url("http://caret.r-forge.r-project.org/Classification_and_Regression_Training_files/exampleModels.RData

resamps <- resamples(list(CART = rpartFit,
                          CondInfTree = ctreeFit,
                          MARS = earthFit))

dotplot(resamps,
         scales =list(x = list(relation = "free")),
         between = list(x = 2))

bwplot(resamps,
        metric = "RMSE")

densityplot(resamps,
             auto.key = list(columns = 3),
             pch = "|")

xyplot(resamps,
        models = c("CART", "MARS"),
        metric = "RMSE")

splom(resamps, metric = "RMSE")
splom(resamps, variables = "metrics")

parallel(resamps, metric = "RMSE")

## End(Not run)
```

Index

*Topic **datasets**

- BloodBrain, 13
- cars, 20
- cox2, 27
- dhfr, 31
- GermanCredit, 44
- mdrr, 57
- oil, 64
- pottery, 82
- segmentationData, 113
- tecator, 120

*Topic **graphs**

- panel.needle, 69

*Topic **hplot**

- calibration, 16
- dotPlot, 33
- dotplot.diff.resamples, 34
- featurePlot, 38
- histogram.train, 45
- lattice.rfe, 51
- lift, 53
- panel.lift2, 67
- plot.train, 72
- plot.varImp.train, 73
- plotClassProbs, 74
- plotObsVsPred, 75
- prcomp.resamples, 82
- resampleHist, 99
- xyplot.resamples, 135

*Topic **manip**

- Alternate Affy Gene Expression
 - Summary Methods., 3
- classDist, 21
- findCorrelation, 41
- findLinearCombos, 42
- normalize.AffyBatch.normalize2Reference, 60
- oneSE, 65
- predict.train, 87

- sensitivity, 114
- spatialSign, 117
- summary.bagEarth, 119

*Topic **models**

- bag.default, 8
- caretFuncs, 18
- caretSBF, 19
- diff.resamples, 31
- dummyVars, 36
- filterVarImp, 40
- format.bagEarth, 43
- normalize2Reference, 62
- nullModel, 63
- plsda, 77
- predictors, 89
- resamples, 100
- rfe, 103
- sbf, 109
- train, 121
- update.train, 130
- varImp, 132

*Topic **multivariate**

- icr.formula, 46
- knn3, 48
- knnreg, 49
- predict.knn3, 85
- predict.knnreg, 86

*Topic **neural**

- avNNet.default, 6
- pcaNNet.default, 70

*Topic **print**

- print.train, 97

*Topic **regression**

- bagEarth, 10
- bagFDA, 12
- predict.bagEarth, 84

*Topic **utilities**

- as.table.confusionMatrix, 4
- BoxCoxTrans.default, 14

- confusionMatrix, 23
- confusionMatrix.train, 25
- createDataPartition, 28
- createGrid, 29
- maxDissim, 55
- modelLookup, 58
- nearZeroVar, 59
- postResample, 80
- preProcess, 94
- print.confusionMatrix, 97
- resampleSummary, 102
- rfeControl, 106
- sbfcControl, 111
- trainControl, 128

- absorp(tecator), 120
- Alternate Affy Gene Expression Summary
 - Methods., 3
- anneal, 41, 42
- anovaScores, 113
- anovaScores(caretSBF), 19
- as.matrix.confusionMatrix, 25
- as.matrix.confusionMatrix
 - (as.table.confusionMatrix), 4
- as.table.confusionMatrix, 4, 25
- avNNet(avNNet.default), 6
- avNNet.default, 6

- bag(bag.default), 8
- bag.default, 8
- bagControl(bag.default), 8
- bagEarth, 10, 44, 85, 94
- bagFDA, 12, 94
- bagging, 94
- barchart, 83, 84
- bbbDescr(BloodBrain), 13
- best, 129
- best(oneSE), 65
- binom.test, 24, 25
- BloodBrain, 13
- boxcox, 14, 15, 96
- BoxCoxTrans, 96
- BoxCoxTrans(BoxCoxTrans.default), 14
- BoxCoxTrans.default, 14
- bwplot, 35, 137
- bwplot.diff.resamples, 33
- bwplot.diff.resamples
 - (dotplot.diff.resamples), 34
- bwplot.resamples, 101

- bwplot.resamples(xyplot.resamples), 135

- calibration, 16
- caretFuncs, 18
- caretSBF, 19, 113
- cars, 20
- cforest, 93, 94
- checkConditionalX(nearZeroVar), 59
- checkResamples(nearZeroVar), 59
- classDist, 21
- cluster.resamples(prcomp.resamples), 82
- confusionMatrix, 5, 23, 27, 97, 116
- confusionMatrix.rfe
 - (confusionMatrix.train), 25
- confusionMatrix.sbf
 - (confusionMatrix.train), 25
- confusionMatrix.train, 25
- contr.dummy(dummyVars), 36
- contr.treatment, 36
- contrasts, 36, 38
- cox2, 27
- cox2Class(cox2), 27
- cox2Descr(cox2), 27
- cox2IC50(cox2), 27
- createDataPartition, 28
- createFolds, 126
- createFolds(createDataPartition), 28
- createGrid, 29, 122, 123, 125, 126
- createMultiFolds(createDataPartition),
 - 28
- createResample(createDataPartition), 28
- ctree, 93, 94
- ctreeBag(bag.default), 8

- defaultSummary, 129
- defaultSummary(postResample), 80
- densityplot, 35, 45, 46, 51, 52, 74, 99, 136,
 - 137
- densityplot.diff.resamples, 33
- densityplot.diff.resamples
 - (dotplot.diff.resamples), 34
- densityplot.resamples, 101
- densityplot.resamples
 - (xyplot.resamples), 135
- densityplot.rfe(lattice.rfe), 51
- densityplot.train, 99
- densityplot.train(histogram.train), 45
- dhfr, 31
- diff.resamples, 31, 35, 101

- dist, 56
- dotPlot, 33
- dotplot, 33–35, 69, 73, 76, 136, 137
- dotplot.diff.resamples, 32, 33, 34
- dotplot.resamples (xyplot.resamples), 135
- dummyVars, 36
- earth, 11, 44, 84, 94
- endpoints (tecatr), 120
- evimp, 134
- extractPrediction, 76
- extractPrediction (predict.train), 87
- extractProb, 74
- extractProb (predict.train), 87
- fastICA, 47, 95, 96
- fattyAcids (oil), 64
- fda, 13, 85, 94
- featurePlot, 38
- filterVarImp, 40
- findCorrelation, 41
- findLinearCombos, 42, 42
- foba, 94
- format.bagEarth, 43
- format.earth, 44
- formula, 38
- gamFuncs (caretFuncs), 18
- gamScores, 113
- gamScores (caretSBF), 19
- generateExprSet-methods, 4
- generateExprVal.method.trimMean (Alternate Affy Gene Expression Summary Methods.), 3
- genetic, 41, 42
- GermanCredit, 44
- glmnet, 93
- hclust, 83, 84
- histogram, 35, 45, 46, 51, 52, 74, 99, 136
- histogram.rfe (lattice.rfe), 51
- histogram.train, 45, 99
- icr (icr.formula), 46
- icr.formula, 46
- ipredbagg, 93, 94
- ipredknn, 49–51
- knn, 49–51, 87
- knn3, 48, 86
- knn3Train (knn3), 48
- knnreg, 49, 86
- knnregTrain (knnreg), 49
- lars, 93
- lattice.options, 16, 53
- lattice.rfe, 51
- ldaBag (bag.default), 8
- ldaFuncs (caretFuncs), 18
- ldaSBF, 113
- ldaSBF (caretSBF), 19
- leaps, 41, 42
- levelplot, 35, 72, 73
- levelplot.diff.resamples, 33
- levelplot.diff.resamples (dotplot.diff.resamples), 34
- lift, 53, 67, 68
- lm, 40, 47
- lmFuncs, 108
- lmFuncs (caretFuncs), 18
- lmSBF, 110, 113
- lmSBF (caretSBF), 19
- loess, 40
- logBBB (BloodBrain), 13
- mahalanobis, 22
- maxDissim, 55
- mcnemar.test, 24
- mdrr, 57
- mdrrClass (mdrr), 57
- mdrrDescr (mdrr), 57
- minDiss (maxDissim), 55
- model.matrix, 36, 38
- modelLookup, 58, 126
- mvrVal, 133
- NaiveBayes, 78
- nbBag (bag.default), 8
- nbFuncs, 108
- nbFuncs (caretFuncs), 18
- nbSBF, 113
- nbSBF (caretSBF), 19
- nearZeroVar, 59
- negPredValue, 25
- negPredValue (sensitivity), 114
- nnet, 7, 70, 71, 94
- nnetBag (bag.default), 8

- normalize.AffyBatch.normalize2Reference, 60
- normalize2Reference, 62
- nullModel, 63
- oil, 64
- oilType (oil), 64
- oneSE, 65
- p.adjust, 32
- pamr.train, 93, 94
- panel.calibration, 17
- panel.calibration (calibration), 16
- panel.dotplot, 69
- panel.lift (panel.lift2), 67
- panel.lift2, 54, 67
- panel.needle, 69, 73
- panel.xyplot, 68
- parallel.resamples (xyplot.resamples), 135
- pcaNNet (pcaNNet.default), 70
- pcaNNet.default, 70
- pickSizeBest, 108
- pickSizeBest (caretFuncs), 18
- pickSizeTolerance, 108
- pickSizeTolerance (caretFuncs), 18
- pickVars (caretFuncs), 18
- plot.prcomp.resamples (prcomp.resamples), 82
- plot.train, 72
- plot.varImp.train, 73
- plotClassProbs, 74, 88
- plotObsVsPred, 75, 88
- plsBag (bag.default), 8
- plsda, 77
- plsr, 78, 79
- posPredValue, 25
- posPredValue (sensitivity), 114
- postResample, 80, 102
- pottery, 82
- potteryClass (pottery), 82
- prcomp, 22, 83, 96
- prcomp.resamples, 82
- predict, 86
- predict.avNNet (avNNet.default), 6
- predict.bag (bag.default), 8
- predict.bagEarth, 11, 84
- predict.bagFDA, 13
- predict.bagFDA (predict.bagEarth), 84
- predict.BoxCoxTrans (BoxCoxTrans.default), 14
- predict.classDist (classDist), 21
- predict.dummyVars (dummyVars), 36
- predict.foba, 94
- predict.icr (icr.formula), 46
- predict.ipredknn, 86, 87
- predict.knn3, 49, 85
- predict.knnreg, 50, 86
- predict.lars, 93
- predict.list (predict.train), 87
- predict.nullModel (nullModel), 63
- predict.pcaNNet (pcaNNet.default), 70
- predict.plsda (plsda), 77
- predict.preProcess (preProcess), 94
- predict.rfe (rfe), 103
- predict.sbf (sbf), 109
- predict.splsda (plsda), 77
- predict.train, 87
- predictors, 89
- preProcess, 7, 15, 47, 71, 94, 122, 126, 130
- print.bagEarth (bagEarth), 10
- print.bagFDA (bagFDA), 12
- print.confusionMatrix, 25, 97
- print.train, 97
- ProbeSet, 3
- R2 (postResample), 80
- randomForest, 94, 109, 122
- resampleHist, 99
- resamples, 33, 35, 82–84, 100, 137
- resampleSummary, 102
- rfe, 19, 26, 27, 51, 52, 100, 103, 108
- rfeControl, 19, 52, 103, 104, 106, 136
- rfeIter (rfe), 103
- rfFuncs, 108
- rfFuncs (caretFuncs), 18
- rfSbf, 113
- rfSbf (caretSbf), 19
- RMSE (postResample), 80
- rpart, 93, 94
- sbf, 20, 26, 27, 100, 109, 113
- sbfControl, 19, 20, 109, 110, 111, 136
- segmentationData, 113
- sensitivity, 25, 114
- spatialSign, 96, 117
- specificity, 25
- specificity (sensitivity), 114

splom, 35, 83, 84, 136, 137
 splom.resamples, 101
 splom.resamples(xyplot.resamples), 135
 spls, 78, 79
 splsda(plsda), 77
 stripplot, 45, 46, 51, 52, 72, 73
 stripplot.rfe(lattice.rfe), 51
 stripplot.train, 99
 stripplot.train(histogram.train), 45
 sumDiss(maxDissim), 55
 summary.bagEarth, 119
 summary.bagFDA(summary.bagEarth), 119
 summary.cubist, 134
 summary.diff.resamples
 (diff.resamples), 31
 summary.gam, 20
 summary.resamples(resamples), 100
 superpc.train, 93, 94
 svmBag(bag.default), 8

t.test, 136
 table, 23
 tecator, 120
 terms.formula, 37
 tolerance(oneSE), 65
 train, 9, 26, 27, 30, 45, 46, 58, 66, 67, 72, 73,
 80, 81, 87, 88, 97–101, 121, 128,
 130, 131
 trainControl, 27, 46, 65–67, 80, 81, 88,
 99–101, 122, 126, 128, 131, 136
 treebagFuncs, 108
 treebagFuncs(caretFuncs), 18
 treebagSBF, 113
 treebagSBF(caretSBF), 19
 trellis.par.set, 17, 54, 68
 trim.matrix, 42, 43
 twoClassSummary(postResample), 80

update.train, 126, 130
 update.trellis, 54

varImp, 34, 132

xyplot, 16, 17, 35, 45, 46, 51–54, 68, 72, 73,
 76, 83, 84, 136, 137
 xyplot.calibration(calibration), 16
 xyplot.lift(lift), 53
 xyplot.resamples, 101, 135
 xyplot.rfe(lattice.rfe), 51
 xyplot.train, 99
 xyplot.train(histogram.train), 45