Package ‘liver’

August 7, 2022

Title  ``Eating the Liver of Data Science''
Version  1.13
Description  Provides a collection of helper functions that make various techniques from data science more user-friendly for non-experts. In this way, our aim is to allow non-experts to become familiar with the techniques with only a minimal level of coding knowledge. Indeed, following an ancient Persian idiom, we refer to this as ``eating the liver of data science'’ which could be interpreted as ```getting intimately close with data science’’. Examples of procedures we include are: data partitioning for out-of-sample testing, computing Mean Squared Error (MSE) for quantifying prediction accuracy, and data transformation (z-score and min-max). Besides such helper functions, the package also includes several interesting datasets that are useful for multivariate analysis.

URL  https://www.uva.nl/profile/a.mohammadi
Depends  R (>= 3.5.0)
Imports  class, ggplot2
Suggests  pROC, skimr, knitr, rmarkdown
VignetteBuilder  knitr
License  GPL (>= 2)
Repository  CRAN
Author  Reza Mohammadi [aut, cre] (<https://orcid.org/0000-0001-9538-0648>),
        Kevin Burke [aut]
Maintainer  Reza Mohammadi <a.mohammadi@uva.nl>
NeedsCompilation  no
Date/Publication  2022-08-07 21:30:02 UTC

R topics documented:

liver-package ................................................................. 2
accuracy ................................................................. 3
adult ................................................................. 4
advertising ............................................................ 5
bank ................................................................. 6
## Description

The liver package provides a collection of helper functions that make various techniques from data science more user-friendly for non-experts. In this way, our aim is to allow non-experts to become familiar with the techniques with only a minimal level of coding knowledge. Indeed, following an ancient Persian idiom, we refer to this as "eating the liver of data science" which could be interpreted as "getting intimately close with data science". Examples of procedures we include are: data partitioning for out-of-sample testing, computing Mean Squared Error (MSE) for quantifying prediction accuracy, and data transformation (z-score and min-max). Besides such helper functions, the package also includes several interesting datasets that are useful for multivariate analysis.

## Author(s)

Reza Mohammadi &lt;a.mohammadi@uva.nl&gt;  
Amsterdam Business School  
University of Amsterdam
**accuracy**

Kevin Burke <kevin.burke@ul.ie>
Department of Statistics
University of Limerick

Maintainer: Reza Mohammadi <a.mohammadi@uva.nl>

---

### Description

Computes average classification accuracy.

### Usage

```r
accuracy( pred, actual, cutoff = NULL, reference = NULL )
```

### Arguments

- **pred**: a numerical vector of estimated values.
- **actual**: a numerical vector of actual values.
- **cutoff**: cutoff value for the case that `pred` is vector of probabilities.
- **reference**: a factor of classes to be used as the true results.

### Value

The computed average classification accuracy (numeric value).

### Author(s)

Reza Mohammadi <a.mohammadi@uva.nl> and Kevin Burke <kevin.burke@ul.ie>

### See Also

- `conf.mat`, `mse`, `mae`

### Examples

```r
pred  = c( "no", "yes", "yes", "no", "no", "yes", "no", "no" )
actual = c( "yes", "no", "yes", "no", "no", "yes", "yes", "yes" )

accuracy( pred, actual )
```
Description

The adult dataset was collected from the US Census Bureau and the primary task is to predict whether a given adult makes more than $50K a year based attributes such as education, hours of work per week, etc. The target feature is income, a factor with levels "<=50K" and ">50K", and the remaining 14 variables are predictors.

Usage

data( adult )

Format

The adult dataset, as a data frame, contains 48598 rows and 15 columns (variables/features). The 15 variables are:

- age: age in years.
- workclass: a factor with 6 levels.
- demogweight: the demographics to describe a person.
- education: a factor with 16 levels.
- education.num: number of years of education.
- marital.status: a factor with 5 levels.
- occupation: a factor with 15 levels.
- relationship: a factor with 6 levels.
- race: a factor with 5 levels.
- gender: a factor with levels "Female","Male".
- capital.gain: capital gains.
- capital.loss: capital losses.
- hours.per.week: number of hours of work per week.
- native.country: a factor with 42 levels.
- income: yearly income as a factor with levels "<=50K" and ">50K".

Details

This dataset can be downloaded from the UCI machine learning repository:
http://www.cs.toronto.edu/~delve/data/adult/desc.html

A detailed description of the dataset can be found in the UCI documentation at:
http://www.cs.toronto.edu/~delve/data/adult/adultDetail.html
References


See Also

*risk, churn, churnTel, bank, advertising, marketing, insurance, cereal, housePrice, house*

Examples

```r
data( adult )
str( adult )
```

<table>
<thead>
<tr>
<th>advertising</th>
<th>advertising data set</th>
</tr>
</thead>
</table>

Description

The dataset is from an anonymous organisation’s social media ad campaign. The advertising dataset contains 11 features and 1143 records.

Usage

```r
data( advertising )
```

Format

The advertising dataset, as a data frame, contains 1143 rows and 11 columns (variables/features). The 11 variables are:

- **ad.id**: an unique ID for each ad.
- **xyz.campaign.id**: an ID associated with each ad campaign of XYZ company.
- **fb.campaign.id**: an ID associated with how Facebook tracks each campaign.
- **age**: age of the person to whom the ad is shown.
- **gender**: gender of the person to whim the add is shown.
- **interest**: a code specifying the category to which the person’s interest belongs (interests are as mentioned in the person’s Facebook public profile).
- **impressions**: the number of times the ad was shown.
- **clicks**: number of clicks on for that ad.
- **spend**: amount paid by company xyz to Facebook, to show that ad.
- **conversion**: total number of people who enquired about the product after seeing the ad.
- **approved**: total number of people who bought the product after seeing the ad.
Details
A detailed description of the dataset can be found:
https://www.kaggle.com/loveall/clicks-conversion-tracking

See Also
adult, risk, churn, churnTel, bank, marketing, insurance, cereal, housePrice, house

Examples

data( advertising )
str( advertising )

Bank marketing data set

Description
The data is related to direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, in order to access if the product (bank term deposit) would be (or not) subscribed. The classification goal is to predict if the client will subscribe a term deposit (variable deposit).

Usage
data( bank )

Format
The bank dataset, as a data frame, contains 4521 rows (customers) and 17 columns (variables/features). The 17 variables are:
Bank client data:
- age: numeric.
- marital: marital status; categorical: "married", "divorced", "single"; note: "divorced" means divorced or widowed.
- education: categorical: "secondary", "primary", "tertiary", "unknown".
- default: has credit in default?; binary: "yes", "no".
- balance: average yearly balance, in euros; numeric.
- housing: has housing loan? binary: "yes", "no".
- loan: has personal loan? binary: "yes", "no".
Related with the last contact of the current campaign:

- **contact**: contact communication type; categorical: "unknown", "telephone", "cellular".
- **day**: last contact day of the month; numeric.
- **month**: last contact month of year; categorical: "jan", "feb", "mar", ..., "nov", "dec".
- **duration**: last contact duration, in seconds; numeric.

Other attributes:

- **campaign**: number of contacts performed during this campaign and for this client; numeric, includes last contact.
- **pdays**: number of days that passed by after the client was last contacted from a previous campaign; numeric. -1 means client was not previously contacted.
- **previous**: number of contacts performed before this campaign and for this client; numeric.
- **poutcome**: outcome of the previous marketing campaign; categorical: "success", "failure", "unknown", "other".

Target variable:

- **deposit**: Indicator of whether the client subscribed a term deposit; binary: "yes" or "no".

**Details**

This dataset can be downloaded from the UCI machine learning repository:

http://archive.ics.uci.edu/ml/datasets/Bank+Marketing

**References**


**See Also**

adult, risk, churn, churnTel, advertising, marketing, insurance, cereal, housePrice, house

**Examples**

```
data(bank)
str(bank)
```
Cereal data set

Description

This dataset contains nutrition information for 77 breakfast cereals and includes 16 variables. The "rating" column is our target as a rating of the cereals (Possibly from Consumer Reports?).

Usage

data(cereal)

Format

The cereal dataset, as a data frame, contains 77 rows (breakfast cereals) and 16 columns (variables/features). The 16 variables are:

- name: Name of cereal.
- manuf: Manufacturer of cereal:
  - A: American Home Food Products;
  - G: General Mills;
  - K: Kelloggs;
  - N: Nabisco;
  - P: Post;
  - Q: Quaker Oats;
  - R: Ralston Purina;
- type: cold or hot.
- calories: calories per serving.
- protein: grams of protein.
- fat: grams of fat.
- sodium: milligrams of sodium.
- fiber: grams of dietary fiber.
- carbo: grams of complex carbohydrates.
- sugars: grams of sugars.
- potass: milligrams of potassium.
- vitamins: vitamins and minerals - 0, 25, or 100, indicating the typical percentage of FDA recommended.
- shelf: display shelf (1, 2, or 3, counting from the floor).
- weight: weight in ounces of one serving.
- cups: number of cups in one serving.
- rating: a rating of the cereals (Possibly from Consumer Reports?).
Details
The original source can be found: https://perso.telecom-paristech.fr/eagan/class/igr204/datasets

See Also
adult, risk, churn, churnTel, bank, advertising, marketing, insurance, housePrice, house

Examples

data( cereal )
str( cereal )

churn

Churn data set

Description
This dataset comes from IBM Sample Data Sets. Customer churn occurs when customers stop doing business with a company, also known as customer attrition. The data set contains 5000 rows (customers) and 20 columns (features). The "Churn" column is our target which indicate whether customer churned (left the company) or not.

Usage
data( churn )

Format
The churn dataset, as a data frame, contains 5000 rows (customers) and 20 columns (variables/features). The 20 variables are:

• state: Categorical, for the 51 states and the District of Columbia.
• area.code: Categorical.
• account.length: count, how long account has been active.
• voice.plan: Categorical, yes or no, voice mail plan.
• voice.messages: Count, number of voice mail messages.
• int1.plan: Categorical, yes or no, international plan.
• int1.mins: Continuous, minutes customer used service to make international calls.
• int1.calls: Count, total number of international calls.
• int1.charge: Continuous, total international charge.
• day.mins: Continuous, minutes customer used service during the day.
• day.calls: Count, total number of calls during the day.
• day.charge: Continuous, total charge during the day.
• eve.mins: Continuous, minutes customer used service during the evening.
• eve.calls: Count, total number of calls during the evening.
• eve.charge: Continuous, total charge during the evening.
• night.mins: Continuous, minutes customer used service during the night.
• night.calls: Count, total number of calls during the night.
• night.charge: Continuous, total charge during the night.
• customer.calls: Count, number of calls to customer service.
• churn: Categorical, yes or no. Indicator of whether the customer has left the company (yes or no).

References


See Also

adult, risk, churnTel, bank, advertising, marketing, insurance, cereal, housePrice, house

Examples

data( churn )

str( churn )

churnCredit  
Churn dataset for Credit Card Customers

Description

Customer churn occurs when customers stop doing business with a company, also known as customer attrition. The data set contains 10127 rows (customers) and 21 columns (features). The "churn" column is our target which indicate whether customer churned (left the company) or not.

Usage

data( churnCredit )
The churnCredit dataset, as a data frame, contains 10127 rows (customers) and 21 columns (variables/features). The 21 variables are:

- **customer.ID**: Customer ID.
- **gender**: Whether the customer is a male or a female.
- **age**: Customer's Age in Years.
- **education**: Educational Qualification of the account holder (example: high school, college graduate, etc.).
- **marital.status**: Married, Single, Divorced, Unknown
- **dependent.counts**: Number of dependent counts.
- **card.category**: Type of Card (Blue, Silver, Gold, Platinum).
- **months.on.book**: Period of relationship with bank.
- **relationship.count**: Total number of products held by the customer.
- **months.inactive**: Number of months inactive in the last 12 months.
- **contacts.count.12**: Number of Contacts in the last 12 months.
- **credit.limit**: Credit Limit on the Credit Card.
- **revolving.balance**: Total Revolving Balance on the Credit Card.
- **open.to.buy**: Open to Buy Credit Line (Average of last 12 months).
- **transaction.amount.Q4.Q1**: Change in Transaction Amount (Q4 over Q1).
- **transaction.amount.12**: Total Transaction Amount (Last 12 months).
- **transaction.count**: Total Transaction Count (Last 12 months).
- **transaction.change**: Change in Transaction Count (Q4 over Q1).
- **utilization.ratio**: Average Card Utilization Ratio.
- **churn**: Whether the customer churned or not (yes or no).

For more information related to the dataset see: [https://www.kaggle.com/sakshigoyal7/credit-card-customers](https://www.kaggle.com/sakshigoyal7/credit-card-customers)

### See Also
- adult.risk
- churn
- churnTel
- bank
- advertising
- marketing
- insurance
- cereal
- housePrice
- house

### Examples
```r
data( churnCredit )

str( churnCredit )```
description

customer churn occurs when customers stop doing business with a company, also known as customer attrition. The data set contains 7043 rows (customers) and 21 columns (features). The "Churn" column is our target which indicate whether customer churned (left the company) or not.

usage

data( churnTel )

format

The churnTel dataset, as a data frame, contains 7043 rows (customers) and 21 columns (variables/features). The 21 variables are:

- customer.ID: Customer ID.
- gender: Whether the customer is a male or a female.
- senior.citizen: Whether the customer is a senior citizen or not (1, 0).
- partner: Whether the customer has a partner or not (yes, no).
- dependent: Whether the customer has dependents or not (yes, no).
- tenure: Number of months the customer has stayed with the company.
- phone.service: Whether the customer has a phone service or not (yes, no).
- multiple.lines: Whether the customer has multiple lines or not (yes, no, no phone service).
- internet.service: Customer's internet service provider (DSL, fiber optic, no).
- online.security: Whether the customer has online security or not (yes, no, no internet service).
- online.backup: Whether the customer has online backup or not (yes, no, no internet service).
- device.protection: Whether the customer has device protection or not (yes, no, no internet service).
- tech.support: Whether the customer has tech support or not (yes, no, no internet service).
- streaming.TV: Whether the customer has streaming TV or not (yes, no, no internet service).
- streaming.movie: Whether the customer has streaming movies or not (yes, no, no internet service).
- contract: The contract term of the customer (month to month, 1 year, 2 year).
- paperless.bill: Whether the customer has paperless billing or not (yes, no).
- payment.method: The customer’s payment method (electronic check, mail check, bank transfer, credit card).
- monthly.charge: The amount charged to the customer monthly.
- total.charges: The total amount charged to the customer.
- churn: Whether the customer churned or not (yes or no).
Details
For more information related to the dataset see: https://www.kaggle.com/blastchar/telco-customer-churn

See Also
adult, risk, churn, bank, advertising, marketing, insurance, cereal, housePrice, house

Examples
data(churnTel)
str(churnTel)

data(churnTel)
str(churnTel)

class data( "Confusion Matrix" )

Description
Create a Confusion Matrix.

Usage
conf.mat( pred, actual, cutoff = NULL, reference = NULL,
proportion = FALSE, dnn = c( "Predict", "Actual" ), ... )

Arguments
pred a vector of estimated values.
actual a vector of actual values.
cutoff cutoff value for the case that pred is vector of probabilities.
reference a factor of classes to be used as the true results.
proportion Logical: FALSE (default) for a confusion matrix with number of cases. TRUE, for a confusion matrix with the proportion of cases.
dnn the names to be given to the dimensions in the result (the dimnames names).
... options to be passed to table.

Value
the results of table on pred and actual.

Author(s)
Reza Mohammadi <a.mohammadi@uva.nl> and Kevin Burke <kevin.burke@ul.ie>

See Also
conf.mat.plot, accuracy
Examples

```r
pred  = c( "no", "yes", "yes", "no", "no", "yes", "no", "no" )
actual = c( "yes", "no", "yes", "no", "no", "no", "yes", "yes" )

conf.mat( pred, actual )
conf.mat( pred, actual, proportion = TRUE )
```

---

conf.mat.plot  
Plot Confusion Matrix

Description

Plot a Confusion Matrix.

Usage

```r
conf.mat.plot( pred, actual, cutoff = NULL, reference = NULL, conf.level = 0,
              margin = 1, color = c( "#ff83a8", "#83ff9b" ), ... )
```

Arguments

- `pred`: a vector of estimated values.
- `actual`: a vector of actual values.
- `cutoff`: cutoff value for the case that `pred` is vector of probabilities.
- `reference`: a factor of classes to be used as the true results.
- `conf.level`: confidence level used for the confidence rings on the odds ratios. Must be a single nonnegative number less than 1; if set to 0 (the default), confidence rings are suppressed.
- `margin`: a numeric vector with the margins to equate. Must be one of 1 (the default), 2, or c(1, 2), which corresponds to standardizing the row, column, or both margins in each 2 by 2 table. Only used if `std` equals "margins".
- `color`: a vector of length 2 specifying the colors to use for the smaller and larger diagonals of each 2 by 2 table.
- `...`: options to be passed to `fourfoldplot`.

Author(s)

Reza Mohammadi <a.mohammadi@uva.nl> and Kevin Burke <kevin.burke@ul.ie>

See Also

`conf.mat`
Examples

pred = c( "no", "yes", "yes", "no", "yes", "no", "yes", "no" )
actual = c( "yes", "no", "yes", "no", "no", "no", "yes", "yes" )

conf.mat.plot( pred, actual )

corona
Corona data set

Description

COVID-19 Coronavirus data - daily (up to 14 December 2020).

Usage

data( corona )

Format

The corona dataset, as a data frame, contains 61900 rows and 12 columns (variables/features).

Details

This dataset can be downloaded from the UCI machine learning repository:


See Also

churn, adult, risk, churnTel, bank, advertising, marketing, insurance, cereal, housePrice, house

Examples

data( corona )

str( corona )
The fertilizer dataset contains 4 features and 96 records. Results from an experiment to compare yields of a crop obtained under three different fertilizers. The target feature is yield.

Usage

```r
data( fertilizer )
```

See Also

```r
adult, risk, churn, churnTel, bank, advertising, marketing, insurance, cereal, housePrice, house
```

Examples

```r
data( fertilizer )
str( fertilizer )
```

Description

Finding missing values.

Usage

```r
find.na( x )
```

Arguments

- `x` a numerical vector, matrix or data.frame.

Value

A numeric matrix with two columns.

Author(s)

Reza Mohammadi <a.mohammadi@uva.nl> and Kevin Burke <kevin.burke@ul.ie>
Examples

```r
x = c( 2.3, NA, -1.4, 0, 3.45 )
find.na( x )
```

---

Description

The house dataset contains 6 features and 414 records. The target feature is `unit.price` and the remaining 5 variables are predictors.

Usage

```r
data( house )
```

Format

The `house` dataset, as a data frame, contains 414 rows and 6 columns (variables/features). The 6 variables are:

- `house.age`: house age (numeric, in year).
- `distance.to.MRT`: distance to the nearest MRT station (numeric).
- `stores.number`: number of convenience stores (numeric).
- `latitude`: latitude (numeric).
- `longitude`: longitude (numeric).
- `unit.price`: house price of unit area (numeric).

Details

A detailed description of the dataset can be found:

https://www.kaggle.com/quantbruce/real-estate-price-prediction

See Also

`adult, risk, churn, churnTel, bank, advertising, marketing, insurance, cereal, housePrice`

Examples

```r
data( house )
str( house )
```
### housePrice dataset

**Description**

This data set contains 1460 rows and 81 columns (features). The "SalePrice" column is the target.

**Usage**

```r
data( housePrice )
```

**Format**

The `housePrice` dataset, as a data frame, contains 1460 rows and 81 columns (variables/features).

**Details**

For more information related to the dataset see:


**See Also**

`adult`, `risk`, `churn`, `churnTel`, `bank`, `advertising`, `marketing`, `insurance`, `cereal`, `house`

**Examples**

```r
data( housePrice )
str( housePrice )
```

### insurance data set

**Description**

The insurance dataset contains 7 features and 1338 records. The target feature is `charge` and the remaining 6 variables are predictors.

**Usage**

```r
data( insurance )
```
Format

The insurance dataset, as a data frame, contains 1338 rows (customers) and 7 columns (variables/features). The 7 variables are:

- **age**: age of primary beneficiary.
- **bmi**: body mass index, providing an understanding of body, weights that are relatively high or low relative to height, objective index of body weight \((\text{kg} / \text{m}^2)\) using the ratio of height to weight, ideally 18.5 to 24.9.
- **children**: Number of children covered by health insurance / Number of dependents.
- **smoker**: Smoking as a factor with 2 levels, yes, no.
- **gender**: insurance contractor gender, female, male.
- **region**: the beneficiary’s residential area in the US, northeast, southeast, southwest, northwest.
- **charge**: individual medical costs billed by health insurance.

Details

A detailed description of the dataset can be found: [https://www.kaggle.com/mirichoi0218/insurance](https://www.kaggle.com/mirichoi0218/insurance)

References


See Also

`adult.risk, churn, churnTel, bank, advertising, marketing, cereal, housePrice, house`

Examples

```r
data( insurance )
str( insurance )
```

---

**kNN**

**k-Nearest Neighbour Classification**

Description

kNN is used to perform k-nearest neighbour classification for test set using training set. For each row of the test set, the k nearest (based on Euclidean distance) training set vectors are found. Then, the classification is done by majority vote (ties broken at random). This function provides a formula interface to the knn function of R package class. In addition, it allows normalization of the given data using the transform function.
Usage

kNN( formula, train, test, k = 1, transform = FALSE, type = "class", l = 0,
use.all = TRUE, na.rm = FALSE )

Arguments

formula a formula, with a response but no interaction terms. For the case of data frame, it is taken as the model frame (see model.frame).
train data frame or matrix of train set cases.
test data frame or matrix of test set cases.
k number of neighbours considered.
transform a character with options FALSE (default), "minmax", and "zscore". Option "minmax" means no transformation. This option allows the users to use normalized version of the train and test sets for the kNN algorithm.
type either "class" (default) for the predicted class or "prob" for model confidence values.
l minimum vote for definite decision, otherwise doubt. (More precisely, less than k-l dissenting votes are allowed, even if k is increased by ties.)
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.
na.rm a logical value indicating whether NA values in x should be stripped before the computation proceeds.

Value

When type = "class" (default), a factor vector is returned, in which the doubt will be returned as NA. When type = "prob", a matrix of confidence values is returned (one column per class).

Author(s)

Reza Mohammadi <a.mohammadi@uva.nl> and Kevin Burke <kevin.burke@ul.ie>

References


See Also

knn, transform
Examples

data( risk )

train = risk[ 1:100, ]
test = risk[ 101, ]

kNN( risk ~ income + age, train = train, test = test )

= Visualizing the Optimal Number of k

Description

Visualizing the Optimal Number of k for k-Nearest Neighbour Classification kNN based on accuracy or Mean Square Error (MSE).

Usage

kNN.plot( formula, train, test, k.max = 10, transform = FALSE, base = "error", report = FALSE, set.seed = NULL, ... )

Arguments

- formula: a formula, with a response but no interaction terms. For the case of data frame, it is taken as the model frame (see model.frame).
- train: data frame or matrix of train set cases.
- test: data frame or matrix of test set cases.
- k.max: the maximum number of number of neighbours to consider, must be at least two.
- transform: a character with options FALSE (default), "minmax", and "zscore". Option "minmax" means no transformation. This option allows the users to use normalized version of the train and test sets for the kNN algorithm.
- base: base measurement: error (default), accuracy, or MSE for Mean Square Error.
- report: a character with options FALSE (default) and TRUE. Option TRUE reports the values of the base measurement.
- set.seed: a single value, interpreted as an integer, or NULL.
- ... options to be passed to kNN().

Author(s)

Reza Mohammadi <a.mohammadi@uva.nl> and Kevin Burke <kevin.burke@ul.ie>

References

See Also

`kNN, transform`

Examples

data( risk )

train = risk[ 1:150, ]
test = risk[ 151:246, ]

kNN.plot( risk ~ income + age, train = train, test = test )
kNN.plot( risk ~ income + age, train = train, test = test, base = "accuracy" )

---

`mae`  
*Mean Absolute Error (MAE)*

Description

Computes mean absolute error.

Usage

`mae( pred, actual, weight = 1, na.rm = FALSE )`

Arguments

- `pred`  
a numerical vector of estimated values.
- `actual`  
a numerical vector of actual values.
- `weight`  
a numerical vector of weights the same length as `pred`.
- `na.rm`  
a logical value indicating whether NA values in `pred` should be stripped before the computation proceeds.

Value

The computed mean squared error (numeric value).

Author(s)

Reza Mohammadi <a.mohammadi@uva.nl> and Kevin Burke <kevin.burke@ul.ie>

See Also

`mse`
Examples

```r
pred = c( 2.3, -1.4, 0, 3.45 )
actual = c( 2.1, -0.9, 0, 2.99 )

mae( pred, actual )
```

Description

The marketing dataset contains 8 features and 40 records as 40 days that report how much we spent, how many clicks, impressions and transactions we got, whether or not a display campaign was running, as well as our revenue, click-through-rate and conversion rate. The target feature is `revenue` and the remaining 7 variables are predictors.

Usage

```r
data( marketing )
```

Format

The `marketing` dataset, as a data frame, contains 40 rows and 8 columns (variables/features). The 8 variables are:

- `spend`: daily spend of money on PPC (apy-per-click).
- `clicks`: number of clicks on for that ad.
- `impressions`: amount of impressions per day.
- `display`: whether or not a display campaign was running.
- `transactions`: number of transactions per day.
- `click.rate`: click-through-rate.
- `conversion.rate`: conversion rate.
- `revenue`: daily revenue.

Details

A detailed description of the dataset can be found:

[https://github.com/chrisBow/marketing-regression-part-one](https://github.com/chrisBow/marketing-regression-part-one)

See Also

- `adult`, `risk`, `churn`, `churnTel`, `bank`, `advertising`, `insurance`, `cereal`, `housePrice`, `house`

Examples

```r
data( marketing )
str( marketing )
```
**minmax**  

---

**Min-Max normalization**

---

**Description**

Performs Min-Max normalization of numerical variables.

**Usage**

```r
minmax( x, columns = NULL, na.rm = FALSE )
```

**Arguments**

- `x`  
  a numerical vector, matrix or data.frame.

- `columns`  
  which columns are going to transfer for the cases that `x` is a matrix or a data.frame. Defaults to all columns.

- `na.rm`  
  a logical value indicating whether NA values in `x` should be stripped before the computation proceeds.

**Value**

transformed version of `x`.

**Author(s)**

Reza Mohammadi <a.mohammadi@uva.nl> and Kevin Burke <kevin.burke@ul.ie>

**See Also**

- `transform`, `zscore`

**Examples**

```r
x = c( 2.3, -1.4, 0, 3.45 )

minmax( x )
```
### Description

Computes mean squared error.

### Usage

```
mse( pred, actual, weight = 1, na.rm = FALSE )
```

### Arguments

- `pred`: a numerical vector of estimated values.
- `actual`: a numerical vector of actual values.
- `weight`: a numerical vector of weights the same length as `pred`.
- `na.rm`: a logical value indicating whether NA values in `pred` should be stripped before the computation proceeds.

### Value

The computed mean squared error (numeric value).

### Author(s)

Reza Mohammadi `<a.mohammadi@uva.nl>` and Kevin Burke `<kevin.burke@ul.ie>`

### See Also

`mae`

### Examples

```
pred  = c( 2.3, -1.4, 0, 3.45 )
actual = c( 2.1, -0.9, 0, 2.99 )

mse( pred, actual )
```
### partition

**Partition the data**

#### Description

Randomly partitions the data (primarily intended to split into "training" and "test" sets) according to the supplied probabilities.

#### Usage

```r
partition( data, prob = c( 0.7, 0.3 ), set.seed = NULL )
```

#### Arguments

- **data**: an \((n \times p)\) matrix or a data.frame.
- **prob**: a numerical vector in \([0, 1]\).
- **set.seed**: a single value, interpreted as an integer, or NULL.

#### Value

a list which includes the data partitions.

#### Author(s)

Reza Mohammadi <a.mohammadi@uva.nl> and Kevin Burke <kevin.burke@ul.ie>

#### Examples

```r
data( iris )
partition( data = iris, prob = c( 0.7, 0.3 ) )
```

### risk

**Risk data set**

#### Description

The risk dataset containing 6 features and 246 records. The target feature is risk, a factor with levels "good risk" and "bad risk" along with 5 predictors.

#### Usage

```r
data( risk )
```
The risk dataset, as a data frame, contains 246 rows (customers) and 6 columns (variables/features). The 6 variables are:

- **age**: age in years.
- **marital**: A factor with levels "single", "married", and "other".
- **income**: yearly income.
- **mortgage**: A factor with levels "yes" and "no".
- **nr_loans**: Number of loans that customers have.
- **risk**: A factor with levels "good risk" and "bad risk".

See Also

- adult, churn, churnTel, bank, advertising, marketing, insurance, cereal, housePrice, house

Examples

```r
data( risk )
str( risk )
```

### skewness

**Description**

Computes the skewness for each field.

**Usage**

```r
skewness( x, na.rm = FALSE )
```

**Arguments**

- `x`: a numerical vector, matrix or data.frame.
- `na.rm`: a logical value indicating whether NA values in `x` should be stripped before the computation proceeds.

**Value**

A numeric vector of skewness values.

**Author(s)**

Reza Mohammadi <a.mohammadi@uva.nl> and Kevin Burke <kevin.burke@ul.ie>
Examples

\[ x = c(2.3, -1.4, 0, 3.45) \]

\[ \text{skewness}(x) \]

---

### skim

**Skim a data frame to get useful summary statistics**

**Description**

`skim()` provides an overview of a data frame as an alternative to `summary()`. This function is a wrapper for the `skimr::skim()` function of R package `skimr`.

**Usage**

```r
skim(data, hist = TRUE, ...)  
```

**Arguments**

- `data` a data frame or matrix.
- `hist` Logical: TRUE (default) to report the histogram of each variable.
- `...` columns to select for skimming. The default is to skim all columns.

**Author(s)**

Reza Mohammadi <a.mohammadi@uva.nl> and Kevin Burke <kevin.burke@ul.ie>

**See Also**

`summary()`

**Examples**

```r
data( risk )  
skim( risk )
```
transform  

Z-score normalization

Description

Performs variable transformation such as Z-score and min-max normalization.

Usage

transform( x, method = c( "minmax", "zscore" ), columns = NULL, na.rm = FALSE )

Arguments

x  
a numerical vector, a matrix or a data.frame.

method  
a method to transfer x.

columns  
which columns are going to transfer for the cases that x is a matrix or a data.frame. Defaults to all columns.

na.rm  
a logical value indicating whether NA values in x should be stripped before the computation proceeds.

Value

transformed version of x.

Author(s)

Reza Mohammadi <a.mohammadi@uva.nl> and Kevin Burke <kevin.burke@ul.ie>

See Also

zscore, minmax

Examples

x = c( 2.3, -1.4, 0, 3.45 )

transform( x, method = "minmax" )

transform( x, method = "zscore" )
Description

Performs Z-score normalization to transform numerical variables.

Usage

\[
\text{zscore}( \ x, \ \text{columns} = \text{NULL}, \ \text{na.rm} = \text{FALSE} )
\]

Arguments

- **x**: a numerical vector, matrix or data.frame.
- **columns**: which columns are going to transfer for the cases that \( x \) is a matrix or a data.frame. Defaults to all columns.
- **na.rm**: a logical value indicating whether NA values in \( x \) should be stripped before the computation proceeds.

Value

transformed version of \( x \).

Author(s)

Reza Mohammadi <a.mohammadi@uva.nl> and Kevin Burke <kevin.burke@ul.ie>

See Also

transform, minmax

Examples

\[
x = c( \ 2.3, \ -1.4, \ 0, \ 3.45 )
\]

\[
zscore( \ x )
\]
Index

* data preprocessing
  - find.na, 16
  - minmax, 24
  - partition, 26
  - skewness, 27
  - transform, 29
  - zscore, 30

* datasets
  - adult, 4
  - advertising, 5
  - bank, 6
  - cereal, 8
  - churn, 9
  - churnCredit, 10
  - churnTel, 12
  - corona, 15
  - fertilizer, 16
  - house, 17
  - housePrice, 18
  - insurance, 18
  - marketing, 23
  - risk, 26

* models
  - kNN, 19
  - knn, 19, 20
  - kNN.plot, 21

* package
  - liver-package, 2

* parameter learning
  - accuracy, 3
  - conf.mat, 13
  - conf.mat.plot, 14
  - mae, 22
  - mse, 25
  - skewness, 27

accuracy, 3, 13
adult, 4, 6, 7, 9–11, 13, 15–19, 23, 27
advertising, 5, 5, 7, 9–11, 13, 15–19, 23, 27
bank, 5, 6, 6, 9–11, 13, 15–19, 23, 27
cereal, 5–7, 8, 10, 11, 13, 15–19, 23, 27
churn, 5–7, 9, 9, 11, 13, 15–19, 23, 27
churnCredit, 10
churnTel, 5–7, 9–11, 12, 15–19, 23, 27
conf.mat, 3, 13, 14
conf.mat.plot, 13, 14
corona, 15
fertilizer, 16
find.na, 16
formula, 20, 21
house, 5–7, 9–11, 13, 15, 16, 17, 18, 19, 23, 27
housePrice, 5–7, 9–11, 13, 15–17, 18, 19, 23, 27
insurance, 5–7, 9–11, 13, 15–18, 18, 23, 27
kNN, 19, 22
knn, 19, 20
kNN.plot, 21
liver-package, 2
mae, 3, 22, 25
marketing, 5–7, 9–11, 13, 15–19, 23, 27
minmax, 24, 29, 30
model.frame, 20, 21
mse, 3, 22, 25
partition, 26
risk, 5–7, 9–11, 13, 15–19, 23, 26
skewness, 27
skim, 28
skimr::skim(), 28
summary(), 28
transform, 19, 20, 22, 24, 29, 30
zscore, 24, 29, 30