Package: optbin (via r-universe)

November 1, 2024

Version 1.4

Date 2024-10-04

Title Optimal Binning of Data

Author Greg Kreider [aut, cre]

Copyright Primordial Machine Vision Systems, Inc.

Maintainer Greg Kreider <support@primachvis.com>

Description Defines thresholds for breaking data into a number of discrete levels, minimizing the (mean) squared error within all bins.

License BSD_3_clause + file LICENSE

NeedsCompilation yes

Repository CRAN

Date/Publication 2024-10-31 18:10:07 UTC

Contents

`assign.optbin` *Bin Assignment*

Description

`assign.optbin` returns an object with the same shape as the input data and values replaced by bin numbers.

Usage

```
assign.optbin(x, binspec, extend.upper=FALSE, by.value=FALSE)
```

Arguments

`x`	numeric data to assign
`binspec`	an optimal binning partition
`extend.upper`	if true then any value in x above the last bin is assigned to that bin, otherwise its bin is set to NA
`by.value`	if true then return average value for bin instead of bin numbers

Details

Replaces the values in a copy of the input data by the bin number it belongs to, or by the bin average value with by.value. The lowest bin always extends to -Inf. The extend.upper argument can open the last bin to +Inf if true. Use this function to get in-place bin assignments for the unsorted data that was passed to optbin.

Value

An object of the same shape as the data.

See Also

`optbin`

Examples

```
d <- c(rnorm(30, mean=10, sd=2), rnorm(40, mean=20, sd=2),
       rnorm(30, mean=30, sd=3))
binned <- optbin(d, 3)
assign.optbin(d, binned)
```

hist.optbin *Histogram with Optimal Bins Marked*

Description

Draw a histogram of the data used to build the optimal binning and mark the extent of the bins.

Usage

```
## S3 method for class 'optbin'
hist(x, bincol=NULL, main=NULL, xlab=NULL, ...)
```

Arguments

x	an object of class optbin.
bincol	vector of colors for showing extent of bins (default uses an internal set)
main	plot title, if not specified will modify the normal histogram title
xlab	x axis label, if not specified will modify the normal histogram label
...	other parameters passed through to hist

Details

The points behind the binning are passed unchanged to the histogram function. Bins are marked with colored bars under the x axis, and lines showing the average value in each are also drawn on top.

Value

None

See Also

optbin, hist

optbin *Optimal Binning of Continuous Variables*

Description

Determines break points in numeric data that minimize the difference between each point in a bin and the average over it.

Usage

```
optbin(x, numbin, metric=c('se', 'mse'), is.sorted=FALSE, max.cache=2^31, na.rm=FALSE)
```

Arguments

x	numeric data
numbin	number of bins to partition vector into
metric	minimize squared error (se) between values and average over bin, or mean squared error (mse) dividing squared error by bin length
is.sorted	set true if x is already in increasing order
max.cache	maximum memory in bytes to use to cache bin metrics; if analysis would need more than use slower calculation without cache
na.rm	drop NA values (which may occur when converting the data to a vector), otherwise cannot proceed with binning

Details

Data is converted into a numeric vector and sorted if necessary. Internally bins are determined by positions within the vector, with the breaks inclusive at the upper end. The bin thresholds are the same, so bin b covers the range thr[b-1] < x <= thr[b], where thr[0] is -Inf. The routine finds the first split found with the best metric, if there is more than one.

The library uses an exhaustive search over all possible breakpoints. It begins by finding the best splits with 2 bins for all pairs of start and endpoints, then adds a third bin, and so on. This rejects most alternatives at each level, leaving an O(nbin * nval * nval) algorithm.

Value

An object of class 'optbin' with components:

x	the original data, sorted
numbins	the number of bins created
call	argument values when function called
metric	cost function used to select best partition
minse	value of SE/MSE metric for all bins
thr	upper threshold of bin range, inclusive
binavg	average of values in each bin
binse	value of SE/MSE metric for each bin
breaks	positions of endpoint (inclusive) of each bin in x

See Also

assign.optbin, print.optbin, summary.optbin, plot.optbin

Examples

```
## Well separated groups
set.seed(17)
d1 <- c(rnorm(75, mean=1, sd=0.2), rnorm(75, mean=3, sd=0.2),
        rnorm(84, mean=6, sd=0.2), rnorm(75, mean=9, sd=0.2),
        rnorm(75, mean=11, sd=0.2), rnorm(150, mean=15, sd=0.2))
## Divides into groups 1+2+3, 4+5, 6, metric is 1176.3
binned3 <- optbin(d1, 3)
summary(binned3)
plot(binned3)
## Divides into groups 1, 2, 3, 4+5, and 6, metric is 169.9
binned5 <- optbin(d1, 5)
plot(binned5)
## Divides into separate groups, metric is 24.4
binned6 <- optbin(d1, 6)
summary(binned6)
plot(binned6)
## Each rnorm group divides roughly in half.
binned12 <- optbin(d1, 12)
plot(binned12)
## A grouping that overlaps, bins near but not at minima between peaks
d2 <- c(rnorm(300, mean=1, sd=0.25), rnorm(400, mean=2, sd=0.25),
        rnorm(300, mean=3, sd=0.25))
binned3b <- optbin(d2, 3)
hist(binned3b, breaks=50, col='yellow')
```

`plot.optbin` *Plotting Optimal Bins*

Description

`plot` method for class `optbin`.

Usage

```
## S3 method for class 'optbin'
plot(x, col=NULL, main="Binned Observations", ...)
```

Arguments

`x`	an object of class `optbin`.
`col`	vector of colors to apply to bins (default uses an internal set)
`main`	title of graph
`...`	other parameters passed through to the underlying plotting routines (do not set xaxt or ann)

Details

The plot will contain the sorted points of the data that generated the bins. Points are color-coded per bin, and the plot contains the average value over the bin as a line. x axis labels are the upper thresholds for each bin.

Value

None

See Also

optbin

`print.optbin` *Printing Optimal Bins*

Description

`print` method for class `optbin`.

Usage

```
## S3 method for class 'optbin'
print(x, ...)
```

Arguments

`x`	an object of class `optbin`.
`...`	generic arguments (ignored)

Details

Shows the upper bounds of each bin, ie. bin b covers threshold[b-1] < x <= threshold[b] where threshold[0] is -Inf. Also prints the total (mean) squared error sum over all bins.

Value

The argument x unchanged, an object of class 'optbin' with components:

`x`	the original data, sorted
`numbins`	the number of bins created
`call`	argument values when function called
`metric`	cost function used to select best partition
`minse`	value of SE/MSE metric for all bins
`thr`	upper threshold of bin range, inclusive
`binavg`	average of values in each bin
`binse`	value of SE/MSE metric for each bin
`breaks`	positions of endpoint (inclusive) of each bin in x

See Also

optbin, summary.optbin

---

summary.optbin *Summarizing Optimal Bins*

Description

summary method for class optbin.

Usage

```
## S3 method for class 'optbin'
summary(object, show.range=FALSE, ...)
```

Arguments

object	an object of class optbin
show.range	if true then print the bin's range of points (endpoint inclusive) in the sorted data
...	generic arguments (ignored)

Details

Prints a table with the upper threshold (inclusive), the average of the data within the bin, and the (mean) squared error sum. show.range also adds a column with the start and end indices of the sorted data belonging to the bin, although this applies to the sorted list and is less useful in general.

Value

Only called for side-effects (printing). There is no return value.

See Also

optbin, print.optbin

Index