

# The caTools Package

November 15, 2005

**Version** 1.5

**Date** Nov 10 2005

**Title** Miscellaneous tools: I/O, moving window statistics, etc.

**Author** Jarek Tuszynski <jaroslav.w.tuszynski@saic.com>

**Maintainer** Jarek Tuszynski <jaroslav.w.tuszynski@saic.com>

**Depends** R (>= 2.2.0), bitops

**Suggests** MASS, rpart, verification, ROC, ROCR, Epi, limma

**Description** Contains several basic utility functions including: moving (rolling, running) window statistic functions, read/write for GIF and ENVI binary files, fast calculation of AUC, LogitBoost classifier, base64 encoder/decoder, round-off error free sum and cumsum, etc.

**License** The caMassClass Software License, Version 1.0 (See COPYING file or “<http://ncicb.nci.nih.gov/download/camassclasslicense.jsp>”)

**URL** <http://ncicb.nci.nih.gov/download/index.jsp>

## R topics documented:

ENVI . . . . .	2
read.gif & write.gif . . . . .	4
LogitBoost . . . . .	8
base64 . . . . .	10
bin2raw & raw2bin . . . . .	12
colAUC . . . . .	14
combs . . . . .	16
predict.LogitBoost . . . . .	17
runfunc . . . . .	18
sample.split . . . . .	23
sum.exact . . . . .	24
trapz . . . . .	26
<b>Index</b>	<b>28</b>

**Description**

Read and write binary data in ENVI format, which is supported by most GIS software.

**Usage**

```
X=read.ENVI(filename, headerfile=paste(filename, ".hdr", sep=""))
write.ENVI (X, filename, interleave = c("bsq", "bil", "bip"))
```

**Arguments**

<code>X</code>	data to be saved in ENVI file. Can be a matrix or 3D array.
<code>filename</code>	character string with name of the file (connection)
<code>headerfile</code>	optional character string with name of the header file
<code>interleave</code>	optional character string specifying interleave to be used

**Details**

ENVI binary files use a generalized raster data format that consists of two parts:

- binary file - flat binary file equivalent to memory dump, as produced by `writeBin` in R or `fwrite` in C/C++.
- header file - small text (ASCII) file containing the metadata associated with the binary file. This file can contain the following fields, followed by equal sign and a variable:
  - `samples` - number of columns
  - `lines` - number of rows
  - `bands` - number of bands (channels, planes)
  - `data type` - following types are supported:
    - \* 1 - 1-byte unsigned integer
    - \* 2 - 2-byte signed integer
    - \* 3 - 4-byte signed integer
    - \* 4 - 4-byte float
    - \* 5 - 8-byte double
    - \* 9 - 2x8-byte complex number made up from 2 doubles
    - \* 12 - 2-byte unsigned integer
  - `header offset` - number of bytes to skip before raster data starts in binary file.
  - `interleave` - Permutations of dimensions in binary data:
    - \* BSQ - Band Sequential (X[col,row,band])
    - \* BIL - Band Interleave by Line (X[col,band,row])
    - \* BIP - Band Interleave by Pixel (X[band,col,row])
  - `byte order` - the endian-ness of the saved data:

- \* 0 - means little-endian byte order, format used on PC/Intel machines
- \* 1 - means big-endian (aka IEEE, aka "network") byte order, format used on UNIX and Macintosh machines

Fields `samples`, `lines`, `bands`, `data_type` are required, while `header_offset`, `interleave`, `byte_order` are optional. All of them are in form of integers except `interleave` which is a string.

This generic format allows reading of many raw file formats, including those with embedded header information. Also it is a handy binary format to exchange data between PC and UNIX/Mac machines, as well as different languages like: C, Fortran, Matlab, etc. Especially since header files are simple enough to edit by hand.

File type supported by most of GIS (geographic information system) software including: ENVI software, Freelook (free file viewer by ENVI), ArcGIS, etc.

### Value

Function `read.ENVI` returns either a matrix or 3D array. Function `write.ENVI` does not return anything.

### Author(s)

Jarek Tuszynski (SAIC) <jaroslav.w.tuszynski@saic.com>

### See Also

Displaying of images can be done through functions: `image`, `image.plot` and `add.image` from **fields** or `plot.im` from **spatstat**.

ENVI files are practically C-style memory-dumps as performed by `readBin` and `writeBin` functions plus separate meta-data header file.

GIF file formats can also store 3D data (see `read.gif` and `write.gif` functions).

Packages related to GIS data: **shapefiles**, **maptools**, **sp**, **spdep**, **adehabitat**, **GRASS**, **PBSmapping**.

### Examples

```
X = array(1:60, 3:5)
write.ENVI(X, "temp.nvi")
Y = read.ENVI("temp.nvi")
stopifnot(X == Y)
readLines("temp.nvi.hdr")

d = c(20,30,40)
X = array(runif(prod(d)), d)
write.ENVI(X, "temp.nvi", interleave="bil")
Y = read.ENVI("temp.nvi")
stopifnot(X == Y)
readLines("temp.nvi.hdr")

file.remove("temp.nvi")
file.remove("temp.nvi.hdr")
```

---

read.gif & write.gif

*Read and Write Images in GIF format*


---

## Description

Read and write files in GIF format. Files can contain single images or multiple frames. Multi-frame images are saved as animated GIF's.

## Usage

```
read.gif(filename, frame=0, flip=FALSE, verbose=FALSE)
write.gif(image, filename, col="gray", scale=c("smart", "never", "always"),
         transparent=NULL, comment=NULL, delay=0, flip=FALSE, interlace=FALSE)
```

## Arguments

filename	Character string with name of the file. In case of <code>read.gif</code> URL's are also allowed.
image	Data to be saved as GIF file. Can be a 2D matrix or 3D array. Allowed formats in order of preference: <ul style="list-style-type: none"> <li>• array of integers in [0:255] range - this is format required by GIF file, and unless <code>scale='always'</code>, numbers will not be rescaled. Each pixel <code>i</code> will have associated color <code>col[image[i]+1]</code>. This is the only format that can be safely used with non-continuous color maps.</li> <li>• array of doubles in [0:1] range - Unless <code>scale='never'</code> the array will be multiplied by 255 and rounded.</li> <li>• array of numbers in any range - will be scaled or clipped depending on <code>scale</code> option.</li> </ul>
frame	Request specific frame from multiframe (i.e., animated) GIF file. By default all frames are read from the file ( <code>frame=0</code> ). Setting <code>frame=1</code> will ensure that output is always a 2D matrix containing the first frame. Some files have to be read frame by frame, for example: files with subimages of different sizes and files with both global and local color-maps (palettes).
col	Color palette definition. Several formats are allowed: <ul style="list-style-type: none"> <li>• array (list) of colors in the same format as output of palette functions like <code>rainbow</code> or <code>heat.colors</code> (ex. <code>'col=rainbow(256)'</code>). Preferred format for precise color control.</li> <li>• palette function itself (ex. <code>'col=rainbow'</code>). Preferred format if not sure how many colors are needed.</li> <li>• character string with name of internally defined palette. At the moment only "gray" and "jet" (Matlab's jet palette) are defined.</li> <li>• character string with name of palette function (ex. <code>'col="rainbow"'</code>)</li> </ul> <p>Usually palette will consist of 256 colors, which is the maximum allowed by GIF format. By default, grayscale will be used.</p>
scale	There are three approaches to rescaling the data to required [0, 255] integer range:

	<ul style="list-style-type: none"> <li>• "smart" - Data is fitted to [0:255] range, only if it falls outside of it. Also, if <code>image</code> is an array of doubles in range [0, 1] than data is multiplied by 255.</li> <li>• "never" - Pixels with intensities outside of the allowed range are clipped to either 0 or 255. Warning is given.</li> <li>• "always" - Data is always rescaled. If <code>image</code> is a array of doubles in range [0, 1] than data is multiplied by 255; otherwise it is scaled to fit to [0:255] range.</li> </ul>
<code>delay</code>	In case of 3D arrays the data will be stored as animated GIF, and <code>delay</code> controls speed of the animation. It is number of hundredths (1/100) of a second of delay between frames.
<code>comment</code>	Comments in text format are allowed in GIF files. Few file viewers can access them.
<code>flip</code>	By default data is stored in the same orientation as data displayed by <code>print</code> function: row 1 is on top, image x-axis corresponds to columns and y-axis corresponds to rows. However function <code>image</code> adopted different standard: column 1 is on the bottom, image x-axis corresponds to rows and y-axis corresponds to columns. Set <code>flip</code> to <code>TRUE</code> to get the orientation used by <code>image</code> .
<code>transparent</code>	Optional color number to be shown as transparent. Has to be an integer in [0:255] range. NA's in the <code>image</code> will be set to transparent.
<code>interlace</code>	GIF files allow image rows to be <code>interlaced</code> , or reordered in such a way as to allow viewer to display image using 4 passes, making image sharper with each pass. Irrelevant feature on fast computers.
<code>verbose</code>	Display details sections encountered while reading GIF file.

## Details

Palettes often contain continuous colors, such that swapping palettes or rescaling of the image data does not affect image appearance in a drastic way. However, when working with non-continuous color-maps one should always provide `image` in [0:255] integer range (and set `scale="never"`), in order to prevent scaling.

If NA or other infinite numbers are found in the `image` by `write.gif`, they will be converted to numbers given by `transparent`. If `transparent` color is not provided than it will be created, possibly after reshuffling.

There are some GIF files not fully supported by `read.gif` function:

- "Plain Text Extension" is not supported, and will be ignored.
- Multi-frame files with unique settings for each frame have to be read frame by frame. Possible settings include: frames with different sizes, frames using local color maps and frames using individual transparency colors.

## Value

Function `write.gif` does not return anything. Function `read.gif` returns a list with following fields:

<code>image</code>	matrix or 3D array of integers in [0:255] range.
<code>col</code>	color palette definitions with number of colors ranging from 1 to 256. In case when <code>frame=0</code> only the first (usually global) color-map (palette) is returned.
<code>comment</code>	Comments imbedded in GIF File

`transparent` color number corresponding to transparent color. If none was stated than NULL, otherwise an integer in [0:255] range. In order for `image` to display transparent colors correctly one should use `y$col[y$transparent+1] = NA`.

### Author(s)

Jarek Tuszynski (SAIC) (jaroslaw.w.tuszynski@saic.com). Encoding Algorithm adapted from code by Christoph Hohmann, which was adapted from code by Michael Mayer. Parts of decoding algorithm adapted from code by David Koblas.

### References

Ziv, J., Lempel, A. (1977) *An Universal Algorithm for Sequential Data Compression*, IEEE Transactions on Information Theory, May 1977.

Copy of official file format description <http://www.danbbs.dk/%7Edino/whirlgif/gif89.html>

Nicely explained file format description <http://semmix.pl/color/exgraf/eeg11.htm>

Christoph Hohmann code and documentation of encoding algorithm <http://members.aol.com/rf21exe/gif.htm>

Michael A. Mayer code <http://www.danbbs.dk/%7Edino/whirlgif/gifcode.html>

Discussion of GIF file legal status can be found in <http://www.cloanto.com/users/mcb/19950127giflzw.html>.

Interesting page on one way of doing animations in R (with help of outside calls) can be found at <http://pinard.progiciels-bpi.ca/plaisirs/animations/index.html>.

### See Also

Displaying of images can be done through functions: `image` (part of R), `image.plot` and `add.image` from **fields** or `plot.im` from **spatstat** package, and possibly many other functions.

Displayed image can be saved in GIF, JPEG or PNG format using several different functions: `GDD` from package **GDD**, `HTMLplot` from package **R2HTML** and functions `jpeg` and `png`.

Functions for directly reading and writing image files:

- `read.pnm` and `write.pnm` from **pixmap** package can process PBM, PGM and PPM images (file types supported by ImageMagic software)
- `read.ENVI` and `write.ENVI` from this package can process files in ENVI format. ENVI files can store 2D images and 3D data (multi-frame images), and are supported by most GIS (Geographic Information System) software including free "freelook".
- `read.jpeg` from **rimage** package can read JPEG files

There are many functions for creating and managing color palettes:

- R provides functions for creating palettes of continuous colors: `rainbow`, `topo.colors`, `heat.colors`, `terrain.colors`, `gray`
- `tim.colors` in package **fields** contains palette similar to Matlab's jet palette (see examples for simpler implementation)
- `rich.colors` in package **gplots** contains two palettes of continuous colors.
- Functions `brewer.pal` from **RColorBrewer** package and `colorbrewer.palette` from **epitools** package contain tools for generating palettes
- `rgb` and `hsv` creates palette from RGB or HSV 3-vectors.
- `col2rgb` translates palette colors to RGB 3-vectors.

**Examples**

```

# visual comparison between image and plot
write.gif( volcano, "volcano.gif", col=terrain.colors, flip=TRUE,
          scale="always", comment="Maunga Whau Volcano")
y = read.gif("volcano.gif", verbose=TRUE, flip=TRUE)
image(y$image, col=y$col, main=y$comment, asp=1)
# browseURL("file://volcano.gif") # inspect GIF file on your hard disk

# test reading & writing
col = heat.colors(256) # choose colormap
trn = 222              # set transparent color
com = "Hello World"    # imbed comment in the file
write.gif( volcano, "volcano.gif", col=col, transparent=trn, comment=com)
y = read.gif("volcano.gif")
stopifnot(volcano==y$image, col==y$col, trn==y$transparent, com==y$comment)
# browseURL("file://volcano.gif") # inspect GIF file on your hard disk

# create simple animated GIF (using image function in a loop is very rough,
# but only way I know of displaying 'animation' in R)
x <- y <- seq(-4*pi, 4*pi, len=200)
r <- sqrt(outer(x^2, y^2, "+"))
image = array(0, c(200, 200, 10))
for(i in 1:10) image[, , i] = cos(r-(2*pi*i/10))/(r^.25)
write.gif(image, "wave.gif", col="rainbow")
y = read.gif("wave.gif")
for(i in 1:10) image(y$image[, , i], col=y$col, breaks=(0:256)-0.5, asp=1)
# browseURL("file://wave.gif") # inspect GIF file on your hard disk

# Another neat animation of Mandelbrot Set
jet.colors = colorRampPalette(c("#00007F", "blue", "#007FFF", "cyan", "#7FFF7F",
                                "yellow", "#FF7F00", "red", "#7F0000")) # define "jet" palette
m = 400
C = complex( real=rep(seq(-1.8,0.6, length.out=m), each=m ),
             imag=rep(seq(-1.2,1.2, length.out=m),      m ) )
C = matrix(C,m,m)
Z = 0
X = array(0, c(m,m,20))
for (k in 1:20) {
  Z = Z^2+C
  X[, , k] = exp(-abs(Z))
}
image(X[, , k], col=jet.colors(256))
write.gif(X, "Mandelbrot.gif", col=jet.colors, delay=100)
# browseURL("file://Mandelbrot.gif") # inspect GIF file on your hard disk
file.remove("wave.gif", "volcano.gif", "Mandelbrot.gif")

# Display interesting images from the web
## Not run:
url = "http://www.ngdc.noaa.gov/seg/cdroms/ged_iib/datasets/b12/gifs/eccnv.gif"
y = read.gif(url, verbose=TRUE, flip=TRUE)
image(y$image, col=y$col, breaks=(0:length(y$col))-0.5, asp=1,
      main="January Potential Evapotranspiration mm/mo")
url = "http://www.ngdc.noaa.gov/seg/cdroms/ged_iib/datasets/b01/gifs/fvvcode.gif"
y = read.gif(url, flip=TRUE)
y$col[y$transparent+1] = NA # mark transparent color in R way
image(y$image, col=y$col[1:87], breaks=(0:87)-0.5, asp=1,

```

```

        main="Vegetation Types")
url = "http://talc.geo.umn.edu/people/grads/hasba002/erosion_vids/run2/r2_dems_5fps(8col
y = read.gif(url, verbose=TRUE, flip=TRUE)
for(i in 2:dim(y$image)[3])
  image(y$image[,i], col=y$col, breaks=(0:length(y$col))-0.5,
        asp=1, main="Erosion in Drainage Basins")
## End(Not run)

```

---

LogitBoost

*LogitBoost Classification Algorithm*


---

## Description

Train logitboost classification algorithm using decision stumps (one node decision trees) as weak learners.

## Usage

```
LogitBoost(xlearn, ylearn, nIter=ncol(xlearn))
```

## Arguments

<code>xlearn</code>	A matrix or data frame with training data. Rows contain samples and columns contain features
<code>ylearn</code>	Class labels for the training data samples. A response vector with one label for each row/component of <code>xlearn</code> . Can be either a factor, string or a numeric vector.
<code>nIter</code>	An integer, describing the number of iterations for which boosting should be run, or number of decision stumps that will be used.

## Details

The function was adapted from `logitboost.R` function written by Marcel Dettling. See references and "See Also" section. The code was modified in order to make it much faster for very large data sets. The speed-up was achieved by implementing a internal version of decision stump classifier instead of using calls to `rpart`. That way, some of the most time consuming operations were precomputed once, instead of performing them at each iteration. Another difference is that training and testing phases of the classification process were split into separate functions.

## Value

An object of class "LogitBoost" including components:

<code>Stump</code>	<p>List of decision stumps (one node decision trees) used:</p> <ul style="list-style-type: none"> <li>column 1: feature numbers or each stump, or which column each stump operates on</li> <li>column 2: threshold to be used for that column</li> <li>column 3: bigger/smaller info: 1 means that if values in the column are above threshold than corresponding samples will be labeled as <code>lablist[1]</code>. Value "-1" means the opposite.</li> </ul> <p>If there are more than two classes, than several "Stumps" will be <code>cbind</code>'ed</p>
<code>lablist</code>	names of each class



**Author(s)**

Jarek Tuszynski (SAIC) <jaroslaw.w.tuszynski@saic.com>

**References**

Dettling and Buhlmann (2002), *Boosting for Tumor Classification of Gene Expression Data*, available on the web page <http://stat.ethz.ch/~dettling/boosting.html>.

<http://www.cs.princeton.edu/~schapire/boost.html>

**See Also**

- `predict.LogitBoost` has prediction half of LogitBoost code
- `logitboost` function from **boost** library
- `logitboost` function from **logitboost** library (not in CRAN or BioConductor but can be found at <http://stat.ethz.ch/~dettling/boosting.html>) is very similar but much slower on very large datasets. It also perform optional cross-validation.

**Examples**

```
data(iris)
Data = iris[,-5]
Label = iris[, 5]

# basic interface
model = LogitBoost(Data, Label, nIter=20)
Lab = predict(model, Data)
Prob = predict(model, Data, type="raw")
t = cbind(Lab, Prob)
t[1:10, ]

# two alternative call syntax
p=predict(model,Data)
q=predict.LogitBoost(model,Data)
pp=p[!is.na(p)]; qq=q[!is.na(q)]
stopifnot(pp == qq)

# accuracy increases with nIter (at least for train set)
table(predict(model, Data, nIter= 2), Label)
table(predict(model, Data, nIter=10), Label)
table(predict(model, Data),
      Label)

# example of splitting the data into train and test set
mask = sample.split(Label)
model = LogitBoost(Data[mask,], Label[mask], nIter=10)
table(predict(model, Data[!mask,], nIter=2), Label[!mask])
table(predict(model, Data[!mask,]),
      Label[!mask])
```

base64

*Convert R vectors to/from the Base64 format***Description**

Convert R vectors of any type to and from the Base64 format for encrypting any binary data as string using alphanumeric subset of ASCII character set.

**Usage**

```
z = base64encode(x, size=NA, endian=.Platform$endian)
x = base64decode(z, what, size=NA, signed = TRUE, endian=.Platform$endian)
```

**Arguments**

x	vector or any structure that can be converted to a vector by <a href="#">as.vector</a> function. Strings are also allowed.
z	String with Base64 code, using [A-Z,a-z,0-9,+,/,=] subset of characters
what	Either an object whose mode will give the mode of the vector to be created, or a character vector of length one describing the mode: one of "numeric", "double", "integer", "int", "logical", "complex", "character", "raw". Same as variable <code>what</code> in <a href="#">readBin</a> functions.
size	integer. The number of bytes per element in the byte stream stored in <code>r</code> . The default, 'NA', uses the natural size. Same as variable <code>size</code> in <a href="#">readBin</a> functions.
signed	logical. Only used for integers of sizes 1 and 2, when it determines if the quantity stored as raw should be regarded as a signed or unsigned integer. Same as variable <code>signed</code> in <a href="#">readBin</a> functions.
endian	If provided, can be used to swap endian-ness. Using "swap" will force swapping of byte order. Use "big" (big-endian, aka IEEE, aka "network") or "little" (little-endian, format used on PC/Intel machines) to indicate type of data encoded in "raw" format. Same as variable <code>endian</code> in <a href="#">readBin</a> functions.

**Details**

The Base64 encoding is designed to encode arbitrary binary information for transmission by electronic mail. It is defined by MIME (Multipurpose Internet Mail Extensions) specification RFC 1341, RFC 1421, RFC 2045 and others. Triplets of 8-bit octets are encoded as groups of four characters, each representing 6 bits of the source 24 bits. Only a 65-character subset ([A-Z,a-z,0-9,+,/,=]) present in all variants of ASCII and EBCDIC is used, enabling 6 bits to be represented per printable character.

Default sizes for different types of `what`: logical - 4, integer - 4, double - 8, complex - 16, character - 2, raw - 1.

**Value**

Function [base64encode](#) returns a string with Base64 code. Function [base64decode](#) returns vector of appropriate mode and length (see `x` above).

**Author(s)**

Jarek Tuszynski (SAIC) <jaroslav.w.tuszynski@saic.com>

**References**

- Base64 description in *Connected: An Internet Encyclopedia* <http://www.freesoft.org/CIE/RFC/1521/7.htm>
- MIME RFC 1341 <http://www.faqs.org/rfcs/rfc1341.html>
- MIME RFC 1421 <http://www.faqs.org/rfcs/rfc1421.html>
- MIME RFC 2045 <http://www.faqs.org/rfcs/rfc2045.html>
- Portions of the code are based on Matlab code by Peter Acklam <http://home.online.no/~pjacklam/matlab/software/util/datautil/>

**See Also**

`xmlValue` from **XML** package reads XML code which sometimes is encoded in Base64 format.  
[`readBin`](#), [`writeBin`](#)

**Examples**

```
x = (10*runif(10)>5) # logical
for (i in c(NA, 1, 2, 4)) {
  y = base64encode(x, size=i)
  z = base64decode(y, typeof(x), size=i)
  stopifnot(x==z)
}
print("Checked base64 for encode/decode logical type")

x = as.integer(1:10) # integer
for (i in c(NA, 1, 2, 4)) {
  y = base64encode(x, size=i)
  z = base64decode(y, typeof(x), size=i)
  stopifnot(x==z)
}
print("Checked base64 encode/decode for integer type")

x = (1:10)*pi # double
for (i in c(NA, 4, 8)) {
  y = base64encode(x, size=i)
  z = base64decode(y, typeof(x), size=i)
  stopifnot(mean(abs(x-z))<1e-5)
}
print("Checked base64 for encode/decode double type")

x = log(as.complex(-(1:10)*pi)) # complex
y = base64encode(x)
z = base64decode(y, typeof(x))
stopifnot(x==z)
print("Checked base64 for encode/decode complex type")

x = "Chance favors the prepared mind" # character
y = base64encode(x)
z = base64decode(y, typeof(x))
stopifnot(x==z)
```

```
print("Checked base64 for encode/decode character type")
```

---

bin2raw & raw2bin *OBSOLETE. Convert R vectors to/from the raw binary format.*

---

## Description

OBSOLETE FUNCTIONS TO BE RETIRED IN THE NEXT VERSION OF THE LIBRARY.  
Convert R vectors of any type to and from the raw binary format, stored as vector of type "raw".

## Usage

```
r = bin2raw(x, size=NA, endian=.Platform$endian)
x = raw2bin(r, what, size=NA, signed = TRUE, endian=.Platform$endian)
```

## Arguments

x	vector or any structure that can be converted to a vector by <a href="#">as.vector</a> function. Strings are also allowed.
r	vector of type "raw"
what	Either an object whose mode will give the mode of the vector to be created, or a character vector of length one describing the mode: one of "numeric", "double", "integer", "int", "logical", "complex", "character", "raw". Same as variable what in <a href="#">readBin</a> and <a href="#">base64decode</a> functions.
size	integer. The number of bytes per element in the byte stream stored in r. The default, 'NA', uses the natural size. See details.
signed	logical. Only used for integers of sizes 1 and 2, when it determines if the quantity stored as raw should be regarded as a signed or unsigned integer.
endian	If provided, can be used to swap endian-ness. Using "swap" will force swapping of byte order. Use "big" (big-endian, aka IEEE, aka "network") or "little" (little-endian, format used on PC/Intel machines) to indicate type of data encoded in "raw" format.

## Details

In R-2.2.0 version [readBin](#) functions [writeBin](#) were modified, making bin2raw and raw2bin functions mostly obsolete. As a result both function will be removed in the next version. Function `writeBin(x, raw(), ...)` does exactly the same as `bin2raw(x, ...)`. Function `raw2bin(r, what, size=size, ...)` is implemented as `readBin(r, what, n=length(r)%/size, size=size, ...)`, assuming size is not NA.

## Value

Function `bin2raw` returns vector of raw values (see `r` above), where each 1-byte raw value correspond to 1-byte of 1-byte of the binary form of other types. Length of the vector is going to be "number of bytes of a single element in array x" times `length(x)`.

Function `raw2bin` returns vector of appropriate mode and length (see `x` above), where each 1-byte raw value correspond to 1-byte of the binary form of other types. Length of the vector is going to be number of bytes per element in array x times `length(x)`. If parameter `what` is equal to "character" than a string (of length 1) is returned instead of vector of characters.

**Author(s)**

Jarek Tuszynski (SAIC) <jaroslaw.w.tuszynski@saic.com>

**See Also**

[readBin](#), [writeBin](#)

**Examples**

```

print(x <- (1:5)*pi)
print(y <- bin2raw(x))
print(z <- raw2bin(y,"double"))

x = (10*runif(10)>5) # logical
for (i in c(NA, 1, 2, 4)) {
  y = bin2raw(x, size=i)
  z = raw2bin(y,typeof(x), size=i)
  stopifnot(x==z)
}
print("Checked bin2raw and raw2bin conversion for logical type")

x = as.integer(1:10) # integer
for (i in c(NA, 1, 2, 4)) {
  y = bin2raw(x, size=i)
  z = raw2bin(y,typeof(x), size=i)
  stopifnot(x==z)
}
print("Checked bin2raw and raw2bin conversion for integer type")

x = (1:10)*pi # double
for (i in c(NA, 4, 8)) {
  y = bin2raw(x, size=i)
  z = raw2bin(y,typeof(x), size=i)
  stopifnot(mean(abs(x-z))<1e-5)
}
print("Checked bin2raw and raw2bin conversion for double type")

x = log(as.complex(-(1:10)*pi)) # complex
y = bin2raw(x)
z = raw2bin(y,typeof(x))
stopifnot(x==z)
print("Checked bin2raw and raw2bin conversion for complex type")

x = "Chance favors the prepared mind" # character
y = bin2raw(x)
z = raw2bin(y,typeof(x))
stopifnot(x==z)
print("Checked bin2raw and raw2bin conversion for character type")

x=(1:10000000)*pi
system.time(raw2bin(bin2raw(x),typeof(x)))
system.time(readBin(writeBin(x, raw()), typeof(x), length(x)))

```

colAUC

*Columnwise Area Under ROC Curve (AUC)***Description**

Calculate Area Under the ROC Curve (AUC) for every column of a matrix. Also, can be used to plot the ROC curves.

**Usage**

```
auc = colAUC(X, y, plotROC=FALSE, alg=c("Wilcoxon", "ROC"))
```

**Arguments**

<code>X</code>	A matrix or data frame. Rows contain samples and columns contain features/variables.
<code>y</code>	Class labels for the X data samples. A response vector with one label for each row/component of X. Can be either a factor, string or a numeric vector.
<code>plotROC</code>	Plot ROC curves. Use only for small number of features. If TRUE, will set <code>alg</code> to "ROC".
<code>alg</code>	algorithm to use: "ROC" integrates ROC curves, while "Wilcoxon" uses Wilcoxon Rank Sum Test to get the same results. Default "Wilcoxon" is faster. This argument is mostly provided for verification.

**Details**

AUC is a very useful measure of similarity between two classes measuring area under "Receiver Operating Characteristic" or ROC curve. In case of data with no ties all sections of ROC curve are either horizontal or vertical, in case of data with ties diagonal sections can also occur. Area under the ROC curve is calculated using `trapz` function. AUC is always in between 0.5 (two classes are statistically identical) and 1.0 (there is a threshold value that can achieve a perfect separation between the classes).

Area under ROC Curve (AUC) measure is very similar to Wilcoxon Rank Sum Test (see [wilcox.test](#)) and Mann-Whitney U Test.

There are numerous other functions for calculating AUC in other packages. Unfortunately none of them had all the properties I needed for use as classification preprocessing, to lower the dimensionality of the data (from tens of thousands to hundreds) before applying standard classification algorithms.

The main properties of this code are:

- Ability to work with multi-dimensional data (X can have many columns).
- Ability to work with multi-class datasets (y can have more than 2 different values).
- Speed - this code was written to calculate AUC's of large number of features, fast.
- Returned AUC is always bigger than 0.5, which is equivalent of testing for each feature `colAUC(x, y)` and `colAUC(-x, y)` and returning the value of the bigger one.

If those properties do not fit your problem, see "See Also" and "Examples" sections for AUC functions in other packages that might be a better fit for your needs.

**Value**

An output is a single matrix with the same number of columns as `X` and "n choose 2" ( $\frac{n!}{(n-2)!2!}$ ) number of rows, where `n` is number of unique labels in `y` list. For example, if `y` contains only two unique class labels (`length(unique(lab))==2`) than output matrix will have a single row containing AUC of each column. If more than two unique labels are present than AUC is calculated for every possible pairing of classes ("n choose 2" of them).

**Author(s)**

Jarek Tuszynski (SAIC) <jaroslav.w.tuszynski@saic.com>

**References**

- Mason, S.J. and Graham, N.E. (1982) *Areas beneath the relative operating characteristics (ROC) and relative operating levels (ROL) curves: Statistical significance and interpretation*, Q. J. R. Meteorol. Soc. textbf30 291-303.
- See <http://www.medicine.mcgill.ca/epidemiology/hanley/software/> to find articles below:
  - Hanley and McNeil (1982), *The Meaning and Use of the Area under a Receiver Operating Characteristic (ROC) Curve*, Radiology 143: 29-36.
  - Hanley and McNeil (1983), *A Method of Comparing the Areas under ROC curves derived from same cases*, Radiology 148: 839-843.
  - McNeil and Hanley (1984), *Statistical Approaches to the Analysis of ROC curves*, Medical Decision Making 4(2): 136-149.
- See [http://rocr.bioinf.mpi-sb.mpg.de/evaluation\\_literature.html](http://rocr.bioinf.mpi-sb.mpg.de/evaluation_literature.html) for bibliography of **ROCR** package.

**See Also**

- [wilcox.test](#) and [pwilcox](#)
- [AUC](#) from **ROC** package
- [performance](#) from **ROCR** package
- [auROC](#) from **limma** package
- [ROC](#) from **Epi** package
- [roc.area](#) from **verification** package
- [rcorr.cens](#) from **Hmisc** package

**Examples**

```
# Load MASS library with "cats" data set that have following columns: sex, body
# weight, hart weight. Calculate how good weights are in predicting sex of cats.
# 2 classes; 2 features; 144 samples
library(MASS); data(cats);
colAUC(cats[,2:3], cats[,1], plotROC=TRUE)

# Load rpart library with "kyphosis" data set that records if kyphosis
# deformation was present after corrective surgery. Calculate how good age,
# number and position of vertebrae are in predicting succesful operation.
# 2 classes; 3 features; 81 samples
library(rpart); data(kyphosis);
colAUC(kyphosis[,2:4], kyphosis[,1], plotROC=TRUE)
```

```

# Example of 3-class 4-feature 150-sample iris data
data(iris)
colAUC(iris[,-5], iris[,5], plotROC=TRUE)

# Compare calAUC with other functions designed for similar purpose
auc = matrix(NA,10,3)
rownames(auc) = c("colAUC(alg='ROC')", "colAUC(alg='Wilcox')", "wilcox.test",
  "sum(rank)", "roc.area", "AUC", "performance", "ROC", "auROC", "rcorr.cens")
colnames(auc) = c("AUC(x)", "AUC(-x)", "AUC(x+noise)")
X = cbind(cats[,2], -cats[,2], cats[,2]+rnorm(nrow(cats)))
y = ifelse(cats[,1]=='F',0,1)
for (i in 1:3) {
  x = X[,i]
  x1 = x[y==1]; n1 = length(x1); # prepare input data ...
  x2 = x[y==0]; n2 = length(x2); # ... into required format
  auc[1,i] = colAUC(x, y, alg="ROC")
  auc[2,i] = colAUC(x, y, alg="Wilcox")
  auc[3,i] = wilcox.test(x1, x2, exact=0)$statistic / (n1*n2)
  r = rank(c(x1,x2))
  auc[4,i] = (sum(r[1:n1]) - n1*(n1+1)/2) / (n1*n2)
  if (require("verification"))
    auc[5,i] = roc.area(y, x)$A.tilda
  if (require("ROC"))
    auc[6,i] = AUC(rocdemo.sca(y, x, dxrule.sca))
  if (require("ROCR"))
    auc[7,i] = performance(prediction( x, y),"auc")@y.values[[1]]
  if (require("Epi")) auc[8,i] = ROC(x,y,grid=0)$AUC
  if (require("limma")) auc[9,i] = auROC(y, x)
  if (require("Hmisc")) auc[10,i] = rcorr.cens(x, y)[1]
}
print(auc)
stopifnot(auc[1, ]==auc[2, ]) # results of 2 alg's in colAUC must be the same
stopifnot(auc[1,1]==auc[3,1]) # compare with wilcox.test results

# time trials
x = matrix(runif(100*1000),100,1000)
y = (runif(100)>0.5)
system.time(colAUC(x,y,alg="ROC" ))
system.time(colAUC(x,y,alg="Wilcox"))

```

combs

*All Combinations of k Elements from Vector v***Description**

Finds all unordered combinations of k elements from vector v.

**Usage**

```
combs(v,k)
```



**Arguments**

<code>v</code>	Any numeric vector
<code>k</code>	Number of elements to choose from vector <code>v</code> . Integer smaller or equal than length of <code>v</code> .

**Value**

`combs(v,k)` (where `v` has length `n`) creates a matrix with  $\frac{n!}{(n-k)!k!}$  (`n choose k`) rows and `k` columns containing all possible combinations of `n` elements taken `k` at a time.

**Author(s)**

Jarek Tuszynski (SAIC) <jaroslav.w.tuszynski@saic.com>

**See Also**

I discovered recently that R packages already have two functions with similar capabilities: [combinations](#) from **gTools** package and [nchoosek](#) from **vsN** package. Also similar to Matlab's `nchoosek` function (<http://www.mathworks.com/access/helpdesk/help/techdoc/ref/nchoosek.html>)

**Examples**

```
combs(2:5, 3) # display examples
combs(c("cats", "dogs", "mice"), 2)

a = combs(1:4, 2)
b = matrix( c(1,1,1,2,2,3,2,3,4,3,4,4), 6, 2)
stopifnot(a==b)
```

---

`predict.LogitBoost` *Prediction Based on LogitBoost Classification Algorithm*

---

**Description**

Prediction or Testing using logitboost classification algorithm

**Usage**

```
predict.LogitBoost(object, xtest, type = c("class", "raw"), nIter=NA, ...)
```

**Arguments**

<code>object</code>	An object of class "LogitBoost" see "Value" section of <a href="#">LogitBoost</a> for details
<code>xtest</code>	A matrix or data frame with test data. Rows contain samples and columns contain features
<code>type</code>	See "Value" section
<code>nIter</code>	An optional integer, used to lower number of iterations (decision stumps) used in the decision making. If not provided than the number will be the same as the one provided in <a href="#">LogitBoost</a> . If provided than the results will be the same as running <a href="#">LogitBoost</a> with fewer iterations.
<code>...</code>	not used but needed for compatibility with generic predict method

## Details

Logitboost algorithm relies on a voting scheme to make classifications. Many (`nIter` of them) weak classifiers are applied to each sample and their findings are used as votes to make the final classification. The class with the most votes "wins". However, with this scheme it is common for two cases have a tie (the same number of votes), especially if number of iterations is even. In that case NA is returned, instead of a label.

## Value

If `type = "class"` (default) label of the class with maximal probability is returned for each sample. If `type = "raw"`, the a-posterior probabilities for each class are returned.

## Author(s)

Jarek Tuszynski (SAIC) ([jaroslaw.w.tuszynski@saic.com](mailto:jaroslaw.w.tuszynski@saic.com))

## See Also

[LogitBoost](#) has training half of LogitBoost code

## Examples

```
# See LogitBoost example
```

---

runfunc

---

Moving Window Analysis of a Vector

---

## Description

A collection of functions to perform moving window (running, rolling window) analysis of vectors

## Usage

```
runmean(x, k, alg=c("C", "R", "exact"),
        endrule=c("NA", "trim", "keep", "constant", "func"))
runmin (x, k, alg=c("C", "R"),
        endrule=c("NA", "trim", "keep", "constant", "func"))
runmax (x, k, alg=c("C", "R"),
        endrule=c("NA", "trim", "keep", "constant", "func"))
runmad (x, k, center=runmed(x,k,endrule="keep"), constant=1.4826,
        endrule=c("NA", "trim", "keep", "constant", "func"))
runquantile(x, k, probs, type=7,
            endrule=c("NA", "trim", "keep", "constant", "func"))
EndRule(x, y, k,
        endrule=c("NA", "trim", "keep", "constant", "func"), Func, ...)
```

## Arguments

x	numeric vector of length n
k	width of moving window; must be an odd integer between three and n
endrule	<p>character string indicating how the values at the beginning and the end, of the data, should be treated. Only first and last <math>k/2</math> values at both ends are affected, where <math>k/2</math> is the half-bandwidth <math>k/2 = k \% 2</math>.</p> <ul style="list-style-type: none"> <li>"trim" - trim the ends output array length is equal to <math>\text{length}(x) - 2 * k/2</math> (<math>\text{out} = \text{out}[(k/2+1):(n-k/2)]</math>). This option mimics output of <code>apply(embed(x,k),1,FUN)</code> and other related functions.</li> <li>"keep" - fill the ends with numbers from x vector (<math>\text{out}[1:k/2] = x[1:k/2]</math>)</li> <li>"constant" - fill the ends with first and last calculated value in output array (<math>\text{out}[1:k/2] = \text{out}[k/2+1]</math>)</li> <li>"NA" - fill the ends with NA's (<math>\text{out}[1:k/2] = \text{NA}</math>)</li> <li>"func" - applies the underlying function to smaller and smaller sections of the array. For example in case of mean: <code>for(i in 1:k/2) out[i]=mean(x[1:i])</code>. This option is not optimized and could be very slow for large windows.</li> </ul> <p>Similar to endrule in <code>runmed</code> function which has the following options: <code>c("median", "keep", "constant")</code>.</p>
alg	an option allowing to choose different algorithms or implementations, if provided. Default is to use of code written in C. Option <code>alg="R"</code> will use slower code written in R. Usefull for debugging and allows extentions in the future.
center	moving window center used by <code>runmad</code> function defaults to running median ( <code>runmed</code> function). Similar to center in <code>mad</code> function.
constant	scale factor used by <code>runmad</code> , such that for gaussian distribution X, <code>mad(X)</code> is the same as <code>sd(X)</code> . Same as constant in <code>mad</code> function.
probs	numeric vector of probabilities with values in [0,1] range used by <code>runquantile</code> . For example <code>Probs=c(0,0.5,1)</code> would be equivalent to running <code>runmin</code> , <code>runmed</code> and <code>runmax</code> . Same as probs in <code>quantile</code> function.
type	an integer between 1 and 9 selecting one of the nine quantile algorithms, same as type in <code>quantile</code> function. Another even more readable description of nine ways to calculate quantiles can be found at <a href="http://mathworld.wolfram.com/Quantile.html">http://mathworld.wolfram.com/Quantile.html</a> .
y	numeric vector of length n, which is partially filled output of one of the run functions. Function <code>EndRule</code> will fill the remaining beginning and end sections using method chosen by endrule argument.
Func	Function name that <code>EndRule</code> will use in case of <code>endrule="func"</code> .
...	Additional parameters to Func that <code>EndRule</code> will use in case of <code>endrule="func"</code> .

## Details

Apart from the end values, the result of `y = runFUN(x, k)` is the same as `"for ( j=(1+k/2):(n-k/2) ) y[j]=FUN(x[(j-k/2):(j+k/2)])"`, where FUN stands for min, max, mean, mad or quantile functions.

The main incentive to write this set of functions was relative slowness of majority of moving window functions available in R and its packages. With exception of `runmed`, a running window median function, all functions listed in "see also" section are slower than very inefficient `"apply(apply(x,k),1,FUN)"` approach. Relative speeds of above functions are as follow:

- `runmin`, `runmax`, `runmean` run at  $O(n)$
- `runmean(..., alg="exact")` can have worst case speed of  $O(n^2)$  for some small data vectors, but average case is still close to  $O(n)$ .
- `runquantile` and `runmad` run at  $O(n*k)$
- `runmed` - related R function run at  $O(n*\log(k))$

Functions `runquantile` and `runmad` are using insertion sort to sort the moving window, but gain speed by remembering results of the previous sort. Since each time the window is moved, only one point changes, all but one points in the window are already sorted. Insertion sort can fix that in  $O(k)$  time.

Function `runquantile` when run in single probability mode automatically recognizes probabilities: 0, 1/2, and 1 as special cases and return output from functions: `runmin`, `runmed` and `runmax` respectively.

All `run*` functions are written in C, but `runmin`, `runmax` and `runmean` also have fast R code versions (see argument `alg="R"`). Those were included for debugging purposes, and as a fallback in hard-to-port situations. See examples.

Function `EndRule` applies one of the five methods (see `endrule` argument) to process end-points of the input array `x`.

In case of `runmean(..., alg="exact")` function a special algorithm is used (see references section) to ensure that round-off errors do not accumulate. As a result `runmean` is more accurate than `filter(x, rep(1/k,k))` and `runmean(..., alg="C")` functions.

All of the functions in this section do not work with infinite numbers (NA, NaN, Inf, -Inf) except for `runmean(..., alg="exact")` which omits them.

## Value

Functions `runmin`, `runmax`, `runmean` and `runmad` return a numeric vector of the same length as `x`. Function `runquantile` returns a matrix of size  $[n \times \text{length(probs)}]$ . In addition `x` contains attribute `k` with (the 'oddified') `k`.

## Note

Function `runmean(..., alg="exact")` is based by code by Vadim Ogranovich, which is based on Python code (see last reference), pointed out by Gabor Grothendieck.

## Author(s)

Jarek Tuszynski (SAIC) <jaroslav.w.tuszynski@saic.com>

## References

- About quantiles: Hyndman, R. J. and Fan, Y. (1996) *Sample quantiles in statistical packages*, *American Statistician*, 50, 361.
- About quantiles: Eric W. Weisstein. *Quantile*. From MathWorld— A Wolfram Web Resource. <http://mathworld.wolfram.com/Quantile.html>
- About insertion sort used in `runmad` and `runquantile`: R. Sedgewick (1988): *Algorithms*. Addison-Wesley (page 99)
- About round-off error correction used in `runmean`: Shewchuk, Jonathan *Adaptive Precision Floating-Point Arithmetic and Fast Robust Geometric Predicates*, <http://www-2.cs.cmu.edu/afs/cs/project/quake/public/papers/robust-arithmetic.ps>

- More on round-off error correction can be found at: <http://aspn.activestate.com/ASPN/Cookbook/Python/Recipe/393090>

## See Also

Links related to each function:

- `runmean` - `mean`, `kernapply`, `filter`, `runsum.exact`, `decompose`, `stl`, `rollMean` from `fSeries` library, `rollmean` from `zoo` library, `subsums` from `magic` library,
- `runmin` - `min`, `rollMin` from `fSeries` library
- `runmax` - `max`, `rollMax` from `fSeries` library, `rollmax` from `zoo` library
- `runquantile` - `quantile`, `runmed`, `smooth`, `rollmedian` from `zoo` library
- `runmad` - `mad`, `rollVar` from `fSeries` library
- generic running window functions: `apply` (`embed(x,k)`, `1`, `FUN`) (fastest), `rollFun` from `fSeries` (slow), `running` from `gtools` package (extremely slow for this purpose), `rapply` from `zoo` library, `subsums` from `magic` library can perform running window operations on data with any dimensions.
- `EndRule` - `smoothEnds(y,k)` function is similar to `EndRule(x,y,k,endrule="func", median)`

## Examples

```
# test runmin, runmax and runmed
k=15; n=200;
x = rnorm(n,sd=30) + abs(seq(n)-n/4)
col = c("black", "red", "green", "blue", "magenta", "cyan")
plot(x, col=col[1], main = "Moving Window Analysis Functions")
lines(runmin(x,k), col=col[2])
lines(runmed(x,k), col=col[3])
lines(runmax(x,k), col=col[4])
legend(0,.9*n, c("data", "runmin", "runmed", "runmax"), col=col, lty=1 )

#test runmean and runquantile
y=runquantile(x, k, probs=c(0, 0.5, 1, 0.25, 0.75), endrule="constant")
plot(x, col=col[1], main = "Moving Window Quantile")
lines(runmean(y[,1],k), col=col[2])
lines(y[,2], col=col[3])
lines(runmean(y[,3],k), col=col[4])
lines(y[,4], col=col[5])
lines(y[,5], col=col[6])
lab = c("data", "runmean(runquantile(0))", "runquantile(0.5)",
"runmean(runquantile(1))", "runquantile(.25)", "runquantile(.75)")
legend(0,0.9*n, lab, col=col, lty=1 )

#test runmean and runquantile
k =25
m=runmed(x, k)
y=runmad(x, k, center=m)
plot(x, col=col[1], main = "Moving Window Analysis Functions")
lines(m, col=col[2])
lines(m-y/2, col=col[3])
lines(m+y/2, col=col[4])
lab = c("data", "runmed", "runmed-runmad/2", "runmed+runmad/2")
legend(0,1.8*n, lab, col=col, lty=1 )
```

```

# numeric comparison between different algorithms
numeric.test = function (n, k) {
  eps = .Machine$double.eps ^ 0.5
  x = rnorm(n,sd=30) + abs(seq(n)-n/4)
  # numeric comparison : runmean
  a = runmean(x,k)
  b = runmean(x,k, alg="R")
  d = runmean(x,k, alg="exact")
  e = filter(x, rep(1/k,k))
  stopifnot(all(abs(a-b)<eps, na.rm=TRUE));
  stopifnot(all(abs(a-d)<eps, na.rm=TRUE));
  stopifnot(all(abs(a-e)<eps, na.rm=TRUE));
  # numeric comparison : runmin
  a = runmin(x,k, endrule="trim")
  b = runmin(x,k, endrule="trim", alg="R")
  c = apply(embed(x,k), 1, min)
  stopifnot(all(a==b, na.rm=TRUE));
  stopifnot(all(a==c, na.rm=TRUE));
  # numeric comparison : runmax
  a = runmax(x,k, endrule="trim")
  b = runmax(x,k, endrule="trim", alg="R")
  c = apply(embed(x,k), 1, max)
  stopifnot(all(a==b, na.rm=TRUE));
  stopifnot(all(a==c, na.rm=TRUE));
  # numeric comparison : runmad
  a = runmad(x,k, endrule="trim")
  b = apply(embed(x,k), 1, mad)
  stopifnot(all(a==b, na.rm=TRUE));
  # numeric comparison : runquantile
  a = runquantile(x,k, c(0.3, 0.7), endrule="trim")
  b = t(apply(embed(x,k), 1, quantile, probs = c(0.3, 0.7)))
  stopifnot(all(abs(a-b)<eps));
}

numeric.test(50, 3) # test different window size vs. vector ...
numeric.test(50,15) # ... length combinations
numeric.test(50,49)
numeric.test(49,49)

# speed comparison
x=runif(100000); k=991;
system.time(runmean(x,k))
system.time(runmean(x,k, alg="R"))
system.time(runmean(x,k, alg="exact"))
system.time(filter(x, rep(1/k,k), sides=2)) #the fastest alternative I know
k=91;
system.time(runmad(x,k))
system.time(apply(embed(x,k), 1, mad)) #the fastest alternative I know

# numerical comparison of round-off error handling
test.runmean = function (x, k) {
  a = k*runmean(x,k, alg="exact")
  b = k*runmean(x,k, alg="C")
  d = k*runmean(x,k, alg="R")
  e = k*filter(x, rep(1/k,k))
  f = k* c(NA, NA, apply(embed(x,k), 1, mean), NA, NA)
  x = cbind(x, a, b, d, e, f)

```

```

    colnames(x) = c("x", "runmean(alg=exact)", "runmean(alg=C)",
                    "runmean(alg=R)", "filter", "apply")
    return(x)
}
a = rep( c(1, 10, -10, -1, 0, 0, 0), 3) # nice-behaving array
b = rep( c(1, 10^20, -10^20, -1, 0, 0, 0), 3) # round-off error prone array
d = rep( c(1, 10^20, 10^40, -10^40, -10^20, -1, 0), 3)
test.runmean(a, 5) #all runmean algorithms give the same result
test.runmean(b, 5) #runmean(alg=R) gives wrong result
test.runmean(d, 5) #only runmean(alg=exact) gives correct result

```

sample.split

*Split Data into Test and Train Set***Description**

Split data from vector Y into two sets in predefined ratio while preserving relative ratios of different labels in Y. Used to split the data used during classification into train and test subsets.

**Usage**

```
sample.split( Y, SplitRatio = 2/3, group = NULL )
```

**Arguments**

Y	Vector of data labels. If there are only a few labels (as is expected) than relative ratio of data in both subsets will be the same.
SplitRatio	Splitting ratio: <ul style="list-style-type: none"> <li>• if (<math>0 \leq \text{SplitRatio} &lt; 1</math>) then SplitRatio fraction of points from Y will be set to TRUE</li> <li>• if (<math>\text{SplitRatio} == 1</math>) then one random point from Y will be set to TRUE</li> <li>• if (<math>\text{SplitRatio} &gt; 1</math>) then SplitRatio number of points from Y will be set to TRUE</li> </ul>
group	Optional vector/list used when multiple copies of each sample are present. In such a case group contains unique sample labels, marking all copies of the same sample with the same label, and the function tries to place all copies in either train or test subset. If provided than has to have the same length as Y.

**Details**

Function `msc.sample.split` is the old name of the `sample.split` function. To be retired soon.

**Value**

Returns logical vector of the same length as Y with random  $\text{SplitRatio} * \text{length}(Y)$  elements set to TRUE.

**Author(s)**

Jarek Tuszynski (SAIC) <jaroslaw.w.tuszynski@saic.com>

**See Also**

- Similar to [sample](#) function.
- Variable `group` is used in the same way as `f` argument in [split](#) and `INDEX` argument in [tapply](#)

**Examples**

```
library(MASS)
data(cats) # load cats data
Y = cats[,1] # extract labels from the data
msk = sample.split(Y, SplitRatio=3/4)
table(Y,msk)
t=sum( msk) # number of elements in one class
f=sum(!msk) # number of elements in the other class
stopifnot( round((t+f)*3/4) == t ) # test ratios

# example of using group variable
g = rep(seq(length(Y)/4), each=4); g[48]=12;
msk = sample.split(Y, SplitRatio=1/2, group=g)
table(Y,msk) # try to get correct split ratios ...
split(msk,g) # ... while keeping samples with the same group label together

# test results
print(paste( "All Labels numbers: total=",t+f,"", train="t,", test="f,
            ", ratio=", t/(t+f) ) )
U = unique(Y) # extract all unique labels
for( i in 1:length(U)) { # check for all labels
  lab = (Y==U[i]) # mask elements that have label U[i]
  t=sum( msk[lab]) # number of elements with label U[i] in one class
  f=sum(!msk[lab]) # number of elements with label U[i] in the other class
  print(paste( "Label",U[i],"numbers: total=",t+f,"", train="t,", test="f,
              ", ratio=", t/(t+f) ) )
}

# use results
train = cats[ msk,2:3] # use output of sample.split to ...
test = cats[!msk,2:3] # create train and test subsets
z = lda(train, Y[msk]) # perform classification
table(predict(z, test)$class, Y[!msk]) # predicted & true labels

# see also LogitBoost example
```

sum.exact

*Basic Sum Operations without Round-off Errors***Description**

Functions for performing basic sum operations without round-off errors

**Usage**

```
sum.exact(..., na.rm = FALSE)
cumsum.exact(x)
runsum.exact(x,k)
```



## Arguments

<code>x</code>	numeric vector
<code>...</code>	numeric vector(s), numbers or other objects to be summed
<code>na.rm</code>	logical. Should missing values be removed?
<code>k</code>	width of moving window; must be an odd integer between one and <code>n</code>

## Details

All three functions use full precision summation using multiple doubles for intermediate values. The sum of numbers `x` & `y` is `a=x+y` with error term `b=error(a+b)`. That way `a+b` is equal exactly `x+y`, so sum of 2 numbers is stored as 2 or fewer values, which when added would under-flow. By extension sum of `n` numbers is calculated with intermediate results stored as array of numbers that can not be added without introducing an error. Only final result is converted to a single number

## Value

Function `sum.exact` returns single number. Function `cumsum.exact` returns vector of the same length as `x`. Function `runsum.exact` returns vector of length `length(x)-k` and attribute "count" containing number of finite (as in `is.finite`) elements in each window.

## Author(s)

Jarek Tuszynski (SAIC) ([jaroslaw.w.tuszynski@saic.com](mailto:jaroslaw.w.tuszynski@saic.com)) based on code by Vadim Ogranovich, which is based on algorithms described in references, pointed out by Gabor Grothendieck.

## References

Round-off error correction is based on: Shewchuk, Jonathan, *Adaptive Precision Floating-Point Arithmetic and Fast Robust Geometric Predicates*, <http://www-2.cs.cmu.edu/afs/cs/project/quake/public/papers/robust-arithmetic.ps> and its implementation found at: <http://aspn.activestate.com/ASPN/Cookbook/Python/Recipe/393090>

McCullough, D.B., (1998) *Assessing the Reliability of Statistical Software, Part I*, The American Statistician, Vol. 52 No 4, <http://www.amstat.org/publications/tas/mccull-1.pdf>

McCullough, D.B., (1999) *Assessing the Reliability of Statistical Software, Part II*, The American Statistician, Vol. 53 No 2 <http://www.amstat.org/publications/tas/mccull.pdf>

NIST Statistical Reference Datasets (StRD) website <http://www.nist.gov/itl/div898/strd>

## See Also

- `sum.exact` - is equivalent to `sum`
- `cumsum.exact` - is equivalent to `cumsum`
- `runsum.exact` - is similar to `runmean(x,k,endrule="trim")`

## Examples

```
x = c(1, 1e20, 1e40, -1e40, -1e20, -1)
a = sum(x);      print(a)
b = sum.exact(x); print(b)
stopifnot(b==0)
```

```
a = cumsum(x);      print(a)
b = cumsum.exact(x); print(b)
stopifnot(b[6]==0)
```

---

trapz

---

*Trapezoid Rule Numerical Integration*


---

## Description

Computes the integral of Y with respect to X using trapezoid rule integration.

## Usage

```
trapz(x, y)
```

## Arguments

x	Sorted vector of x-axis values.
y	Vector of y-axis values.

## Details

The function has only two lines:

```
idx = 2:length(x)
return (as.double( (x[idx] - x[idx-1]) %*% (y[idx] + y[idx-1])) / 2)
```

## Value

Integral of Y with respect to X or area under the Y curve.

## Note

Trapezoid rule is not the most accurate way of calculating integrals (it is exact for linear functions), for example Simpson's rule (exact for linear and quadratic functions) is more accurate.

## Author(s)

Jarek Tuszynski (SAIC) <jaroslaw.w.tuszynski@saic.com>

## References

D. Kincaid & W. Chaney (1991), *Numerical Analysis*, p.445

## See Also

- [intg](#) from **PROcess** package
- [trapezint](#) from **ROC** package
- [integrate](#)
- Matlab's trapz function (<http://www.mathworks.com/access/helpdesk/help/techdoc/ref/trapz.html>)

**Examples**

```
# integral of sine function in [0, pi] range suppose to be exactly 2.  
# lets calculate it using 10 samples:  
x = (1:10)*pi/10  
trapz(x, sin(x))  
# now lets calculate it using 1000 samples:  
x = (1:1000)*pi/1000  
trapz(x, sin(x))
```

# Index

- \*Topic **array**
  - runfunc, 18
  - sum.exact, 24
- \*Topic **classif**
  - LogitBoost, 7
  - predict.LogitBoost, 17
  - sample.split, 22
- \*Topic **file**
  - base64, 9
  - bin2raw & raw2bin, 11
  - ENVI, 1
  - read.gif & write.gif, 3
- \*Topic **math**
  - trapz, 25
- \*Topic **models**
  - combs, 16
- \*Topic **smooth**
  - runfunc, 18
  - sum.exact, 24
- \*Topic **ts**
  - runfunc, 18
  - sum.exact, 24
- \*Topic **univar**
  - colAUC, 13
- \*Topic **utilities**
  - runfunc, 18
  - sum.exact, 24
- add.image, 3, 6
- apply, 18–20
- as.vector, 10, 12
- AUC, 15
- auROC, 15
- base64, 9
- base64decode, 10, 12
- base64decode (base64), 9
- base64encode, 10
- base64encode (base64), 9
- bin2raw(bin2raw & raw2bin), 11
- bin2raw & raw2bin, 11
- brewer.pal, 6
- col2rgb, 6
- colAUC, 13
- colorbrewer.palette, 6
- combinations, 16
- combs, 16
- cumsum, 25
- cumsum.exact (sum.exact), 24
- decompose, 20
- embed, 18, 20
- EndRule (runfunc), 18
- ENVI, 1
- filter, 19, 20
- GDD, 6
- gray, 6
- heat.colors, 4, 6
- hsv, 6
- HTMLplot, 6
- image, 3–6
- image.plot, 3, 6
- integrate, 26
- intg, 26
- is.finite, 24
- jpeg, 6
- kernapply, 20
- LogitBoost, 7, 17
- logitboost, 9
- mad, 19, 20
- max, 20
- mean, 20
- min, 20
- msc.sample.split (sample.split), 22
- nchoosek, 16
- performance, 15
- plot.im, 3, 6

png, 6  
predict.LogitBoost, 9, 17  
print, 4  
pwilcox, 15  
  
quantile, 19, 20  
  
rainbow, 4, 6  
rapply, 20  
raw2bin(bin2raw & raw2bin), 11  
rcorr.cens, 15  
read.ENVI, 6  
read.ENVI(ENVI), 1  
read.gif, 3  
read.gif(read.gif & write.gif), 3  
read.gif & write.gif, 3  
read.jpeg, 6  
read.pnm, 6  
readBin, 3, 10–12  
rgb, 6  
rich.colors, 6  
ROC, 15  
roc.area, 15  
rollFun, 20  
rollMax, 20  
rollmax, 20  
rollMean, 20  
rollmean, 20  
rollmedian, 20  
rollMin, 20  
rollVar, 20  
rpart, 8  
runfunc, 18  
runmad(runfunc), 18  
runmax(runfunc), 18  
runmean, 25  
runmean(runfunc), 18  
runmed, 18–20  
runmin(runfunc), 18  
running, 20  
runquantile(runfunc), 18  
runsum.exact, 20  
runsum.exact(sum.exact), 24  
  
sample, 23  
sample.split, 22  
sd, 19  
smooth, 20  
smoothEnds, 20  
split, 23  
stl, 20  
subsums, 20  
sum, 25  
  
sum.exact, 24  
  
tapply, 23  
terrain.colors.colors, 6  
tim.colors, 6  
topo.colors, 6  
trapezint, 26  
trapz, 14, 25  
  
wilcox.test, 14, 15  
write.ENVI, 6  
write.ENVI(ENVI), 1  
write.gif, 3  
write.gif(read.gif & write.gif), 3  
write.pnm, 6  
writeBin, 2, 3, 11, 12  
  
xmlValue, 11