Package ‘kknn’

February 20, 2015

Title Weighted k-Nearest Neighbors
Version 1.2-5
Date 2014-02-11
Author Klaus Schliep & Klaus Hechenbichler
Description Weighted k-Nearest Neighbors Classification, Regression and Clustering
Maintainer Klaus Schliep <klaus.schliep@gmail.com>
Depends R (>= 2.10)
Imports igraph (>= 0.6), Matrix, stats
License GPL (>= 2)
NeedsCompilation yes
Repository CRAN
Date/Publication 2014-02-11 18:13:13

R topics documented:

kknn-package .......................................................... 2
contr.dummy ......................................................... 2
glass ................................................................. 3
ionosphere ............................................................ 4
kknn ................................................................. 5
miete ................................................................. 7
simulation ........................................................... 9
specClust ............................................................ 10
train.kknn ........................................................... 12

Index 15
Weighted k-Nearest Neighbors Classification and Clustering

Description

Weighted k-Nearest Neighbors Classification, Regression and spectral Clustering

The complete list of functions can be displayed with library(help = kknn).

Author(s)

Klaus Schliep
Klaus Hechenbichler
Maintainer: Klaus Schliep <klaus.schliep@gmail.com>

References


Contrast Matrices

Description

Returns a matrix of contrasts.

Usage

contr.dummy(n, contrasts = TRUE)
contr.ordinal(n, contrasts = TRUE)
contr.metric(n, contrasts = TRUE)

Arguments

n A vector containing levels of a factor, or the number of levels.
contrasts A logical value indicating whether contrasts should be computed.

Details

contr.dummy is standard dummy-coding, contr.metric has the same effect like as.numeric (makes sense of course only for ordered variables). contr.ordinal computes contrasts for ordinal variables.
Value

A matrix with \( n \) rows and \( n-1 \) columns for contr.ordinal, a matrix with \( n \) rows and \( n \) columns for contr.dummy and a vector of length \( n \) for contr.metric.

Author(s)

Klaus P. Schliep <klaus.schliep@gmail.com>

References


See Also

contrasts, contr.poly and contr.sdif

Examples

contr.metric(5)
contr.ordinal(5)
contr.dummy(5)

glass

Glass Identification Database

Description

A data frame with 214 observations, where the problem is to predict the type of glass in terms of their oxide content (i.e. Na, Fe, K, etc). The study of classification of types of glass was motivated by criminological investigation. At the scene of the crime, the glass left can be used as evidence... if it is correctly identified!

Usage

data(glass)

Format

A data frame with 214 observations on the following 11 variables.

- **Id** Id number.
- **RI** Refractive index.
- **Na** Sodium (unit measurement: weight percent in corresponding oxide, as are attributes 4-10).
- **Mg** Magnesium.
- **Al** Aluminum.
Si  Silicon.
K  Potassium.
Ca  Calcium.
Ba  Barium.
Fe  Iron.

**Type**  Type of glass: (class attribute)
1 building windows float processed
2 building windows non float processed
3 vehicle windows float processed
4 vehicle windows non float processed (none in this database)
5 containers
6 tableware
7 headlamps

**Source**

- Creator: B. German, Central Research Establishment, Home Office Forensic Science Service, Aldermaston, Reading, Berkshire RG7 4PN
- Donor: Vina Spiehler, Ph.D., DABFT, Diagnostic Products Corporation

The data have been taken from the UCI Machine Learning Database Repository [http://www.ics.uci.edu/~mlearn/MLRepository.html](http://www.ics.uci.edu/~mlearn/MLRepository.html) and were converted to R format by <klaus.schliep@gmail.com>.

**Examples**

```r
data(glass)
str(glass)
```

---

**Description**

This radar data was collected by a system in Goose Bay, Labrador. This system consists of a phased array of 16 high-frequency antennas with a total transmitted power on the order of 6.4 kilowatts. See the paper for more details. The targets were free electrons in the ionosphere. "Good" radar returns are those showing evidence of some type of structure in the ionosphere. "Bad" returns are those that do not; their signals pass through the ionosphere.

Received signals were processed using an autocorrelation function whose arguments are the time of a pulse and the pulse number. There were 17 pulse numbers for the Goose Bay system. Instances in this database are described by 2 attributes per pulse number, corresponding to the complex values returned by the function resulting from the complex electromagnetic signal.

**Usage**

```r
data(ionosphere)
```
**knn**

**Format**

A data frame with 351 observations on the following 35 variables. The first 34 continuous covariates are used for the prediction. The 35th attribute is either g ("good") or b ("bad") according to the definition summarized above. This is a binary classification task.

**Source**

Vince Sigillito (vgs@aplacen.apl.jhu.edu), Space Physics Group, Applied Physics Laboratory, Johns Hopkins University, Johns Hopkins Road, Laurel, MD 20723

The data have been taken from the UCI Machine Learning Database Repository [http://www.ics.uci.edu/~mlearn/MLRepository.html](http://www.ics.uci.edu/~mlearn/MLRepository.html) and were converted to R format by <klaus.schliep@gmail.com>.

**Examples**

data(ionosphere)

---

**Description**

Performs k-nearest neighbor classification of a test set using a training set. For each row of the test set, the k nearest training set vectors (according to Minkowski distance) are found, and the classification is done via the maximum of summed kernel densities. In addition even ordinal and continuous variables can be predicted.

**Usage**

```r
knn(formula = formula(train), train, test, na.action = na.omit(),
k = 7, distance = 2, kernel = "optimal", ykernel = NULL, scale=TRUE,
contrasts = c('unordered' = "contr.dummy", ordered = "contr.ordinal"))
knn.dist(learn, valid, k = 10, distance = 2)
```

**Arguments**

- `formula`: A formula object.
- `train`: Matrix or data frame of training set cases.
- `test`: Matrix or data frame of test set cases.
- `learn`: Matrix or data frame of training set cases.
- `valid`: Matrix or data frame of test set cases.
- `na.action`: A function which indicates what should happen when the data contain 'NA's.
- `k`: Number of neighbors considered.
- `distance`: Parameter of Minkowski distance.
kernel

Kernel to use. Possible choices are "rectangular" (which is standard unweighted knn), "triangular", "epanechnikov" (or beta(2,2)), "biweight" (or beta(3,3)), "triweight" (or beta(4,4)), "cos", "inv", "gaussian", "rank" and "optimal".

ykernel

Window width of an y-kernel, especially for prediction of ordinal classes.

scale

logical, scale variable to have equal sd.

contrasts

A vector containing the 'unordered' and 'ordered' contrasts to use.

Details

This nearest neighbor method expands knn in several directions. First it can be used not only for classification, but also for regression and ordinal classification. Second it uses kernel functions to weight the neighbors according to their distances. In fact, not only kernel functions but every monotonic decreasing function \( f(x) \forall x > 0 \) will work fine.

The number of neighbours used for the "optimal" kernel should be \( \left\lfloor \frac{(2(d+4)/(d+2))(d/(d+4))k}{f(x)} \right\rfloor \), where \( k \) is the number that would be used for unweighted knn classification, i.e. kernel="rectangular". This factor \( (2(d+4)/(d+2))(d/(d+4)) \) is between 1.2 and 2 (see Samworth (2012) for more details).

Value

\( \text{kknn} \) returns a list-object of class \( \text{kknn} \) including the components

- \text{fitted.values} Vector of predictions.
- \text{CL} Matrix of classes of the k nearest neighbors.
- \text{W} Matrix of weights of the k nearest neighbors.
- \text{D} Matrix of distances of the k nearest neighbors.
- \text{C} Matrix of indices of the k nearest neighbors.
- \text{prob} Matrix of predicted class probabilities.
- \text{response} Type of response variable, one of continuous, nominal or ordinal.
- \text{distance} Parameter of Minkowski distance.
- \text{call} The matched call.
- \text{terms} The 'terms' object used.

Author(s)

Klaus P. Schliep <klaus.schliep@gmail.com>
Klaus Hechenbichler

References


miete

See Also

train.kknn, simulation, knn and knn1

Examples

library(kknn)

data(iris)

m <- dim(iris)[1]
val <- sample(1:m, size = round(m/3), replace = FALSE, prob = rep(1/m, m))
iris.learn <- iris[-val,]
iris.valid <- iris[val,]
iris.kknn <- kknn(Species~., iris.learn, iris.valid, distance = 1,
kernel = "triangular")
summary(iris.kknn)

fit <- fitted(iris.kknn)
table(iris.valid$Species, fit)
pcol <- as.character(as.numeric(iris.valid$Species))
pairs(iris.valid[,4], pch = pcol, col = c("green3", "red")
[(iris.valid$Species != fit)+1])

data(ionosphere)

ionosphere.learn <- ionosphere[1:200,]
ionosphere.valid <- ionosphere[-c(1:200),]
fit.kknn <- kknn(class ~ ., ionosphere.learn, ionosphere.valid)
table(ionosphere.valid$class, fit.kknn$fit)

(fit.train1 <- train.kknn(class ~ ., ionosphere.learn, kmax = 15,
kernel = c("triangular", "rectangular", "epanechnikov", "optimal"), distance = 1))
table(predict(fit.train1, ionosphere.valid), ionosphere.valid$class)

(fit.train2 <- train.kknn(class ~ ., ionosphere.learn, kmax = 15,
kernel = c("triangular", "rectangular", "epanechnikov", "optimal"), distance = 2))
table(predict(fit.train2, ionosphere.valid), ionosphere.valid$class)

miete

Munich Rent Standard Database (1994)

Description

Many german cities compose so-called rent standards to make a decision making instrument available to tenants, landlords, renting advisory boards and experts. The rent standards are used in particular for the determination of the local comparative rent (i.e. net rent as a function of household size, equipment, year of construction, etc.). For the composition of the rent standards, a representative random sample is drawn from all relevant households, and the interesting data are determined by interviewers by means of questionnaires. The dataset contains the data of 1082 households interviewed for the munich rent standard 1994.

Usage

data(miete)
Format

A data frame with 1082 observations on the following 18 variables.

**nm** Net rent in DM.

**wfl** Floor space in sqm.

**bj** Year of construction.

**bado** Bathroom in apartment?
   1: no
   0: yes

**zh** Central heating?
   1: yes
   0: no

**ww0** Hot water supply?
   1: no
   0: yes

**badkach** Tiled bathroom?
   1: yes
   0: no

**fenster** Window type:
   1: plain windows
   0: state-of-the-art windows

**kueche** Kitchen type
   1: well equipped kitchen
   0: plain kitchen

**mvdauer** Lease duration in years.

**bijkat** Age category of the building (bj categorized)
   1: built before 1919
   2: built between 1919 and 1948
   3: built between 1949 and 1965
   4: built between 1966 and 1977
   5: built between 1978 and 1983
   6: built after 1983

**wflkat** Floor space category (wfl categorized):
   1: less than 50 sqm
   2: between 51 sqm and 80 sqm
   3: at least 81 sqm

**nmqm** Net rent per sqm.

**rooms** Number of rooms in household.

**nmkat** Net rent category (nm categorized):
   1: less than 500 DM
   2: between 500 DM and 675 DM
   3: between 675 DM and 850 DM
   4: between 850 DM and 1150 DM
   5: at least 1150 DM
simulation

adr  Address type:
1: bad
2: average
3: good

wohn Residential type:
1: bad
2: average
3: good

Source
The data were converted to R format by <klaus.schliep@gmail.com>.

Examples
data(miete)
str(miete)

<table>
<thead>
<tr>
<th>simulation</th>
<th>Crossvalidation procedure to test prediction accuracy</th>
</tr>
</thead>
</table>

Description
simulation tests prediction accuracy of regression and/or classification techniques via simulation of different test sets.

Usage
simulation(formula, data, runs = 10, train = TRUE, k = 11, ...)

Arguments
- formula A formula object.
- data Matrix or data frame.
- runs Number of crossvalidation runs.
- train A logical value. If TRUE the training procedure for selecting optimal values of k and kernel is performed.
- k Number or maximal number of neighbors considered, dependent of choice for train.
- ... Further arguments passed to or from other methods.

Value
A matrix, containing the mean and variance of the misclassification error, the absolute and the squared distances.
**Author(s)**

Klaus P. Schliep <klaus.schliep@gmail.com>

**References**


**See Also**

`kknn` and `train.kknn`

**Examples**

```r
library(kknn)
data(miete)
simulation(nmqm ~ wfl + bjkat + zh, data = miete, runs = 5, kernel = "triangular", k = 15)
simulation(wflkat ~ nm + bjkat + zh, data = miete, runs = 5)
simulation(zh ~ wfl + bjkat + nmqm, data = miete, runs = 5)
```

---

**Description**

Spectral clustering based on k-nearest neighbor graph.

**Usage**

```r
specClust(data, centers=NULL, nn = 7, method = "symmetric", gmax=NULL, ...)  
## S3 method for class 'specClust'  
plot(x, ...)
```

**Arguments**

- `data`  
  Matrix or data frame.
- `centers`  
  number of clusters to estimate, if NULL the number is chosen automatical.
- `nn`  
  Number of neighbors considered.
- `method`  
  Normalisation of the Laplacian ("none", "symmetric" or "random-walk").
- `gmax`  
  maximal number of connected components.
- `x`  
  an object of class specClust
- `...`  
  Further arguments passed to or from other methods.
specClust

Details

specClust allows to estimate several popular spectral clustering algorithms, for an overview see von Luxburg (2007).

The Laplacian is constructed from a from nearest neighbors and there are several kernels available. The eigenvalues and eigenvectors are computed using the binding in igraph to arpack. This should ensure that this algorithm is also feasible for larger datasets as the the distances used have dimension n*m, where n is the number of observations and m the number of nearest neighbors. The Laplacian is sparse and has roughly n*m elements and only k eigenvectors are computed, where k is the number of centers.

Value

specClust returns a kmeans object or in case of k being a vector a list of kmeans objects.

Author(s)

Klaus P. Schliep <klaus.schliep@gmail.com>

References


See Also

kknn, arpack, kmeans

Examples

data(iris)
cl <- specClust(iris[,1:4], 3, nn=5)
pcol <- as.character(as.numeric(iris$Species))
pairs(iris[,1:4], pch = pcol, col = c("green", "red", "blue")[cl$cluster])
table(iris[,5], cl$cluster)
train.kknn

**Training kknn**

**Description**

Training of kknn method via leave-one-out crossvalidation.

**Usage**

```r
train.kknn(formula, data, kmax = 11, distance = 2, kernel = "optimal",
ykernel = NULL, scale = TRUE, contrasts = c('unordered' = "contr.dummy",
ordered = "contr.ordinal"), ...)
cv.kknn(formula, data, kcv = 10, ...)
```

**Arguments**

- `formula`: A formula object.
- `data`: Matrix or data frame.
- `kmax`: Maximum number of `k`.
- `distance`: Parameter of Minkowski distance.
- `kernel`: Kernel to use. Possible choices are "rectangular" (which is standard unweighted knn), "triangular", "epanechnikov" (or beta(2,2)), "biweight" (or beta(3,3)), "triweight" (or beta(4,4)), "cos", "inv", "gaussian" and "optimal".
- `ykernel`: Window width of an y-kernel, especially for prediction of ordinal classes.
- `scale`: Logical, scale variable to have equal sd.
- `contrasts`: A vector containing the 'unordered' and 'ordered' contrasts to use.
- `...`: Further arguments passed to or from other methods.
- `kcv`: Number of partitions for k-fold cross validation.

**Details**

`train.kknn` performs leave-one-out crossvalidation and is computationally very efficient. `cv.kknn` performs k-fold crossvalidation and is generally slower and does not yet contain the test of different models yet.

**Value**

`train.kknn` returns a list-object of class `train.kknn` including the components.

- `misclass`: Matrix of misclassification errors.
- `mean.abs`: Matrix of mean absolute errors.
- `mean.squ`: Matrix of mean squared errors.
- `fitted.values`: List of predictions for all combinations of kernel and `k`.
train.kknn  13

- **best.parameters**
  List containing the best parameter value for kernel and k.

- **response**
  Type of response variable, one of *continuous, nominal* or *ordinal*.

- **distance**
  Parameter of Minkowski distance.

- **call**
  The matched call.

- **terms**
  The 'terms' object used.

**Author(s)**
Klaus P. Schliep <klaus.schliep@gmail.com>

**References**


**See Also**

- `kknn` and `simulation`

**Examples**

```r
library(kknn)
## Not run:
data(miete)
(train.con <- train.kknn(nnqm ~ wfl + bjkat + zh, data = miete, kmax = 25, kernel = c("rectangular", "triangular", "epanechnikov", "gaussian", "rank", "optimal")))
plot(train.con)

(train.ord <- train.kknn(wflkat ~ nm + bjkat + zh, miete, kmax = 25, kernel = c("rectangular", "triangular", "epanechnikov", "gaussian", "rank", "optimal")))
plot(train.ord)

(train.nom <- train.kknn(zh ~ wfl + bjkat + nnqm, miete, kmax = 25, kernel = c("rectangular", "triangular", "epanechnikov", "gaussian", "rank", "optimal")))
plot(train.nom)

## End(Not run)
data(glass)
glass <- glass[-1]
(fit.glass1 <- train.kknn(Type ~ ., glass, kmax = 15, kernel = c("triangular", "rectangular", "epanechnikov", "optimal"), distance = 1))
(fit.glass2 <- train.kknn(Type ~ ., glass, kmax = 15, kernel = c("triangular", "rectangular", "epanechnikov", "optimal"), distance = 2))
plot(fit.glass1)
```
plot(fit.glass2)
Index

*Topic **classif**
  contr.dummy, 2
  kknn, 5
  simulation, 9
  train.kknn, 12

*Topic **cluster**
  specclust, 10

*Topic **datasets**
  glass, 3
  ionosphere, 4
  miete, 7

*Topic **design**
  contr.dummy, 2

*Topic **package**
  kknn-package, 2

arpack, 11

contr.dummy, 2
contr.metric (contr.dummy), 2
contr.ordinal (contr.dummy), 2
contr.poly, 3
contr.sdif, 3
contrasts, 3
cv.kknn (train.kknn), 12
glass, 3
ionosphere, 4
kknn, 5, 10, 11, 13
kknn-package, 2
kmeans, 11
knn, 7
knn1, 7
miete, 7

plot.specClust (specClust), 10
plot.train.kknn (train.kknn), 12
predict.kknn (kknn), 5
predict.train.kknn (train.kknn), 12
print.kknn (kknn), 5
print.train.kknn (train.kknn), 12

simulation, 7, 9, 13
specClust, 10
summary.kknn (kknn), 5
summary.train.kknn (train.kknn), 12

train.kknn, 7, 10, 12