Package 'probout'

January 5, 2018

Version 1.1.1

Date 2018-01-05

Title Unsupervised Multivariate Outlier Probabilities for Large Datasets

Author Chris Fraley [aut, cre]

Maintainer Chris Fraley <cfraley@tableau.com>

Depends R (>= 3.1.0), FNN, mclust, MASS

Suggests mvtnorm, GDAdata, bibtex, knitr, rmarkdown

VignetteBuilder knitr

Description Estimates unsupervised outlier probabilities for multivariate numeric data with many observations from a nonparametric outlier statistic.

License MIT + file LICENSE

URL https://www.r-project.org

NeedsCompilation yes

Repository CRAN

Date/Publication 2018-01-05 17:41:37 UTC

R topics documented:

allProb *Outlier probabilities for all observations*

Description

Outlier probabilities for all of the data, obtained by assigning to each observation the probabilty of the its associated leader partition.

Usage

```
allProb( leaderInstance, partprob)
```

Arguments

leaderInstance	A single component from a call to leader, giving leader algorithm results for one value of the partitioning radius.
partprob	A vector of probabilities for each partition in leaderInstance.

Value

A vector of probabilities for each observation in the data underlying leaderInstance. Each observation inherits the probability of its associated partition.

See Also

leader, partProb

Examples

```
 set.seed(0)

 lead <- leader(faithful)
 nlead <- length(lead[[1]]$partitions)

# repeat multiple times to account for randomness
 ntimes <- 100
 probs <- matrix( NA, nlead, ntimes)
 for (i in 1:ntimes) {
    probs[,i] <- partProb( simData(lead[[1]]), method = "distance")
 }

# median probability for each partition
 partprobs <- apply( probs, 1, median)

 quantile(partprobs)

# plot leaders with outlier probability > .95
 plot( faithful[,1], faithful[,2], pch = 16, cex = .5,
```

```
      main = "red : instances with outlier probability > .95")
allprobs <- allProb( lead[[1]], partprobs)
out <- allprobs > .95
points( faithful[out,1], faithful[out,2], pch = 8, cex = 1, col = "red")
```

`leader` *Leader Algorithm for Data Partitioning*

Description

Partitions the data according to Hartigan's leader algorithm, and provides ranges, centroids, and variances for the partitions.

Usage

```
leader(data, radius = NULL, scale = T)
```

Arguments

`data`	A numeric vector or matrix of observations. If a matrix, rows correspond to observations and columns correspond to variables.
`radius`	A vector of values for the partitioning radius. Wilkinson's default radius is used if `radius` is left unspecified (see function `LWradius`).
`scale`	A logical variable indicating whether or not the data should be mapped to the unit hypercube. The default is to scale the data. Values of the radius will not be scaled; they should be specifed relative to the unit hypercube unless `scale = F`.

Details

Given a partitioning radius `r`, the leader algorithm makes one pass through the data, designating an observation as a new leader if it is not within `r` of an existing leader, and otherwise assigning it to the partition associated with the nearest existing leader. The set of leaders typically depends on the order of the data observations.

If `radius = 0`, then all of the data observations are leaders, and only `radius` and `leaders` are returned as output components.

This implementation does a completely new nearest-neighbor search for each observation and for each radius. A more efficient approach would be to maintain, for each radius, a data structure (such as a kd-tree) allowing fast nearest-neighbor search. These data structures could then be updated to account for new observations. Currently, there doesn't seem to be a way to do this in R.

Value

A list with one component for each value of `radius`, each having the following sub-components:

`radius`	The value of the radius associated with the partitioning.

partitions	A list with one component for each partition, giving the indexes (as observations in the data) of the members of the partition. The first index is that of the associated *leader* (sometimes called *exemplar*).
leaders	The indexes of the leaders for each partition.
centroids	The centroids for each partition, as a matrix with rows corresponding to the partitions and columns corresponding to variables if multidimensional. These will be the data if `radius == 0`.
variances	The variances for each partition, as a matrix with rows corresponding to the partitions and columns corresponding to variables if multidimensional.
ranges	A list with two components: `min` and `max` giving the minimum and maximum values for each variable for each partition. These `range` components are given as a matrix with rows corresponding to the partitions and columns corresponding to variables if multidimensional.
maxdist	A vector with one value for each partition, giving the largest distance from each leader to any member of its partition.

References

J. A. Hartigan, *Clustering Algorithms*, Wiley, 1975.

L. Wilkinson, Visualizing Outliers, Technical Report, University of Illinois at Chicago, 2016. https://www.cs.uic.edu/~wilkinson/Publications/outliers.pdf

See Also

LWradius

Examples

```
 radius.default <- LWradius(nrow(faithful),ncol(faithful))
 lead <- leader(faithful, radius = c(0,radius.default))

# number of partitions for each radius
 sapply(lead, function(x) length(x$partitions))

# plot the leaders for the non-zero radius
 plot( faithful[,1], faithful[,2],
       main = "blue indicates leaders (default radius)",
       pch = 16, cex = .5)
 ldrs <- lead[[2]]$leaders
 points( faithful[ldrs,1], faithful[ldrs,2],
         pch = 8, col = "dodgerblue", cex = .5)
```

logdens *Log Density for Gaussian Mixture Model*

Description

Computes the log density for observations in a univariate or multivariate Gaussian mixture model with spherical or diagonal (co)variance that varies across components.

Usage

```
logdens( x, simData, shrink = 1)
```

Arguments

x	A numeric vector or matrix for which the log density is to be computed.
simData	Observations from a call to simData, which includes the partition centroids and variance information for the underlying simulation model.
shrink	Shrinkage parameter for the mixture model variance. To be consistent with the shrinkage as described in partProb, the variance is scaled by the *square* of shrink. The default value is shrink = 1, so that no shrinkage is applied to the variance.

Details

If either radius = 0, or simData returns only centroids (nsim = 0), then no density estimate is attempted.

Value

A vector giving the log density of x in the model as specified by simData, with optional shrinkage applied to the variance.

References

G. Celeux and G. Govaert, Gaussian Parsimonious Mixture Models, *Pattern Recognition*, 1995.

G. J. McLachlan and D. Peel, *Finite Mixture Models*, Wiley, 2000.

C. Fraley and A. E. Raftery, Model-based clustering, discriminant analysis and density estimation, *Journal of the American Statistical Association*, 2002.

See Also

partProb

Examples

```
lead <- leader(faithful)
sim <- simData( lead)

logdens( faithful, sim)
```

`LWradius` *Wilkinson's default leader-partitioning radius*

Description

Wilkinson's default leader-partitioning radius.

Usage

```
LWradius( n, p)
```

Arguments

`n`	The number of observations (rows) in the data.
`p`	The number of variables (columns) in the data; `p = 1` if univariate.

Value

Wilkinson's default leader partitioning radius `0.1/(log(n)^(1/p))`.

References

L. Wilkinson (2016), Visualizing Outliers, Technical Report, University of Illinois at Chicago, `https://www.cs.uic.edu/~wilkinson/Publications/outliers.pdf`.

See Also

`leader`

Examples

```
x1 <- rnorm(10000)
LWradius(length(x1),1)

LWradius(nrow(faithful),ncol(faithful))
```

`OutlierStatistic` *Nonparametric Outlier Statistic*

Description

Robust nonparametric outlier statistic for univariate or multivariate data.

Usage

```
OutlierStatistic( x, nproj=1000, prior=NULL, seed=NULL)
```

Arguments

`x`	A numeric vector or matrix for which the outlier statistic is to be determined.
`nproj`	If `x` is multivariate, the number of random projections to be used in computing the statistic.
`prior`	If `x` is multivariate, a prior estimate of the statistic for each observation in `x`, to be used as a base line for maximization relative to new random projections.
`seed`	An optional integer argument to `set.seed` for reproducible simulations. By default the current seed will be used. Reproducibility can also be achieved by calling `set.seed` before calling `OutlierStatistic`.

Value

A vector giving the maximum value of the outlier statistic for each observation over all projections.

References

W. A. Stahel, *Breakdown of Covariance Estimators*, doctoral thesis, Fachgruppe Fur Statistik, Eidgenossische Technische Hochshule (ETH), 1981.

D. L. Donoho, *Breakdown Properties of Multivariate Location Estimators*, doctoral thesis, Department of Statistics, Harvard University, 1982.

Note

Note that partition probabilities are computed from an exponential distribution fit to the outlier statistic, rather than from the empirical distribution of the outlier statistic.

See Also

`partProb`

Examples

```
stat <- OutlierStatistic(faithful)
q.99 <- quantile(stat,.99)
out <- stat > q.99

plot( faithful[,1], faithful[,2],
      main="red : .99 quantile for outlier statistic", cex=.5)
points( faithful[out,1], faithful[out,2],
        pch = 4, col = "red", lwd = 1, cex = .5)

require(mvtnorm)

set.seed(0)
Sigma <- crossprod(matrix(rnorm(2*2),2,2))
x <- rmvt( 10000, sigma = Sigma, df = 2)

stat <- OutlierStatistic(x)
q.95 <- quantile(stat,.95)

hist(x, main = "gray : .95 quantile for outlier statistic", col = "black")
abline( v = x[stat > q.95], col = "gray")
hist(x, col = "black", add = TRUE)
```

`partProb` *Partition outlier probabilities*

Description

Assigns outlier probabilities to the partitions by fitting an exponential distribution to a nonparametric outlier statistic for simulated data or partition centroids.

Usage

```
partProb( simData, method = c("intrinsic","distance","logdensity","distdens",
          "density"), shrink = 1, nproj = 1000, seed = NULL)
```

Arguments

`simData`	Observations from a call to `simData`, which includes the partition centroids and (optionally) simulated data as well.
`method`	One of the following options:

`"intrinsic"`	:	outlier statistic applied to simulation data (centroids if no simulation)
`"distance"`	:	outlier statistic applied to distances between NN partitions
`"logdensity"`	:	outlier statistic applied to differences in log density between NN partitions
`"distdens"`	:	outlier statistic applied to a matrix consisting of the `"distance"` and `"logdensity"` values
`"density"`	:	outlier statistic applied to smallest/largest ratios of density between NN partitions

The default is to use the "intrinsic" method.

shrink — Shrinkage parameter for outlier detection data. The offsets from simData are scaled by this factor before adding them to the partition centroids as data for outlier detection. The default value is shrink = 1, so that no shrinkage is applied to simulation offsets.

nproj — If the data is multivariate or method = "distdens", the number of random projections to be used to obtain the outlier statistic.

seed — An optional integer argument to set.seed for reproducible outlier statistics. By default the current seed will be used. Reproducibility can also be achieved by calling set.seed before calling partProb.

Details

"logdensity" is generally prefered over "density", because negative values that are large in magniude of the logarithm of the density will not be numerically distinguishable as density values.

Value

A vector of probabilities for each partition, obtained by fitting an exponential distribution to the outlier statistic.

References

C. Fraley, Estimating Outlier Probabilities for Large Datasets, 2017.

See Also

simData, OutlierStatistic, allProb

Examples

```
 set.seed(0)

 lead <- leader(faithful)
 nlead <- length(lead[[1]]$partitions)

# repeat multiple times to account for randomness
 ntimes <- 100
 probs <- matrix( NA, nlead, ntimes)
 for (i in 1:ntimes) {
    probs[,i] <- partProb( simData(lead[[1]]), method = "distance")
 }

# median probability for each partition
 partprobs <- apply( probs, 1, median)

 quantile(probs)

# plot leaders with outlier probability > .95
```

```
plot( faithful[,1], faithful[,2], pch = 16, cex = .5,
      main = "red : leaders with outlier probability > .95")
out <- partprobs > .95
l <- lead[[1]]$leaders
points( faithful[l[out],1], faithful[l[out],2], pch = 8, cex = 1, col = "red")
```

`simData` *Simulates observations for outlier determination.*

Description

Simulates observations from a mixture model based on information on partitions from the `leader` function.

Usage

```
simData( leaderInstance, nsim=NULL, model=c("diagonal","spherical"), seed=NULL)
```

Arguments

`leaderInstance`	A single component from a call to `leader`, giving Leader Algorithm results for one value of the partitioning radius.
`nsim`	The number of observations to be simulated. Only the radius and centroids are returned of `nsim = 0` or `leaderInstance$radius == 0`) — no observations are simulated. Default: `min(# observations,max(# partitions, 1000))`.
`model`	For multivariate data, a vector of character strings indicating the type of Gaussian mixture model covariance to be used in generating the simulated observations (see `details`). For univariate data, the observations are generated from a model in which the variances may vary across components.
`seed`	An optional integer argument to `set.seed` for reproducible simulations. By default the current seed will be used. Reproducibility can also be achieved by calling `set.seed` before calling `simData`.

Details

The following models are available for multivariate data:

`"spherical"`	:	spherical, varying volume
`"diagonal"`	:	diagonal, varying volume and shape

An ellipsoidal model is also possible, but has not yet been implemented.
If `nsim = 0` or `leaderInstance$radius == 0`, no observations are simulated, and only the radius and partition centroids are returned.

Value

A list with the following components:

`radius`	The value of the radius associated with `leaderInstance`.
`location`	The vector or matrix of centroids of the partitions. If a matrix, rows correspond to the partitions and columns to the variables.
`index`	A vector of integer values giving the index of the partition associated with each simulated observation.
`offset`	A vector of numeric values giving offset for the simulated observations from their associated centroids.
`weight`	A vector of numeric values between 0 and 1 giving the proportion of data observations in each partition.
`scale`	The scale (variance) of the mixture components in a univariate or spherical model. Set to 1 for each component in the diagonal model.
`shape`	A matrix giving the variances of the mixture component in a diagonal model. The rows correspond to the dimensions of the data, while the columns correspond to the mixture components (partitions).

References

C. Fraley, Estimating Outlier Probabilities for Large Datasets, 2017.

See Also

`leader`, `partProb`

Examples

```
 radius.default <- LWradius(nrow(faithful),ncol(faithful))
 lead <- leader(faithful, radius = c(0,radius.default))

# (simulated) data for outlier statistic (no simulation for radius = 0)
 sim <- lapply( lead, simData)

# components of simData output
 lapply( sim, names)
```

Index