Package ‘OutlierDetection’

June 15, 2019

Type Package
Title Outlier Detection
Version 0.1.1
Author Vinay Tiwari, Akanksha Kashikar
Maintainer Vinay Tiwari <vinaystiwari786@gmail.com>
Imports ggplot2, DDoutlier, depth, depthTools, ldbod, spatstat, plotly
License GPL-2
Encoding UTF-8
LazyData true
RoxygenNote 6.1.1
NeedsCompilation no
Repository CRAN
Date/Publication 2019-06-15 16:10:03 UTC

R topics documented:
dens .......................................................... 2
depthout .................................................... 3
disp .......................................................... 4
maha ........................................................ 5
Outlier detection using Robust Kernal-based Outlier Factor (RKOF) algorithm

Description

Takes a dataset and finds its outliers using Robust Kernal-based Outlier Factor (RKOF) algorithm.

Usage

dens(x, k = 0.05 * nrow(x), C = 1, alpha = 1, sigma2 = 1,
cutoff = 0.95, rnames = F, boottimes = 100)

Arguments

- **x**: dataset for which outliers are to be found.
- **k**: No. of nearest neighbours to be used, default value is 0.05*nrow(x).
- **C**: Multiplication parameter for k-distance of neighboring observations. Act as bandwidth increaser. Default is 1 such that k-distance is used for the gaussian kernel.
- **alpha**: Sensitivity parameter for k-distance/bandwidth. Small alpha creates small variance in RKOF and vice versa. Default is 1.
- **sigma2**: Variance parameter for weighting of neighboring observations.
- **cutoff**: Percentile threshold used for distance, default value is 0.95.
- **rnames**: Logical value indicating whether the dataset has rownames, default value is False.
- **boottimes**: Number of bootstrap samples to find the cutoff, default is 100 samples.

Details

dens computes outlier score of an observation using DDoutlier package (based on RKOF algorithm) and based on the bootstrapped cutoff, labels an observation as outlier. Outlierliness of the labelled 'Outlier' is also reported and it is the bootstrap estimate of probability of the observation being an outlier. For bivariate data, it also shows the scatterplot of the data with labelled outliers.
Value

Outlier Observations: A matrix of outlier observations
Location of Outlier: Vector of Sr. no. of outliers
Outlier probability: Vector of proportion of times an outlier exceeds local bootstrap cutoff

Author(s)

Vinay Tiwari, Akanksha Kashikar

References


Examples

```r
#Create dataset
x <- iris[,1:4]
#Outlier detection
dens(x, k=4, c=1)
```

Description

Takes a dataset and finds its outliers using depth-based method

Usage

```r
depthout(x, rnames = FALSE, cutoff = 0.05, boottimes = 100)
```

Arguments

- `x`: dataset for which outliers are to be found
- `rnames`: Logical value indicating whether the dataset has rownames, default value is False
- `cutoff`: Percentile threshold used for depth, default value is 0.05
- `boottimes`: Number of bootstrap samples to find the cutoff, default is 100 samples

Details

depthout computes depth of an observation using depthTools package and based on the bootstrapped cutoff, label an observation as outlier. Outlierliness of the labelled `Outlier` is also reported and it is the bootstrap estimate of probability of the observation being an outlier. For bivariate data, it also shows the scatterplot of the data with labelled outliers.
**Value**

Outlier Observations: A matrix of outlier observations
Location of Outlier: Vector of Sr. no. of outliers
Outlier probability: Vector of proportion of times an outlier exceeds local bootstrap cutoff

**Author(s)**

Vinay Tiwari, Akanksha Kashikar

**References**

Johnson, T., Kwok, I., and Ng, R.T. 1998. Fast computation of 2-dimensional depth contours. In Proc. Int. Conf. on Knowledge Discovery and Data Mining (KDD), New York, NY. Kn

**Examples**

Create dataset
```
x=iris[,1:4]
```
#Outlier detection
```
depthout(x,cutoff=0.05)
```

**Description**

Takes a dataset and finds its outliers using dispersion-based method

**Usage**

```
disp(x, cutoff = 0.95, rnames = FALSE, boottimes = 100)
```

**Arguments**

- `x`: dataset for which outliers are to be found
- `cutoff`: Percentile threshold used for distance, default value is 0.95
- `rnames`: Logical value indicating whether the dataset has rownames, default value is False
- `boottimes`: Number of bootstrap samples to find the cutoff, default is 100 samples

**Details**

disp computes LOO dispersion matrix for each observation (dispersion matrix without considering the current observation) and based on the bootstrapped cutoff for score (difference between determinant of LOO dispersion matrix and det of actual dispersion matrix), labels an observation as outlier. Outlierliness of the labelled 'Outlier' is also reported and it is the bootstrap estimate of probability of the observation being an outlier. For bivariate data, it also shows the scatterplot of the data with labelled outliers.
Value

Outlier Observations: A matrix of outlier observations
Location of Outlier: Vector of Sr. no. of outliers
Outlier probability: Vector of proportion of times an outlier exceeds local bootstrap cutoff

Author(s)

Vinay Tiwari, Akanksha Kashikar

References

Jin, W., Tung, A., and Han, J. 2001. Mining top-n local outliers in large databases. In Proc. ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (SIGKDD), San Francisco, CA.

Examples

```r
#Create dataset
x = iris[,1:4]
#Outlier detection
disp(x,cutoff=0.99)
```

---

maha

*Outlier detection using Mahalanobis Distance*

Description

Takes a dataset and finds its outliers using modelbased method

Usage

```r
maha(x, cutoff = 0.95, rnames = FALSE)
```

Arguments

- `x` dataset for which outliers are to be found
- `cutoff` Percentile threshold used for distance, default value is 0.95
- `rnames` Logical value indicating whether the dataset has rownames, default value is False

Details

maha computes Mahalanobis distance an observation and based on the Chi square cutoff, labels an observation as outlier. Outlierliness of the labelled 'Outlier' is also reported based on its p values. For bivariate data, it also shows the scatterplot of the data with labelled outliers.
Value

Outlier Observations: A matrix of outlier observations
Location of Outlier: vector of Sr. no. of outliers
Outlier probability: vector of (1-p value) of outlier observations

Author(s)

Vinay Tiwari, Akanksha Kashikar

References


Examples

#Create dataset
X=iris[,1:4]
#Outlier detection
maha(X,cutoff=0.9)

nn

Outlier detection using k Nearest Neighbours Distance method

Description

Takes a dataset and finds its outliers using distance-based method

Usage

nn(x, k = 0.05 * nrow(x), cutoff = 0.95, Method = "euclidean",
    rnames = FALSE, boottimes = 100)

Arguments

x dataset for which outliers are to be found
k No. of nearest neighbours to be used, default value is 0.05*nrow(x)
cutoff Percentile threshold used for distance, default value is 0.95
Method Distance method, default is Euclidean
rnames Logical value indicating whether the dataset has rownames, default value is False
boottimes Number of bootstrap samples to find the cutoff, default is 100 samples
nnk

Details

nnk computes average knn distance of observation and based on the bootstrapped cutoff, labels anobservation as outlier. Outlierliness of the labelled 'Outlier' is also reported and it is the bootstrapestimate of probability of the observation being an outlier. For bivariate data, it also shows thescatterplot of the data with labelled outliers.

Value

Outlier Observations: A matrix of outlier observations
Location of Outlier: Vector of Sr. no. of outliers
Outlier probability: Vector of proportion of times an outlier exceeds local bootstrap cutoff

Author(s)

Vinay Tiwari, Akanksha Kashikar

References


Examples

#Create dataset
X=iris[,1:4]
#Outlier detection
nn(X,k=4)

nnk (Outlier detection using kth Nearest Neighbour Distance method)

Description

Takes a dataset and finds its outliers using distance-based method

Usage

nnk(x, k = 0.05 * nrow(x), cutoff = 0.95, Method = "euclidean",
    rnames = FALSE, boottimes = 100)
Arguments

x  dataset for which outliers are to be found
k  No. of nearest neighbours to be used, default value is 0.05*nrow(x)
cutoff  Percentile threshold used for distance, default value is 0.95
Method  Distance method, default is Euclidean
rnames  Logical value indicating whether the dataset has rownames, default value is False
boottimes  Number of bootstrap samples to find the cutoff, default is 100 samples

Details

nnk computes kth nearest neighbour distance of an observation and based on the bootstrapped cut-off, labels an observation as outlier. Outlierliness of the labelled 'Outlier' is also reported and it is the bootstrap estimate of probability of the observation being an outlier. For bivariate data, it also shows the scatterplot of the data with labelled outliers.

Value

Outlier Observations: A matrix of outlier observations
Location of Outlier: Vector of Sr. no. of outliers
Outlier probability: Vector of proportion of times an outlier exceeds local bootstrap cutoff

Author(s)

Vinay Tiwari, Akanksha Kashikar

References


Examples

#Create dataset
X=iris[,1:4]
#Outlier detection
nnk(X,k=4)
Outlier Detection

Description

Takes a dataset and finds its outliers using combination of different methods.

Usage

\[
\text{OutlierDetection}(x, k = 0.05 \times \text{nrow}(x), \text{cutoff} = 0.95, \\
\text{Method} = \"euclidean\", \text{rnames} = \text{FALSE}, \text{depth} = \text{FALSE}, \\
\text{dense} = \text{FALSE}, \text{distance} = \text{FALSE}, \text{dispersion} = \text{FALSE})
\]

Arguments

- **x**: dataset for which outliers are to be found
- **k**: No. of nearest neighbours to be used for outlier detection using bootstrapping, default value is 0.05*\text{nrow}(x)
- **cutoff**: Percentile threshold used for distance, default value is 0.95
- **Method**: Distance method, default is Euclidean
- **rnames**: Logical value indicating whether the dataset has rownames, default value is False
- **depth**: Logical value indicating whether depth based method should be used or not, default is False
- **dense**: Logical value indicating whether density based method should be used or not, default is False
- **distance**: Logical value indicating whether distance based methods should be used or not, default is False
- **dispersion**: Logical value indicating whether dispersion based methods should be used or not, default is False

Details

OutlierDetection finds outlier observations for the data using different methods and based on all the methods considered, labels an observation as outlier(intersection of all the methods). For bivariate data, it also shows the scatterplot of the data with labelled outliers.

Value

- Outlier Observations: A matrix of outlier observations
- Location of Outlier: Vector of Sr. no. of outliers

Author(s)

Vinay Tiwari, Akanksha Kashikar
Examples

outlierdetection(iris[,,-5])

Description

Takes a dataset, and finds its outliers based on principal components using combination of different methods.

Usage

PCOutlierDetection(xL k = 0.05 * nrow(x), cutoff = 0.95, Method = "euclidean", rnames = FALSE, depth = FALSE, dense = FALSE, distance = FALSE, dispersion = FALSE, infocut = 0.9)

Arguments

x         dataset for which outliers are to be found
k         No. of nearest neighbours to be used for outlier detection using bootstrapping, default value is 0.05*nrow(x)
cutoff    Percentile threshold used for distance, default value is 0.95
Method     Distance method, default is Euclidean
rnames     Logical value indicating whether the dataset has rownames, default value is False
depth      Logical value indicating whether depth based method should be used or not, default is False
dense      Logical value indicating whether density based method should be used or not, default is False
distance   Logical value indicating whether distance based methods should be used or not, default is False
dispersion Logical value indicating whether dispersion based methods should be used or not, default is False
infocut    Amount of variation for deciding the no. of principal components to be retained in the analysis, default is 0.9

Details

OutlierDetection finds outlier observations for the principal component space using different methods and based on all the methods considered, labels an observation as outlier(intersection of all the methods). For bivariate data, it also shows the scatterplot of the data with labelled outliers.
**Value**

Outlier Observations: A matrix of outlier observations
Location of Outlier: Vector of Sr. no. of outliers

**Author(s)**

Vinay Tiwari, Akanksha Kashikar

**Examples**

```
PCOutlierDetection(iris[,,-5])
```

---

**UnivariateOutlierDetection**

*Univariate Outlier Detection (Intersection of all the methods)*

**Description**

Takes a vector and finds its outliers using combination of different methods

**Usage**

```
univariateoutlierdetection(x, k = 0.05 * length(x), cutoff = 0.95, 
dist = FALSE, dens = FALSE, depth = FALSE, method = "euclidean", 
rnames = FALSE)
```

**Arguments**

- **x**: vector for which outliers are to be found
- **k**: No. of nearest neighbours to be used for distance methods, default value is 0.05*length(x)
- **cutoff**: Percentile threshold used for outlier detection using bootstrapping, default value is 0.95
- **dist**: Logical value indicating whether distance based methods should be used or not, default is False
- **dens**: Logical value indicating whether density based method should be used or not, default is False
- **depth**: Logical value indicating whether depth based method should be used or not, default is False
- **method**: Distance method, default is euclidean
- **rnames**: Logical value indicating whether the dataset has rownames, default value is False
Details
UnivariateOutlierDetection finds outlier observations for an univariate data using different methods and based on all the methods, labels an observation as outlier(intersection of all the methods). It also shows the scatterplot of the data with labelled outliers with observation no. as x-axis.

Value
Outlier Observations: A vector of outlier observations
Location of Outlier: Vector of Sr. no. of outliers

Author(s)
Vinay Tiwari, Akanksha Kashikar

Examples
#Create dataset
X=iris[,1:4]
#Outlier detection
dethout(X,cutoff=0.05)
UnivariateOutlierDetection(iris[,1],cutoff=.95,Method=“euclidean”,rnames=FALSE)
Index

dens, 2
depthout, 3
disp, 4

maha, 5

nn, 6
nnk, 7

OutlierDetection, 9
PCOutlierDetection, 10
UnivariateOutlierDetection, 11