Package ‘lookout’

February 12, 2021

Type Package
Title Leave One Out Kernel Density Estimates for Outlier Detection
Version 0.1.0
Maintainer Sevvandi Kandanaarachchi <sevvandik@gmail.com>
Description Outlier detection using leave-one-out kernel density estimates and extreme value theory. The bandwidth for kernel density estimates is computed using persistent homology, a technique in topological data analysis. Using peak-over-threshold method, a generalized Pareto distribution is fitted to the log of leave-one-out kde values to identify outliers.
License GPL-3
Encoding UTF-8
LazyData true
RoxygenNote 7.1.1
Imports TDAstats, evd, RANN, ggplot2, tidyr
Suggests knitr, rmarkdown
URL https://sevvandi.github.io/lookout/
NeedsCompilation no
Author Sevvandi Kandanaarachchi [aut, cre]
Rob Hyndman [aut] (<https://orcid.org/0000-0002-2140-5352>)
Repository CRAN
Date/Publication 2021-02-12 10:20:02 UTC

R topics documented:

autplot.lookoutliers .................................................. 2
autplot.persistingoutliers ......................................... 2
lookout ................................................................. 3
persisting_outliers ..................................................... 4

Index 6
autoplot.lookoutliers  
*Plots outliers identified by lookout algorithm.*

**Description**

Scatterplot of two columns from the data set with outliers highlighted.

**Usage**

```r
## S3 method for class 'lookoutliers'
autoplot(object, columns = 1:2, ...)
```

**Arguments**

- `object`  : The output of the function ‘lookout’.
- `columns` : Which columns of the original data to plot (specified as either numbers or strings)
- `...`    : Other arguments currently ignored.

**Value**

A ggplot object.

**Examples**

```r
X <- rbind(
  data.frame(x = rnorm(500),
              y = rnorm(500)),
  data.frame(x = rnorm(5, mean = 10, sd = 0.2),
              y = rnorm(5, mean = 10, sd = 0.2))
)
lo <- lookout(X)
autoplot(lo)
```

autoplot.persistingoutliers  
*Plots outlier persistence for a range of significance levels.*

**Description**

This function plots outlier persistence for a range of significance levels using the algorithm lookout, an outlier detection method that uses leave-one-out kernel density estimates and generalized Pareto distributions to find outliers.

**Usage**

```r
## S3 method for class 'persistingoutliers'
autoplot(object, alpha = object$alpha, ...)
```
**lookout**  

Identifies outliers using the algorithm lookout.

**Description**  
This function identifies outliers using the algorithm lookout, an outlier detection method that uses leave-one-out kernel density estimates and generalized Pareto distributions to find outliers.

**Usage**  
```r  
lookout(X, alpha = 0.05, unitize = TRUE, bw = NULL, gpd = NULL)  
```

**Arguments**  
- `X` The input data in a dataframe, matrix or tibble format.
- `alpha` The level of significance. Default is `0.05`.
- `unitize` An option to normalize the data. Default is `TRUE`, which normalizes each column to `[0, 1]`.
- `bw` Bandwidth parameter. Default is `NULL` as the bandwidth is found using Persistent Homology.
- `gpd` Generalized Pareto distribution parameters. If ‘NULL’ (the default), these are estimated from the data.

**Examples**  
```r  
X <- rbind(  
data.frame(  
  x = rnorm(500),  
  y = rnorm(500)  
),  
data.frame(  
  x = rnorm(5, mean = 10, sd = 0.2),  
  y = rnorm(5, mean = 10, sd = 0.2)  
)  
)  
plot(X, pch = 19)  
outliers <- persisting_outliers(X, unitize = FALSE)  
autoplot(outliers)  
```
persisting_outliers

Value

A list with the following components:

- **outliers**: The set of outliers.
- **outlier_probability**: The GPD probability of the data.
- **bandwidth**: The bandwidth selected using persistent homology.
- **kde**: The kernel density estimate values.
- **lookdke**: The leave-one-out kde values.
- **gpd**: The fitted GPD parameters.

Examples

```r
X <- rbind(
  data.frame(x = rnorm(500), y = rnorm(500)),
  data.frame(x = rnorm(5, mean = 10, sd = 0.2), y = rnorm(5, mean = 10, sd = 0.2))
)
lo <- lookout(X)
lo
autoplot(lo)
```

**persisting_outliers**  Computes outlier persistence for a range of significance values.

Description

This function computes outlier persistence for a range of significance values, using the algorithm lookout, an outlier detection method that uses leave-one-out kernel density estimates and generalized Pareto distributions to find outliers.

Usage

```r
persisting_outliers(x, alpha = seq(0.01, 0.1, by = 0.01), st_qq = 0.9, unitize = TRUE, num_steps = 20)
```
Arguments

X  The input data in a matrix, data.frame, or tibble format. All columns should be numeric.

alpha  Grid of significance levels.

st_qq  The starting quantile for death radii sequence. This will be used to compute the starting bandwidth value.

unitize  An option to normalize the data. Default is TRUE, which normalizes each column to $[0,1]$.

num_steps  The length of the bandwidth sequence.

Value

A list with the following components:

out  A 3D array of $N \times \text{num_steps} \times \text{num_alpha}$ where $N$ denotes the number of observations, num_steps denote the length of the bandwidth sequence and num_alpha denotes the number of significance levels. This is a binary array and the entries are set to 1 if that observation is an outlier for that particular bandwidth and significance level.

bw  The set of bandwidth values.

gpdparas  The GPD parameters used.

lookoutbw  The bandwidth chosen by the algorithm lookout using persistent homology.

Examples

X <- rbind(
  data.frame(x = rnorm(500),
              y = rnorm(500)),
  data.frame(x = rnorm(5, mean = 10, sd = 0.2),
              y = rnorm(5, mean = 10, sd = 0.2))
)
plot(X, pch = 19)
outliers <- persisting_outliers(X, unitize = FALSE)
outliers
autoplot(outliers)
Index

autoplot.lookoutliers, 2
autoplot.persistingoutliers, 2

lookout, 3

persisting_outliers, 4