Package 'hopkins'

August 20, 2023

Type Package

Title Calculate Hopkins Statistic for Clustering

Version 1.1

Description Calculate Hopkins statistic to assess the clusterability of data. See Wright (2023) <doi:10.32614/RJ-2022-055>.

URL https://kwstat.github.io/hopkins/

BugReports https://github.com/kwstat/hopkins/

License MIT + file LICENSE

NeedsCompilation no

RoxygenNote 7.2.3

Imports donut, pdist, RANN

Suggests knitr, rmarkdown, spatstat.data, testthat (>= 3.0.0)

VignetteBuilder knitr

Config/testthat/edition 3

Encoding UTF-8

Author Kevin Wright [aut, cre] (<https://orcid.org/0000-0002-0617-8673>)

Maintainer Kevin Wright <kw.stat@gmail.com>

Repository CRAN

Date/Publication 2023-08-20 04:32:31 UTC

R topics documented:

hopkins *Hopkins statistics for clustering tendency*

Description

Calculate Hopkins statistic for given data.

Usage

```
hopkins(
  X,
  m = ceiling(nrow(X)/10),
  d = ncol(X),
  k = 1,
  U = NULL,
  method = "simple"
)
```

Arguments

X	Data (matrix or data.frame) to check clusterability.
m	Number of rows to sample from X. Default is 1/10th the number of rows of X.
d	Dimension of the data (number of columns of X).
k	kth nearest neighbor to find.
U	Data containing m uniformly-sampled points.
method	Either "simple" or "torus".

Details

Calculated values 0-0.3 indicate regularly-spaced data. Values around 0.5 indicate random data. Values 0.7-1 indicate clustered data.

CAUTION: This function does NOT center and scale the columns of X. You may need to do this manually before using this function.

You should NOT set The parameter 'd'. It is included here to allow for comparisons of `hopkins::hopkins()` and `clustertend::hopkins()`.

The data U is also not normally set by the user. It is included here to allow for unit testing and also for customization of the uniformly-sampled points (e.g. enlarged by 5 percent as suggested by some authors).

Some authors suggest sampling less than 10 percent of points. Others suggest m>10 points to avoid small-sample problems. The distribution of Hopkins statistic requires that nearest neighbors to the selected points be mutually independent, so that only a few of the points can be marked. The distribution of Hopkins statistic is Beta(m,m), independent of the dimensionality of the data d.

Cross & Jain say "The m sampling points are few enough in number, relative to n (the number of events), that their presence does not materially affect the overall density. Ratios of at least 10 to 1

and preferably 20 to 1 are used in the literature. On the other hand, it seems that m should be at least 10 in order to avoid any small sample problems with the distributions of the statistics. This effectively limits the methods to problems with at least 100 events. In high dimensions, very little can be said about data sets that are sparser than that."

Note:

Comparison of `hopkins::hopkins()` and `clustertend::hopkins()`.

The 'hopkins::hopkins()' function uses distances^d (where "distance" is the Euclidean distance between points and "d" is the number of columns in the data). The value returned is: Hopkins statistic.

The 'clustertend::hopkins()' function uses distances^1. The value returned is: 1 - Hopkins statistic.

Value

The value of Hopkins statistic.

Author(s)

Kevin Wright

References

Hopkins, B. and Skellam, J.G., 1954. A new method for determining the type of distribution of plant individuals. Annals of Botany, 18(2), pp.213-227.

Cross, G. R., and A. K. Jain. (1982). Measurement of clustering tendency. Theory and Application of Digital Control. Pergamon, 1982. 315-320.

Examples

```
set.seed(1)
hopkins(iris[, 1:4], m=15) # .9952293

hop <- rep(NA, 100)
for(i in 1:100){
  hop[i] <- hopkins(iris[,1:4], m=8)
}
mean(hop)
```

---

hopkins.pval *Calculate the p-value for Hopkins statistic*

Description

Calculate the p-value for Hopkins statistic

Usage

```
hopkins.pval(x, n)
```

Arguments

x	Observed value of Hopkins statistic
n	Number of events/points sampled.

Details

Under null hypothesis of spatial randomness, Hopkins statistic has a Beta(m,m) distribution, where 'm' is the number of events/points sampled. This function calculates the p-value for the statistic.

Value

A p-value between 0 and 1.

Author(s)

Kevin Wright

References

Michael T. Gastner (2005). Spatial distributions: Density-equalizing map projections, facility location, and two-dimensional networks. Ph.D. dissertation, Univ. Michigan (Ann Arbor, 2005). http://hdl.handle.net/2027.42/125368

Examples

```
hopkins.pval(0.21, 10) # .00466205
```

Index