Package ‘PACBO’

July 5, 2016

Type Package

Title Clustering Online Datasets

Version 0.1.0

Date 2016-06-29

Author Le Li

Maintainer Le Li <le@iadvize.com>

Description A function for clustering online datasets. The number of cells is data-driven which need not to be chosen in advance by the user. The method is introduced and fully described in Le Li, Benjamin Guedj and Sebastien Loustat (2016), “PAC-Bayesian Online Clustering” (arXiv preprint: <https://arxiv.org/abs/1602.00522>).

License GPL (>= 2)


NeedsCompilation yes

Depends mnormt

Repository CRAN

RoxygenNote 5.0.1

Date/Publication 2016-07-05 11:46:43

R topics documented:

PACBO ................................................................. 2
runiform_ball ....................................................... 3

Index 5
Description

This function performs clustering on online datasets. The number of cells is data-driven and need not to be chosen in advance by the user.

Usage

\[
\text{PACBO}(\text{mydata}, R, \text{coeff} = 2, \text{K\_max} = 50, \text{scaling} = \text{FALSE}, \\
\text{var\_ind} = \text{FALSE}, \text{N\_iterations} = 500, \text{plot\_ind} = \text{FALSE}, \text{axis\_ind} = c(1, 2))
\]

Arguments

- **mydata**: a matrix where each row corresponds to an observation of length \(d\).
- **R**: a positive real value that should be larger than the maximum Euclidean distance of all the observations in \(\text{mydata}\). We recommend to set \(R\) equaling to this maximum Euclidean distance.
- **coeff**: a positive real value, enforcing large number of cells. The default, 2, should be convenient for most users. A larger value brings more cells for the clustering.
- **K\_max**: a positive integer indicating the maximum number of cells allowed for the clustering.
- **scaling**: logical indicating whether the matrix \(\text{mydata}\) should be centered and scaled. The centering is done by subtracting the column means of \(\text{mydata}\) from their corresponding columns; the scaling is done by dividing the (centered) columns of \(\text{mydata}\) by their standard deviations. We recommend to set it to \text{TRUE} only when the maximum Euclidean distance of all the observations in \(\text{mydata}\) is smaller than 1.
- **var\_ind**: logical indicating whether predicted centers of cells will be calculated sequentially. If \text{TRUE}, at each round, predicted centers of cells will be calculated on the basis of the past observations and past predicted centers. Setting this to \text{FALSE} will largely save execution time.
- **N\_iterations**: a positive integer indicating the number of iterations of algorithm.
- **plot\_ind**: logical indicating whether clusters should be plotted.
- **axis\_ind**: numeric indicating which axes are to be plotted if \(d \geq 2\). The default is the first two coordinates of observations.

Details

The PACBO algorithm is introduced and fully described in Le Li, Benjamin Guedj, Sebastien Loustau (2016), "PAC-Bayesian Online Clustering" (https://arxiv.org/abs/1602.00522). It relies on PAC-Bayesian approach, allowing for a dynamic (i.e., time-dependent) estimation of the number of clusters, up to \(\text{K\_max}\) clusters. Its implementation is done via an RJMCMC-flavored algorithm.
**runiform_ball**

**Value**

Returns a list including

- predicted_centers
  - a matrix of predicted centers of cells, where each row corresponds to a center.

- nb_of_clusters
  - positive integer indicating the estimation of the number of cells for the dataset.

- labels
  - labels for observations in mydata.

**Author(s)**

Le Li <le@iadvize.com>

**References**


**Examples**

```r
## generating 4 clusters of 100 points in \( \mathbb{R}^{5} \).
set.seed(100)
Nb <- 4
d <- 5
T <- 100
proportion = rep(1/Nb, Nb)
Mean_vectors <- matrix(runif(d*Nb, min=-10, max=10), nrow=Nb, byrow=TRUE)
mydata <- matrix(replicate(T, rmnorm(1, mean= Mean_vectors[sample(1:Nb, 1, prob = proportion), ],
varcov = diag(1,d))), nrow = T, byrow=TRUE)
R <- max(sqrt(rowSums(mydata^T)))
## run the algorithm.
result <- PACBO(mydata, R, plot_ind = TRUE)
```

---

**runiform_ball**

*Multivariate Uniform Distribution on a ball*

**Description**

This function generates random samples from multivariate uniform distribution on a ball in \( \mathbb{R}^{d} \), equipped with \( L^2 \) norm (*i.e.*, Euclidean distance), centered in \( \mathbf{0} \), with radius \( R \).

**Usage**

```r
runiform_ball(n, d, R)
```

**Arguments**

- \( n \)
  - number of desired samples.

- \( d \)
  - positive integer, representing the dimension of the observations.

- \( R \)
  - positive real value, the radius of the ball in \( \mathbb{R}^{d} \).
runiform_ball

Details
This function generates samples from the multivariate uniform distribution whose density is

\[ \pi(c, R) = \frac{\Gamma(d/2 + 1)}{\pi^{d/2}} \frac{1}{R^d} 1_{B_d(R)}(c), \]

where \( 1_{B_d(R)} \) is a centered \( \ell^2 \) ball with radius R.

Value
a matrix of \( n \) samples of length \( d \).

Examples
```r
# generating 10000 samples from uniform distribution on a unit ball in \( \text{\textbf{R}} \)^2
result <- runiform_ball(10000, 2, 1)
plot(result)
```
Index

PACBO, 2
runiform_ball, 3