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 Loustau (2016), ``PAC-Bayesian Online Clustering'' (arXiv preprint: <https://arxiv.org/abs/1602.00522>).

License GPL (>= 2)

URL https://arxiv.org/abs/1602.00522

NeedsCompilation yes

Depends mnormt

Repository CRAN

RoxygenNote 5.0.1

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

R topics documented:

PACBO *PACBO*

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

```
PACBO(mydata, R, coeff = 2, K_max = 50, scaling = FALSE,
  var_ind = FALSE, N_iterations = 500, plot_ind = FALSE, 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 `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 `mydata` should be centered and scaled. The centering is done by subtracting the column means of `mydata` from their corresponding columns; the scaling is done by dividing the (centered) columns of `mydata` by their standard deviations. We recommend to set it to `TRUE` only when the maximum Euclidean distance of all the observations in `mydata` is smaller than 1.
`var_ind`	logical indicating whether predicted centers of cells will be calculated sequentially. If `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 `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 >= 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 `K_max` clusters. Its implementation is done via an RJMCMC-flavored algorithm.

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

Le Li, Benjamin Guedj and Sebastien Loustau (2016), PAC-Bayesian Online Clustering, arXiv preprint: https://arxiv.org/abs/1602.00522.

Examples

```
## generating 4 clusters of 100 points in \strong{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,ncol=d, byrow=TRUE)
mydata <- matrix(replicate(T, rmnorm(1, mean= Mean_vectors[sample(1:Nb, 1, prob = proportion),],
varcov = diag(1,d))), nrow = T, byrow=T)
R <- max(sqrt(rowSums(mydata^2)))
##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 **R**^d, equipped with L^2 norm (*i.e.,* Euclidean distance), centered in **0**, with radius R.

Usage

```
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 **R**^d.

Details

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

$$\pi(c, R) = \Gamma(d/2 + 1)/\pi^{d/2} * 1/(R)^d 1_{B_d(R)}(c),$$

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

Value

a matrix of n samples of length d.

Examples

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

Index