Package 'Ckmeans.1d.dp'

February 19, 2015

Type Package

Title Optimal k-Means Clustering for One-Dimensional Data

Version 3.3.1

Date 2015-02-10

Author Joe Song and Haizhou Wang

Maintainer Joe Song <joemsong@cs.nmsu.edu>

Depends R (>= 2.10.0)

Description A dynamic programming algorithm for optimal one-dimensional k-means clustering. The algorithm minimizes the sum of squares of within-cluster distances. As an alternative to the standard heuristic k-means algorithm, this algorithm guarantees optimality and repeatability.

License LGPL (>= 3)

NeedsCompilation yes

Suggests testthat

Repository CRAN

Date/Publication 2015-02-11 00:41:47

R topics documented:

`Ckmeans.1d.dp` *Optimal K-means Clustering in One-dimension by Dynamic Programming*

Description

Perform optimal k-means clustering on one-dimensional data.

Usage

```
Ckmeans.1d.dp(x, k=c(1,9))
```

Arguments

x	a one-dimensional array containing input data to be clustered.
k	the number of clusters, or an array of required min and max numbers of clusters. The default is c(1,9). When a rannge is provided, the number of clusters will be determined within the range by Bayesian information criterion.

Details

Distance-based k-means clustering assigns all elements in the input vector x into k clusters to minimize the sum of squares of within-cluster distances (*withinss*) from each element to its corresponding cluster centre (mean). When a rannge is provided for k, the exact number of clusters will be determined within the range by Bayesian information criterion. The Ckmeans.1d.dp algorithm groups 1-D data given by x into k cluster by dynamic programming (Wang and Song, 2011). It guarantees the optimality of clustering – the sum of *withinss* for each cluster is always the minimum. In contrast, heuristic k-means algorithms may be inconsistent or non-optimal from run to run. The run time of the algorithm is $O($ max(k) n^2).

Value

An object of class "Ckmeans.1d.dp" which has a print method and is a list with components:

cluster	a vector of cluster indices assigned to each element in x. Each cluster is indexed by an integer from 1 to k.
centers	a vector of cluster centres.
withinss	the within-cluster sum of squares for each cluster.
size	a vector of the number of points in each cluster.

Author(s)

Joe Song and Haizhou Wang

References

Wang, H. and Song, M. (2011) Ckmeans.1d.dp: optimal k-means clustering in one dimension by dynamic programming. *The R Journal* **3**(2), 29–33. Retrieved from http://journal.r-project.org/archive/2011-2/RJournal_2011-2_Wang+Song.pdf

Examples

```
# Ex. 1 The number of clusters is provided.
# Generate data from a Gaussian mixture model of two components
x <- c(rnorm(50, sd=0.3), rnorm(50, mean=1, sd=0.3))
# Divide x into 2 clusters
k <- 2
result <- Ckmeans.1d.dp(x, k)
```

```
plot(x, col=result$cluster, pch=result$cluster, cex=1.5,
     main="Optimal k-means clustering",
     sub=paste("Number of clusters given:", k))
abline(h=result$centers, col=1:k, lty="dashed", lwd=2)
legend("bottomright", paste("Cluster", 1:k), col=1:k, pch=1:k, cex=1.5)

# Ex. 2 The number of clusters is determined by Bayesian information criterion
# Generate data from a Gaussian mixture model of two components
x <- c(rnorm(50, mean=-1, sd=0.3), rnorm(50, mean=1, sd=1))
# Divide x into k clusters, k automatically selected (default: 1~9)
result <- Ckmeans.1d.dp(x)
k <- max(result$cluster)
plot(x, col=result$cluster, pch=result$cluster, cex=1.5,
     main="Optimal k-means clustering",
     sub=paste("Number of clusters is estimated to be", k))
abline(h=result$centers, col=1:k, lty="dashed", lwd=2)
legend("topleft", paste("Cluster", 1:k), col=1:k, pch=1:k, cex=1.5)
```

`print.Ckmeans.1d.dp` *Print Results from Ckmeans.1d.dp*

Description

Print the result returned by calling Ckmeans.1d.dp

Usage

```
## S3 method for class 'Ckmeans.1d.dp'
print(x, ...)
```

Arguments

`x`	object returned by calling Ckmeans.1d.dp
`...`	Ignored arguments

Value

An object of class "`Ckmeans.1d.dp`" which has a print method and is a list with components:

`cluster`	a vector of integers (`1:k`) indicating the cluster to which each point is allocated.
`centers`	a vector of cluster centres.
`withinss`	the within-cluster sum of squares for each cluster.
`size`	a vector of the number of points in each cluster.

Author(s)

Joe Song and Haizhou Wang

References

Wang, H. and Song, M. (2011) Ckmeans.1d.dp: optimal *k*-means clustering in one dimension by dynamic programming. *The R Journal* **3**(2), 29–33. Retrieved from http://journal.r-project.org/archive/2011-2/RJournal_2011-2_Wang+Song.pdf

Examples

```
# Example: clustering data generated from a Gaussian mixture model of two components
x <- rnorm(50, mean=-1, sd=0.3)
x <- append(x, rnorm(50, mean=1, sd=0.3) )
res <- Ckmeans.1d.dp(x)
print(res)
```

Index