Package ‘LARkmeans’

November 15, 2016

Type Package

Title K-means Clustering

Version 1.0

Date 2016-11-09

Author Lasse Rintakumpu

Maintainer Lasse Rintakumpu <rintakumpu@gmail.com>

Description Provides a custom class for k-means clustering.

Suggests knitr, rmarkdown

VignetteBuilder knitr

License GPL (>= 2)

R topics documented:

`LARkmeans-package` *K-means Clustering*

Description

Provides a custom class for k-means clustering.

Details

This package was not yet installed at build time.

Index: This package was not yet installed at build time.

Author(s)

Lasse Rintakumpu

Maintainer: Lasse Rintakumpu <rintakumpu@gmail.com>

References

Forgy, E. W. (1965) Cluster analysis of multivariate data: efficiency vs interpretability of classifications. Biometrics **21**, 768–769.

Hamerly, G.; Elkan, C. (2002) Alternatives to the k-means algorithm that find better clusterings (PDF). Proceedings of the eleventh international conference on Information and knowledge management (CIKM).

Lloyd, S. P. (1957, 1982) Least squares quantization in PCM. Technical Note, Bell Laboratories. Published in 1982 in IEEE Transactions on Information Theory **28**, 128–137.

---

distEuclidean *Euclidean Distance*

Description

Calculates the Euclidean distance between two matrices with k columns.

Usage

```
distEuclidean(x, y)
```

Arguments

x	a matrix of size n x k.
y	a matrix of size m x k.

Value

a distance matrix of size n x m.

See Also

dist

Examples

```
## The function is currently defined as
function (x, y)
{
    z <- matrix(0, nrow = nrow(x), ncol = nrow(y))
    for (k in 1:nrow(y)) {
        z[, k] <- sqrt(colSums((t(x) - y[k, ])^2))
    }
    return(z)
  }
```

kMeans *K-Means Clustering*

Description

Performs k-means clustering `m` times on a data matrix. `fitted` returns a vector with the class labels of the best run.

Usage

```
kMeans(X, k, m=10, ind, max.iter=50, ...)

## S3 method for class 'kMeans'
fitted(object)
```

Arguments

`X`	a numeric matrix of data.
`k`	the desired number of clusters.
`m`	the number of times to run the clustering algorithm. The default is 10.
`ind`	a numeric vector of columns indicating the variables used in the clustering.
`max.iter`	the maximum number of iterations for a single run of the clustering algorithm. The default is 50.
`...`	not used.

Details

The matrix data given by X is clustered by the standard k-means method, also known as Lloyd-Forgy method (1957 & 1965). This method aims at minimizing the within-cluster sum of squares objective and thus assigns the clusters by the smallest Euclidean distance of observation to the cluster center.

The Random Partition method as described by Hamerly and Elkan (2002) is used for computing the initial cluster means.

Value

kMeans returns an object of class `kMeans` which has a print, summary, predict, plot and a fitted method. It is a list with the following components:

`Cbest`	the vector of the best group labels.
`ObjBest`	the value of the objective function for the best solution.
`CentroidsBest`	the matrix containing the centroids of the best solution.
`m`	the number of repetitions.
`k`	the number of groups.
`Xname`	name of the data set used for the clustering.
`Ind`	the value of input `ind`.
`Y`	the data used for the clustering.

`Best` value of which of the runs was the best.

`Call` a matrix with `n` rows and `m` columns giving the cluster assignments for all `m` runs.

`ObjAll` a vector having the objective functions of all runs.

`StatusAll` a vector having the status from all runs.

References

Forgy, E. W. (1965) Cluster analysis of multivariate data: efficiency vs interpretability of classifications. Biometrics **21**, 768–769.

Hamerly, G.; Elkan, C. (2002) Alternatives to the k-means algorithm that find better clusterings (PDF). Proceedings of the eleventh international conference on Information and knowledge management (CIKM).

Lloyd, S. P. (1957, 1982) Least squares quantization in PCM. Technical Note, Bell Laboratories. Published in 1982 in IEEE Transactions on Information Theory **28**, 128–137.

See Also

`kmeans`

Examples

```
# Example using random data from three different populations
# with one variable

set.seed(63555)
exampleData <- matrix(nrow=90, ncol=1)
exampleData[1:30, 1] <- rnorm(30, mean=3, sd=1)
exampleData[31:60, 1] <- rnorm(30, mean=6, sd=1)
exampleData[61:90, 1] <- rnorm(30, mean=9, sd=1)
kMeansResult <- kMeans(exampleData, k=3)

kMeansResult
# K-Means clustering for iris
# Number of runs: 10
# Status of best run: converged

fitted(kMeansResult)
# [1] 2 1 2 2 2 1 2 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 1 1 1 1 1 1 1 1 1 3 1 1 1 1 1
# [47] 1 1 3 1 1 1 1 1 1 1 1 1 3 1 3 3 3 3 3 3 3 3 1 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3

summary(kMeansResult)
#K-Means clustering for exampleData

#Clusters to be detected: 3
#Cluster sizes detected: 30 28 32

#Number of runs: 10
#Status of best run: converged

#Criterion value: 3522.284
#Summary of criterion values:
#Min: 3522.284
#Q1: 3546.504
```

```
#Mean: 3543.455
#Q3: 3546.504
#Max: 3546.504
```

plot.kMeans *Plot Method for K-Means Clustering*

Description

Plot method for objects of class "kMeans". Calls plot function to produce a scatter plot of the data values and cluster assignments provided by the given kMeans object.

Usage

```
plot.kMeans(X, col, pch, ...)
```

Arguments

X	an object of the class kMeans.
col	a specification for the plotting color.
pch	either an integer specifying a symbol or a single character to be used as the default in plotting points. See points for possible values and their interpretation.
...	further arguments to be passed to or from methods.

See Also

par

Examples

```
set.seed(63555)
exampleData <- matrix(nrow=90, ncol=1)
exampleData[1:30, 1] <- rnorm(30, mean=3, sd=1)
exampleData[31:60, 1] <- rnorm(30, mean=6, sd=1)
exampleData[61:90, 1] <- rnorm(30, mean=9, sd=1)
kMeansResult <- kMeans(exampleData, k=3)
plot(kMeansResult)
```

predict.kMeans *Predict Method for K-Means Clustering*

Description

This function assigns observations in the data matrix newData the most likeliest clusters using the best solution from a kMeans object.

Usage

```
predict.kMeans(X, newData)
```

Arguments

X object of class kMeans.

newData a data matrix or data frame having the same columns as the original X when kMeans was called.

Value

Returns a vector of cluster assignments for newData based on the kMeans object.

See Also

kMeans

Examples

```
set.seed(63555)
exampleData <- matrix(nrow=90, ncol=1)
exampleData[1:30, 1] <- rnorm(30, mean=3, sd=1)
exampleData[31:60, 1] <- rnorm(30, mean=6, sd=1)
exampleData[61:90, 1] <- rnorm(30, mean=9, sd=1)
kMeansResult <- kMeans(exampleData, k=3)

fitted(kMeansResult)
# [1] 2 1 2 2 2 1 2 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 1 1 1 1 1 1 1 1 1 3 1 1 1 1 1
# [47] 1 1 3 1 1 1 1 1 1 1 1 1 3 1 3 3 3 3 3 3 3 3 1 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3

newData <- matrix(nro=30, ncol=1)
newData[1:30, 1] <- rnorm(30, mean=6, sd=1)
predict(kMeansResult, newData)
# [1] 1 1 1 1 1 1 2 1 1 1 1 1 1 1 2 1 1 3 1 1 1 1 1 1 1 1 1 1 1 1
```

Index