Cluster Analysis: nagdmc_kmeans

Purpose

nagdmc_kmeans computes a k-means cluster analysis.

Declaration

```
#include <nagdmc.h>
void nagdmc_kmeans(long rec1, long nvar, long nrec, long dblk, double data[],
                   long nxvar, long xvar[], long iwts, long k, long ic[],
                   double c[], long maxit, int *info);
```

Parameters

1: **rec1** – `long` *Input*

On entry: the index in the data of the first data record used in the analysis.

Constraint: **rec1** ≥ 0.

2: **nvar** – `long` *Input*

On entry: the number of variables in the data.

Constraint: **nvar** ≥ 1.

3: **nrec** – `long` *Input*

On entry: the number of consecutive records, beginning at **rec1**, used in the analysis.

Constraint: **nrec** > 1.

4: **dblk** – `long` *Input*

On entry: the total number of records in the data block.

Constraint: **dblk** $\geq$ **rec1** $+$ **nrec**.

5: **data**[**dblk** $*$ **nvar**] – `double` *Input*

On entry: the data values for the jth variable (for $j = 0, 1, \ldots,$ **nvar**-1) are stored in **data**[$i*$**nvar**$+j$], for $i = 0, 1, \ldots,$ **dblk** $- 1$.

6: **nxvar** – `long` *Input*

On entry: the number of variables in the analysis. If **nxvar** $= 0$, all variables in the data, excluding **iwts** (if **iwts** ≥ 0), are used in the analysis.

Constraint: $0 \leq$ **nxvar** $\leq$ **nvar**.

7: **xvar**[**nxvar**] – `long` *Input*

On entry: the indices indicating the position in **data** in which the variables are stored. If **nxvar** $= 0$ then **xvar** must be 0, and the indices of variables are given by $j = 0, 1, \ldots,$ **nvar** $- 1$.

Constraints: if **nxvar** > 0, $0 \leq$ **xvar**[i] $<$ **nvar**, for $i = 0, 1, \ldots,$ **nxvar** $- 1$; otherwise **xvar** must be 0.

8: **iwts** – `long` *Input*

On entry: the index in **data** in which the weights are stored. If **iwts** $= -1$, no weights are used.

Constraints: $-1 \leq$ **iwts** $<$ **nvar**; if **nxvar** > 0, **iwts** $\neq$ **xvar**[i], for $i = 0, 1, \ldots,$ **nxvar** $- 1$.

9: **k** – `long` *Input*

On entry: the number of clusters to be formed.

Constraint: $1 <$ **k** $<$ **nrec**.

10: **ic**[**nrec**] – `long` *Input/Output*

On entry: the initial allocation of the **nrec** data records to clusters.
On exit: the final allocation of data records to clusters.

Constraints: $0 \leq$ **ic**[i] $<$ **k**, for $i = 0, 1, \ldots,$ **nrec** $- 1$.

11: **c[k∗nvar]** – `double` *Output*

On exit: the element $\mathbf{c}[i * \mathbf{nvar} + j]$ contains the mean of the jth variable for the ith cluster, for $i = 0, 1, \ldots, \mathbf{k} - 1$; for $j = 0, 1, \ldots, \mathbf{nvar} - 1$.

12: **maxit** – `long` *Input*

On entry: the maximum number of iterations used to compute the cluster analysis.

Constraint: $\mathbf{maxit} > 0$.

13: **info** – `int *` *Output*

On exit: **info** gives information on the success of the function call:

0: the function successfully completed its task.

-1: the computations have not converged in **maxit** iterations.

i; $i = 1, 2, 3, 4, 6, 7, \ldots, 10, 12$: the specification of the ith formal parameter was incorrect.

51: a weight is negative.

52: a cluster has no members in the initial allocation.

99: the function failed to allocate enough memory.

> 100: an error occurred in a function specified by the user.

Notation

nrec	the number of data records, n.
nxvar	the number of variables, p.
xvar	the variables that take the values in X.
iwts	if $\mathbf{iwts} \geq 0$, **iwts** is the index in the data that defines the weights, w_i, for $i = 1, 2, \ldots, n$.
k	the number of clusters, k.

Description

Let X be a data matrix of n data records on p variables and let $x_{ij} \in X$ denote the ith value of on the jth variable, for $j = 1, 2, \ldots, p$; for $i = 1, 2, \ldots, n$.

k-means clustering allocates each data record to one of k groups or clusters to minimise the within-cluster sum of squares:

$$\sum_{l=1}^{k} \sum_{i \in S_l} \sum_{j=1}^{p} (x_{ij} - \bar{x}_{lj})^2,$$

where S_l is the set of data records in the lth cluster and $\bar{x}_{lj}$ is the mean for the variable j over cluster l.

In addition to the data matrix, a k by p matrix giving the initial cluster centres for the k clusters is required. The objects are then initially allocated to the cluster with the nearest cluster mean. Given the initial allocation, the procedure is to search iteratively for the k-partition with locally optimal within-cluster sum of squares by moving points from one cluster to another.

Optionally, weights for each object, w_i, can be used so that the clustering is based on within-cluster weighted sums of squares:

$$\sum_{l=1}^{k} \sum_{i \in S_l} \sum_{j=1}^{p} w_i (x_{ij} - \tilde{x}_{lj})^2,$$

where $\tilde{x}_{lj}$ is the weighted mean for variable j over cluster l.

The algorithm used is based on the algorithm of Hartigan and Wong [3].

References and Further Reading

Everitt B S (1974) *Cluster Analysis* Heinemann.

Krzanowski W J (1990) *Principles of Multivariate Analysis* Oxford University Press.

Hartigan J A and Wong M A (1979) Algorithm AS136: A k-means clustering algorithm *Applied Statistics* **28** 100–108.

See Also

nagdmc_nrgp	allocates data records to the nearest cluster.
nagdmc_rints	can be used to form intial cluster centres at random.
nagdmc_wcss	computes the within-cluster sum of squares following a clustering.
kmeans_ex.c	the example calling program.