Cluster Analysis: nagdmc_nrgp

Purpose

nagdmc_nrgp finds the nearest group to a data record given the group centroids.

Declaration

```
#include <nagdmc.h>
void nagdmc_nrgp(long rec1, long nvar, long nrec, long dblk, double data[],
                 void (*dfun)(long, long, double [], char *, int *), char *comm,
                 long chunksize, long nxvar, long xvar[], long ng, double g[],
                 long ing[], 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** $- 1$. When the data function is used, **data** is not referenced.

6: **dfun** – function supplied by user *External Procedure*

On entry: the pointer to a data function supplied by the user.

Constraint: if **dfun** is a valid pointer, **data** must be 0.

The specification of **dfun** is:

```
void dfun(long irec, long chunksize, double x[], char *comm, int *ierr)
```

1: **irec** – `long` *Input*

On entry: the index in the data of the first record returned.

2: **chunksize** – `long` *Input*

On entry: the number of consecutive records returned.

3: **x**[**chunksize**$*$**nvar**] – `double` *Output*

On exit: data values for the jth variable (for $j = 0, 1, \ldots,$ **nvar** $- 1$) must be returned in **x**[$i *$ **nvar** $+ j$], for $i = 0, 1, \ldots,$ **chunksize** $- 1$.

4: **comm** – `char *` *Input*

On entry: a communication parameter allowing additional information to be passed to **dfun**. This parameter is passed 'as is' through the calling function.

5: **ierr** – `int *` *Output*
On exit: if the value pointed to by **ierr** on return is greater than 100, the NAG DMC function will terminate immediately and **info** will point to this value.

7: **comm** – `char *` *Input*
On entry: a communication parameter allowing additional information to be passed to **dfun**. This parameter is passed 'as is' through the calling function.

8: **chunksize** – `long` *Input*
On entry: if the data function is used, the function inputs no more than **chunksize** data records at a time; otherwise **chunksize** is not referenced.
Constraint: if **dfun** $\neq 0$, **chunksize** ≥ 1.

9: **nxvar** – `long` *Input*
On entry: the number of variables in the analysis. If **nxvar** $= 0$, all variables in the data are used in the analysis.
Constraint: $0 \leq$ **nxvar** $\leq$ **nvar**.

10: **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.

11: **ng** – `long` *Input*
On entry: the number of groups in the clustering.
Constraint: **ng** > 1.

12: **g[ng**$*$**nvar]** – `double` *Input*
On entry: **g**$[i *$ **nvar** $+ j]$ contains the mean value for the jth variable of the ith group, for $j = 0, 1, \ldots,$ **nvar** $- 1$; for $i = 0, 1, \ldots,$ **ng** $- 1$. Note that the value corresponding to the weights, if any, will be ignored.

13: **ing[nrec]** – `long` *Output*
On exit: **ing**$[i]$ is the nearest group to the ith data record in the analysis, for $i = 0, 1, \ldots,$ **nrec** $- 1$.

14: **info** – `int *` *Output*
On exit: **info** gives information on the success of the function call:

0: the function successfully completed its task.
i; $i = 1, 2, 3, 4, 6, 8, 9, \ldots, 11$: the specification of the ith formal parameter was incorrect.
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.
data the data set X.
nxvar determines the number of variables in the analysis.
ng the number of groups in the clustering.
g the vectors of group centroids c_k, for $k = 1, 2, \ldots, l$.
ing the allocation, $a_i - 1$, of data records to groups, for $i = 1, 2, \ldots, n$.

Description

Let X be a set of n data records x_i on p variables, for $i = 1, 2, \ldots, n$, and c_k be a user-supplied vector of p elements that defines the centroid of group k. Given the centroids of a clustering containing l

groups, the Euclidean distance, d_{ik}, from the ith data record to the kth centroid is:

$$d_{ik} = \left[\sum_{k=1}^{l}(x_{ij} - c_{kj})^2\right]^{\frac{1}{2}}, \quad i = 1, 2, \ldots, n,$$

where x_{ij} and c_{kj} are the values of the ith data record and kth centroid on variable j, respectively.

The ith data record is allocated to the group number a_i with the minimum distance in d_{ik}, for $k = 1, 2, \ldots, l$.

References and Further Reading

None.

See Also

None.