Nearest Neighbours: nagdmc\_knnc

Purpose

*nagdmc\_knnc* computes $k$-nearest neighbour classifications given a binary tree computed by *nagdmc\_kdtree* using training data.

Declaration

```c
#include <nagdmc.h>
void nagdmc_knnc(long rec1, long nvar, long nrec, long dblk, double data[],
                  long iproot, double prior[], double rho, long uc, int norm,
                  long k, long res[], long nn[], double dist[], int *info);
```

Parameters

1: \texttt{rec1} – long \hspace{1cm} Input
   \textit{On entry:} the index in the data of the first data record used in the analysis.
   \textit{Constraint:} \texttt{rec1} $\geq$ 0.

2: \texttt{nvar} – long \hspace{1cm} Input
   \textit{On entry:} the number of variables in the data.
   \textit{Constraint:} \texttt{nvar} > 1.

3: \texttt{nrec} – long \hspace{1cm} Input
   \textit{On entry:} the number of consecutive records, beginning at \texttt{rec1}, used in the analysis.
   \textit{Constraint:} \texttt{nrec} > 1.

4: \texttt{dblk} – long \hspace{1cm} Input
   \textit{On entry:} the total number of records in the data block.
   \textit{Constraint:} \texttt{dblk} $\geq$ \texttt{rec1} + \texttt{nrec}.

5: \texttt{data[dblk*nvar]} – double \hspace{1cm} Input
   \textit{On entry:} the data values for the $j$th variable (for $j = 0, 1, \ldots, \texttt{nvar} - 1$) are stored in \texttt{data[i*nvar+j]}, for $i = 0, 1, \ldots, \texttt{dblk} - 1$.

6: \texttt{iproot} – long \hspace{1cm} Input
   \textit{On entry:} the integer value of the root node of a binary tree as returned by *nagdmc\_kdtree*.

7: \texttt{prior[c]} – double \hspace{1cm} Input
   \textit{On entry:} if \texttt{prior} is set to 0, uniform priors are used; otherwise \texttt{prior[i]} gives the prior probability for the $i$th of $c$ categories on the dependent variable in the analysis, for $i = 0, 1, \ldots, c - 1$.
   \textit{Constraints:} if \texttt{prior} is not 0, \texttt{prior[i]} $\geq$ 0, for $i = 0, 1, \ldots, c - 1$, and the elements in \texttt{prior} must sum equal to one.

8: \texttt{rho} – double \hspace{1cm} Input
   \textit{On entry:} the value of maximum probability of group membership that must be exceeded for classification. Each data record with a maximum probability of group membership less than or equal to \texttt{rho} is classified as \texttt{uc}.
   \textit{Constraint:} 0 $\leq$ \texttt{rho} < 1.

9: \texttt{uc} – double \hspace{1cm} Input
   \textit{On entry:} the value that should be assigned to data records if the value of \texttt{rho} is not exceeded.

10: \texttt{norm} – int \hspace{1cm} Input
    \textit{On entry:} the norm used to compute distances. If \texttt{norm} = 1, the $\ell_1$-norm (or Manhattan distance) is used; otherwise \texttt{norm} = 2 and the $\ell_2$-norm (or Euclidean distance) is used.
    \textit{Constraint:} \texttt{norm} $\in \{1, 2\}$.
11: \( k \) — long
\( \text{Input} \)
On entry: the number of nearest neighbours used in the computation.
Constraint: \( 0 < k < \text{nrec} \).

12: \( \text{res}[\text{nrec}] \) — long
\( \text{Output} \)
On exit: \( \text{res}[i] \) contains the \( k \)-nearest neighbour classification of the \( i \)th data record, for \( i = 0, 1, \ldots, \text{nrec} - 1 \).

13: \( \text{nn}[\text{nrec} \times k] \) — long
\( \text{Output} \)
On exit: if \( \text{nn} \) is set to 0, it is not referenced; otherwise \( \text{nn}[i \times k + j] \) contains the index in the training data for the \( j \)th nearest neighbour to the \( i \)th data record, for \( j = 0, 1, \ldots, k - 1 \); for \( i = 0, 1, \ldots, \text{nrec} - 1 \).

14: \( \text{dist}[\text{nrec} \times k] \) — double
\( \text{Output} \)
On exit: if \( \text{dist} \) is set to 0, it is not referenced; otherwise \( \text{dist}[i \times k + j] \) contains the distance from the \( i \)th data record to its \( j \)th nearest neighbour, for \( j = 0, 1, \ldots, k - 1 \); for \( i = 0, 1, \ldots, \text{nrec} - 1 \).

15: \( \text{info} \) — int *
\( \text{Output} \)
On exit: \( \text{info} \) gives information on the success of the function call:
0: the function successfully completed its task.
1; \( i = i \); \( i = 1 \); \( i = 2 \); \( i = 3 \); \( i = 4 \); \( i = 7 \); \( i = 8 \); \( i = 10 \); \( i = 11 \): the specification of the \( i \)th formal parameter was incorrect.
57: information in the binary tree has been corrupted.
99: the function failed to allocate enough memory.
100: an internal error occurred during the execution of the function.

Notation
\( \text{nrec} \) the number of data records to classify, \( n \).
\( \text{data} \) the data values, \( X \).
\( \text{prior} \) the prior probabilities \( p_l \), for \( l = 1, 2, \ldots, c \).
\( \rho \) the threshold for accepting classifications, \( \rho \).
\( \text{uc} \) the dummy value representing unclassified data records, \( z \).
\( k \) the number of nearest neighbours used in the calculations, \( k \).
\( \text{res} \) the nearest neighbour classifications \( \hat{y}_i \), for \( i = 1, 2, \ldots, n \).

Description
Let \( X \) be a set of \( n \) data records \( x_i \), for \( i = 1, 2, \ldots, n \), on \( p \) independent variables and a categorical dependent variable \( y \). The \( j \)th value of the \( i \)th data record is denoted by \( x_{ij} \). Each member of \( X \) is to be classified into one of \( c \) categories where the prior probability of the \( l \)th category is \( p_l \), for \( l = 1, 2, \ldots, c \).

The \( k \)-nearest neighbour approach searches a set of training data records \( T \) (i.e., data records with known categories for \( y \)) to find the \( k \)-nearest data records to \( x_i \). Nearest neighbours are found by using a binary tree search, e.g., see Bentley (1975). The proximity of \( x_i \) to a member \( t \) of \( T \) is defined by a distance calculated over the independent variables and can be defined by using one of:

(a) the \( \ell_1 \)-norm or Manhattan distance:
\[
\sum_{j=1}^{p} |x_{ij} - t_j|,
\]
where \( | \cdot | \) denotes the modulus operator;
(b) the \( \ell_2 \)-norm or Euclidean distance:
\[
\left( \sum_{j=1}^{p} (x_{ij} - t_j)^2 \right)^{1/2}.
\]
Let \( S_i \) be a set containing the \( k \)-nearest neighbours in \( T \) to \( x_i \), and \( h_{il} \) be the number of members of \( S_i \) belonging to the \( l \)th category. The posterior probability \( \theta_{il} \) of \( x_i \) belonging to the \( l \)th category is given by,

\[
\theta_{il} = \frac{p_l h_{il}}{\sum_{m=1}^{c} p_m h_{im}}.
\]

Let \( q \) denote the index of the maximum value in \( \theta_{il} \), for \( l = 1, 2, \ldots, c \). Given a user-supplied value for \( \rho \), \( x_i \) is classified by setting the \( i \)th value of the dependent variable, \( \hat{y}_i \), to category value \( q \) if \( \theta_{iq} > \rho \); otherwise \( x_i \) is unclassified and \( \hat{y}_i \) is assigned a dummy value, say \( z \).

References and Further Reading

Bentley J L (1975) Multi-dimensional binary search trees used for associative searching Communications of the ACM 18(9) 509–517.


See Also

- nagdmc_kdtree computes a binary tree for a nearest neighbour analysis.
- nagdmc_free_kdtree frees the memory containing a binary tree.
- nagdmc_load_kdtree loads a binary tree from a file into memory.
- nagdmc_save_kdtree writes a binary tree to file.
- knnc_ex.c the example calling program.