Package 'OkNNE'

January 20, 2025

Type Package

Title A k-Nearest Neighbours Ensemble via Optimal Model Selection for Regression

Version 1.0.1

Description Optimal k Nearest Neighbours Ensemble is an ensemble of base k nearest neighbour models each constructed on a bootstrap sample with a random subset of features. k closest observations are identified for a test point ``x'' (say), in each base k nearest neighbour model to fit a stepwise regression to predict the output value of ``x''. The final predicted value of ``x'' is the mean of estimates given by all the models. The implemented model takes training and test datasets and trains the model on training data to predict the test data. Ali, A., Hamraz, M., Kumam, P., Khan, D.M., Khalil, U., Sulaiman, M. and Khan, Z. (2020) <DOI:10.1109/ACCESS.2020.3010099>.

Depends R (>= 3.5.0)

License GPL (>= 3)

Imports FNN,stats

NeedsCompilation no

Author Amjad Ali [aut, cre, cph],
Zardad Khan [aut, ths],
Muhammad Hamraz [aut]

Maintainer Amjad Ali <aalistat1@gmail.com>

Repository CRAN

Date/Publication 2022-12-19 15:00:09 UTC

Contents

1

| OkNNE-package | *A k-Nearest Neighbours Ensemble via Optimal Model Selection for Regression* |

Description

Optimal k-Nearest Neighbours Ensemble "OkNNE" is an ensemble of base k-NN models each constructed on a bootstrap sample with a random subset of features. k closest observations are identified for a test point "x" (say), in each base k-NN model to fit a stepwise regression to predict the output value of "x". The final predicted value of "x" is the mean of estimates given by all the models. OkNNE takes training and test datasets and trains the model on training data to predict the test data.

Details

Package:	OkNNE
Type:	Package
Version:	1.0.0
Date:	2020-07-22
License:	GPL-3

Author(s)

Amjad Ali, Muhammad Hamraz, Zardad Khan

Maintainer: Amjad Ali <aalistat1@gmail.com>

References

A. Ali et al., "A k-Nearest Nieghbours Based Ensemble Via Optimal Model Selection For Regression," in IEEE Access, doi: 10.1109/ACCESS.2020.3010099.

Li, S. (2009). Random KNN modeling and variable selection for high dimensional data.

Shengqiao Li, E James Harner and Donald A Adjeroh. (2011). Random KNN feature selection- a fast and stable alternative to Random Forests. BMC Bioinformatics , 12:450.

Alina Beygelzimer, Sham Kakadet, John Langford, Sunil Arya, David Mount and Shengqiao Li (2019). FNN: Fast Nearest Neighbor Search Algorithms and Applications. R package version 1.1.3.

Venables, W. N. and Ripley, B. D. (2002). Modern Applied Statistics with S. New York: Springer (4th ed).

OKNNE	*Optimal k-Nearest Neighbours Ensemble*

Description

Optimal k-Nearest Neighbours Ensemble "OkNNE" is an ensemble of base k-NN models each constructed on a bootstrap sample with a random subset of features. k closest observations are identified for a test point "x" (say), in each base k-NN model to fit a stepwise regression to predict the output value of "x". The final predicted value of "x" is the mean of estimates given by all the models. OkNNE takes training and test datasets and trains the model on training data to predict the test data.

Usage

```
OKNNE(xtrain, ytrain, xtest = NULL, ytest = NULL, k = 10, B = 100,
direction = "forward", q = trunc(sqrt(ncol(xtrain))), algorithm =
c("kd_tree", "cover_tree", "CR", "brute"))
```

Arguments

xtrain	The features space of the training dataset.
ytrain	The response variable of training dataset.
xtest	The test dataset to be predicted.
ytest	The response variable of test dataset.
k	The maximum number of nearest neighbors to search. The default value is set to 10.
B	The number of bootstrap samples.
direction	Method used to fit stepwise models. By default forward procedure is used.
q	The number of features to be selected for each base k-NN model.
algorithm	Method used for searching nearest neighbors.

Value

PREDICTIONS	Predicted values for test data response variable
RMSE	Root mean square error estimate based on test data
R.SQUARE	Coefficient of determination estimate based on test data
CORRELATION	Correlation estimate based on test data

Author(s)

Amjad Ali, Muhammad Hamraz, Zardad Khan

Maintainer: Amjad Ali <aalistat1@gmail.com>

References

A. Ali et al., "A k-Nearest Nieghbours Based Ensemble Via Optimal Model Selection For Regression," in IEEE Access, doi: 10.1109/ACCESS.2020.3010099.

Li, S. (2009). Random KNN modeling and variable selection for high dimensional data.

Shengqiao Li, E James Harner and Donald A Adjeroh. (2011). Random KNN feature selection - a fast and stable alternative to Random Forests. BMC Bioinformatics , 12:450.

Alina Beygelzimer, Sham Kakadet, John Langford, Sunil Arya, David Mount and Shengqiao Li (2019). FNN: Fast Nearest Neighbor Search Algorithms and Applications. R package version 1.1.3.

Venables, W. N. and Ripley, B. D. (2002). Modern Applied Statistics with S. New York: Springer (4th ed).

Examples

```
data(SMSA)

anyNA(SMSA)
#[1] FALSE

dim(SMSA)
#[1] 59 15

n=nrow(SMSA)

X <- SMSA[names(SMSA)!="NOx"]
Y <- SMSA[names(SMSA)=="NOx"]

set.seed(25)
train.obs <- sample(1:n, 0.7*n, replace = FALSE)
test.obs <- (1:n)[-train.obs]
xtrain <- X[train.obs,]; ytrain <- Y[train.obs,];
xtest <- X[test.obs,]; ytest <- Y[test.obs,]

OkNNE.MODEL = OKNNE(xtrain = xtrain, ytrain = ytrain, xtest = xtest, ytest
= ytest, k = 10, B = 5, q = trunc(sqrt(ncol(xtrain))), direction = "both",
algorithm=c("kd_tree", "cover_tree", "CR", "brute"))

OkNNE.MODEL
```

SMSA   *Standard Metropolitan Statistical Areas*

Description

The properties of Standard Metropolitan Statistical Areas (a standard Census Bureau designation of the region around a city) in the United States, collected from a variety of sources. The data include information on the social and economic conditions in these areas, on their climate, and some indices of air pollution potentials. The dataset has 59 observations on 15 variables.

Usage

```
data("SMSA")
```

Format

A data frame with 59 observations on the following 15 variables.

JanTemp Mean January temperature (in degrees Farenheit)

JulyTemp Mean July temperature (in degrees Farenheit)

RelHum Relative Humidity

Rain Annual rainfall (in inches)

Mortality Age adjusted mortality

Education Median education

PopDensity Population density

PerNonWhite Percentage of non whites

PerWC Percentage of white collour workers

pop Population

popPerhouse Population per household

income Median income

HCPot HC pollution potential

S02Pot Sulfur Dioxide pollution potential

NOx Nitrous Oxide (target variable)

Source

https://www.openml.org/d/1091

References

U.S. Department of Labour Statistics Authorization: free use

Examples

```
data(SMSA)
## maybe str(SMSA) ; plot(SMSA) ...
```

Index