Package 'NPNP'

November 25, 2016

Title An optimal semi-parametric diagnosis test using multiple
   biomarkers

Version 1.0

Author Ting Lyu and Hong Zhang

Maintainer Hong Zhang <zhanghfd@fudan.edu.cn>

Description NPNP is an R package used for diagnosis test through combining multiple biomarkers.

Depends mnormt, R (>= 2.14)

License GPL (>= 2)

Encoding UTF-8

LazyData true

URL http://github.com/zhanghfd/NPNP

BugReports http://github.com/zhanghfd/NPNP/issues

RoxygenNote 5.0.1

R topics documented:

NPNP-package   *An optimal semi-parametric diagnosis test using multiple biomarkers*

Description

NPNP is a semi-parametric diagnosis test that combines multiple biomarkers. The biomarker measures are not necessary to be normally distributed, and a data dependent non-linear transformation function is applied to each biomarker using training data. NPNP is based on the Neyman-Pearson lemma so that is statistically efficient.

This package includes one main function (i.e., NPNP) and two simulated datasets (i.e., train and test) for illustration. Using NPNP, a diagnosis model is trained using training samples and new subject(s) can be diagnosed. Two candidate estimation methods can be used to estimate the unknown transformation function, one is 'simultaneous' and the other is 'sequential'.

Details

Package:	NPNP
Type:	Package
Version:	1.0
Date:	2016-11-15
License:	GPL (>= 2)

Author(s)

Ting Lyu and Hong Zhang

Maintainer: Hong Zhang <zhanghfd@fudan.edu.cn>

References

Lyu T, Ying Z, Zhang H. (2016) An optimal semi-parametric diagnostic test using multiple biomarkers.

NPNP   *Disease diagnosing through combining multiple biomarkers.*

Description

This function returns false positive rates (FPRs), false negative rates (FNRs), area under ROC curve (AUC), and possibly predicted disease status if test samples are provided. The major inputs of this function include (1) training data information (both biomarker values and disease statuses) and (2) testing data information (only biomarker values) if possible.

Usage

```
NPNP(train_validation,test_biomarker=NULL,tv_ratio=1,n.validation=5,fpr_thre=NULL,transformation
```

Arguments

train_validation

> A matrix containing the sample information on both disease status and biomarkers that is used for training and validation. The first column is for disease statuses and the other columns are for biomarker values.

test_biomarker   A matrix containing the test sample information on biomarkers. The biomarkers in this matrix should correspond to those in 'train_validation'.

tv_ratio   Training versus validation sample size ratio, with the default value of 1.

n.validation   Number of random validation experiments, with the default value of 5.

fpr_thre   False positive rate that determines the threshold for predicting disease status.

transformation Method for estimating transformation function, which is either 'simultaneous' (default) or 'sequential'.

Value

fpr	A vector containing false positive rates 0.01, 0.02, ..., 0.99.
fnr	A vector containing false negative rates corresponding to false positive rates.
auc	Area under ROC curve.
predict_d	A vector containing predicted disease statuses, which is provided if 'test_biomarker' is available. The threshold used for predicting the disease statuses either is determined by 'fpr_thre' (if available) or corresponds to the minimal value of fpr+fnr.

Examples

```
data(train);
data(test);

train_validation=as.matrix(train);

test_biomarker=as.matrix(test)[,-1];

res = NPNP(train_validation,test_biomarker=test_biomarker,transformation=simultaneous);

plot(res$fpr,1-res$fnr,xlab=1-Specificity,ylab=Sensitivity,type=l); # ROC curve

auc = res$auc;  # Area under ROC curve

predict_d = res$predict_d; # Predicted disease statuses
```

---

test   *A simulated testing dataset.*

Description

The dataset contains the information on disease status and values from three biomarkers. This dataset is used as testing sample.

Usage

```
data(test)
```

Format

A data frame containing the following variables for each of 300 observations: status (disease status coded as 1 for case or 0 for control), values from three biomarkers.

Examples

```
data(test);
```

---

train   *A simulated training dataset.*

Description

The dataset contains the information on disease status and values from three biomarkers. This dataset is used as training sample.

Usage

```
data(train)
```

Format

A data frame containing the following variables for each of 300 observations: status (disease status coded as 1 for case or 0 for control), values from three biomarkers.

Examples

```
data(train);
```

Index