Package ‘diproperm’

May 14, 2021

Type Package
Title Conduct Direction-Projection-Permutation Tests and Display Plots
Version 0.2.0
License GPL-3
Encoding UTF-8
LazyData true
RoxygenNote 7.1.0
Suggests knitr, rmarkdown
Imports usethis, ggplot2, lemon, gridExtra, parallel, dplyr, DWDLargeR, e1071, Matrix, SparseM, sampling
Depends R (>= 2.10)
NeedsCompilation no
Author Andrew G. Allmon [aut, cre], J.S. Marron [aut], Michael G. Hudgens [aut]
Maintainer Andrew G. Allmon <allmondrew@yahoo.com>
Repository CRAN
Date/Publication 2021-05-14 20:02:12 UTC

R topics documented:

DiProPerm ......................... 2
loadings ................................ 4
mushrooms ................................ 5
plotdpp ................................ 5

Index 7
DiProPerm conducts a DiProPerm test

Description

This package conducts a Direction-Projection-Permutation (DiProPerm) test. DiProPerm is a two-sample hypothesis test for comparing two high-dimensional distributions. The DiProPerm test is exact, i.e., the type I error is guaranteed to be controlled at the nominal level for any sample size. For more details see Wei et al. (2016).

Usage

DiProPerm(
  X, 
  y, 
  B = 1000, 
  classifier = "dwd", 
  univ.stat = "md", 
  balance = TRUE, 
  alpha = 0.05, 
  cores = 2 
)

Arguments

x                     An n x p data matrix.
y                     A vector of n binary class labels -1 and 1.
B                     The number of permutations for the DiProPerm test. The default is 1000.
classifier            A string designating the binary linear classifier. classifier="dwd", distance weighted discrimination (DWD), is the default. classifier="dwd" implements a generalized DWD model from the genDWD function in the DWDLargeR package. The penalty parameter, C, in the genDWD function is calculated using the penaltyParameter function in DWDLargeR. The genDWD and penaltyParameter functions have several arguments which are set to recommended default values. More details on the algorithm used to calculate the DWD solution can be found in Lam et al. (2018). Other options for the binary classifier include the "md", mean difference direction, and "svm", support vector machine. The "svm" option uses the default implementation from svm.
univ.stat             A string indicating the univariate statistic used for the projection step. univ.stat="md", the mean difference, is the default.
balance               A logical indicator for whether a balanced permutation design should be implemented. The default is TRUE.
alpha                 An integer indicating the level of significance. The default is 0.05.
cores                 An integer indicating the number of cores to be used for parallel processing. The default is 2. Note, parallel processing is only available on MacOS and Ubuntu operating systems at this time. Windows users will default to using 1 core.
Value

A list containing:

- \( X \) The observed \( n \times p \) data matrix.
- \( y \) The observed vector of \( n \) binary class labels -1 and 1.
- \( \text{obs.teststat} \) The observed univariate test statistic.
- \( x_w \) Projection scores used to compute the specified univariate statistic.
- \( w \) The loadings of the binary classification.
- \( Z \) The \( Z \) score of the observed test statistic.
- \( \text{cutoff.value} \) The cutoff value to achieve an alpha level of significance.
- \( \text{pvalue} \) The pvalue from the permutation test.
- \( \text{perm.dist} \) A list containing the permuted projection scores and permuted class labels for each permutation.
- \( \text{perm.stats} \) A \( B \) dimensional vector of univariate test statistics.

Author(s)

Andrew G. Allmon, J.S. Marron, Michael G. Hudgens

References


Examples

data(mushrooms)
\`
X <- Matrix::t(mushrooms$X)
y <- mushrooms$y
dpp <- DiProPerm(X=X, y=y, B=10)
\`
loadings

Returns the loadings of the binary linear classifier (e.g. DWD)

Description

Returns the variable indexes who had the highest loadings in the binary classification. Higher loading values indicate a variable’s contribution toward the separation between the two classes. The loadings vector is a unit vector; thus the individual loadings range from -1 to 1, and the sum of the squares of the loadings equals one. The loadings direction vector points from the negative to positive class. Thus, positive entries correspond to variables that tend to be larger for the positive class.

Usage

loadings(dpp, loadnum = length(dpp$w))

Arguments

dpp A DiProPerm list.
loadnum An integer indicating the number of variables to display. For example, if loadnum=5 then the indexes for the five variables who contributed most toward the separation of the two classes are displayed. The default is to print out all the loadings.

Value

Returns the indexes and loadings for variables which contributed the most toward the separation of the binary classifier.

Author(s)

Andrew G. Allmon, J.S. Marron, Michael G. Hudgens

Examples

data(mushrooms)
X <- Matrix::t(mushrooms$X)
y <- mushrooms$y
dpp <- DiProPerm(X=X, y=y, B=10)
loadings(dpp, loadnum=3)
**mushrooms**

*Classification data from Audobon Society Field Guide (1981)*

**Description**

This data set includes descriptions of hypothetical samples corresponding to 23 species of gilled mushrooms in the Agaricus and Lepiota family. Each species is identified as definitely edible, definitely poisonous, or of unknown edibility and not recommended. The unknown class was combined with the poisonous class such that there were two classes: definitely edible and poisonous/unknown.

**Usage**

mushrooms

**Format**

A list (X) containing a 112x8124 matrix of 8124 mushrooms with 112 features; and an outcome vector (y) of length 8124 containing the class information (-1 = "edible", 1 = "poisonous/unknown"). The 112 features are 0/1 dummy variables for 22 different categorical attributes. All 22 attributes and their categories are displayed at the source link below.

**Source**

This data can be found at the UCI Machine Learning Data Repository website. [https://archive.ics.uci.edu/ml/datasets/Mushroom](https://archive.ics.uci.edu/ml/datasets/Mushroom)

**References**


---

**plotdpp**

*Plots diagnostics from DiProPerm test*

**Description**

This function plots the diagnostics of a DiProPerm test including the projection scores for the observed data, projection scores for the permutations with the smallest and largest univariate statistic values, and permutation distribution for B univariate statistics.

**Usage**

plotdpp(dpp, plots = "all", w = 0.001, h = 0.001)
Arguments

\begin{itemize}
\item \texttt{dpp}  
A DiProPerm object.
\item \texttt{plots}  
A string designating the desired plots to be displayed: "obs" displays the projection scores for the observed data, "min" displays the projection scores for the permutation with the smallest univariate statistic value, "max" displays the projection scores for the permutation with the largest univariate statistic value, "permdist" displays the permutation distribution for \( B \) univariate statistics, and "all" displays all 4 diagnostic plots in one plot. Additionally, one can specify "perm1" to display the projection scores for the first permutation and "perm2" to display the projection scores for the second permutation.
\item \texttt{w}  
An integer indicating the width of the jitter. The default is 0.001.
\item \texttt{h}  
An integer indicating the height of the jitter. The default is 0.001.
\end{itemize}

Value

A ggplot

Author(s)

Andrew G. Allmon, J.S. Marron, Michael G. Hudgens

Examples

\begin{verbatim}
data(mushrooms) X <- Matrix::t(mushrooms$X) y <- mushrooms$y dpp <- DiProPerm(X=X, y=y, B=10) plotdpp(dpp)
\end{verbatim}
Index

* datasets
  mushrooms, 5
DiProPerm, 2
genDWD, 2
loadings, 4
mushrooms, 5
penaltyParameter, 2
plotdpp, 5
svm, 2