Package ‘HSTree’

January 7, 2023

Type Package

Title Hierarchical Shrinkage: improving the accuracy and interpretability of tree-based methods

Version 0.8.0

Author Haoxue Wang, Abhineet Agarwal, Yan Shuo Tan, Omer Ronen, Chandan Singh, Bin Yu

Maintainer Haoxue Wang <hw613@cam.ac.uk>

Description Hierarchical shrinkage is an extremely fast post-hoc regularization method which works on any decision tree (or tree-based ensemble, such as Random Forest). It does not modify the tree structure, and instead regularizes the tree by shrinking the prediction over each node towards the sample means of its ancestors (using a single regularization parameter). Experiments over a wide variety of datasets show that hierarchical shrinkage substantially increases the predictive performance of individual decision trees and decision-tree ensembles.``Hierarchical shrinkage" is first defined in Agarwal et al. (2022) <https://proceedings.mlr.press/v162/agarwal22b.html>

Imports rpart, randomForest, gbm

License MIT + file LICENSE

URL https://github.com/wanghaoxue0/HSTree

BugReports https://github.com/wanghaoxue0/HSTree/issues

Encoding UTF-8

RoxygenNote 7.2.2

R topics documented:

HSTreeClassifier *HSTree fit for classification*

Description

modify the decision tree(or each tree in ensemble) structure based on hierarchical shrinkage regularization

Usage

```
HSTreeClassifier(X, y, reg_param=1, max_leaf_nodes=20, interaction.depth=1, estimator="CART", shri
```

Arguments

`X` the design matrix

`y` the response vector

`reg_param` Higher is more regularization (can be arbitrarily large, should not be < 0)

`max_leaf_nodes` the maximum number of leaf nodes, the default is 20

`interaction.depth` Integer specifying the maximum depth of each tree (i.e., the highest level of variable interactions allowed). A value of 1 implies an additive model, a value of 2 implies a model with up to 2-way interactions, etc. Default is 1.

`estimator` CART decision tree or tree ensemble model (e.g. RandomForest or Gradient-Boosting) Defaults to CART Classification Tree with 20 max leaf nodes Note: this estimator will be directly modified and keep its original functions

`shrinkage` shrinkage methods, default is "node_based", options are: 1. node_based shrinks based on number of samples in parent node 2. leaf_based only shrinks leaf nodes based on number of leaf samples 3. constant shrinks every node by a constant lambda

Value

call: the input setting

fit: keep all the output as the orginal estimator, only replace the value in each node based on shrinkage methods

regularization: chr "HSTree"

shrinkage: the hierarchical shrinkage method used in this model

class: the estimator class

HSTreeClassifierCV *HSTree fit for classification with cross validation*

Description

modify the decision tree(or each tree in ensemble) structure with the best regularization parameter chosen by cross validation

Usage

```
HSTreeClassifierCV(X, y, cv=4, verbose=FALSE, reg_param=c(0.1, 1, 10, 50, 100, 500), max_leaf_nodes
```

Arguments

`X`	the design matrix
`y`	the response vector
`cv`	cross validation level, default is 4 (divide the data into 4 portions, train:test=3:1)
`verbose`	whether to print the cross validation process and result
`reg_param`	numerical array of regularization parameter, default is c(0.1, 1, 10, 50, 100, 500)
`max_leaf_nodes`	the maximum number of leaf nodes, the default is 20
`interaction.depth`	Integer specifying the maximum depth of each tree (i.e., the highest level of variable interactions allowed). A value of 1 implies an additive model, a value of 2 implies a model with up to 2-way interactions, etc. Default is 1.
`estimator`	CART decision tree or tree ensemble model (e.g. RandomForest or GradientBoosting) Defaults to CART Classification Tree with 20 max leaf nodes Note: this estimator will be directly modified and keep its original functions
`shrinkage`	shrinkage methods, default is "node_based", options are: 1. node_based shrinks based on number of samples in parent node 2. leaf_based only shrinks leaf nodes based on number of leaf samples 3. constant shrinks every node by a constant lambda

Value

call: the input setting

fit: keep all the output as the orginal estimator, only replace the value in each node based on shrinkage methods

regularization: chr "HSTree"

shrinkage: the hierarchical shrinkage method used in this model

class: the estimator class

HSTreeRegressor *HSTree fit for regression*

Description

modify the decision tree(or each tree in ensemble) structure based on hierarchical shrinkage regularization

Usage

```
HSTreeRegressor(X, y, reg_param=1, max_leaf_nodes=20, interaction.depth=1, estimator="CART", shrin
```

Arguments

`X` the design matrix

`y` the response vector

`reg_param` Higher is more regularization (can be arbitrarily large, should not be < 0)

`max_leaf_nodes` the maximum number of leaf nodes, the default is 20

`interaction.depth` Integer specifying the maximum depth of each tree (i.e., the highest level of variable interactions allowed). A value of 1 implies an additive model, a value of 2 implies a model with up to 2-way interactions, etc. Default is 1.

`estimator` CART decision tree or tree ensemble model (e.g. RandomForest or GradientBoosting) Defaults to CART Classification Tree with 20 max leaf nodes Note: this estimator will be directly modified and keep its original functions

`shrinkage` shrinkage methods, default is "node_based", options are: 1. node_based shrinks based on number of samples in parent node 2. leaf_based only shrinks leaf nodes based on number of leaf samples 3. constant shrinks every node by a constant lambda

Value

call: the input setting

fit: keep all the output as the orginal estimator, only replace the value in each node based on shrinkage methods

regularization: chr "HSTree"

shrinkage: the hierarchical shrinkage method used in this model

class: the estimator class

HSTreeRegressorCV *HSTree fit for regression with cross validation*

Description

modify the decision tree(or each tree in ensemble) structure with the best regularization parameter chosen by cross validation

Usage

```
HSTreeRegressorCV(X, y, cv=4, verbose=FALSE, reg_param=c(0.1, 1, 10, 50, 100, 500), max_leaf_nodes=
```

Arguments

`X`	the design matrix
`y`	the response vector
`cv`	cross validation level, default is 4 (divide the data into 4 portions, train:test=3:1)
`verbose`	whether to print the cross validation process and result
`reg_param`	numerical array of regularization parameter, default is c(0.1, 1, 10, 50, 100, 500)
`max_leaf_nodes`	the maximum number of leaf nodes, the default is 20
`interaction.depth`	Integer specifying the maximum depth of each tree (i.e., the highest level of variable interactions allowed). A value of 1 implies an additive model, a value of 2 implies a model with up to 2-way interactions, etc. Default is 1.
`estimator`	CART decision tree or tree ensemble model (e.g. RandomForest or GradientBoosting) Defaults to CART Classification Tree with 20 max leaf nodes Note: this estimator will be directly modified and keep its original functions
`shrinkage`	shrinkage methods, default is "node_based", options are: 1. node_based shrinks based on number of samples in parent node 2. leaf_based only shrinks leaf nodes based on number of leaf samples 3. constant shrinks every node by a constant lambda

Value

call: the input setting

fit: keep all the output as the orginal estimator, only replace the value in each node based on shrinkage methods

regularization: chr "HSTree"

shrinkage: the hierarchical shrinkage method used in this model

class: the estimator class

Index