

hhcart: An implementation of HHCART(G) - A reflected feature space for CART

Phil Davies

02/05/2020

hhcart: An implementation of HHCART(G) - A reflected feature space for CART.

This vignette is intended to provide details of the functions and methods in the hhcarr package. The hhcarr programs build classification or regression models using the HHCART(G) algorithm (?) to induce oblique decision trees.

This vignette is divided into the following sections:

The HHCART(G) Algorithm

In this section, we will describe HHCART(G) (?) an oblique classification tree algorithm. But first, to place it in context we will review both the HHCART algorithm (?) and the Geometric decision tree (GDT) algorithm (?) as HHCART(G) uses a modified GDT angle bisector to define an alternative reflected feature space for the HHCART algorithm.

HHCART

HHCART is a comprehensive extension of the work of ?, the HHCART algorithm finds oblique splits which can be linear combinations of both quantitative and qualitative features (?). We will use a two-class classification problem to explain the concept of HHCART. Although we look at a two-class problem here, the algorithm does generalise to multi-class problems. To split a node, the author's approach is to find a separating hyperplane for each class and to examine the resulting orientation of each. This can be taken as the most stretched direction as represented by the dominant eigenvector from the covariance matrix for each class.

For a set of training examples, the estimated covariance matrix can be defined as follows. Let x_1, x_2, \dots, x_n be p dimensional feature vectors, where $x_i = (x_{i1}, x_{i2}, \dots, x_{ip})^T$ and p the number of feature variables. The estimated covariance matrix can then be written as:

$$S = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(x_i - \bar{x})^T$$

Where $\bar{x} = (\bar{x}_1, \bar{x}_2, \dots, \bar{x}_p)^T$ is a vector of feature column means and n , the number of training examples. Given there are two classes, two dominant eigenvectors, d^1 and d^2 can be found, one for each class.

A hyperplane, parallel to either d^1 or d^2 can be a candidate direction for a class separating hyperplane. If one of the eigenvectors, say d^1 is reflected such that it becomes parallel to one of the coordinate axes, e_1 , then the orientation of the separating hyperplane will also be parallel to e_1 in the reflected space. The separating hyperplane can now be found by performing axis-parallel splits along the e_2 direction in the reflected space (?).

Householder Reflection

The concept of the HHCART algorithm is reflecting the set of training examples using a householder matrix, which was used by ? to induce space partitions. For a p -dimensional space, a householder matrix H can be defined as follows:

Let d^1 be the normalised dominant eigenvector for the class 1 training examples with $e_{1_{p \times 1}} = (1, 0, \dots, 0)^T$ a basis vector. Both vectors, d^1 and e_1 , will have the same norm. Then there exists an orthogonal symmetric matrix $H_{p \times p}$, with p the number of features, such that:

$$\begin{aligned} e_1 &= Hd^1 \\ \text{where } H &= I - 2uu^T \\ \text{and } u &= \frac{e_1 - d^1}{\|e_1 - d^1\|_2} \end{aligned} \tag{1}$$

Construction of the householder matrix

This section will explain the construction of the householder matrix. Let d^1 be a vector to be reflected using the matrix H such that $e_1 = Hd^1$. Vectors d^1 and e_1 are shown in Figure-??.

Vector e_1 can be written as the addition of two vectors, $e_1 = d^1 + \tilde{u}$, where $\tilde{u} = e_1 - d^1$. From Figure-?? as e_1 is the reflection of d^1 , we can see that $AB = \|d^1\| \cos \theta$ and $AC = 2AB$. Therefore:

$$\begin{aligned} \tilde{u} &= 2\|d^1\| \frac{\tilde{u}}{\|\tilde{u}\|} \cos \theta \\ \text{hence, } e_1 &= d^1 + 2\|d^1\| \frac{\tilde{u}}{\|\tilde{u}\|} \cos \theta \\ \text{since, } d^1 \tilde{u} &= \|d^1\| \|\tilde{u}\| \cos(\pi - \theta) \\ \text{we have: } \cos \theta &= \frac{d^1 \cdot \tilde{u}}{\|d^1\| \|\tilde{u}\|} \\ \text{and } e_1 &= d^1 - 2\|d^1\| \frac{\tilde{u}}{\|\tilde{u}\|} \frac{d^1 \cdot \tilde{u}}{\|d^1\| \|\tilde{u}\|} \\ e_1 &= d^1 - 2\tilde{u} \frac{\tilde{u}^T d^1}{\|\tilde{u}\|_2} \end{aligned} \tag{2}$$

Since $\tilde{u} = e_1 - d^1$, equation (??) gives $\tilde{u} = \|\tilde{u}\|u$. Therefore, substituting \tilde{u} in equation (??) by $\|\tilde{u}\|u$ gives:

$$\begin{aligned} e_1 &= (I - 2uu^T)d^1 \\ \text{that is, } H &= I - 2uu^T \end{aligned}$$

Since the householder matrix, $H_{p \times p}$ is both symmetric and orthogonal, a point in the transformed space can be mapped back to the original space at minimal cost. Let x be a point in the original space, if the transformed point is defined as $\hat{x} = Hx$, multiplying both sides by H gives $H\hat{x} = HHx$. Since H is orthogonal and symmetric, $H^T = H^{-1}$, therefore, $HH = I$, where I is the identity matrix. Therefore, $H\hat{x} = x$.

The householder reflection

Denoting the full set of training examples as $\mathfrak{D}_{n \times p}$, the reflected example set $\hat{\mathfrak{D}}_{n \times p}$, is obtained by multiplying $\mathfrak{D}_{n \times p}$ by the householder matrix H .

$$\hat{\mathfrak{D}} = \mathfrak{D}H$$

The mechanism of the householder reflection is that it makes a vector d^1 parallel to e_1 by a reflection through the plane perpendicular to vector $e_1 - d^1$. Where $d^i, i = 1, \dots, p$ is the i^{th} component of d . The resulting householder matrix is given in Figure 2:

$$H = \begin{bmatrix} d_1^1 & d_2^1 & d_3^1 & \dots & d_p^1 \\ d_2^1 & 1 - \frac{(d_2^1)^2}{1-d_1^1} & \frac{-d_2^1 d_3^1}{1-d_1^1} & \dots & \frac{-d_2^1 d_p^1}{1-d_1^1} \\ d_3^1 & \frac{-d_3^1 d_2^1}{1-d_1^1} & 1 - \frac{(d_3^1)^2}{1-d_1^1} & \dots & \frac{-d_3^1 d_p^1}{1-d_1^1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ d_p^1 & \frac{-d_p^1 d_2^1}{1-d_1^1} & \frac{-d_p^1 d_3^1}{1-d_1^1} & \dots & 1 - \frac{(d_p^1)^2}{1-d_1^1} \end{bmatrix}$$

Figure 2: The householder matrix.

Each column of H represents the direction of a coordinate axis in the reflected space. Axis-parallel splits are searched along these axes, the best split found will be oblique in the original feature space. HHCART, by using all possible eigenvectors for reflections creates an enriched axis-parallel search space, which in turn may offer up the opportunity to find better splits.

For a p -dimensional classification problem with C classes, there will be Cp eigenvectors to be considered for the householder reflection. This leads to an increase in the time complexity for tree induction but gives an opportunity to produce more accurate and compact trees.

In some instances, the orientation of an eigenvector may be parallel to a feature axis in the original feature space. When this happens, the householder transformation is not required, and hence, the best separating hyperplane is found by performing axis-parallel splits in the original space.

Householder reflection for multiclass problems

In the case of a multiclass classification problem, the HHCART algorithm works much in the same way as it does for a two-class problem. At a node, eigenvectors are computed for each class. For each eigenvector, a householder matrix is computed, and the reflection performed. Axis-parallel splits are then carried out in each of the reflected spaces, thereby making it possible to find the best split in a reflected space created from a non-dominant eigenvector.

HHCART(G)

Contrasting with other HHCART tree algorithms, the HHCART(G) algorithm considers only one reflected feature space when splitting nodes. Searching a single feature space ensures HHCART(G) is much less computationally expensive compared to other HHCART tree-based algorithms. For example, to split each node using the HHCART(D) algorithm, C generalised eigenvalue problems need to be solved, as the training examples are reflected C times and Cp splitting dimensions need be searched to find the best split (?). Contrast this with HHCART(G) where only one generalized eigenvalue problem need be solved, the training examples are reflected just the once, thus allowing the best split to be found after searching only p splitting dimensions.

? introduced the modified angle bisector which is used by HHCART(G) to define its reflected feature space. Returning to our two-class classification problem, with training examples, \mathfrak{D} :

$$\mathfrak{D} = \{(x_i, y_i) : x_i \in \mathbb{R}^p, y_i \in \{-1, 1\} \text{ and } i = 1, 2, \dots, n\}$$

At node t , let $A \in \mathbb{R}^{n_A \times p}$ be a matrix containing training examples where $y_i = 1$, and let $B \in \mathbb{R}^{n_B \times p}$ be the matrix containing training examples where $y_i = -1$. Using a clustering hyperplane, one for each class, the angle bisectors can be defined using the form:

$$w^T x + b = 0 \quad (3)$$

Equation (3) can be written as $\tilde{w}^T x + \bar{b}$, where $\tilde{w} = (w^T, b)^T$ and $\bar{b} = (x^T, 1)^T$. These clustering hyperplanes are chosen such that each hyperplane is closest to all points of one class and is farthest from all points of the other class (?). Specifically, they are solutions to the following optimization problems:

$$\begin{aligned} \tilde{w}_1 &= \underset{\tilde{w} \neq 0}{\operatorname{argmax}} \frac{\tilde{w}^T M \tilde{w}}{\tilde{w}^T G \tilde{w}} \\ \text{and } \tilde{w}_2 &= \underset{\tilde{w} \neq 0}{\operatorname{argmin}} \frac{\tilde{w}^T M \tilde{w}}{\tilde{w}^T G \tilde{w}} \quad \#eq : eqn4 \end{aligned} \quad (4)$$

where $M = \frac{1}{n_B} (B, \mathcal{K})^T (B, \mathcal{K})$
and $G = \frac{1}{n_A} (A, \mathcal{K})^T (A, \mathcal{K})$
and $\mathcal{K} =$ a column vector of ones

If matrix G has full column rank, the solution to these optimization problems can be calculated using an LU-decomposition method (?). The solutions are the eigenvectors corresponding to the maximum and minimum eigenvalues from the following generalised eigenvalue problem (?).

$$M \tilde{w} = \lambda G \tilde{w} \quad (5)$$

? have shown for any \tilde{w} that is a local solution of the optimization problem given by equations (5) and (6) will satisfy equation (5), and the eigenvalue λ gives the value of the corresponding objective function. Finding the clustering hyperplane parameters, (w_1, b_1) and (w_2, b_2) becomes a problem of finding eigenvectors corresponding to the maximum and minimum eigenvalues of the generalised eigenvalue problem of Eqn (5). If \tilde{w}_1 is a solution to Eqn (5), $k\tilde{w}_1$ will be a solution for any k . Choosing $k = \frac{1}{\|\tilde{w}_1\|}$ gives clustering hyperplanes $\tilde{w}_1 = (w_1^T, b_1)^T$ and $\tilde{w}_2 = (w_2^T, b_2)^T$ and are unit scaled such that $\|w_1\| = \|w_2\| = 1$. Upon finding the clustering hyperplanes, the hyperplane associated with the current node will be an angle bisector of this pair of hyperplanes. If $\tilde{w}_3^T x + b_3 = 0$ and $\tilde{w}_4^T x + b_4 = 0$ are the angle bisectors of $\tilde{w}_1^T x + b = 0$ and $\tilde{w}_2^T x + b = 0$, it can be shown when $w_1 \neq w_2$ the clustering hyperplanes are given by:

$$\tilde{w}_3 = \tilde{w}_1 + \tilde{w}_2 \quad (6)$$

$$\tilde{w}_4 = \tilde{w}_1 - \tilde{w}_2 \quad (7)$$

In the case when $w_1 = w_2$ the clustering hyperplanes are parallel, the chosen angle bisector will be midway between the two parallel hyperplanes:

$$\tilde{w}_3 = (\tilde{w}_1^T, (b_1 + b_2)/2)^T \quad (8)$$

The hyperplane Gini index is used to evaluate both hyperplanes at node t ; it's defined as:

$$\operatorname{Gini}(\tilde{w}) = 2L(1 - L_A)L_A + 2(1 - L)(1 - R_A)R_A \quad (9)$$

Where L is the fraction of points at the left descendent node when using split \tilde{w} and $L_A(R_A)$ is the fraction of samples at the left(right) node from Matrix A. The angle bisector with the least impurity as measured by the Gini index is chosen to split the node (?). The best angle bisector to split node t is then used to define

a reflected feature space; specifically, the training examples at node t are reflected using the householder matrix so that the normal vector of the angle bisector w is parallel to the first coordinate axis e .

Using the reflected training examples \hat{D}^t , we find the axis-parallel split s^t that minimises the split Gini index:

$$Gini(\tilde{s}) = L \sum_{k=1}^C P_{lk}(1 - P_{lk}) + (1 - L) \sum_{k=1}^C P_{rk}(1 - P_{rk}) \quad (10)$$

Where L is the fraction of points that will go to the left descendent node after the split and $P_{lk}(P_{rk})$ is the fraction of points at the left(right) node from class k (?).

An exhaustive search over all p dimensions is performed to find the best split. This can be done quickly and efficiently as each dimension is independent and as such, can be treated separately, thereby making the search embarrassingly parallel (?). The best axis-parallel split in the reflected feature space can be denoted $z_j \leq s^t$ (the j^{th} coordinate direction in the reflected feature space). In the original feature space, this split is oblique and given by $h_j^T x \leq s^t$, where h_j is the j^{th} column of the householder matrix H .

This method is run recursively on all descendent nodes until no further splitting is possible or specific stopping conditions are satisfied. These stopping conditions relate to two user-specified parameters used by the HHCART(G) algorithm. Specifically, these are, **n_min** the minimum number of training examples, implemented in the **hcartr** package as the **n_min** parameter and ϵ , the minimum node impurity, the equivalent parameter in the **hcartr** package is **min_node_impurity**. The default values as implemented in **hcartr** are **n_min = 2** and **min_node_impurity = 0.2**. In our experiments, we use ten-fold cross-validation to choose an appropriate value for ϵ as suggested by ?.

During tree induction, if there are fewer than n_{min} training examples at a node or if the node's Gini index is less than ϵ , the node becomes a terminal or leaf node, the majority class becomes the class label for the node. In the event of a tie, the majority class is chosen at random from among the tied classes.

Small sample sizes

During tree induction, the number of training examples at each node reduces with each split, which can result in several issues as this number becomes small. The most apparent issue is one of computational cost; would an axis-parallel split be more efficient when the number of training examples is small? rather than creating and searching an oblique split? This cost is a common problem for any oblique decision tree, for instance, to address this issue in the OC1 algorithm, the authors ? only use oblique splits when the number of training examples at a node exceeds the number of feature variables by a factor of two.

Another issue that is less obvious and is wholly dependent upon the nature of the data in the training examples. A small number of training examples at a node can result in matrix G becoming rank deficient and as such LU-decomposition will not work. In this case, ? address this issue by defining the angle bisector as the eigenvector corresponding to the largest eigenvalue of:

$$\tilde{M} = QQ^T M QQ^T$$

Where Q is a matrix whose columns are an orthogonal basis for the $null(G)$. However, if $null(G) \in null(M)$, \tilde{M} will be the zero matrix, and the method fails (?). ? provide a solution for this special case, they define $\tilde{w} = u + v$, where $u \in range(G)$ and $v \in null(G)$, then the left optimization in equation (??) \@ref{eq:eqn4} \@ref{eq:eqn4} (?) becomes:

$$\begin{aligned} \tilde{w}_1 &= \underset{\tilde{w} \neq 0}{argmax} \frac{(u+v)^T M (u+v)}{(u+v)^T G (u+v)} \\ &= \underset{u \neq 0}{argmax} \frac{u^T M u}{u^T G u} \end{aligned}$$

because $v \in \text{null}(G)$ with $\text{null}(G) \subseteq \text{null}(M)$ and values of u giving $u^T G u = 0$ are avoided. The right optimization in equation (??) is solved in similar fashion. Define $\tilde{w} = u + v$, where $u \in \text{range}(M)$ and $v \in \text{null}(M)$. Then, the right optimization in equation (??) (?) becomes:

$$\begin{aligned}\tilde{w}_2 &= \underset{\tilde{w} \neq 0}{\operatorname{argmax}} \frac{(u+v)^T G (u+v)}{(u+v)^T M (u+v)} \\ &= \underset{u \neq 0}{\operatorname{argmax}} \frac{u^T G u}{u^T M u}\end{aligned}$$

because $v \in \text{null}(M)$ with $\text{null}(M) \supseteq \text{null}(G)$ and values of u giving $u^T M u = 0$ are avoided. The angle bisectors \tilde{w}_1 and \tilde{w}_2 are calculated using (??), and the hyperplane with the lower hyperplane Gini index is chosen to split the node. If \tilde{w}_1 and \tilde{w}_2 are parallel, the angle bisector becomes (??).

Usage of `hhcartr`

We now demonstrate the use of `hhcartr` by illustrating the use of the `HHDecisionTreeClassifier()` function. We will use the supplied dataset `hhcart_segment` from `hhcartr` to ensure the examples are readily repeatable. See the help page of `hhcartr` in **R** for more details on the dataset. We begin by loading the required packages into **R**, assuming they have been installed (using `install.packages()`).

```
library("hhcartr")
library("ggplot2")
```

The first step is to load the dataset, then the feature variables (`X`) and target variable (`y`) can be initialised. In the case of the `hhcart_segment` dataset, the `X` variable is created from columns 2-20 from the original UCI dataset and `y` is created from column 1 of the same. The `hhcart_segment` dataset also provides it's own test dataset.

```
data("hhcartr_cancer", package = "hhcartr")
X      <- hhcartr_segment$X
y      <- hhcartr_segment$y
test_data <- hhcartr_segment$test_data
```

Once you have established the dataset, it is advisable to summarize the basic dataset information.

```
dim(X)
```

```
## [1] 210 19
```

```
names(X)
```

```
## [1] "REGION.CENTROID.ROW" "REGION.PIXEL.COUNT" "SHORT.LINE.DENSITY.5"
## [4] "SHORT.LINE.DENSITY.2" "VEDGE.MEAN" "VEDGE.SD"
## [7] "HEDGE.MEAN" "HEDGE.SD" "INTENSITY.MEAN"
## [10] "RAWRED.MEAN" "RAWBLUE.MEAN" "RAWGREEN.MEAN"
## [13] "EXRED.MEAN" "EXBLUE.MEAN" "EXGREEN.MEAN"
## [16] "VALUE.MEAN" "SATURATION.MEAN" "HUE.MEAN"
## [19] "X"
```

It is also useful to review the output of the `summary()`, `head()`, `tail()` and `str()` functions on the dataset but we will skip this here for the sake of brevity. Before building the model we need to prepare the training dataset, specifically identifying the variables to ignore in the modelling. Identifiers and any output variables should not be used (as independent variables) for modelling. In the case of datasets supplied with `hhcartr` all appropriate variables are assigned to the `X` and `y` objects of the dataset. Next we deal with missing values.

Currently this implementation does not support data with missing values. In the supplied datasets with **hncartr** there are no missing values. It is useful to review the distribution of the target variable.

```
table(y)
```

```
## y
## BRICKFACE    CEMENT    FOLIAGE    GRASS    PATH    SKY    WINDOW
##           30         30         30         30         30         30         30
```

Demonstrate Model **HHDecisionTreeClassifier**

We now instantiate the chosen model. The aim of the model is to predict the target based on all other variables.

```
model <- HHDecisionTreeClassifier(n_folds=10, min_node_impurity = 0.22)
```

We use function **setDataDescription()** to provide the data source in all subsequent display command outputs. The **dataDescription** parameter of the model **HHDecisionTreeClassifier()** can also be used to provide a descriptive name for the data source.

```
setDataDescription("Segmentation")
```

We now set a seed so that the results can be replicated exactly.

```
set.seed(2020)
```

We are now ready to train the model, initiate training by passing the feature variables (X object) and the target variable (y object) to the model **fit()** function.

```
model.output <- model$fit(X, y)
```

```
## [1] "HHDecisionTreeClassifier() tree inference starts. Fold size=[21]."  
## [1] "***** Starting Fold 1 of 10. Tree 1 of 1."  
  
## Warning in if (new_threshold <= xnode$node_threshold) {: the condition has  
## length > 1 and only the first element will be used  
  
## [1] "***** Starting Fold 2 of 10. Tree 1 of 1."  
## [1] "***** Starting Fold 3 of 10. Tree 1 of 1."  
  
## Warning in if (new_threshold <= xnode$node_threshold) {: the condition has  
## length > 1 and only the first element will be used  
  
## Warning in if (new_threshold <= xnode$node_threshold) {: the condition has  
## length > 1 and only the first element will be used  
  
## [1] "***** Starting Fold 4 of 10. Tree 1 of 1."  
## [1] "***** Starting Fold 5 of 10. Tree 1 of 1."  
  
## Warning in if (new_threshold <= xnode$node_threshold) {: the condition has  
## length > 1 and only the first element will be used  
  
## [1] "***** Starting Fold 6 of 10. Tree 1 of 1."
```

```

## Warning in if (new_threshold <= xnode$node_threshold) {: the condition has
## length > 1 and only the first element will be used

## Warning in if (new_threshold <= xnode$node_threshold) {: the condition has
## length > 1 and only the first element will be used

## [1] "***** Starting Fold 7 of 10. Tree 1 of 1."

## Warning in if (new_threshold <= xnode$node_threshold) {: the condition has
## length > 1 and only the first element will be used

## Warning in if (new_threshold <= xnode$node_threshold) {: the condition has
## length > 1 and only the first element will be used

## [1] "***** Starting Fold 8 of 10. Tree 1 of 1."

## Warning in if (new_threshold <= xnode$node_threshold) {: the condition has
## length > 1 and only the first element will be used

## [1] "***** Starting Fold 9 of 10. Tree 1 of 1."

## Warning in if (new_threshold <= xnode$node_threshold) {: the condition has
## length > 1 and only the first element will be used

## Warning in if (new_threshold <= xnode$node_threshold) {: the condition has
## length > 1 and only the first element will be used

## Warning in if (new_threshold <= xnode$node_threshold) {: the condition has
## length > 1 and only the first element will be used

## Warning in if (new_threshold <= xnode$node_threshold) {: the condition has
## length > 1 and only the first element will be used

## [1] "***** Starting Fold 10 of 10. Tree 1 of 1."
## [1] "HHDDecisionTreeClassifier() tree inference complete."

```

After the tree building process has completed, results detailing accuracy, the number of nodes and the number of leaves per tree can be displayed for each fold by using the **results()** function. The output of this function can also be saved in a variable for future processing such as plotting.

```

res <- results(model.output)

## [1] "HHDDecisionTreeClassifier() : Run Statistics for Dataset: Segmentation"
## [1] "using parameters:"
## [1] "n_folds -[10] n_trees -[1] n_min -[2] min_node_impurity -[0.22]"
## [1] "sampleWithReplacement -[FALSE] useIdentity -[FALSE] testSize -[0.2]"
## [1] "Mean Accuracy -[81.4285714285714] Mean Nodes -[39.2] Mean Leaves -[20.1]"

```

The **accuracy()** method of the results object can be used to display the accuracy, the number of nodes and the number of leaves per tree for each fold. These are saved in a list and are accessible for future plotting. The accuracy, the mean over all **n_folds**, reported here is the accuracy achieved on each folds test dataset, **res\$accuracy()**

```
## [[1]]
## Accuracy Number_of_Nodes Number_of_Leaves
## 1 76.19048 29 15
## 2 90.47619 41 21
## 3 90.47619 39 20
## 4 90.47619 33 17
## 5 71.42857 39 20
## 6 71.42857 39 20
## 7 85.71429 41 21
## 8 66.66667 53 27
## 9 90.47619 37 19
## 10 80.95238 41 21
```

The **margin()** method of the results object can be used to display the margin for each tree on each fold. `?margin` defines the margin for a random forest as:

$$mg(\underline{X}, Y) = \frac{\sum_{k=1}^K I(h_K(\underline{X}) = Y)}{K} - \max_{j \neq Y} \left[\frac{\sum_{k=1}^K I(h_K(\underline{X}) = j)}{K} \right]$$

Where $I(\cdot)$ is the indicator function, $h_1, \dots, h_K(\underline{x})$ is a collection of classifiers, with Y, \underline{X} a random vector sampled from the training data. The margin of a single tree is defined as the proportion of correctly classified samples minus the proportion of the most misclassified class. Thus for a single tree, the margin can be written as:

$$mg(\underline{X}, Y) = \frac{I(h_K(\underline{X}) = Y)}{K} - \max_{j \neq Y} \left[\frac{I(h_K(\underline{X}) = j)}{K} \right]$$

```
res$margin()
```

```
## [1, ] 0.6190476
## [2, ] 0.8095238
## [3, ] 0.8571429
## [4, ] 0.8095238
## [5, ] 0.6190476
## [6, ] 0.5714286
## [7, ] 0.8095238
## [8, ] 0.5238095
## [9, ] 0.8571429
## [10, ] 0.7142857
```

The **print()** function is used to visualise training accuracy results on a histogram plot, the plot is rendered using **ggplot2** (?). A 95% confidence interval about the mean is also displayed.

```
print(model.output)
```

The standard **predict()** function is available to apply new data to the model. The Segmentation dataset provides a 2,100 sample test dataset. Here we apply this test dataset loaded previously (see above) to obtain the mean prediction accuracy for each fold. The object returned by the **predict()** function supports both an **accuracy** and **predictions** method.

```
preds <- predict(model.output, test_data = test_data)
```

```
## [1] "Predicting on Test Dataset..."
```

HHCART(G) Classification Accuracy Distribution for dataset: Segmentation.

95.0 confidence interval 67.7% and 90.5%; Mean 81.4%

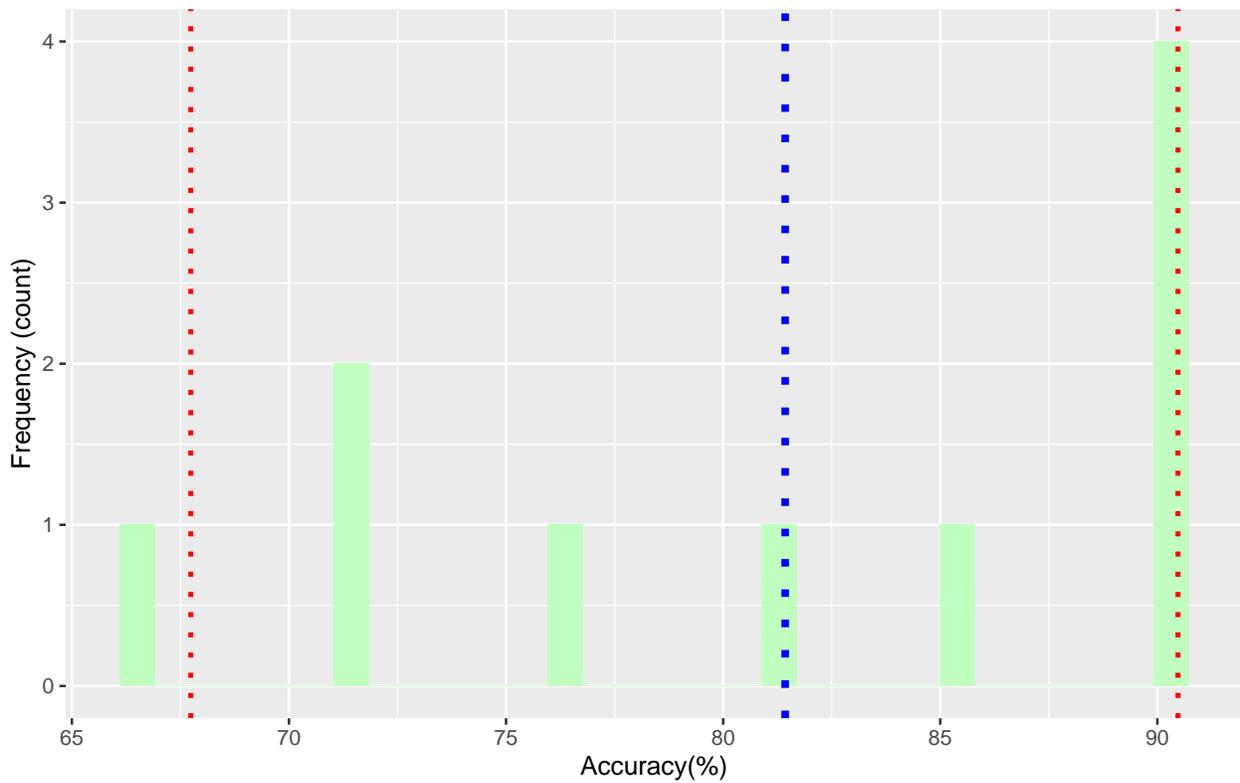


Figure 1: Classification accuracy for 10-folds after training on the Segmentation dataset.


```
## [1] "Test Data Accuracy: Mean accuracy —[83.0571428571429]"
```

The **accuracy** method returns a data frame containing the accuracy values for each fold, this object can be saved as a variable. The prediction accuracy returned is displayed as a percentage.

```
preds$accuracy()
```

```
##      Fold Accuracy
## 1         1 87.42857
## 2         2 83.09524
## 3         3 83.33333
## 4         4 87.90476
## 5         5 81.33333
## 6         6 82.28571
## 7         7 81.76190
## 8         8 79.80952
## 9         9 80.76190
## 10        10 82.85714
```

The **predictions** method returns a data frame containing the predictions for each row of the **test_data** object, each column provides the prediction for each fold for each respective row, this object can be saved as a variable. The **options(max.print=50)** limits the output display of predictions to the first few rows.

```
options(max.print=50)
preds$predictions()
```

```
##           [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
## [1,]         4    4    4    4    4    4    4    4    4    4
## [2,]         4    4    4    4    4    4    4    4    4    4
## [3,]         4    4    4    4    4    4    4    4    4    4
## [4,]         4    4    4    4    4    4    4    4    4    4
## [5,]         4    4    4    4    4    4    4    4    4    4
## [ reached getOption("max.print") — omitted 2095 rows ]
```

A single tree, or indeed all trees induced by the model can be visualized using the **displayTree()** function, it takes a single parameter **n**, the **n th** tree to display. The **hhcartr** package generates **DOT** (?) language statements to describe the structure of the selected tree, the **DOT** language statements are written to a temporary dataset and displayed using the **grViz()** function from the **DiagrammeR** package (?).

```
library(DiagrammeRsvg)
library(rsvg)
outp <- displayTree(1)
```

```
## Loading required package: DiagrammeR
```

We have a model that is close to **84.0%** accurate on the unseen test dataset. In the invocations we have demonstrated above, all parameters except **min_node_impurity** use their default values.

Models

The R package **hhcartr** supports the following models.

HHDecisionTreeClassifier

The **hhcartr** model **HHDecisionTreeClassifier** is used for classification problems. is used to induce regression trees.

When a **HHDecisionTreeClassifier** model is first instantiated the package **checkmate** (?) is used to parse all provided parameter values; the function will terminate with an appropriate error message on the first parameter to fail validation. Input is passed to the **HHDecisionTreeClassifier** model by way of feature variables (X) and target variable (y) as separate objects to the **fit** method. The feature variables (X) can be either a dataframe or a matrix of numeric data, categorical feature variables are not yet supported, whereas the target variable (y) must be of type factor. The model **fit** method will terminate with an appropriate error message if input type validation fails. The **hhcartr** package does not yet support NA's in the input data.

HHDecisionTreeRegressor

The **hhcartr** model **HHDecisionTreeRegressor** is used to induce regression trees. These predict a real number rather than a class. Splits are looked for that minimise the prediction squared error, the prediction at each leaf node being the mean for that node. Internally the model uses parameter **useIdentity=TRUE** which forces all splits to be axis-parallel.

When a **HHDecisionTreeRegressor** model is first instantiated the package **checkmate** (?) is used to parse all provided parameter values; the function will terminate with an appropriate error message on the first parameter to fail validation. Input is passed to the **HHDecisionTreeRegressor** model by way of feature variables (X) and target variable (y) as separate objects to the **fit** method. The feature variables (X) can be either a dataframe or a matrix of numeric data, categorical feature variables are not yet supported, whereas the target variable (y) must be of type numeric. The model **fit** method will terminate with an appropriate error message if input type validation fails. The **hhcartr** package does not yet support NA's in the input data.

Model Parameters

The following section describes the common model parameters i.e. parameters that can be specified for any of the *hhcartr* supported models, namely:

- **HHDecisionTreeClassifier**
- **HHDecisionTreeRegressor**

Common Model Parameters

The following parameters may be specified on any *hhcartr* model.

1. **n_min** This parameter is used to stop splitting a node when a minimum number of samples at that node has been reached. The default value is 2.
2. **n_trees** The number of trees to induce(grow) per fold or trial. The default value is 1.
3. **n_folds** The **n_folds** parameter is used to specify the number of folds to use i.e. split the input data into **n_folds** equal amounts. For **n_folds** times, use one portion of the input data as a test dataset, and the remaining **n_folds-1** portion as the training dataset. The model is then trained using these training and test datasets, once training is complete the next fold or portion of the input dataset is treated as the test dataset and the remainder the training dataset, the model is then trained again. This process is repeated until all portions or folds of the input dataset have been used as a test dataset. When **n_folds=1** the **testSize** parameter determines the size of the test dataset. The default value is 5.
4. **testSize** The **testSize** parameter determines how much of the input dataset is to reserved for use as an internal test dataset. The remainder is used as the training dataset. This parameter is only used when the parameter **n_folds=1**. For values of **n_folds** greater than one, the computed fold size will

govern the test dataset size used (see the `n_folds` parameter for more details). The value can have a range of 0.0 to 1.0, the default value is 0.2.

5. **dataDescription** The **dataDescription** parameter is a short description used to describe the dataset being modelled. It is used in output displays and plots as documentation. The default value is **Unknown**.

HHDecisionTreeClassifier Model Parameters

The model HHDecisionTreeClassifier model supports the common model parameters as described above, as well as the following model specific parameters (with default values).

1. **min_node_impurity** Node splitting stops if the node impurity falls below this threshold. The node impurity is calculated using the hyperplane Gini index. The default value is 0.2.
2. **useIdentity** The **useIdentity** parameter when set **TRUE** will direct *hhcartr* to transform the training data using the identity matrix (no actual transform takes place) prior to finding the optimal splits rather than using the householder transform to reflect the data. The default value is **FALSE**.

Table 1 provides a summary of the default parameters for model HHDecisionTreeClassifier.

Table 1: Table 1: Default parameters for model HHDecisionTreeClassifier.

Parameter	Default value
n_min	2
min_node_impurity	0.2
n_trees	1
n_folds	5
testSize	0.2
useIdentity	FALSE
dataDescription	Unknown

HHDecisionTreeRegressor Model Parameters

The model HHDecisionTreeRegressor model supports just the common model parameters as described above.

Table 4 provides a summary of the default parameters for model HHDecisionTreeClassifier.

Table 2: Table 4: Default parameters for model HHDecisionTreeRegressor.

Parameter	Default value
n_min	2
n_trees	1
n_folds	5
testSize	0.2
dataDescription	Unknown

General Commands

The following functions, demonstrated above, can be applied to all **hhcartr** models, **fit()**, **results()**, **print()**, and **predict()**. The **results()** function provides the **accuracy()** and **margin()** methods, the function **predict()** supports the **accuracy()** and **predictions()** methods. General functions **setDataDescription()** and **displayTree()** are also part of the **hhcartr** package. The **HHDecisionTreeRegressor** model does

not yet support the `margin()` method.

HHDecisionTreeRegressor()

As demonstrated above the `print()` function on the results from the `HHDecisionTreeClassifier` model returns the model accuracy, whereas for the `HHDecisionTreeRegressor` model the r^2 and root mean square error (RMSE) are returned along with the number of nodes and the number of leaves for each tree.

The r^2 or coefficient of determination value is the ratio of explained variation over the total variation, it will have a value between 0.0 and 1.0. The r^2 value is computed as follows:

$$R^2 \equiv 1 - \frac{SS_{res}}{SS_{tot}}$$

where:

$$SS_{tot} = \sum_i (y_i - \bar{y})^2$$

$$SS_{res} = \sum_i (y_i - f_i)^2$$

with: \bar{y} = mean of observed data, and f_i = predicted values.

Visualizing the Tree Structure

As the `displayTree()` function can be used to display the structure of an induced tree, `R`'s `View()` function can be used to display the internal structure of all tree's returned by the `fit()` function. This internal tree structure is comprised of a binary tree of S3 objects, this structure can be accessed programmatically by the user.

hhcartr Supplied Datasets

The `hhcartr` package provides a number of datasets, the same as those used by the authors of the `HHCART(G)` package (?) and includes several bench-mark datasets downloaded from the UCI Machine Learning Repository (?) as used by ?. See the help page of `hhcartr` in `R` for a description of each dataset using the appropriate `? hhcartr dataset name` command. Table 2 and Table 3 provide an overview of the datasets, with the size of the corresponding CSV (comma separated values) file of each. They also provide the name of the dataset as supplied in the `hhcartr` package.

Table 3: Table 2: Overview of the datasets used by ? used in the experiments detailed below. They have all been downloaded from the UCI repository (?).

Dataset	Features	Classes	Examples	hhcartr dataset name	
Balance scale (BS)		4	3	625	hhcartr_balance
Boston housing (BH)		13	2	506	hhcartr_boston
Breast cancer (BC)		9	2	638	hhcartr_cancer
BUPA		6	2	345	hhcartr_bupa
Glass (GLS)		9	7	214	hhcartr_glass
Pima Indians (PIMA)		8	2	768	hhcartr_pima
Wine (WINE)		13	3	178	hhcartr_wine
Survival (SUR)		3	2	306	hhcartr_survival
Heart (HRT)		13	2	270	hhcartr_heart

Dataset	Features	Classes	Examples	hhcartr dataset name	
Letter (LET)		16	26	20,000	hhcartr_letters

Table 4: Table 3: Overview of the datasets from ? used in the experiments detailed below. They have all been downloaded from the UCI repository (?).

Dataset	Features	Classes	Examples	Test Examples	hhcartr dataset name	
Landsat (LS)		36	7	4435	2000	hhcartr_landsat
Pendigits (PEN)		16	10	7494	3498	hhcartr_pendigits
Segmentation (SEG)		19	7	210	2100	hhcartr_segment
Shuttle (SHUT)		9	7	43,500	14,500	hhcartr_shuttle
Vehicle (VEH)		18	2	846	-	hhcartr_vehicle

The data is not preprocessed, filtered or normalised, that is, all datasets are provided as-is after downloading from the UCI machine learning repository, with the following exceptions. In the glass identification (GLS) dataset (?) the Id number column is not provided, as is the case with the original Wisconsin breast cancer (BC) dataset (?). Traditionally, the Boston housing (BH) dataset (?) is used with regression algorithms, to turn it into a binary classification problem, we created a new target variable using the condition $MEDV > 20$. As supplied, the landsat (LS) dataset (?) does not have any examples of class six. Currently this implementation does not support data with missing values. In the supplied datasets with **hhcartr** there are no missing values. Test data sets are provided for the following datasets: **segment**, **pendigits**, **landsat** and **shuttle**.

A Tree Node

In **hhcartr**, each tree node is represented by a hash object, from the **hash** package (?). Thus each generated tree is represented as a tree structure of hash objects. Each hash object or tree node contains the following information, see Table 5 below:

Table 5: Table 5: hhcartr: Tree Node variables and descriptions.

Variable	Description
node_gini	The gini-index for this node.
node_tot_samples	The total number of samples assigned to this node.
node_num_samples_per_class	The number of samples for each class at this node. This will sum to node_tot_samples .
node_predicted_class	The predicted class for this node.
node_feature_index	The column number in either the householder matrix or original dataset used to determine the split for this node.
node_threshold	The value of the feature used to determine the split at the current node. The feature column used to make the split can be found in node variable node_feature_index .
node_householder_matrix	The householder matrix associated with this node. This need only be the node_feature_index column of the householder matrix. This will be updated at a later date.
node_children_left	The node containing the samples when the parent node is split i.e. those samples that have values in the node_feature_index column that are less than the node_threshold_value .

Variable	Description
node_children_right	The node containing the samples when the parent node is split i.e. those samples that have values in the node_feature_index column that are equal or greater than the node_threshold value .
node_parent	The parent node
node_parentid	The objectid of the parent node.
node_objectid	The nodes objectid. Nodes are number sequentially as they are created, with the root node having an objectid of zero.
node_type	The type of node, can take values of “DECISION” or “TERMINAL”. A “DECISION” node has a splitting rule and a “TERMINAL” or leaf node is a node that will not be split any further as it is either pure (all the same class) or the splitting stop criteria has been meet i.e. the n_min value and min_node_impurity values are less than the specified parameter values. See below for further discussion on terminal node creation.
node_using_householder	A flag indicating that this node was split using the householder matrix. This field can have a value of either TRUE or FALSE.
node_children_left_NA	A flag indicating whether or not the node_children_left field contains “NA”. This field can have a value of either TRUE or FALSE. A value of TRUE indicates this node has no children. If node_children_left_NA and node_children_right_NA are both TRUE this implies that this node is a TERMINAL node.
node_children_right_NA	A flag indicating whether or not the node_children_right field contains “NA”. This field can have a value of either TRUE or FALSE. A value of TRUE indicates this node has no children. If node_children_left_NA and node_children_right_NA are both TRUE this implies that this node is a TERMINAL node.
node_oob_training_indices	List of indices of the out-of-bag samples in the training set used to create this tree. These samples have not been used to grow the tree, but if parameter OOBEE (available in the HHRandomForestClassifier model, which will be available in the next release of hhcart) is TRUE they will be used to calculate the out-of-bag error estimate. Only the root node of a tree will contain valid values; for all other nodes in the tree this field will contain a value of “NA”. This field is only used by the HHRandomForestClassifier model.
node_depth	The depth in the tree at which this node occurs. At the time of writing this field is not referenced by any hhcart functionality.

Terminal Node Creation

In **hhcart**, terminal nodes are created during the tree induction process in either of two ways, firstly through user specified parameters or by forced terminal node creation.

User parameters

The **hhcart** parameters that control the creation of a tree terminal node are **n_min** and **min_node_impurity**. These parameters have been previously defined above.

Forced node creation

During the node splitting process, if either of the following conditions arise the splitting process on the current node terminates and a terminal node is created, irrespective of the values of **n_min** and **min_node_impurity**. These conditions are as follows:

1. **Feature values the same** In the situation where all values in a feature column are the same and this pattern is observed for each of the other feature columns, thus making it impossible to find a valid split on any feature column.
2. **Single sample** In the situation where there are two classes in the node under consideration with one of the classes having only one sample. This situation generally occurs when there are a small number of samples at the node, but **min_node_impurity** has not quite been reached.