C50

データサイエンス関連基礎調査WG

2025-01-16

Table of Contents

## パッケージの概要

C50パッケージではC5.0アルゴリズムを用いた決定木モデルを使用できます。

## 参考URL

公式ドキュメント <https://cran.r-project.org/web/packages/C50/C50.pdf>

## 使用例：irisデータの分類

irisデータを用いて、がく弁・花弁の長さ・幅の情報からアヤメを分類するモデルをC50パッケージを用いて構築します。

### irisデータセットを読み込む

irisデータを読み込み、データの先頭を表示します。

* Sepal.Length：がく弁の長さ
* Sepal.Width：がく弁の幅
* Petal.Length：花弁の長さ
* Petal.Width：花弁の幅

アヤメの種類はsetosa(1)、versicolor(2)、virginica(3)の3種類です。

data(iris)  
head(iris)

## Sepal.Length Sepal.Width Petal.Length Petal.Width Species  
## 1 5.1 3.5 1.4 0.2 setosa  
## 2 4.9 3.0 1.4 0.2 setosa  
## 3 4.7 3.2 1.3 0.2 setosa  
## 4 4.6 3.1 1.5 0.2 setosa  
## 5 5.0 3.6 1.4 0.2 setosa  
## 6 5.4 3.9 1.7 0.4 setosa

### irisデータの構造

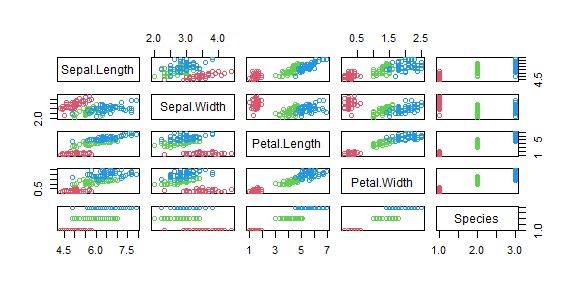
irisデータの各種構造を確認します。

str(iris)

## 'data.frame': 150 obs. of 5 variables:  
## $ Sepal.Length: num 5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...  
## $ Sepal.Width : num 3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...  
## $ Petal.Length: num 1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...  
## $ Petal.Width : num 0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...  
## $ Species : Factor w/ 3 levels "setosa","versicolor",..: 1 1 1 1 1 1 1 1 1 1 ...

また、データを散布図にプロットして確認します。

plot(iris, col=c(2, 3, 4)[iris$Species])



### モデル構築

irisデータをモデル生成のための訓練データと、モデル評価のためのテストデータに分割します。データ割合は訓練データを7割、テストデータを3割とします。確認のため、データサイズを出力します。

# 再現性のためにシードを設定  
set.seed(123)  
  
# データの分割  
sample\_indices <- sample(1:nrow(iris), 0.7 \* nrow(iris))   
train\_data <- iris[sample\_indices, ]  
test\_data <- iris[-sample\_indices, ]  
  
# データサイズの確認  
c(nrow(iris), nrow(train\_data), nrow(test\_data))

## [1] 150 105 45

モデルを生成します。

model <- C5.0(  
 x = train\_data[-5],   
 y = train\_data$Species,   
)

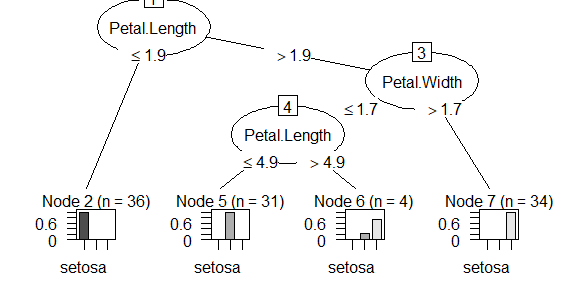
モデルのサマリを表示します。

summary(model)

##   
## Call:  
## C5.0.default(x = train\_data[-5], y = train\_data$Species)  
##   
##   
## C5.0 [Release 2.07 GPL Edition] Thu Jan 16 18:17:23 2025  
## -------------------------------  
##   
## Class specified by attribute `outcome'  
##   
## Read 105 cases (5 attributes) from undefined.data  
##   
## Decision tree:  
##   
## Petal.Length <= 1.9: setosa (36)  
## Petal.Length > 1.9:  
## :...Petal.Width > 1.7: virginica (34/1)  
## Petal.Width <= 1.7:  
## :...Petal.Length <= 4.9: versicolor (31/1)  
## Petal.Length > 4.9: virginica (4/1)  
##   
##   
## Evaluation on training data (105 cases):  
##   
## Decision Tree   
## ----------------   
## Size Errors   
##   
## 4 3( 2.9%) <<  
##   
##   
## (a) (b) (c) <-classified as  
## ---- ---- ----  
## 36 (a): class setosa  
## 30 2 (b): class versicolor  
## 1 36 (c): class virginica  
##   
##   
## Attribute usage:  
##   
## 100.00% Petal.Length  
## 65.71% Petal.Width  
##   
##   
## Time: 0.0 secs

plot関数を用いると、決定木の様子が可視化されます。

plot(model)



構築したモデルを使用して予測を行います。

res <- predict(model, test\_data[-5])

結果を確認します。予測が概ねあっていることが分かります。

(table(res, test\_data$Species))

##   
## res setosa versicolor virginica  
## setosa 14 0 0  
## versicolor 0 17 0  
## virginica 0 1 13