Indicadores de doença cardíaca

João Pedro Albino

1/9/2022

### Introdução

Este é um documento elaborado utilizando o pacote R Markdown. Markdown é uma sintaxe de formatação simples para a criação de documentos HTML, PDF e MS Word. Para obter mais detalhes sobre o uso de R Markdown, consulte <http://rmarkdown.rstudio.com>.

Quando você clica no botão \*\* Knit \*\*, um documento é gerado, incluindo tanto o conteúdo quanto a saída de qualquer fragmento de código R embutido no documento.

### Sinopse

O objetivo deste documento é apresentar uma análise simples no conjunto de dados da UCI - Heart Disease Data Set, de forma a ajudar a encontrar os *três atributos mais promissores* para prever possível estreitamento no diâmetro das veias cardíacas.

Os dados utilizados estão disponíveis no UCI Machine Learning Repository (<http://archive.ics.uci.edu/ml/datasets/Heart+Disease>) e foram doados por:

* Instituto Húngaro de Cardiologia. Budapeste: Andras Janosi, M.D.
* Hospital Universitário, Zurique, Suíça: William Steinbrunn, M.D.
* Hospital Universitário, Basel, Suíça: Matthias Pfisterer, M.D.
* V.A. Medical Center, Long Beach e Cleveland Clinic Foundation: Robert Detrano, M.D., Ph.D.

O dataset aqui utilizado foi pré-processado pela Universidade de Cleaveland

### Baixando os dados

if (!file.exists("./data/processed.cleveland.data")) {  
 download.file(url = "http://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/processed.cleveland.data",   
 destfile = "./data/processed.cleveland.data")  
}

Verificando o “checksum” do arquivo armazenado localmente

### Carregando o pacote necessário: tools  
require(tools)

## Carregando pacotes exigidos: tools

md5sum("./data/processed.cleveland.data")

## ./data/processed.cleveland.data   
## "cf81c26cdc2bac254552bf959cc3ecb2"

### Carregando dados no dataframe heart.data

heart.data <- read.csv("./data/processed.cleveland.data", header = FALSE)

A página da fonte da fonte de dados original afirma que devemos ter 303 instâncias e 75 atributos.

Porém, o arquivo de dados processados pela Cleveland deve ter 14 atributos.

Vamos verificar se temos os dados de forma adequada:

nrow(heart.data)

## [1] 303

ncol(heart.data)

## [1] 14

head(heart.data)

## V1 V2 V3 V4 V5 V6 V7 V8 V9 V10 V11 V12 V13 V14  
## 1 63 1 1 145 233 1 2 150 0 2.3 3 0.0 6.0 0  
## 2 67 1 4 160 286 0 2 108 1 1.5 2 3.0 3.0 2  
## 3 67 1 4 120 229 0 2 129 1 2.6 2 2.0 7.0 1  
## 4 37 1 3 130 250 0 0 187 0 3.5 3 0.0 3.0 0  
## 5 41 0 2 130 204 0 2 172 0 1.4 1 0.0 3.0 0  
## 6 56 1 2 120 236 0 0 178 0 0.8 1 0.0 3.0 0

Os dados parecem corretos, então podemos prosseguir com a análise.

A descrição dos atributos do conjunto de dados na página da web são os seguintes:

1. age - age in years
2. sex - sex (1 = male; 0 = female)
3. cp - chest pain type (1 = typical angina; 2 = atypical angina; 3 = non-anginal pain; 4 = asymptomatic)
4. trestbps - resting blood pressure (in mm Hg on admission to the hospital)
5. chol - serum cholestoral in mg/dl
6. fbs - fasting blood sugar > 120 mg/dl (1 = true; 0 = false)
7. restecg - resting electrocardiographic results (0 = normal; 1 = having ST-T; 2 = hypertrophy)
8. thalach - maximum heart rate achieved
9. exang - exercise induced angina (1 = yes; 0 = no)
10. oldpeak - ST depression induced by exercise relative to rest
11. slope - the slope of the peak exercise ST segment (1 = upsloping; 2 = flat; 3 = downsloping)
12. ca - number of major vessels (0-3) colored by flourosopy
13. thal - 3 = normal; 6 = fixed defect; 7 = reversable defect
14. num - the predicted attribute - diagnosis of heart disease (angiographic disease status) (Value 0 = < 50% diameter narrowing; Value 1 = > 50% diameter narrowing)

### Preparando os dados

Para pré-processar os dados, vamos ajustar os nomes das colunas de maneira adequada:

names(heart.data) <- c("age", "sex", "cp", "trestbps", "chol", "fbs", "restecg", "thalach", "exang", "oldpeak", "slope", "ca", "thal", "num")

Ascolunas *ca* e *thal* possuem observações com valores ausentes (missing values) indicados por “?”.

Para que possamos tratá-las adequadamente na nossa análise, iremos preenchê-las com NA (Not Available). NA é uma constante lógica de comprimento 1 que contém um indicador de valor ausente. NA pode ser forçado a qualquer outro tipo de vetor, exceto *raw*.

heart.data$ca[heart.data$ca == "?"] <- NA  
heart.data$thal[heart.data$thal == "?"] <- NA

Iremos também corrigir os tipos de algumas das variáveis, alterando-os para o tipo *fator*.

**Fator** em R é um tipo de variável usada para categorizar e armazenar os dados, tendo um número limitado de valores diferentes. Um fator armazena os dados como um vetor de valores inteiros. Fator em R também é conhecido como uma variável categórica que armazena valores de dados de string e inteiros como níveis.

heart.data$sex <- factor(heart.data$sex)  
levels(heart.data$sex) <- c("female", "male")  
heart.data$cp <- factor(heart.data$cp)  
levels(heart.data$cp) <- c("typical","atypical","non-anginal","asymptomatic")  
heart.data$fbs <- factor(heart.data$fbs)  
levels(heart.data$fbs) <- c("false", "true")  
heart.data$restecg <- factor(heart.data$restecg)  
levels(heart.data$restecg) <- c("normal","stt","hypertrophy")  
heart.data$exang <- factor(heart.data$exang)  
levels(heart.data$exang) <- c("no","yes")  
heart.data$slope <- factor(heart.data$slope)  
levels(heart.data$slope) <- c("upsloping","flat","downsloping")  
heart.data$ca <- factor(heart.data$ca) # não convertendo o level porque não é necessário  
heart.data$thal <- factor(heart.data$thal)  
levels(heart.data$thal) <- c("normal","fixed","reversable")  
heart.data$num <- factor(heart.data$num) # não convertendo o nível porque não é necessário

### Analisando os dados

Primeira etapa, resumindo os dados transformados:

summary(heart.data)

## age sex cp trestbps   
## Min. :29.00 female: 97 typical : 23 Min. : 94.0   
## 1st Qu.:48.00 male :206 atypical : 50 1st Qu.:120.0   
## Median :56.00 non-anginal : 86 Median :130.0   
## Mean :54.44 asymptomatic:144 Mean :131.7   
## 3rd Qu.:61.00 3rd Qu.:140.0   
## Max. :77.00 Max. :200.0   
## chol fbs restecg thalach exang   
## Min. :126.0 false:258 normal :151 Min. : 71.0 no :204   
## 1st Qu.:211.0 true : 45 stt : 4 1st Qu.:133.5 yes: 99   
## Median :241.0 hypertrophy:148 Median :153.0   
## Mean :246.7 Mean :149.6   
## 3rd Qu.:275.0 3rd Qu.:166.0   
## Max. :564.0 Max. :202.0   
## oldpeak slope ca thal num   
## Min. :0.00 upsloping :142 0.0 :176 normal :166 0:164   
## 1st Qu.:0.00 flat :140 1.0 : 65 fixed : 18 1: 55   
## Median :0.80 downsloping: 21 2.0 : 38 reversable:117 2: 36   
## Mean :1.04 3.0 : 20 NA's : 2 3: 35   
## 3rd Qu.:1.60 NA's: 4 4: 13   
## Max. :6.20

### Selecionado os dados

O conteúdo da variável *num* possui uma ampla faixa de valores (0 - 4). Entretanto, apenas as instâncias com “0” e “1” estão efetivamente relacionadas à possibilidade de estreitamento do diâmetro dos vasos. Portanto, iremos selecionar apenas as observações relacionados a esses dois resultados.

heart.data <- heart.data[heart.data$num == "0" | heart.data$num == "1", ]

Verificando novamente se temos os dados de forma adequada:

nrow(heart.data)

## [1] 219

ncol(heart.data)

## [1] 14

head(heart.data)

## age sex cp trestbps chol fbs restecg thalach exang oldpeak  
## 1 63 male typical 145 233 true hypertrophy 150 no 2.3  
## 3 67 male asymptomatic 120 229 false hypertrophy 129 yes 2.6  
## 4 37 male non-anginal 130 250 false normal 187 no 3.5  
## 5 41 female atypical 130 204 false hypertrophy 172 no 1.4  
## 6 56 male atypical 120 236 false normal 178 no 0.8  
## 8 57 female asymptomatic 120 354 false normal 163 yes 0.6  
## slope ca thal num  
## 1 downsloping 0.0 fixed 0  
## 3 flat 2.0 reversable 1  
## 4 downsloping 0.0 normal 0  
## 5 upsloping 0.0 normal 0  
## 6 upsloping 0.0 normal 0  
## 8 upsloping 0.0 normal 0

### Árvore de Decisão

**Árvores de decisão** usadas para problemas de classificação são chamadas de *Árvores de Classificação*. Nas árvores de classificação, cada nó terminal ou folha contém um rótulo que indica a classe predita para um determinado conjunto de dados. Neste tipo de árvore pode existir dois ou mais nós terminais com a mesma classe.

*Árvores de Classificação/Decisão* também podem ser definidas como uma representação de uma tabela de decisão sob a forma de árvore. Trata-se de uma forma alternativa de expressar as mesmas regras que são obtidas quando se constrói a tabela.

Para a construção da árvore, será utilizada a biblioteca *rpart*. Rpart é uma biblioteca de aprendizado de máquina em R que é usada para construir árvores de classificação e regressão. Esta biblioteca implementa particionamento recursivo.

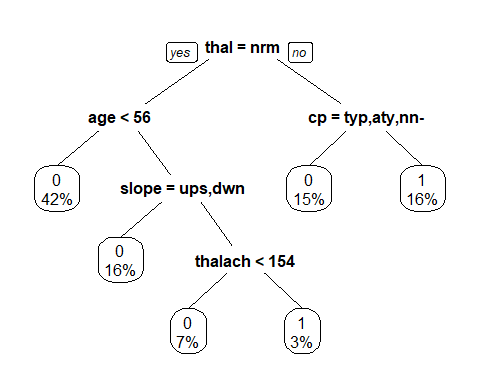
library(rpart)

Montando a árvore:

heart.tree <- rpart(num ~ age + sex + cp + trestbps + chol + fbs + restecg +   
 thalach + exang + oldpeak + slope + ca + thal,  
 method = "class",   
 data = heart.data)

Plotando a árvore:

library(rpart.plot)  
prp(heart.tree, extra = 100)



### Conslusão

Baseado no gráfico da árvore, os três melhores atributos que podem prever o possível estreitamento do diâmetro são as variáveis / colunas: tal, cp e idade.

### R e informações dos pacotes utilizados

As seguintes versões de R e pacotes foram usadas para a geração deste documento:

sessionInfo()

## R version 4.1.2 (2021-11-01)  
## Platform: x86\_64-w64-mingw32/x64 (64-bit)  
## Running under: Windows 10 x64 (build 19044)  
##   
## Matrix products: default  
##   
## locale:  
## [1] LC\_COLLATE=Portuguese\_Brazil.1252 LC\_CTYPE=Portuguese\_Brazil.1252   
## [3] LC\_MONETARY=Portuguese\_Brazil.1252 LC\_NUMERIC=C   
## [5] LC\_TIME=Portuguese\_Brazil.1252   
##   
## attached base packages:  
## [1] tools stats graphics grDevices utils datasets methods   
## [8] base   
##   
## other attached packages:  
## [1] rpart.plot\_3.1.0 rpart\_4.1-15   
##   
## loaded via a namespace (and not attached):  
## [1] compiler\_4.1.2 magrittr\_2.0.1 fastmap\_1.1.0 htmltools\_0.5.2  
## [5] yaml\_2.2.1 stringi\_1.7.6 rmarkdown\_2.11 highr\_0.9   
## [9] knitr\_1.37 stringr\_1.4.0 xfun\_0.29 digest\_0.6.29   
## [13] rlang\_0.4.12 evaluate\_0.14