Heart disease indicators

Synopsis

Simple analysis which should help to find three most promising attributes for predicting possible diameter narrowing. I will use data from UCI Machine Learning Repository donated by:

1. Hungarian Institute of Cardiology. Budapest: Andras Janosi, M.D.
2. University Hospital, Zurich, Switzerland: William Steinbrunn, M.D.
3. University Hospital, Basel, Switzerland: Matthias Pfisterer, M.D.
4. V.A. Medical Center, Long Beach and Cleveland Clinic Foundation:Robert Detrano, M.D., Ph.D.

Md5sum of file used for analysis: 2d91a8ff69cfd9616aa47b59d6f843db. If you download file with differed sum, results might be different.

Downloading data

```
if (!file.exists("processed.cleveland.data")) {
  download.file(url = "http://archive.ics.uci.edu/ml/machine-learning-
databases/heart-disease/processed.cleveland.data", destfile =
"processed.cleveland.data")
  }
require(tools)
## Loading required package: tools
md5sum("processed.cleveland.data")
##           processed.cleveland.data
## "2d91a8ff69cfd9616aa47b59d6f843db"
```

Loading data into data frame

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

Data source webpage claims that we should have 303 instances and 75 attributes. But processed data file for Cleveland should have 14 attributes. Lets check if we have proper data

```
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
```

Data looks OK, so I can go further with analysis. Decryption on attributes from data source webpage:

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)

Lets adjust names accordingly:

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

ca and *thal* have missing values indicated by "?" lets treat them properly

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

And also lets fix variable types:

```
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) # not doing level conversion
because its not necessary
heart.data$thal <- factor(heart.data$thal)
levels(heart.data$thal) <- c("normal","fixed","reversable")
heart.data$num <- factor(heart.data$num) # not doing level conversion
because its not necessary
```

Summary of prepared data:

```
summary(heart.data)
##       age          sex                    cp         trestbps
##  Min.   :29.0   female: 97   typical     : 23   Min.   : 94
##  1st Qu.:48.0   male  :206   atypical    : 50   1st Qu.:120
##  Median :56.0                non-anginal : 86   Median :130
##  Mean   :54.4                asymptomatic:144   Mean   :132
##  3rd Qu.:61.0                                   3rd Qu.:140
##  Max.   :77.0                                   Max.   :200
##       chol        fbs             restecg       thalach    exang
##  Min.   :126   false:258   normal     :151   Min.   : 71   no :204
##  1st Qu.:211   true : 45   stt        :  4   1st Qu.:134   yes: 99
##  Median :241               hypertrophy:148   Median :153
##  Mean   :247                                 Mean   :150
##  3rd Qu.:275                                 3rd Qu.:166
##  Max.   :564                                 Max.   :202
##     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
```

Selecting data

Results which “0” and “1” are related to possibility of diameter narrowing. Lets select only data ralated to those two results.

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

Classification tree

I will use *rpart* package for classification tree.

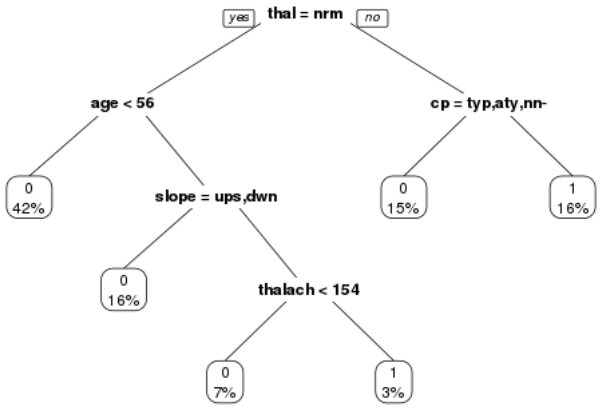
```
library(rpart)
```

Growing the tree

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

Plotting the tree

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



Results

If I would have to pick three best attributes that can predict possible diameter narrowing I would select **thal**, **cp** and **age**.

R and packages information

Following versions of R and packages were used.

```
sessionInfo()
## R version 3.1.1 (2014-07-10)
## Platform: x86_64-pc-linux-gnu (64-bit)
##
## locale:
##  [1] LC_CTYPE=pl_PL.UTF-8       LC_NUMERIC=C
##  [3] LC_TIME=pl_PL.UTF-8        LC_COLLATE=pl_PL.UTF-8
##  [5] LC_MONETARY=pl_PL.UTF-8    LC_MESSAGES=pl_PL.UTF-8
##  [7] LC_PAPER=pl_PL.UTF-8       LC_NAME=C
##  [9] LC_ADDRESS=C               LC_TELEPHONE=C
## [11] LC_MEASUREMENT=pl_PL.UTF-8 LC_IDENTIFICATION=C
##
## attached base packages:
## [1] tools     stats     graphics  grDevices utils     datasets
methods
## [8] base
##
## other attached packages:
## [1] rpart.plot_1.4-4 rpart_4.1-8      knitr_1.6
##
## loaded via a namespace (and not attached):
## [1] evaluate_0.5.5 formatR_0.10   stringr_0.6.2
```