Hands-on R - Pré-processamento dos dados para análises posteriores

Rodrigo Heldt - rodrigoheldt@gmail.com

IMED - HOPEAD - Nov 13, 2017

Pré-processamento dos dados para análises posteriores

1. Limpando os objetos do ambiente

```
rm(list=ls())
```

2. Carregando (e instalando) pacotes de funções necessários

```
# install.packages('caTools')
# Esse comando deve ser rodado apenas na primeira vez que você usar o pacote para
# instalá-lo em seu computador. Após instalado, a cada vez que for utilizá-lo basta
# carregar o pacote usando a função library

library(caTools)
```

3. Definindo o diretório de trabalho

```
# Usar a funcao setwd

setwd("C:/Users/Rodrigo/Dropbox/Hands-on R Workshop/Parte 2 - Pre-processamento dos dados")
```

4. Importando a base de dados de um arquivo csv

```
# Usar a funcao read.csv

base = read.csv(file = 'Data.csv', sep = ",")

# Visualizar as primeiras linhas da base de dados importada. Usar a funcao head

head(base)
```

```
##   SalarioPorHora LogSalarioPorHora Mulher Idade Educacao PartTime
## 1             66             4.190      0    49        1        1
## 2             34             3.526      1    42        1        1
## 3             70             4.248      1    42        1        1
## 4             47             3.850      0    38        1        0
## 5            107             4.673      1    54        1        1
## 6            188             5.236      1    54        1        0
```

5. Codificando variáveis categóricas

```
# Usar a função as.factor para que o R transforme as variáveis desejadas na classe tipo fator (categóri

base$Mulher = as.factor(base$Mulher)
base$Educacao = as.factor(base$Educacao)
base$PartTime = as.factor(base$PartTime)
```

6. Definindo variaveis numéricas

```
# Usar a função as.numeric para que o R transforme as variáveis desejadas na classe
# tipo numeric (número)

base$SalarioPorHora = as.numeric(base$SalarioPorHora)
base$LogSalarioPorHora = as.numeric(base$LogSalarioPorHora)
base$Idade = as.numeric(base$Idade)
```

7. Divindindo a base em uma parte para treinamento dos modelos e outra para teste das previsões do modelo treinado

```
set.seed(123)
split = sample.split(base$SalarioPorHora, SplitRatio = 0.8)
baseTrain = subset(base, split == TRUE)
baseTest = subset(base, split == FALSE)
```

8. Extra: Exportando objetos para o seu diretório de trabalho

```
# Caso eu queira exportar em csv as novas bases criadas, posso salvar no meu diretório
# de trabalho os objetos do meu environment

write.csv(x = base, file="Data_new.csv", row.names= F)
```