Pesquisa Reprodutível

Estruturea de uma Análise de Dados

Delermando Branquinho Filho

Etapas em uma análise de dados

- Definir a pergunta
- Definir o conjunto de dados ideal
- Determinar quais dados você pode acessar
- Obter os dados
- Limpe os dados
- Análise exploratória de dados
- Previsão / modelagem estatística
- Interpretar os resultados
- Resultados do desafio
- Sintetizar / escrever resultados
- Criar código reproduzível

Etapas em uma análise de dados

- Definir a pergunta
- Definir o conjunto de dados ideal
- Determinar quais dados você pode acessar
- Obter os dados
- Limpe os dados
- Análise exploratória de dados
- Previsão estatística / modelagem
- Interpretar resultados
- Resultados do desafio
- Sintetizar / escrever resultados
-

Criar código reprodutível

Um exemplo

Iniciar com uma pergunta geral

Posso detectar automaticamente os e-mails que são SPAM ou não?

Faz concreta

Posso usar características quantitativas dos e-mails para classificá-los como SPAM / HAM?

Sub amostra de nosso conjunto de dados

Precisamos gerar um conjunto de testes e treinamento (previsão)

```
# Se não estiver instalado, instale o pacote kernlab
library(kernlab)
data(spam)
# Executrando o subsampling
set.seed(3435)
trainIndicator = rbinom(4601,size=1,prob=0.5)
table(trainIndicator)
```

```
## trainIndicator
##    0    1
## 2314 2287
```

```
trainSpam = spam[trainIndicator==1,]
testSpam = spam[trainIndicator==0,]
```

Análise exploratória de dados

- Veja resumos dos dados
- Verificar dados em falta
- Criar gráficos exploratórios
- Executar análises exploratórias (por exemplo, clustering)

Names

```
names(trainSpam)
```

```
##  [1] "make"              "address"           "all"
##  [4] "num3d"             "our"               "over"
##  [7] "remove"            "internet"          "order"
## [10] "mail"              "receive"           "will"
## [13] "people"            "report"            "addresses"
## [16] "free"              "business"          "email"
## [19] "you"               "credit"            "your"
## [22] "font"              "num000"            "money"
## [25] "hp"                "hpl"               "george"
## [28] "num650"            "lab"               "labs"
## [31] "telnet"            "num857"            "data"
## [34] "num415"            "num85"             "technology"
## [37] "num1999"           "parts"             "pm"
## [40] "direct"            "cs"                "meeting"
## [43] "original"          "project"           "re"
## [46] "edu"               "table"             "conference"
## [49] "charSemicolon"     "charRoundbracket"  "charSquarebracket"
```

```
## [52] "charExclamation"   "charDollar"        "charHash"
## [55] "capitalAve"        "capitalLong"       "capitalTotal"
## [58] "type"
```

Head

```
head(trainSpam)
```

```
##    make address  all num3d  our over remove internet order mail receive
## 1  0.00    0.64 0.64     0 0.32 0.00   0.00        0  0.00 0.00    0.00
## 7  0.00    0.00 0.00     0 1.92 0.00   0.00        0  0.00 0.64    0.96
## 9  0.15    0.00 0.46     0 0.61 0.00   0.30        0  0.92 0.76    0.76
## 12 0.00    0.00 0.25     0 0.38 0.25   0.25        0  0.00 0.00    0.12
## 14 0.00    0.00 0.00     0 0.90 0.00   0.90        0  0.00 0.90    0.90
## 16 0.00    0.42 0.42     0 1.27 0.00   0.42        0  0.00 1.27    0.00
##    will people report addresses free business email  you credit your font
## 1  0.64   0.00      0         0 0.32        0  1.29 1.93   0.00 0.96    0
## 7  1.28   0.00      0         0 0.96        0  0.32 3.85   0.00 0.64    0
## 9  0.92   0.00      0         0 0.00        0  0.15 1.23   3.53 2.00    0
## 12 0.12   0.12      0         0 0.00        0  0.00 1.16   0.00 0.77    0
## 14 0.00   0.90      0         0 0.00        0  0.00 2.72   0.00 0.90    0
## 16 0.00   0.00      0         0 1.27        0  0.00 1.70   0.42 1.27    0
##    num000 money hp hpl george num650 lab labs telnet num857 data num415
## 1       0  0.00  0   0      0      0   0    0      0      0 0.00      0
## 7       0  0.00  0   0      0      0   0    0      0      0 0.00      0
## 9       0  0.15  0   0      0      0   0    0      0      0 0.15      0
## 12      0  0.00  0   0      0      0   0    0      0      0 0.00      0
## 14      0  0.00  0   0      0      0   0    0      0      0 0.00      0
## 16      0  0.42  0   0      0      0   0    0      0      0 0.00      0
##    num85 technology num1999 parts pm direct cs meeting original project re
## 1      0          0    0.00     0  0   0.00  0       0      0.0       0  0
## 7      0          0    0.00     0  0   0.00  0       0      0.0       0  0
## 9      0          0    0.00     0  0   0.00  0       0      0.3       0  0
## 12     0          0    0.00     0  0   0.00  0       0      0.0       0  0
## 14     0          0    0.00     0  0   0.00  0       0      0.0       0  0
## 16     0          0    1.27     0  0   0.42  0       0      0.0       0  0
##    edu table conference charSemicolon charRoundbracket charSquarebracket
## 1    0     0          0         0.000            0.000                 0
## 7    0     0          0         0.000            0.054                 0
## 9    0     0          0         0.000            0.271                 0
## 12   0     0          0         0.022            0.044                 0
## 14   0     0          0         0.000            0.000                 0
## 16   0     0          0         0.000            0.063                 0
##    charExclamation charDollar charHash capitalAve capitalLong capitalTotal
## 1            0.778      0.000    0.000      3.756          61          278
## 7            0.164      0.054    0.000      1.671           4          112
## 9            0.181      0.203    0.022      9.744         445         1257
## 12           0.663      0.000    0.000      1.243          11          184
## 14           0.000      0.000    0.000      2.083           7           25
## 16           0.572      0.063    0.000      5.659          55          249
##    type
```

```
## 1  spam
## 7  spam
## 9  spam
## 12 spam
## 14 spam
## 16 spam
```

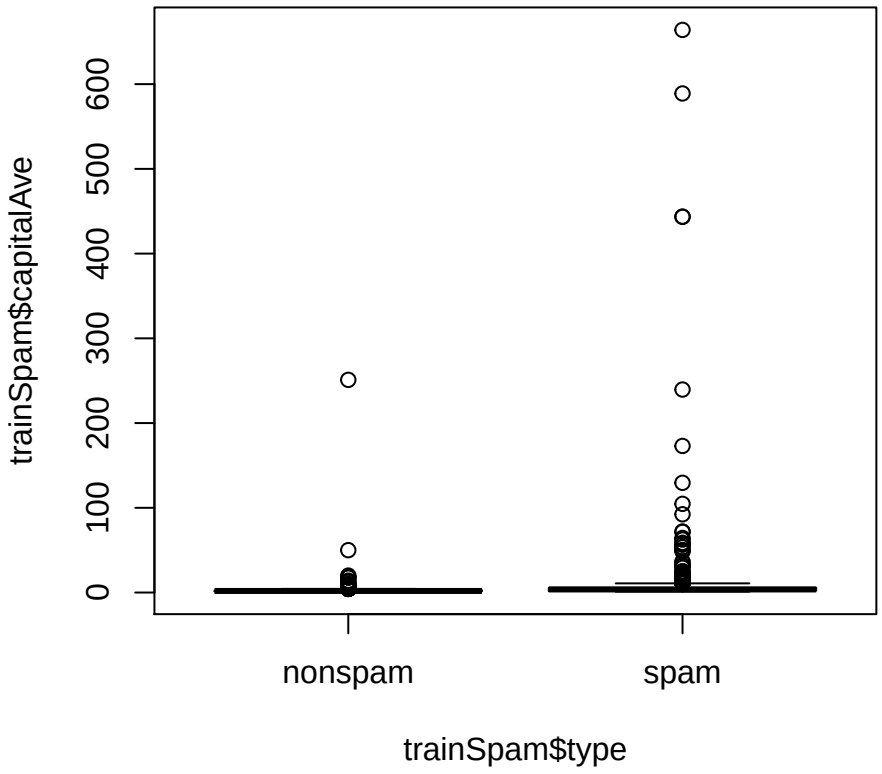
Summaries

```
table(trainSpam$type)
```

```
##
## nonspam    spam
##    1381     906
```

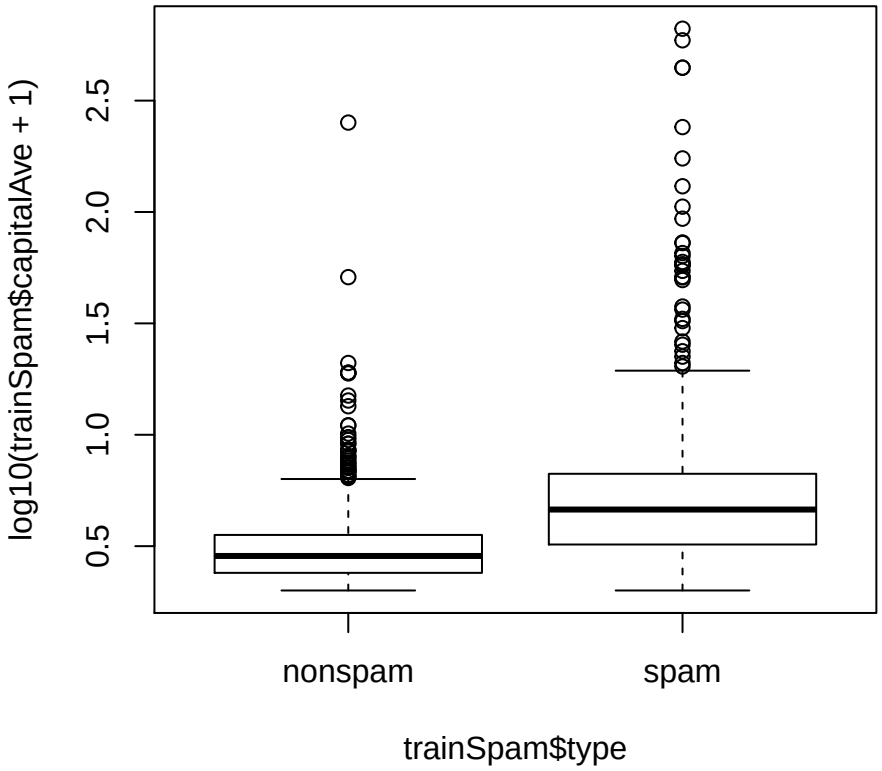
Plots

```
plot(trainSpam$capitalAve ~ trainSpam$type)
```



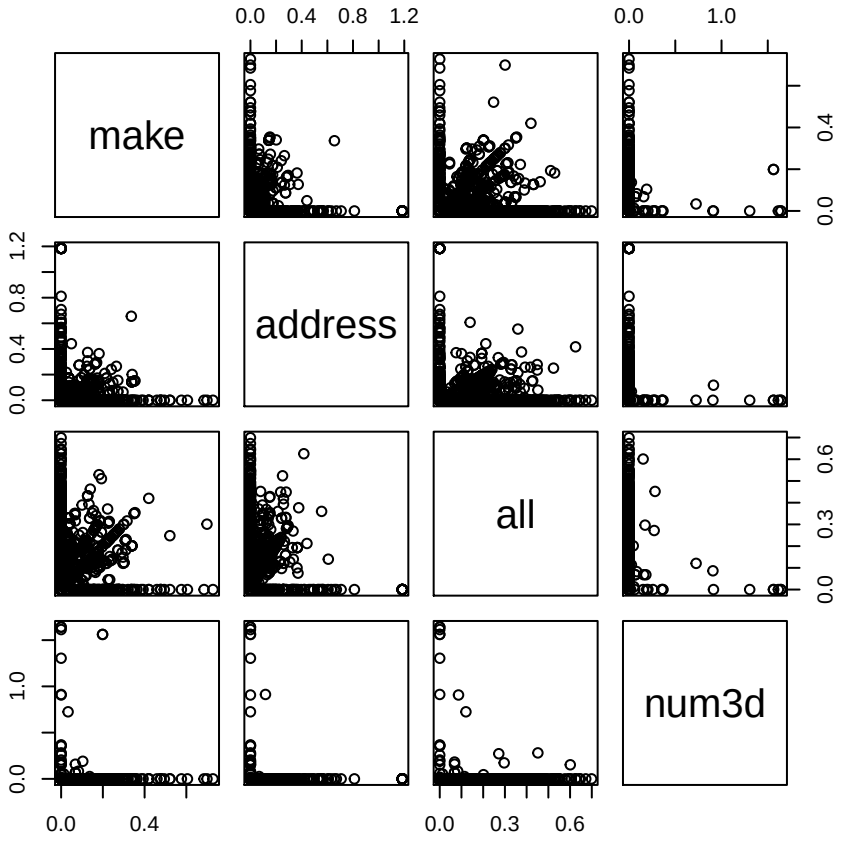
Plots

```
plot(log10(trainSpam$capitalAve + 1) ~ trainSpam$type)
```



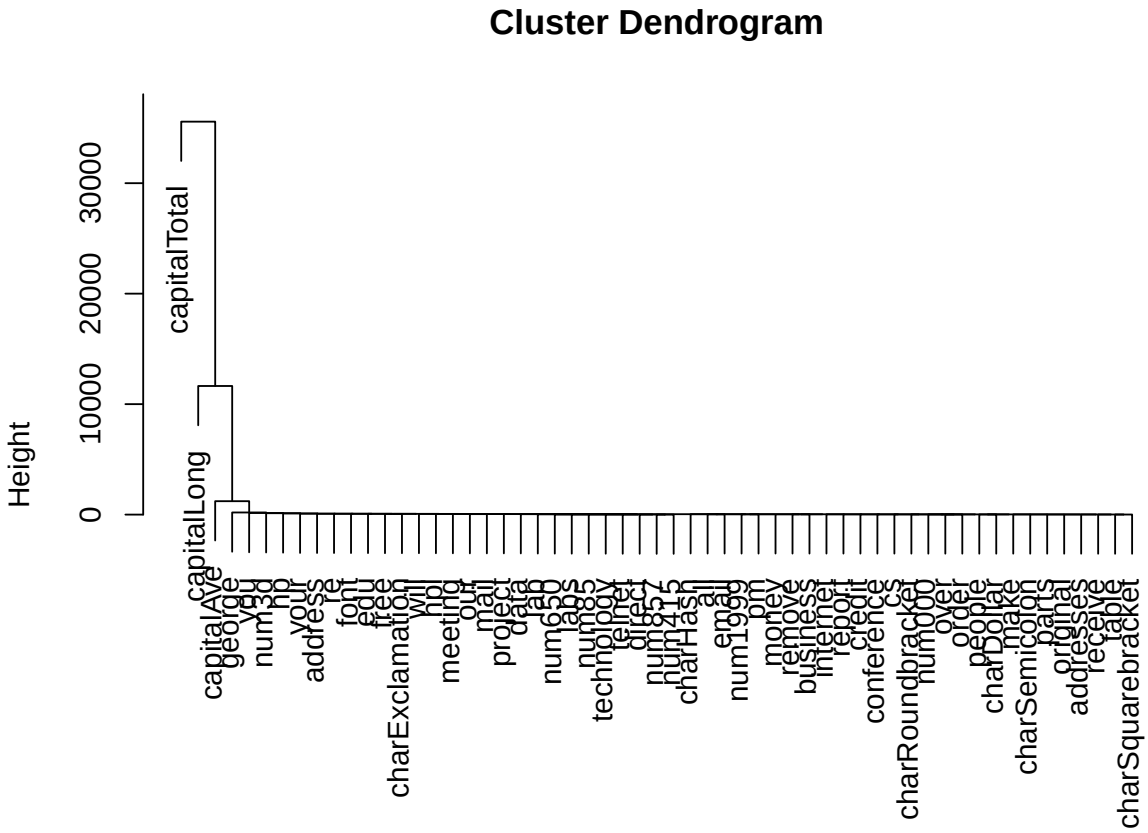
Relationships between predictors

```
plot(log10(trainSpam[,1:4]+1))
```



Clustering

```
hCluster = hclust(dist(t(trainSpam[,1:57])))
plot(hCluster)
```



Previsão / modelagem estatística

- Deve ser informado pelos resultados de sua análise exploratória
- Métodos exatos dependem da questão de interesse
- As transformações / processamento devem ser contabilizadas quando necessário
- Medidas de incerteza devem ser relatadas

```
trainSpam$numType = as.numeric(trainSpam$type)-1
costFunction = function(x,y) sum(x!=(y > 0.5))
cvError = rep(NA,55)
library(boot)
for(i in 1:55){
  lmFormula = reformulate(names(trainSpam)[i], response = "numType")
  glmFit = glm(lmFormula,family="binomial",data=trainSpam)
  cvError[i] = cv.glm(trainSpam,glmFit,costFunction,2)$delta[2]
}
```

```
## Qual predictor tem erro de validação cruzada mínimo?
names(trainSpam)[which.min(cvError)]
```

```
## [1] "charDollar"
```

Obtenha uma medida de incerteza

```
## Use the best model from the group
predictionModel = glm(numType ~ charDollar,family="binomial",data=trainSpam)

## Obter previsões no conjunto de teste
predictionTest = predict(predictionModel,testSpam)
predictedSpam = rep("nonspam",dim(testSpam)[1])

## Classificar como "spam" para aqueles com prob > 0.5
predictedSpam[predictionModel$fitted > 0.5] = "spam"
```

Get a measure of uncertainty

```
## Classification table
table(predictedSpam,testSpam$type)
```

```
##
## predictedSpam nonspam spam
##       nonspam    1346  458
##       spam         61  449
```

```
## Error rate
(61+458)/(1346+458 + 61 + 449)
```

```
## [1] 0.2242869
```

Nosso exemplo

- A fração de charcters que são sinais de dólar pode ser usado para prever se um e-mail é Spam
- Qualquer coisa com mais de 6,6% de sinais de dólar é classificado como Spam
- Mais sinais de dólar sempre significa mais Spam sob nossa previsão
- Nossa taxa de erro de teste foi de 22,4%

Resultados do desafio

Desafie todas as etapas: * Pergunta * Fonte de dados * Em processamento * Análise * Conclusões * Desafio de medidas de incerteza * Desafie escolhas de termos a incluir nos modelos * Pense em análises alternativas em potencial

---

Sintetizar / escrever resultados

- Conduzir com a pergunta
- Resumir as análises na história
- Não incluir todas as análises, incluí-lo
- Se for necessário para a história
- Se for necessário para enfrentar um desafio
- Ordem de análise de acordo com a história, em vez de cronologicamente
- Incluir figuras “bonitas” que contribuam para a história

No nosso exemplo

- Conduzir com a pergunta
- Posso usar características quantitativas dos e-mails para classificá-los como SPAM / HAM?
- Descrever a abordagem
- Dados coletados de UCI -> criado treinamento / conjuntos de teste
- Relações exploradas
- Escolha modelo logístico no treinamento definido por validação cruzada
- Aplicado ao teste, 78% de precisão do conjunto de teste
- Interpretar os resultados
- O número de sinais de dólar parece razoável, p. “Ganhar dinheiro com Viagra \$ \$ \$ \$!”
- Resultados do desafio
- 78% não é tão grande
- Eu poderia usar mais variáveis
- Por que regressão logística?