ML\_Bagging\_in\_R

Juliano Sarnes Longo

12/03/2022

## ML Bagging em R

### Introdução

Quando construímos uma *árvore de decisão* para tentar predizer algum resultado utilizamos somente um único conjunto de dados para treinar o modelo, sendo assim, existe uma certa desvantagem ao utilizar uma única árvore de decisão, o problema é que ela tende a sofrer uma *alta variância* e por consequência tende a ter um *baixo viés*.

Se dividirmos o conjunto de dados em duas metades e aplicarmos a árvore de decisão e aplicarmos a árvore de decisão nessas metades, os resultados podem ser bem diferentes.

Um método que pode ser utilizado para reduzir a variação de uma única árvore de decisão é conhecido como **bagging** (ensacamento), também chamado de **bootstrap aggregation**.

### Entendo o funcionamento do bagging “ensacamento”

1. Pegue *b* amostras de bootstrap do conjunto de dados original;
2. Construia uma árvore de decisão para cada amostra bootstrap;
3. Faça a média das previsões de cada árvore para chegar a um modelo final;

Ao realizar as operações acima criaremos centenas ou até mesmo milhares de árvores de decisão individuais e obteremos as previsões médias de todas as árvores, assim muitas vezes acabaremos com um modelo ensacado ajustado que produz uma taxa de erro de teste muito menor em comparação com uma única árvore.

### Criando o Bagging com R

1. Bibliotecas do R para construção do Bagging.

library(caret) # para construcao da árvore.

## Warning: package 'caret' was built under R version 4.1.2

## Carregando pacotes exigidos: ggplot2

## Carregando pacotes exigidos: lattice

library(e1071) # para calcular a importância da variável.

## Warning: package 'e1071' was built under R version 4.1.2

library(rpart) # para ajustar árvores de decisão.   
library(ipred) # para ajustar bagging.

## Warning: package 'ipred' was built under R version 4.1.2

library(dplyr) # para aplicação de data wrangling

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

1. Analisando a o dataset Airquality.

Para a construção desse mini projeto usarei o dataset **airquality** que já faz parte dos pacotes do R Studio, este conjunto de dados contém as medições da qualidade do ar feitas na cidade de New York em 153 dias.

class(airquality)

## [1] "data.frame"

#head(airquality)  
  
# substitui NAs por 0.  
df\_airquality <- data.frame(airquality) %>% replace(is.na(.), 0)  
head(df\_airquality)

## Ozone Solar.R Wind Temp Month Day  
## 1 41 190 7.4 67 5 1  
## 2 36 118 8.0 72 5 2  
## 3 12 149 12.6 74 5 3  
## 4 18 313 11.5 62 5 4  
## 5 0 0 14.3 56 5 5  
## 6 28 0 14.9 66 5 6

O dataset airquality é composto por 153 observações e 6 variáveis, sendo que nossa variável resposta é **Ozone**.

dim(df\_airquality)

## [1] 153 6

1. Criação do modelo ensacado “bagged”.

modeloEnsacado <- ipred::bagging(formula = Ozone ~ .,   
 data = df\_airquality,   
 nbagg = 150, # número de amostras bootstrap.  
 coob = TRUE, # indica se uma estimativa de taxa de erro fora do ensacamento (   
 # classificação incorreta,  
 # erro quadrático médio ou   
 # pontuação de Brier) deve ser computada.   
 control = rpart.control(minsplit = 2, cp = 0) # hiperparâmetros.  
 )  
  
modeloEnsacado

##   
## Bagging regression trees with 150 bootstrap replications   
##   
## Call: bagging.data.frame(formula = Ozone ~ ., data = df\_airquality,   
## nbagg = 150, coob = TRUE, control = rpart.control(minsplit = 2,   
## cp = 0))  
##   
## Out-of-bag estimate of root mean squared error: 24.2891

Observe que criamos o bagging com as seguintes definições: 1. Usamos 150 amostras bootstrap para construir o modelo ensacado; 2. Definimos o parâmetro coob como TRUE para obter o erro estimado fora do bagg. 3. Hiperparêmtros: - minsplit = 2 informa ao modelo para exigir apenas 2 observações em um nó para dividir. - cp = 0 o parâmetro de complexidade. definido como 0 indica que não exigimos que o modelo seja capaz de melhorar o ajuste geral em qualquer valor para realizar uma divisão.

## Conclusão:

Analisando os resultados.

Bagging regression trees with 150 bootstrap replications

Call: bagging.data.frame(formula = Ozone ~ ., data = df\_airquality, nbagg = 150, coob = TRUE, control = rpart.control(minsplit = 2, cp = 0))

Out-of-bag estimate of root mean squared error: 24.4453

Por fim verificamos o modelo nos diz que a RMSE (erro quadrático médio ou MSE) estimado fora do ensacamento é de 24.4453, ou seja, esta é a diferença média entre o valor previsto para o Ozone e o valor real observado.