ML\_VALIDACAO\_CRUZADA\_K-FOLD

Juliano Sarnes Longo

25/03/2022

## Objetivo

O objeto deste script é mostrar como avaliar o desempenho de um modelo preditivo usando Cross Validation K-Fold, uma idéia simples e muito poderosa, usada para garantir que a acurácia do modelo preditivo desenvolvido é realmente eficiente, ou seja, as predições resultantes geradas pelo modelo correspondem aos dados observados.

## Validação cruzada K-Fold ou Cross Validation K-Fold

Como mencionado acima a validação cruzada k-fold tem como objetivo garantir que o modelo desenvolvido consegue gerar previsões de forma eficiente em relação aos dados da observação.

A validação cruzada estende a idéia de se ter uma única amostra de retenção ou teste, para múltiplas amostras de retenção sequenciais, onde o algoritmo básico aplica o seguinte esquema funcional:

1. Divide um conjunto de dados de forma aleatória em k agrupamentos também chamado de “dobra” ou “fold”, sendo que esses agrupamentos possuem aproximadamente o mesmo de tamanho.
2. O algoritmo escolhe UM dos agrupamentos ou “folds” para utilizar como conjunto de retenção ou teste e os demais folds “k-1” são aplicados ao modelo como conjunto de treinamento.
3. Usa o conjunto de retenção ou teste para validar o modelo, geralmente é nessa etapa que é calculado erro quadrático médio (MSE), que será usado para validação do modelo.

Quanto menor for o valor do **MSE** mais próximas as predições estão das observações.

1. O algoritmo repete o processo K vezes.

## Aplicando a idéia de Validação Cruzada ou Cross Validation K-Fold usando R.

Para aplicar o Cross Validation K-Fold usando a linguagem R ou criar um dataframe com números inteiros.

# criação do dataframe df.input  
df.input <- data.frame(y = c(14, 12, 12, 13, 7, 8, 7, 4, 6, 5),   
 x1 = c(6, 8, 12, 14, 14, 15, 17, 22, 24, 23),   
 x2 = c(2, 5, 4, 3, 4, 6, 7, 5, 8, 9))  
  
df.input

## y x1 x2  
## 1 14 6 2  
## 2 12 8 5  
## 3 12 12 4  
## 4 13 14 3  
## 5 7 14 4  
## 6 8 15 6  
## 7 7 17 7  
## 8 4 22 5  
## 9 6 24 8  
## 10 5 23 9

Utilizarei um modelo de **Regressão Linear Múltipla** para aplicar o conjunto de dados e efetuar a validação cruzada K-fold com k=5 folds e assim poder avaliar a eficiência do modelo.

library(caret)

## Warning: package 'caret' was built under R version 4.1.3

## Carregando pacotes exigidos: ggplot2

## Carregando pacotes exigidos: lattice

## Warning: package 'lattice' was built under R version 4.1.3

# definição do cross validation.  
controle <- trainControl(method = "cv" # método validação Cross Validation.  
 , number = 5 # total de dobras ou folds 5.  
 )  
  
# criação do modelo de regressão linear múltipla para avaliar a eficiência deste  
# modelo usando o Cross Validation K-Fold.  
modeloRLM <- train(  
 y ~ x1 + x2, # formula  
 data=df.input, # dataframe  
 method = "lm", # aplica a regressão linear múltipla.  
 trControl = controle # aplica a validação cruzada no conjunto de treino.  
 )

## Warning in nominalTrainWorkflow(x = x, y = y, wts = weights, info = trainInfo, :  
## There were missing values in resampled performance measures.

print(modeloRLM)

## Linear Regression   
##   
## 10 samples  
## 2 predictor  
##   
## No pre-processing  
## Resampling: Cross-Validated (5 fold)   
## Summary of sample sizes: 8, 8, 8, 8, 8   
## Resampling results:  
##   
## RMSE Rsquared MAE   
## 1.897228 1 1.684835  
##   
## Tuning parameter 'intercept' was held constant at a value of TRUE

## Interpretando as informações geradas pelo modelo de Regressão Linear Múltipla com Cross Validation K-Fold.

1. **10 samples:** Existem 10 observações.
2. **2 predictor:** Existem 2 variáveis preditoras.
3. **No pre-processing:** Não ocorreu nenhum tipo de pré-processamento do conjunto de dados.
4. **Resampling: Cross-Validated (5 fold):** o método de reamostragem que foi aplicado para avaliação do modelo foi o Cross Validation com 5 Folds.
5. **Summary of sample sizes: 8, 8, 8, 8, 8:** foram criados 5 fold de tamanho 8.
6. **Resampling results:**

* **RMSE:** A raiz do erro quadrático médio, mede a diferença média entre as previsões feitas pelo modelo e as observações reais, quanto menor for o RMSE, mais próximo um modelo pode prever as observações reais.
* **Rsquared:** É uma medida da correlação entre as previsões feitas pelo modelo e as observações reais, quanto maior for o R-quadrado, mais próximo um modelo pode prever as observações reais.
* **MAE:** é o erro absoluto médio, esta é a diferença média absoluta entre as previsões geradas pelo modelo e as observações reais, quanto menor for o MAE mais próximos um modelo pode prever as observações reais.

As métricas de saída (RMSE, Rsquared e MAE) nos permite ter uma ideia de como está o desempenho do modelo, geralmente ajustamos os diferentes modelos e fazemos a comparação das três métricas fornecidas na saída, o modelo que produzir as menores taxas de erro é o melhor modelo.

# examinando o modelo final ajustado.  
modeloRLM$finalModel

##   
## Call:  
## lm(formula = .outcome ~ ., data = dat)  
##   
## Coefficients:  
## (Intercept) x1 x2   
## 17.0594 -0.4446 -0.2582

## O modelo final ajustado.

Visualização das predições do modelo feita em cada um dos 5 folds:

#predição em cada fold.  
modeloRLM$resample

## RMSE Rsquared MAE Resample  
## 1 0.258375 1 0.2094519 Fold1  
## 2 3.060175 1 2.5739982 Fold2  
## 3 2.489796 1 2.3121197 Fold3  
## 4 1.218724 1 1.1668567 Fold4  
## 5 2.459067 NA 2.1617471 Fold5

Em algumas literaturas encontramos a seguinte indicação: Geralmente escolha entre 5 e 10 folds, por que esse é o número ideal de folds capazes de produzir taxas de erros e testes confiáveis, por oferecer um equilíbrio ideal entre viés e variância.

**-Page 184, An Introduction to Statistical Learning**