Cross Validation

Zakaria RIDADARAJAT

02/02/2021

Entête :

Nous allons explorer dans ce document la méthode Cross Validation faite par Rindra LUTZ et Marko ARSIC.

Voici le lien vers leur GitHub est: **https://github.com/ARSICMrk/ARSIC_PSBx/tree/main/R_Travail_Sup**

Synthèse :

Ce document est basé sur essentiellement sur la méthode de le cross validation.Cette méthode nous permet de mesurer la capacité de généralisation d'un modèle sans introduire de biais, ni fuites de données. Le plus commun consiste à diviser son dataset en données d'entrainement, de validation et de test.

Pour mieux comprendre cette méthode, nous allons se base sur l'exemple mentionné sur le rapport de Marko et Rindra.

Extrait commenté :

Tout d'abord, ils ont commencé par installer et charger le package tidyverse et caret qui est indispensable pour la construction du modèle prédictif, après ils sont fait appel à la table swiss.

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --

## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.4     v dplyr   1.0.2
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.0

## Warning: package 'ggplot2' was built under R version 3.6.2

## Warning: package 'tibble' was built under R version 3.6.2

## Warning: package 'tidyr' was built under R version 3.6.2

## Warning: package 'readr' was built under R version 3.6.2

## Warning: package 'purrr' was built under R version 3.6.2
```

```
## Warning: package 'dplyr' was built under R version 3.6.2
```

```
## -- Conflicts --------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(caret)
```

```
## Loading required package: lattice
```

```
##
## Attaching package: 'caret'
```

```
## The following object is masked from 'package:purrr':
##
##     lift
```

```
data("swiss")

sample_n(swiss, 3)
```

```
##              Fertility Agriculture Examination Education Catholic
## Payerne           74.2        58.1          14         8     5.23
## Rive Droite       44.7        46.6          16        29    50.43
## V. De Geneve      35.0         1.2          37        53    42.34
##              Infant.Mortality
## Payerne                  23.8
## Rive Droite              18.2
## V. De Geneve             18.0
```

Ils sont choisis un cross validation k-fold de 10, puis ils entrainent le modèle.

```
## Linear Regression
##
## 47 samples
##  5 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 42, 42, 42, 42, 42, 44, ...
## Resampling results:
##
##   RMSE      Rsquared   MAE
##   7.424916  0.6922072  6.31218
##
## Tuning parameter 'intercept' was held constant at a value of TRUE
```

Evaluation :

1) Exécution du fichier rmd:

Le fichier rmd s'exécute parfaitement sur la totalité du script.

2) Qualité de rédaction :

Il est très agréable sur le plan visuel, ils bien respectes les règles de base d'un rapport.

3) Aspect didactique

Le document est parfaitement clair car ils ont bien détaillé toutes les étapes nécessaires pour appliquer la méthode du cross validation. Ils sont aussi montres l'importance de cette technique pour les projets data.

4) Lisibilite du rmarkdown:

Le code est bien organisé et expliqué et le rmarkdown est bien fait.

5) Bibliographie :

Effectivement, les auteurs ont mentionné la bibliographie à la fin du rapport pour les gens qui veulent approfondir un peu plus sur le sujet.

Conclusion :

En globale, le travail est bien fait et limpide. Rindra LUTZ et Marko ARSIC m'ont bien assisté indirectement à travers leur travail à connaitre les étapes essentielles de la méthode cross validation et son utilité.