

# Evaluation des dossiers sur la Programmation en R

Les « 12 travaux d'AstéRix »

Daif Hakim

31/01/2021

## Cross Validation

### 1. Critères d'évaluation

Dans ce dossier, nous allons parcourir et faire une évaluation des travaux de mes camarades sur la Programmation en R:

Nos critères seront :

- 1. Visuel et l'organisation du dossier:** Es ce que la mise en page est propre ? Le travail est-il organisé (chapitres, sections, sous-sections, Parties) ?
- 2. Qualité de la rédaction du dossier:** le document est-il agréable à lire? es ce que l'auteur a respecté les consignes de rédaction d'un article scientifique ou d'une thèse... ?
- 3. Compréhension de l'idée générale:**Les Auteurs arrivent-ils à transmettre l'idée générale ?
- 4. lisibilité et fonctionnement du code (explication et commentaires):** Es ce que le code fonctionne parfaitement sans erreurs ? Es ce que l'auteur a mis des commentaires pour expliquer chaque ligne de commande ?
- 5. L'utilisation d'exemples et de cas concrets (cas pratique):** l'Auteur a t'il utilisé des exemples et des cas pratique lors de sa rédaction ?

### 2. Lien vers le document commenté

Pour ce dossier, nous avons étudié le travail de Marko ARSIC et Rindra LUTZ qui porte sur la cross validation le lien (github) vers leur dossier est : [https://github.com/ARSICMrk/ARSIC\\_PSBx/tree/main/R\\_Travail\\_Sup](https://github.com/ARSICMrk/ARSIC_PSBx/tree/main/R_Travail_Sup)

Le dossier en question est le dossier **Cross Validation.Rmd**

### 3. Synthèse du document

La cross validation est une Régression logistique (méthodes de régression, méthodes prédictives) et qui est une méthode statistique qui permet d'évaluer la capacité de généralisation d'un modèle.

La crosse validation vise à : construire un modèle permettant de prédire/expliciter les valeurs prises par une variable cible qualitative , à partir d'un ensemble de variables explicatives quantitatives ou qualitatives. Une fois que le modèle a été établi, il est alors nécessaire de valider sa fiabilité.

Marko et Rindra nous ont expliqué la cross validation d'une façon très détaillée, en commençant par expliquer les problématiques liées du data puis les bases du data mining.

Ils nous ont présenté également la méthode descriptive et la méthode prédictive, c'est à partir de cette méthode prédictive que nous entameront la cross validation qui est une régression logique, et aussi la partie principale dans ce document, ils nous ont montré quelques exemples pour nous faire comprendre mieux cross validation.

les auteurs n'ont pas oublié de faire un point sur les différentes méthodes utilisées dans le cross validation :

- LOOCV (leave-one-out cross-validation) : 1 Observation à prédire et M-1 pour l'entraînement
- LKOCV (leave-k-out cross-validation): K Observation à prédire même principe que LOOCV
- k-fold cross-validation : les observations sont aléatoirement divisées en K sous-échantillons de tailles égales

#### 4. Extrait commenté :

DATA Table swiss, représente la mesure de la fécondité et des indicateurs socio-économiques standardisés pour chacune des 47 provinces francophones de Suisse vers l'an 1888 -Téléchargement des données : `data("swiss")` -Inspecter les données : `sample_n(swiss, 3)`

Nous allons commenter l'exemple utilisé dans ce document, Marko et Rindra ont d'abord installé et chargé le package `tidyverse` et `caret` qui est indispensable pour construire un modèle prédictif. Ensuite ils utilisent le data frame `swiss` pour effectuer les analyses.

```
library(tidyverse)
library(caret)
data("swiss")
sample_n(swiss, 3)
```

Et puis, ils utilisent la méthode cross validation pour définir l'échantillon d'entraînement.

```
set.seed(123)
train.control <- trainControl(method = "cv", number = 10)
```

Enfin, ils ont entraîné le modèle :

```
model <- train(Fertility ~., data = swiss, method = "lm",
               trControl = train.control)
print(model)
```

Pendant tout l'exemple, les auteurs ont très bien expliqué le processus et chaque ligne de code, ils ont même comparé juste après les résultats en utilisant le cross validation avec 3 répétitions .

### 5. Evaluation du travail suivant les 5 critères précités

#### 1. Visuel et l'organisation du dossier :

c'est un dossier propre, Visuel agréable, organisation bonne mais peut être fallait mieux diviser le travail en sections ou en partie pour une parfaite organisation.

#### 2. Qualité de la rédaction du dossier :

le document est très bien rédigé, des phrases simples à comprendre, agréable à la lecture

### **3. compréhension de l'idée générale :**

j'en déduis que les auteurs maîtrisent très bien le sujet d'une façon générale, je pense aussi que les lecteurs qui maîtrisent déjà les notions de base de la programmation en R, n'auront aucun problème à la compréhension de l'idée générale, et vu qu'ils ont consacré des exemples explicatifs à chaque partie cela facilite énormément la compréhension de chaque section.

### **4. L'utilisation d'exemples et de cas concrets (cas pratique) :**

Les auteurs vont utiliser un exemple pour chaque section. Ils vont détailler au maximum pour qu'on puisse comprendre.

### **5. lisibilité, explication (commentaires) et fonctionnement du code :**

Le code est parfaitement lisible, L'exécution du RMarkdown se réalise parfaitement, les commandes sont très bien expliquées d'une façon générale, chaque commande a son commentaire ce qui implique une compréhension parfaite des commandes R et de l'algorithme.

## **6. Conclusion**

Il s'agit d'un bon travail, Les deux auteurs ont dans l'ensemble bien documenté leur dossier. C'est un travail bien construit, recherché, a visée didactique, avec un effort réel pour le rendre accessible au plus grand nombre, vu qu'ils ont commencé par expliquer les problématiques liées du data et les bases du data mining... puis ils ont présenté les différentes méthodes utilisées dans la cross validation...

Tout ça montre à quel point le travail est très bien fait.

Très bien.