Validation Croisée

Benjamin GUIGON

January 2021

1 Lien

Article écrit par Rindra LUTZ Nicolas ALLIX

https://github.com/Nicolas-all/PSB1

2 Méthodes de machine learning

En machine learning il existe 2 grands types de méthode pour traiter les données : la méthode descriptive et la méthode prédictive. Aussi appelées respectivement, méthodes non supervisées et méthodes supervisées.

1 - *Méthodes Descriptives : recherche de structure des données*

Utilisées pour :

- Permet de mettre en évidence des informations non visibles simplement
- Permet de résumer, synthétiser les données
- Sans variable ou phénomène à expliquer a priori

Types :

- les analyses factorielles
- les analyses typologiques
- modèles combiantoires
- les modèles à base de règles

2 - *Méthodes Prédictives : modélisation et prédiction*

Utilisées pour :

- Permet de définir un pattern (un modèle/une relation) pour expliquer un événement
- Permet d'extrapoler la cible

• Avec une variable/un événement à expliquer

Types :

• Régressions (linéaire, logistique, etc)
• Arbres de décision
• Réseaux de neurones
• Analyse discriminante
• SVM - Support Vecteur Machine

3 Validation Croisée

La validation croisée est une méthode d'estimation de fiabilité d'un modèle fondé sur une technique d'échantillonnage.

En machine learning, il y a une manipulation courrante qui vise à diviser les données en deux datasets afin de s'entrainer sur l'un et tester le modèle sur l'autre. Quand on utilise la validation croisée, il convient d'ajouter un 3ème d'échantillons :

- Train
- Valid
- Test

Ce nouvel échantillon permet de tester plusieurs modèles avec plusieurs paramètres afin d'avoir le meilleur résultat possible.

Les 3 méthodes de validation croisée sont :

- LOOCV (leave-one-out cross-validation)
- LKOCV (leave-k-out cross-validation)
- k-fold cross-validation : on divise l'échantillon original en k échantillons, puis on sélectionne un des k échantillons comme ensemble de validation pendant que les k-1 autres échantillons constituent l'ensemble d'apprentissage. Après apprentissage, on peut calculer une performance de validation, puis on répète l'opération en sélectionnant un autre échantillon de validation parmi les blocs de départ. La moyenne et l'écart type des k scores de performances peuvent être calculés pour estimer le biais et la variance de la performance de validation.

4 Sur-apprentissage

Il n'y a pas sur-apprentissage lorsque la performance du modèle en Test est légèrement plus faible que celle en Train. Un écart trop grand est signe de sur-apprentissage.

5 Avis

Cet article, amène bien le sujet en expliquant les différents types d'algorithmes et de méthodes utilisées. Il explique ensuite l'interêt de la validation croisée. L'explication pourrait être accompagné d'un exemple pour un peu mieux comprendre les étapes et les manipulations de différents échantillons.

Cet article m'a appris le concept de validation croisée que je ne connaissais que de nom. C'est une méthode qu'il est indispensable de connaitre pour pouvoir rendre les modèles les plus justes à la partie "métier" de l'entreprise. Il n'est jamais trop d'une seule validation. Il manque peut être un exemple concret sur python, qui aurait donné une belle profondeur et une encore plus grande pertinance.