Validation croisée en Machine Learning

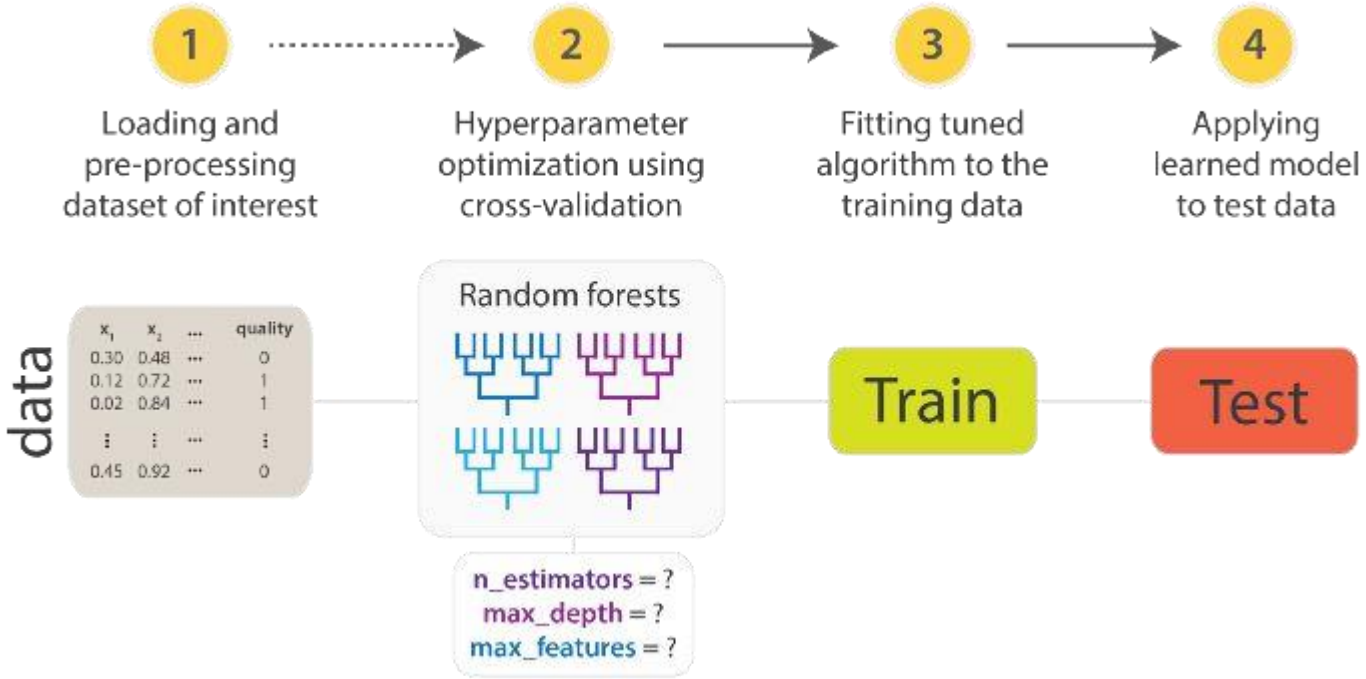
Bassam Kurdy Ph.D
<bassam.kurdy@apinum.fr>

Introduction

? Pourquoi divisons-nous l'ensemble de données ?

- Pour évaluer les performances du modèle
  - déterminer le pouvoir prédictif du modèle

Introduction



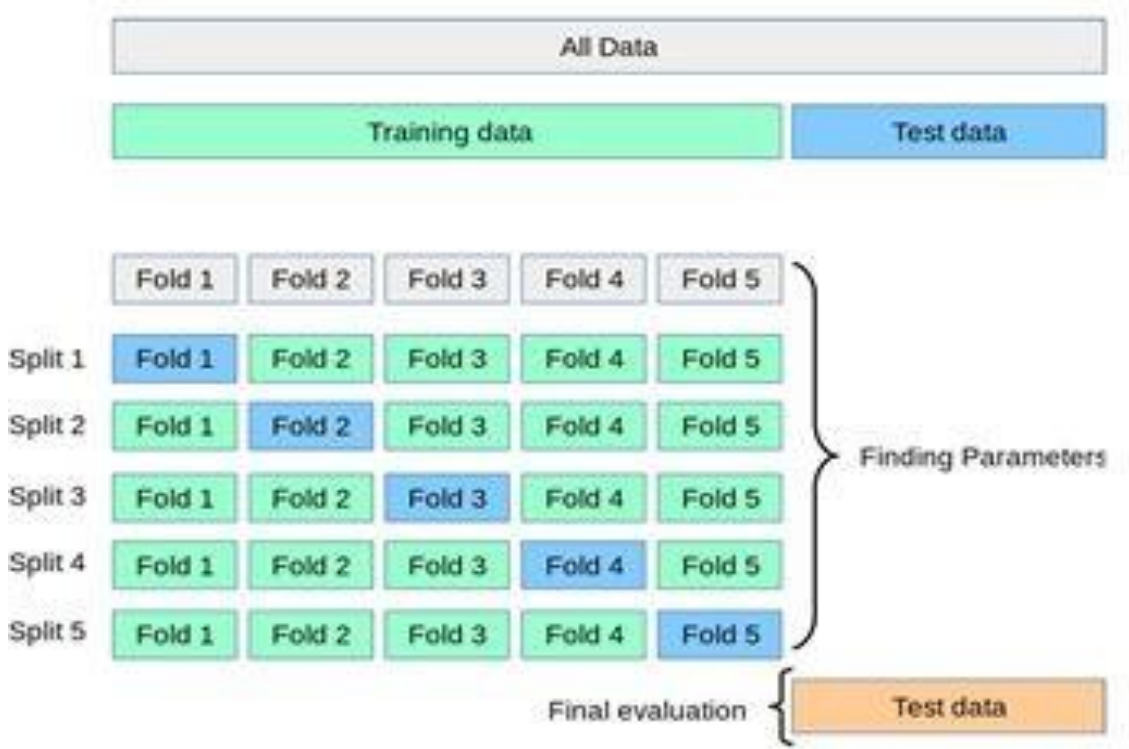
Validation croisée | Quésaco ?

- La **validation croisée** (ou **cross-validation**) est une méthode statistique qui permet d'évaluer la capacité de généralisation d'un modèle.
- La validation croisée est une méthode **stable** et **fiable** pour **évaluer la performance** d'un modèle.

Validation croisée | Quésaco ?

- Un ensemble de tests est conservé pour l'évaluation finale
- Avec les données restantes, les données sont divisées en k fold.
  - Le modèle est ensuite entraîné en utilisant le k-1 des folds (données d'apprentissage).
  - En utilisant le kème ensemble (ensemble de validation), la prédiction est effectuée à l'aide de différentes valeurs d'hyperparamètres et sélectionnez les hyperparamètres qui donnent le meilleur score de validation.
- Le modèle est évalué avec l'ensemble de test en utilisant les meilleurs hyperparamètres.

Validation croisée | k-fold cross validation



Validation croisée : Stratified Cross-Validation

- Supposons que nous avons un jeu de données avec des classes ordonnées.
- La technique de stratification permet de s'assurer que chaque classe est représentée dans chaque partie (fold) lors d'une validation croisée.
- Deux implémentations avec Scikit Learn :
  - StratifiedKFold
  - StratifiedShuffleSplit

Validation croisée | StratifiedKFold

- Variation de KFold.
- Les données sont mélangées et divisées en respectant la pourcentage des observations pour chaque classe.

Validation croisée | StratifiedShuffleSplit

- Fusion de StratifiedKFold et ShuffleSplit.
- Les données sont mélangées et divisées en n parties (folds).
- L'opération est réitérée plusieurs fois.
- Probabilité de chevauchement des données d'entraînement et de test.

Validation croisée | A retenir !

- La méthode de la validation croisée à choisir dépend du problème.
  - compromis entre le temps de calcul et la métrique d'évaluation.
- Pour la méthode k-fold, il est recommandé de choisir k=10
  - démontré expérimentalement que c'est la valeur optimale.
- Pour les problèmes de classification, l'utilisation des méthode stratifiées est recommandée.

Ressources

Train, Validation, Test Set in Machine Learning– How to understand

A Gentle Introduction to k-fold Cross-Validation

Cross-Validation for Parameter Tuning, Model Selection, and Feature Selection

5 Reasons why you should use Cross-Validation in your Data Science Projects

Cross validation and model selection - Scikit learn