



Engénierie des caractéristiques (Feature Engineering)

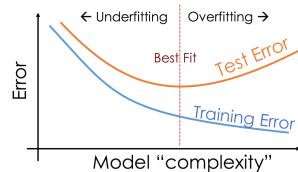
Engénierie des caractéristiques (Feature Engineering)

- Processus de sélection, modification ou création de caractéristiques d'entrée
- Rappel des points clés
 - Sélection des caractéristiques - identifier les caractéristiques les plus pertinentes
 - Lasso, Ridge, Elastic Net, RFE, PCA, ...
 - Transformation des caractéristiques
 - Standardisation (standardscaler), one-hot encoding, label encoding, normalisation (minmaxscaler), ...
 - Création de caractéristiques - générer des variables à partir de la connaissance métier ou en combinant des variables existantes
 - Expansion polynomiale, (polynomialfeatures), ...

Concepts liés

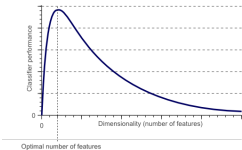
3

- Sur/sous apprentissage



Split train-test

- Fléau de la surdimensionnalité



Sélection de caractéristiques (Lasso, Ridge, ...)

- Biais dans les échantillons
 - Les données ne représentent pas fidèlement la population réelle

Stratified train/test split, ...



Validation Croisée

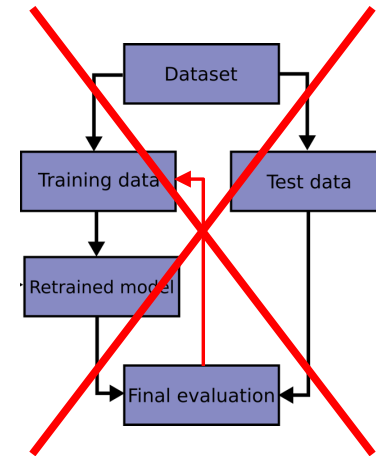
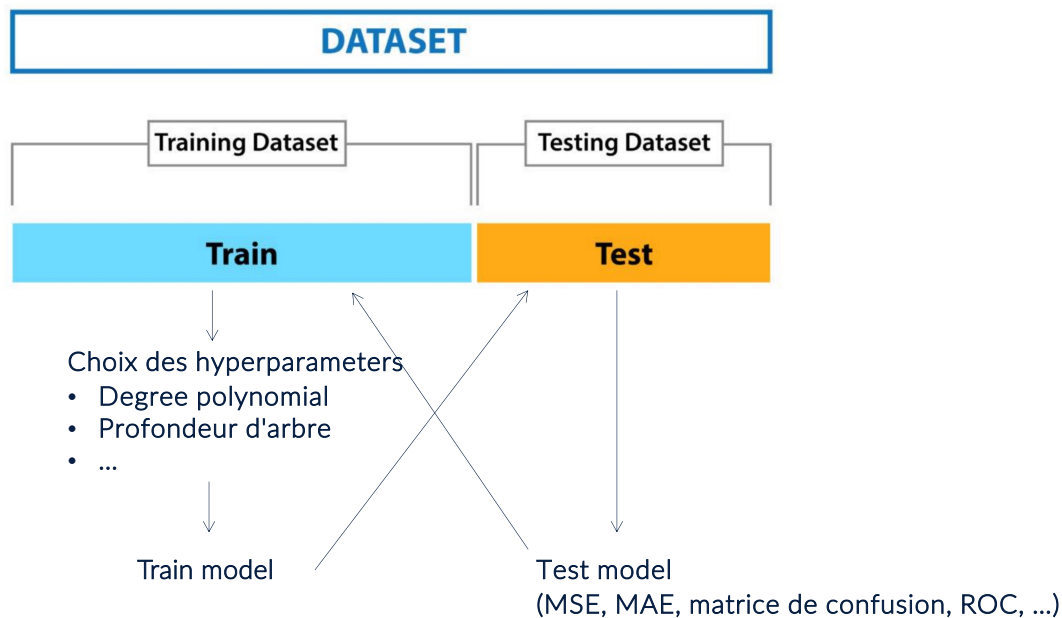
(Cross-validation)

Validation Croisée (Cross Validation)

- Définition
 - Évaluer la capacité d'un modèle à généraliser sur des données non vues à l'entraînement
 - En divisant le jeu de données en sous-ensembles d'entraînement et de test plusieurs fois

Validation Croisée

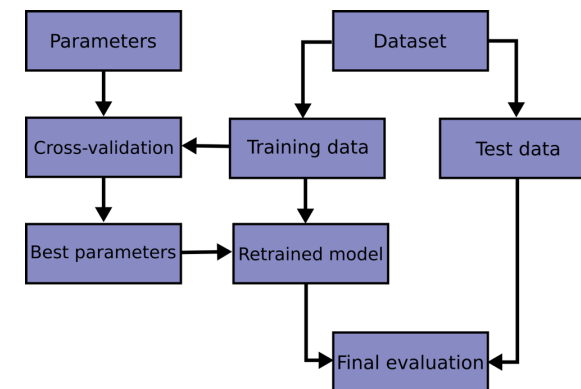
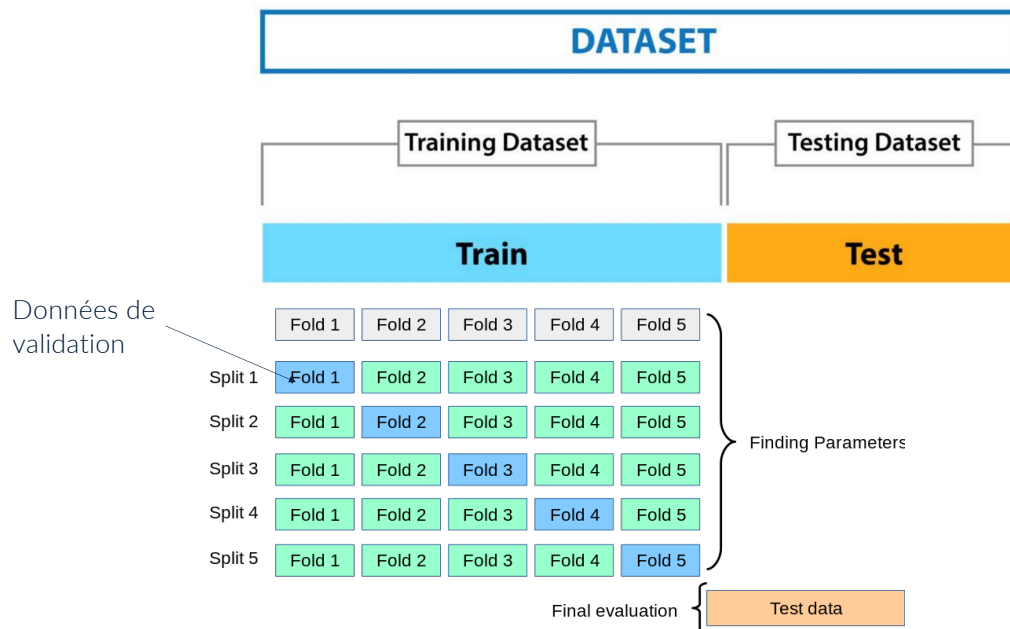
- Exemple d'une approche naïve et inexacte d'entraînement machine



⚠ Fuite d'information des données de test dans les données d'entraînement! - modèle trop optimiste par choix des hyperparamètres

Validation Croisée

- Méthode de validation croisée



- Validation croisée en k blocs (k-Fold Cross-Validation)
- Validation croisée "Leave-One-Out" (LOOCV)
- k-Fold stratifié
- Division temporelle (Time-Series Split)

Validation Croisée

- Importance de la validation croisée
 - Réduit le risque de sur/sous-estimer les performances d'un modèle
 - Pour ajuster les hyperparamètres et évaluer les performances, en gardant le jeu de test totalelement séparé pour l'évaluation finale

Fonctions scikit-learn utiles

- `sklearn.pipeline.Pipeline`
- `sklearn.model_selection.cross_val_score` – methode de validation croisée
- `sklearn.model_selection.GridSearchCV` – validation croisée avec selection des hyperparamètres optimums
- `sklearn.feature_selection.RFE` – Recursive Feature Elimination (RFECV), pour réduire l'espace des caractéristiques

Ref: https://scikit-learn.org/stable/modules/cross_validation.html

QUESTIONS ?

