Échantillonage

Comment évaluer la performance des algorithmes?

Atelier #8

Retour sur les notions de base de l’apprentissage automatique

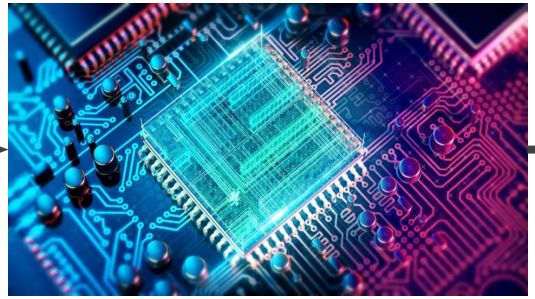
Qu'est-ce que l'**APPRENTISSAGE AUTOMATIQUE** dans un contexte de programmation?

Contexte : Création d'un modèle des données pour répondre à une question précise.

But : Entraîner un algorithme pour faire des prédictions sur un jeu de données.

DONNÉES

ALGORITHME



MODÈLE

Qu’est-ce qu’un **ALGORITHME**?

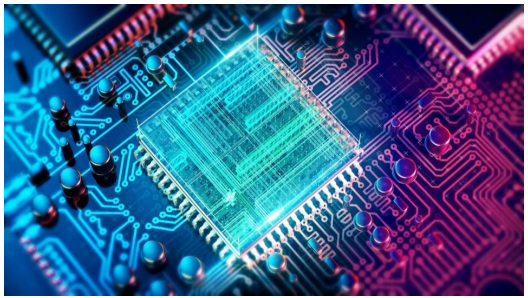
Définition : C’est une méthode statistique qui apprend des données qu’on lui donne pour créer un modèle.

3 principaux types d’algorithmes :

- Régression
- Classification
- Regroupement - (“*clustering*”)

Exemples d’algorithmes d’apprentissage automatique :

- Régression linéaire - (“*linear regression*”)
- Régression logistique - (“*logistic regression*”)
- Arbres de décisions - (“*decision tree*”)
- Réseaux de neurones - (“*neural network*”)



Qu’est-ce qu’un **MODÈLE**?

Définition : C’est le résultat de ce qui a été appris par l’algorithme entraîné sur un jeu de données. Il est unique au jeu de données précis à partir duquel il a été entraîner.

Analogie
Le modèle d’apprentissage automatique est équivalent à un **programme** informatique.

Évaluation d’un modèle

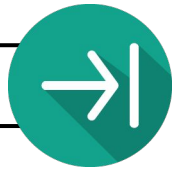
Pourquoi doit-on évaluer un modèle?

Trouver une façon de sélectionner un modèle d’apprentissage automatique parmis plusieurs.

Comment évaluer la performance d’un modèle?

Le but ultime est d’estimer comment un modèle va performer sur de nouvelles valeurs qui n’étaient pas connues dans le jeu de données initial.

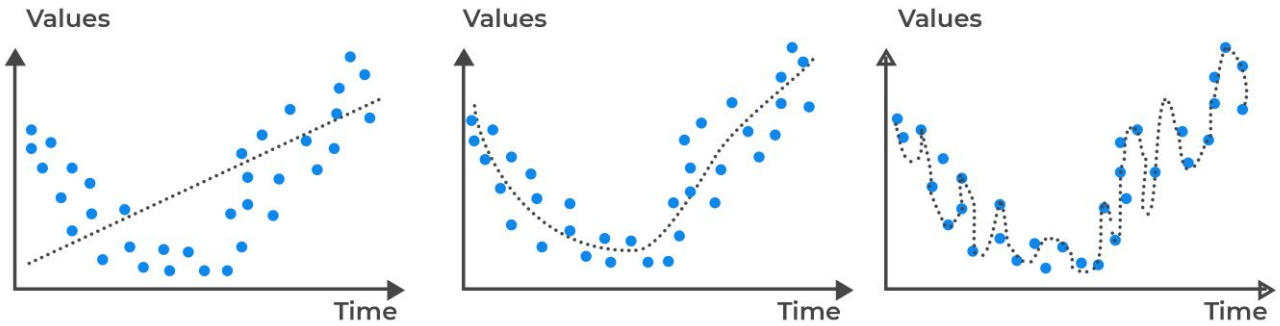
Faire attention au sur-ajustement!



Entraînement et sur-ajustement

IDÉE INITIALE : Entraîner et tester un modèle sur les mêmes données.

PROBLÈME : Maximiser la précision d'un modèle résulte en la création de modèles complexes qui sur-ajustent les données d'entraînement.



Généralisé **Équilibré** **Sur-ajusté**

SOLUTION : Pour éviter d'avoir un algorithme qui sur-ajuste un modèle, il est essentiel de le tester et d'évaluer sa performance sur un jeu de données inconnu du modèle dont on connaît les valeurs à prédire.

Méthodes d’ échantillonnage

Partitionner en ensemble de données d'entraînement et de test

Ensemble de données d'entraînement = "*training set*"

- Données utilisées pour entraîner l'algorithme d'apprentissage automatique.

Ensemble de données test = "*testing set*"

- Données utilisées pour tester la performance de l'algorithme d'apprentissage automatique.

IMPORTANT

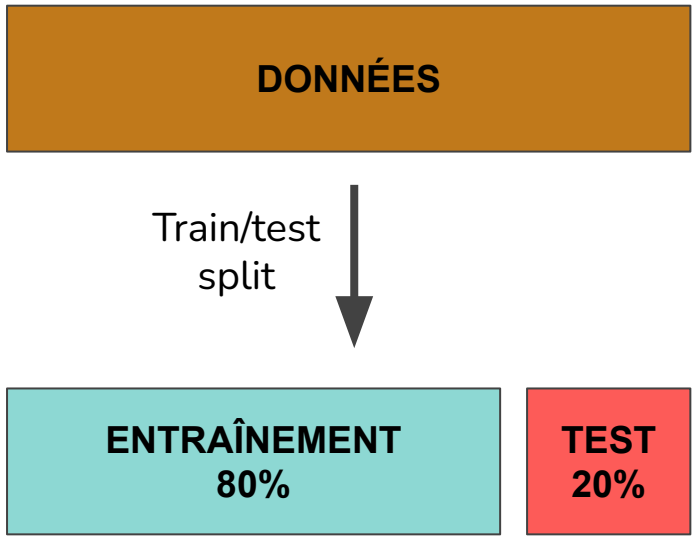
La **séparation** du jeu de données en ensemble d'entraînement et de test doit se faire de façon **aléatoire**.

Comment séparer les données d'entraînement des données test?

La règle standard du 80-20

La pratique commune est d'utiliser

- **80 %** des données pour l'entraînement
- **20%** des données pour tester.



Pourquoi séparer sous-échantillonner les données en entraînement / test?

Avantages de la méthode entraînement / test :

1. Méthode d'évaluation très rapide à exécuter.
2. L'interprétation des résultats est plus intuitive.
3. Idéale pour les grands jeux de données (plus que 1 000 000 000 de lignes).
4. Facilite et accélère l'entraînement d'algorithmes complexes.

Inconvénients de la méthode entraînement / test :

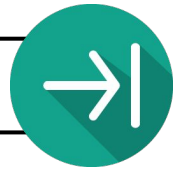
1. La variabilité (variance) entre les ensembles de données d'entraînement et de test peut être grande et causer un manque de précision.

Étapes de la validation croisée

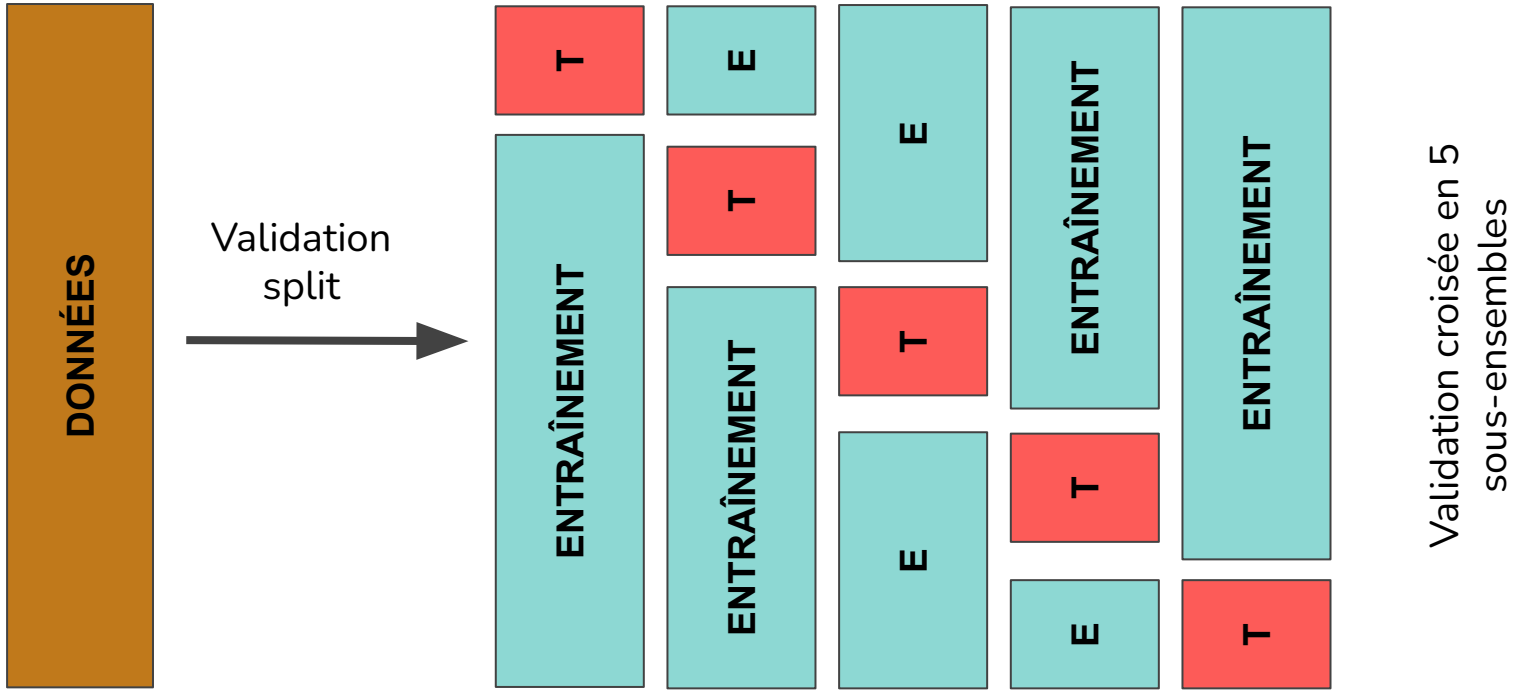
Supposons qu'on veuille faire une validation croisée en 5 ensembles (K)

1. Séparer le jeu de données en 5 ensembles égaux.
2. Utiliser 1 ensemble en tant que données tests et les 4 autres pour l'entraînement.
3. Calculer la précision pour cette première itération.
4. Répéter les étapes 2 et 3 pour toutes les *ensembles (K =5), donc 5 fois au total.*
5. Utiliser la moyenne des 5 itérations comme précision du modèle.

Nous pouvons aussi visualiser ces étapes!



Validation croisée (*"cross validation"*)



Pourquoi utiliser la validation croisée?

Avantages de la validation croisée :

1. Méthode d'évaluation plus précise et moins variable (variance plus faible).
2. Permet de faire des prédictions plus fiables sur des données nouvelles.
3. Réduire la chance d'obtenir une précision dans les extrêmes.
4. C'est la méthode la plus fréquemment utilisée ("*Gold standard*")

Inconvénients de la validation croisée :

1. Le temps d'entraînement peut être très long sur des gros jeux de données.

Validation croisée (*"cross validation"*)