Formation Machine Learning

Introduction au Machine Learning

Bassem Ben Hamed

ENET'Com, Université de Sfax

17 février 2026

Plan du Chapitre

Panorama de l'Intelligence Artificielle

Intelligence Artificielle : Hiérarchie des concepts

Artificial Intelligence

Is the field of study

Machine Learning

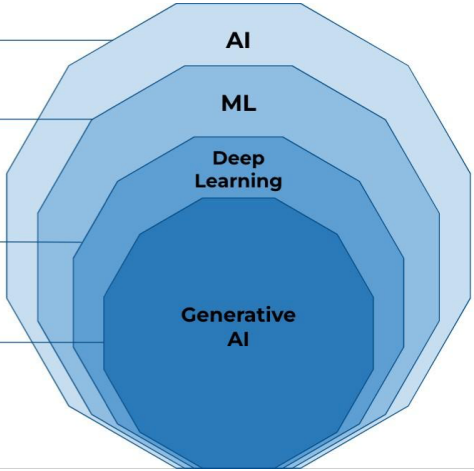
Is a branch of AI that focus on the creation of intelligent machines that learn from data. Another very well know branch inside AI is **Optimization**.

Deep Learning

Is a subset of Machine Learning methods, based on **Artificial Neural Networks.**
Examples: CNNs, RNNs

Generative AI

A type of ANNs that generate data that is similar to the data it was trained on.
Examples: GANs, LLMs



Définitions Formelles

Intelligence Artificielle (IA)

Discipline visant à créer des systèmes capables d'effectuer des tâches qui nécessiteraient normalement l'intelligence humaine.

Machine Learning (ML)

Sous-domaine de l'IA où les systèmes apprennent à partir de données sans être explicitement programmés. Le ML utilise des algorithmes pour identifier des **motifs** et prendre des **décisions**.

Deep Learning (DL)

Sous-domaine du ML utilisant des réseaux de neurones artificiels profonds (multicouches) pour modéliser des représentations complexes.

Types d'Apprentissage

Apprentissage Supervisé

- Données étiquetées : $(x_i, y_i)_{i=1}^N$
- Objectif : apprendre $f : X \to Y$
- Exemples :
  - Classification
  - Régression

Apprentissage Non Supervisé

- Données non étiquetées : $(x_i)_{i=1}^N$
- Découverte de structure
- Exemples :
  - Clustering
  - Réduction de dimension

Apprentissage Semi-Supervisé

- Combinaison de données étiquetées et non étiquetées
- Utile quand l'étiquetage est coûteux

Apprentissage par Renforcement

- Agent apprenant par interaction
- Système de récompenses
- Applications : jeux, robotique

Notion de Modèle

Composantes d'un Modèle ML

1. **Données** : Ensemble d'observations $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$
2. **Hypothèses** : Espace des fonctions possibles $\mathcal{H}$
3. **Fonction de perte** : $\mathcal{L}(y, \hat{y})$ mesure l'erreur de prédiction
4. **Objectif** : Trouver $f^* \in \mathcal{H}$ qui minimise le risque

Risque Empirique

$$\hat{R}(f) = \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}(y_i, f(x_i))$$

Empirical Risk Minimization (ERM)

Principe Fondamental

Le principe de minimisation du risque empirique consiste à chercher le modèle qui minimise l'erreur sur les données d'entraînement :

$$f^* = \arg\min_{f \in \mathcal{H}} \hat{R}(f) = \arg\min_{f \in \mathcal{H}} \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}(y_i, f(x_i))$$

Exemple : Régression Linéaire

- Hypothèse : $f(\mathrm{x}) = \mathrm{x}^T \boldsymbol{\beta}$
- Perte : $\mathcal{L}(y, \hat{y}) = (y - \hat{y})^2$ (MSE)
- ERM : $\boldsymbol{\beta}^* = \arg\min_{\boldsymbol{\beta}} \frac{1}{N} \sum_{i=1}^{N} (y_i - \mathrm{x}_i^T \boldsymbol{\beta})^2$

Risque Réel vs Risque Empirique

- **Risque réel** : $R(f) = \mathbb{E}_{(X,Y)}[\mathcal{L}(Y, f(X))]$
- **Risque empirique** : $\hat{R}(f) = \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}(y_i, f(x_i))$
- **Objectif** : $\hat{R}(f) \approx R(f)$ pour un bon modèle

Écart de Généralisation

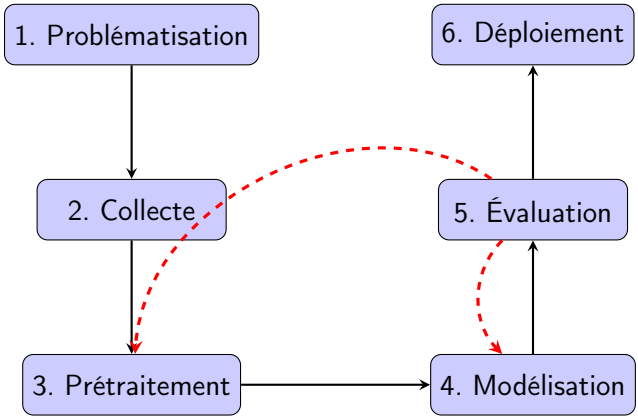
L'écart entre le risque réel et le risque empirique :

$$|R(f) - \hat{R}(f)|$$

Dépend de la **complexité du modèle** et de la **taille de l'échantillon**.

Cycle de vie d'un projet Data Science

Les Étapes d'un Projet Data Science



Itérations

- Fine-tuning hyperparamètres
- Feature engineering
- Sélection de features
- Essai nouveaux modèles
- Retour au preprocessing

1. Problématisation Mathématique

Questions Clés

- Quel est le **problème** à résoudre ?
- S'agit-il de **classification**, **régression**, **clustering**, etc. ?
- Quelles sont les **variables d'entrée** X et de **sortie** Y ?
- Quelle **métrique** d'évaluation utiliser ?

Exemples

- **Classification** : prédire si un email est spam ($y \in \{0, 1\}$)
- **Régression** : prédire le prix d'une maison ($y \in \mathbb{R}$)
- **Clustering** : segmenter les clients (pas de y)

2. Collecte et Compréhension des Données

Sources de Données

- Bases de données internes
- APIs et web scraping
- Fichiers (CSV, JSON, XML, etc.)
- Capteurs et IoT
- Open data

Compréhension des Données

- Analyse exploratoire (EDA - Exploratory Data Analysis)
- Statistiques descriptives : moyenne, médiane, variance
- Visualisations : histogrammes, box plots, scatter plots
- Identification des problèmes : valeurs manquantes, outliers

3. Prétraitement et Nettoyage

Tâches Principales

1. **Gestion des valeurs manquantes**
   - Suppression
   - Imputation (moyenne, médiane, modèle)
2. **Détection et traitement des outliers**
   - Z-score, IQR
   - Transformation robuste
3. **Encodage des variables catégorielles**
   - One-hot encoding
   - Label encoding
   - Target encoding
4. **Normalisation/Standardisation**
   - Z-score : $z = \frac{x-\mu}{\sigma}$
   - Min-Max : $x' = \frac{x-\min}{\max-\min}$

4. Modélisation

Choix du Modèle

Le choix dépend de :

- Type de problème (classification, régression, etc.)
- Nature des données (linéaire, non-linéaire)
- Taille du dataset
- Interprétabilité requise
- Contraintes computationnelles

Exemples d'Algorithmes

- **Régression** : Régression linéaire, Ridge, Lasso, Elastic Net
- **Classification** : Logistic Regression, SVM, Random Forest, XGBoost
- **Clustering** : k-Means, DBSCAN, Hierarchical Clustering
- **Deep Learning** : CNN, RNN, Transformers

5. Évaluation

Métriques - Régression

- **MSE** : $\frac{1}{N}\sum_i (y_i - \hat{y}_i)^2$
- **RMSE** : $\sqrt{\text{MSE}}$
- **MAE** : $\frac{1}{N}\sum_i |y_i - \hat{y}_i|$
- **R^2** : coefficient de détermination

Métriques - Classification

- **Accuracy** : $\frac{TP+TN}{TP+TN+FP+FN}$
- **Precision** : $\frac{TP}{TP+FP}$
- **Recall** : $\frac{TP}{TP+FN}$
- **F1-Score** : $2 \cdot \frac{P \cdot R}{P+R}$
- **AUC-ROC**

Validation Croisée

K-fold cross-validation pour une évaluation robuste

6. Déploiement et Monitoring

Déploiement

- Mise en production du modèle
- Création d'une API (REST, gRPC)
- Containerisation (Docker, Kubernetes)
- Scalabilité et optimisation

Monitoring

- Suivi des performances en production
- Détection de la **dérive des données** (data drift)
- Détection de la **dérive du modèle** (model drift)
- Ré-entraînement périodique
- Logging et alertes

Problèmes Courants en ML

1. Surapprentissage (Overfitting)

- Le modèle apprend le **bruit** au lieu du **signal**
- Performance excellente sur train, mauvaise sur test
- **Solutions** : régularisation, cross-validation, early stopping

2. Sous-apprentissage (Underfitting)

- Le modèle est trop simple
- Mauvaises performances sur train et test
- **Solutions** : modèle plus complexe, plus de features

3. Biais dans les Données

Données non représentatives $\Rightarrow$ prédictions biaisées

Compromis Biais-Variance (1/2)

Biais (Bias)

Erreur due à un modèle trop simple qui ne capture pas les vrais motifs des données.

Formule mathématique :

$$\text{Biais}^2 = \Big(\mathbb{E}[\hat{f}(x)] - f(x)\Big)^2$$

- $\hat{f}(x)$: valeur prédite par le modèle
- $f(x)$: valeur réelle
- $\mathbb{E}[\hat{f}(x)]$: prédiction moyenne sur différents ensembles

Biais élevé $\Rightarrow$ Sous-apprentissage (underfitting)

Variance

Sensibilité du modèle aux fluctuations des données d'entraînement.

Formule mathématique :

$$\text{Variance} = \mathbb{E}\Big[\big(\hat{f}(x) - \mathbb{E}[\hat{f}(x)]\big)^2\Big]$$

- $\hat{f}(x)$: valeur prédite
- $\mathbb{E}[\hat{f}(x)]$: moyenne des prédictions

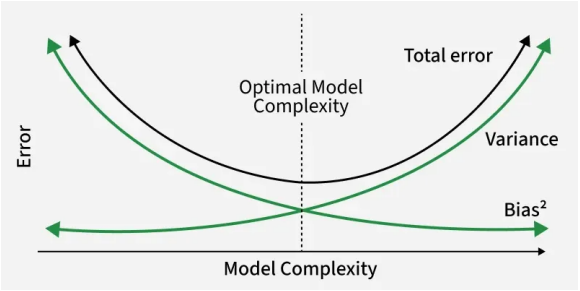
Variance élevée $\Rightarrow$ Sur-apprentissage (overfitting)

Erreur Totale

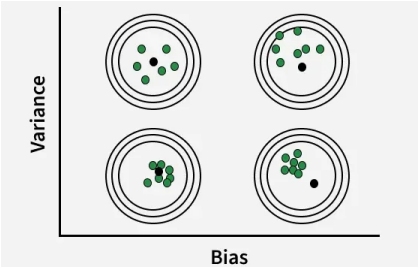
$$\text{Erreur Totale} = \text{Biais}^2 + \text{Variance} + \text{Bruit Irréductible}$$

Compromis Biais-Variance (2/2)

Évolution selon la complexité



Compromis (Trade-off)



Balance Optimale

- **Underfitting** : Biais élevé + Variance faible ⇒ Modèle trop simple
- **Optimal** : Biais modéré + Variance modérée ⇒ Meilleure généralisation
- **Overfitting** : Biais faible + Variance élevée ⇒ Modèle trop complexe

Dérive des Données (Data Drift)

Définition

La distribution des données d'entrée change au fil du temps :

$$P_{\text{train}}(X) \neq P_{\text{production}}(X)$$

Exemples

- Changement de comportement des utilisateurs
- Évolution des tendances du marché
- Modifications des capteurs
- Événements exceptionnels (COVID-19, crise économique)

Détection

Tests statistiques : Kolmogorov-Smirnov, Jensen-Shannon divergence

Résumé du Chapitre

Points Clés

1. L'IA, le ML et le DL forment une hiérarchie de concepts
2. Il existe 4 types d'apprentissage : supervisé, non supervisé, semi-supervisé, par renforcement
3. Un projet Data Science suit un cycle structuré en 6 étapes
4. L'ERM est le principe fondamental de l'apprentissage
5. Le compromis biais-variance est crucial pour la généralisation
6. Le monitoring en production est essentiel (drift detection)

Programme de la Formation

- Regression Linéaire
- Regression Logistique
- Arbre de Décision
- Méthodes Ensembliste (Bagging, Boosting, Stacking)
- Machine à Vecteurs de Support
- Méthodes de Sélection des Caractéristiques
- Réduction de Dimensionnalité
- K-Means, Regroupement Hiérarchique, Modèle de Mélanges Gaussiens (GMM)

Merci pour votre attention !

Questions ?

Bassem Ben Hamed
ENET'Com, Université de Sfax