Chapitre 1 : Panorama du Machine Learning

23 février 2026

Plan du chapitre

Introduction au Machine Learning

- **Branche de l'IA :** Apprentissage statistique (premiers travaux dans les années 70).
- **Premières applications historiques :**
  - Reconnaissance optique de caractères (OCR) dans les années 50.
  - Filtres anti-spam dans les années 90.
- **Omniprésence actuelle :** Reconnaissance vocale, d'images, systèmes de recommandation...

Définitions fondamentales

Arthur Samuel (1959)

« Apprendre des données sans programmation explicite »

Tom Mitchell (1997)

« Une machine apprend d'une expérience E au sujet d'une tâche T si sa performance P sur T augmente avec E. »

Le principe : Transformer des expériences (les données) en une expertise (l'algorithme/modèle) pour prédire ou classer.

Exemples d'application classiques

- **Filtre anti-spam :** Apprentissage à partir de données taguées, plus facile à maintenir qu'une longue liste de règles manuelles.
- **Problèmes trop complexes (sans solution classique) :**
  - Exemple : Reconnaissance vocale à grande échelle. L'approche statistique surpasse l'analyse des caractéristiques vocales pures.
- **Fouille de données (Data Mining) :** Découvrir des relations cachées dans d'immenses volumes de données.

Autres applications courantes

- Classement d'images par catégories.
- Diagnostic médical (tumeurs via IRM).
- Classification d'articles de presse.
- Prédictions de prix (ex : immobilier).
- Détection d'anomalies (séquences temporelles, sécurité).
- Segmentation client.

Supervision de l'apprentissage (1/2)

1. Apprentissage Supervisé

Apprentissage à partir d'exemples **étiquetés** par un instructeur.

- But : Généraliser à de nouveaux cas (ex : spam).
- Modèles : k-NN, Régressions, SVM, Arbres, Réseaux de neurones.

2. Apprentissage Non Supervisé

Apprentissage sur des données **sans étiquette**.

- But : Trouver une structure, grouper (Clustering), réduire la dimension, PCA.

Supervision de l'apprentissage (2/2)

3. Apprentissage Semi-supervisé

Mélange des deux : grouper (non supervisé) puis étiqueter un groupe entier avec une seule étiquette (ex : tri de photos).

4. Apprentissage par Renforcement

L'agent observe, agit et reçoit une **récompense** ou une pénalité.

- But : Maximiser la récompense à terme (ex : Jeu d'échecs).

Traitement des données et Modélisation

Hors ligne vs En ligne :

- **Hors ligne (Batch) :** Traitement de toutes les données d'un coup. Exigeant en ressources, lent à mettre à jour.
- **En ligne :** Traitement en flux (ou mini-batches). Idéal pour les données rapides. Risque lié à la mauvaise qualité soudaine des données (nécessite un monitoring).

Basé sur les cas vs Basé sur un modèle :

- **Cas :** Mémorise et compare (mesure de similarité).
- **Modèle :** Construit un modèle global, l'entraîne, puis infère.

Les deux grands pièges du ML

1. Les mauvaises données

- Volume insuffisant ou non représentatif.
- Mauvaise qualité (bruit, données aberrantes, incomplètes).
- Caractéristiques (*features*) non pertinentes.
- **Conséquence :** Le nettoyage des données occupe la majorité du temps du Data Scientist.

2. Le mauvais algorithme

- **Surapprentissage (Overfitting) :** Généralisation abusive, apprentissage "par cœur".
- **Sous-apprentissage (Underfitting) :** Modèle trop simple pour capter la complexité du problème.

De la déduction à l'induction

- La mémorisation pure empêche la généralisation.
- Il faut choisir les **bonnes caractéristiques** pour induire une règle.
- **Le piège des corrélations illusoires :**
  - *Skinner (1947)* : Les pigeons superstitieux associent un jouet aléatoire à la nourriture.
  - *Garcia (1996)* : Les rats ignorent la sonnette comme indicateur de poison (caractéristique non pertinente pour eux).

Règle d'or de l'information

Moins on a d'information *a priori*, plus il faut de données. Avec des règles *a priori* strictes (ex : systèmes experts), les données ne sont même plus nécessaires !

Besoins et Liens interdisciplinaires

Pourquoi a-t-on besoin du ML ?

- **Complexité :** Impossible d'écrire les règles à la main (ex : conduite autonome).
- **Volumes :** Dépasse les capacités humaines (ex : génétique).
- **Adaptivité :** S'ajuste aux changements d'environnement.

Disciplines connectées :

- **IA :** Le ML est une sous-catégorie.
- **Statistiques :** Mais le ML travaille sur des échantillons finis et sans connaître la loi de distribution initiale.
- **Mais aussi :** Algorithmique, Algèbre linéaire, Optimisation.

Méthodes Classiques vs Deep Learning (1/2)

Bien que le Deep Learning domine la perception (vision, audio, NLP), les méthodes classiques restent reines en entreprise :

- **Données tabulaires :** XGBoost, LightGBM ou Random Forest gèrent mieux les données hétérogènes que les réseaux de neurones.
- **Small Data :** Avec peu de données, le Deep Learning sur-apprend. SVM ou Régressions régularisées sont plus robustes.

Méthodes Classiques vs Deep Learning (2/2)

- **Explicabilité :** Banque, santé, justice exigent des modèles transparents (Arbres de décision, Régression logistique).
- **Edge AI (Ressources limitées) :** IoT et embarqué nécessitent des modèles frugaux en énergie (SVM, Random Forest).
- **Cas spécifiques :** Séries temporelles simples (ARIMA) ou détection d'anomalies (Isolation Forest).

Le Rasoir d'Ockham

Toujours commencer par une "baseline" simple. Si une Random Forest donne 95% en 10 secondes, inutile de configurer un réseau de neurones complexe pour 95,2%.