# Machine Leraning

*27.11.2024*

*Session de formation*

Expert : Andrei Kucharavy
Mémoire : Zotrim Uka
Présent : David Guillaume, Dasek Joiakim, Fairon Hugo, Uka Zotrim, Cardoso Rafael, Laurent Térence

# Table des matières

# *Introduction*

La formation dispensée par Andrei, doctorant en Machine Learning, a couvert les bases fondamentales de ce domaine, en partant des concepts historiques jusqu'à leurs applications modernes. L'objectif principal était de fournir une compréhension approfondie des approches, méthodes et défis liés à l'apprentissage automatique, tout en abordant les problématiques concrètes que posent ces technologies.

# *Les bases historiques et théoriques du Machine Learning*

Le Machine Learning trouve ses origines dans les travaux d'Alan Turing en 1945. Il avait anticipé que, pour résoudre des problèmes d'une grande complexité, il deviendrait nécessaire de créer des programmes capables de s'auto-générer. Ces programmes autonomes, aujourd'hui appelés modèles de Machine Learning, prennent leurs décisions à partir de données sans intervention humaine directe. À cette époque, les approches de programmation étaient limitées par la complexité cyclomatique des codes. Cela désigne le nombre de chemins décisionnels que le programme peut suivre lors de son exécution. Une complexité cyclomatique excessive rendait le code difficilement compréhensible et maintenable.

Les langages de programmation ont évolué pour réduire cette complexité en introduisant des abstractions telles que les fonctions et les procédures. Ces dernières permettent de masquer des niveaux élevés de complexité en les encapsulant dans des modules simples. Aujourd'hui, on considère qu'une fonction avec une complexité cyclomatique inférieure à 10 est lisible et maintenable, alors qu'une complexité supérieure à 25 la rend pratiquement inintelligible.

Alan Turing avait également observé que, bien que certains problèmes comme les mathématiques puissent être résolus avec une complexité raisonnable, d'autres plus vastes nécessiteraient des milliers de programmeurs pendant plusieurs années pour explorer toutes les possibilités. Il en conclut que le seul moyen réaliste de progresser dans ces domaines serait de permettre aux programmes d'évoluer et de prendre des décisions par eux-mêmes. C'est l'idée qui sous-tend le Machine Learning.

# *Transition vers le Machine Learning*

Le Machine Learning repose sur des modèles capables d'apprendre à partir des données. Contrairement aux programmes traditionnels écrits par des humains, les modèles de ML ajustent leurs comportements en fonction des informations qu'ils reçoivent. Les premières applications pratiques de cette discipline ont été orientées vers des tâches de classification. La classification consiste à organiser des données en catégories distinctes, en utilisant des règles simples comme des instructions conditionnelles du type « si… alors… sinon ». Bien que la classification semble rudimentaire, elle constitue la base de nombreuses tâches plus complexes.

Le principe clé du Machine Learning est que tout problème peut être réduit à une série de décisions ou classifications. Même les tâches de régression, qui consistent à prédire des valeurs continues, peuvent être interprétées comme des classifications déguisées. Par exemple, la prédiction d'une valeur dans une série temporelle peut être vue comme le choix parmi plusieurs intervalles prédéfinis.

## *Méthodes et algorithmes de Machine Learning*

L'apprentissage automatique se divise en trois grandes catégories : supervisé, non supervisé et semi-supervisé. Ces approches diffèrent selon le niveau d'intervention humaine nécessaire et les types de données utilisées.

### Apprentissage supervisé

Dans ce cadre, le modèle est formé à partir de données étiquetées, c'est-à-dire des paires d'entrée-sortie connues. Cette méthode repose sur un ensemble d'exemples pour entraîner le modèle à prédire correctement les sorties pour de nouvelles entrées. Par exemple, dans un contexte de conduite autonome, un modèle peut être entraîné à distinguer entre un feu rouge et un feu vert.

### Apprentissage non supervisé

Contrairement à l'apprentissage supervisé, cette approche ne repose pas sur des étiquettes explicites. Le modèle cherche plutôt à identifier des motifs ou des clusters dans les données. Les algorithmes de clustering, tels que K-Means, DBSCAN et Watershed, sont couramment utilisés pour regrouper des données similaires en fonction de leur densité ou de leur proximité.

### Apprentissage semi-supervisé

Cette méthode combine des éléments supervisés et non supervisés. Elle est utilisée lorsque seules une partie des données sont étiquetées. Les modèles apprennent à combler les lacunes en utilisant les données non étiquetées, souvent par des techniques telles que le masquage partiel des données d'entraînement pour prédire les valeurs manquantes.

## *Concepts fondamentaux du Machine Learning*

Plusieurs notions clés structurent le domaine du Machine Learning et sont essentielles pour comprendre les modèles et leur fonctionnement.

### Fonction de perte (Loss Function)

La fonction de perte mesure l'écart entre les prédictions du modèle et les résultats réels. Elle sert à évaluer la qualité des décisions prises par le modèle. La descente de gradient, une méthode d'optimisation, permet de minimiser cette fonction en ajustant progressivement les paramètres du modèle.

### Shattering Dimension

Ce concept, également appelé dimension de Vapnik-Chervonenkis (VC), décrit la capacité d'un modèle à séparer les données en fonction des dimensions pertinentes. Plus la dimension est élevée, plus le modèle doit être complexe. Une complexité excessive peut toutefois entraîner des phénomènes d'overfitting, où le modèle devient trop spécifique aux données d'entraînement et échoue à généraliser.

### Embedding

Les embeddings sont des représentations compactes de données complexes, telles que des images ou du texte, dans des espaces de dimensions réduites. Ces transformations permettent aux modèles d'identifier plus facilement les motifs significatifs dans les données.

## *Applications pratiques*

La formation a exploré plusieurs applications du Machine Learning, illustrant comment les concepts théoriques se traduisent en outils pratiques :

- Réseaux de neurones : Ces modèles, inspirés de la biologie, sont capables d'apprendre des fonctions complexes et de généraliser des relations non linéaires dans les données. Ils constituent la base du deep learning, où des couches multiples de neurones traitent les données de manière hiérarchique.

- Clustering : Les algorithmes tels que K-Means, DBSCAN ou Watershed sont utilisés pour regrouper des données similaires. Par exemple, ils permettent d'identifier des comportements différents dans des séries temporelles.

- Analyse de séries temporelles : Cette technique est particulièrement utile pour des tâches prédictives, telles que la prévision de ventes ou d'événements météorologiques.

## *Limitations et défis*

Un des défis majeurs du Machine Learning est de trouver un équilibre entre complexité et généralisabilité. Les modèles trop simples ne capturent pas suffisamment les nuances des données, tandis que les modèles trop complexes risquent de mémoriser les données d'entraînement sans être capables de généraliser.

L'explicabilité des modèles, en particulier des réseaux de neurones profonds, reste une problématique ouverte. Bien que des approches telles que les matrices d'attention permettent de visualiser certains aspects des décisions prises, il est souvent difficile de comprendre entièrement le fonctionnement interne des modèles.

## *Conclusion*

La formation a offert une vue d'ensemble complète des concepts, algorithmes et défis du Machine Learning. Elle a également insisté sur l'importance de bien comprendre les données et leur contexte avant d'appliquer un modèle. Les outils modernes, comme la bibliothèque Scikit-Learn, permettent d'expérimenter facilement avec différents algorithmes sur des données simples, facilitant ainsi l'apprentissage et la mise en pratique.

Ce résumé met en lumière l'évolution du domaine, tout en offrant une base solide pour explorer des applications spécifiques et avancer vers des projets plus complexes.