

Apprentissage Automatique

l'apprentissage automatique

- La définition de l'apprentissage automatique en anglais machine learning, apprentissage artificiel ou apprentissage statistique est un champ d'étude de l'intelligence artificielle qui se fonde sur des approches mathématiques et statistiques pour donner aux ordinateurs la capacité d' « apprendre » à partir de données, c'est-à-dire d'améliorer leurs performances à résoudre des tâches sans être explicitement programmés pour chacune. Plus largement, il concerne la conception, l'analyse, l'optimisation, le développement et l'implémentation de telles méthodes. On parle d'apprentissage statistique car l'apprentissage consiste à créer un modèle dont l'erreur statistique moyenne est la plus faible possible. »

- **Exemple 1:**

Supposons que l'on dispose d'une collection d'articles de journaux. Comment identifier des groupes d'articles portant sur un même sujet ?

- **Exemple 2:**

Supposons que l'on dispose d'un certain nombre d'images représentant des chiens, et d'autres représentant des chats. Comment classer automatiquement une nouvelle image dans une des catégories « chien » ou « chat » ?

- **Exemple 3:**

Supposons que l'on dispose d'une base de données regroupant les caractéristiques de logements dans une ville : superficie, quartier, étage, prix, année de construction, nombre d'occupants, montant des frais de chauffage. Comment prédire la facture de chauffage à partir des autres caractéristiques pour un logement qui n'appartiendrait pas à cette base ?

Trois grandes approches relèvent de l'apprentissage automatique : l'apprentissage supervisé, l'apprentissage non-supervisé, et l'apprentissage par renforcement. Bien entendu,

cette classification est sujette à discussion, l'apprentissage semi-supervisé ou l'apprentissage faiblement supervisé (par exemple) apparaissant aux interfaces de ces approches. Dans l'exemple 1, on cherche à regrouper les articles portant sur un même sujet, sans disposer d'exemples d'articles dont on sait a priori qu'ils portent sur ce sujet, et sans connaître à l'avance les sujets à identifier. On parlera donc de problème d'apprentissage non-supervisé. Dans les exemples 2 et 3, on cherche à prédire une caractéristique qui est soit une catégorie (exemple 2), soit un montant de facture (exemple 3), à partir d'exemples pour lesquels on connaît la valeur de cette caractéristique. Il s'agit de problèmes d'apprentissage supervisé.

Apprentissage non-supervisé

- L'apprentissage non-supervisé (unsupervised learning) traite des données non-étiquetées.
- L'objectif est d'identifier automatiquement des caractéristiques communes aux observations.

Apprentissage non-supervisé

- Les méthodes de réduction de dimension, comme l'analyse en composantes principales, ou les méthodes d'estimation de densités de probabilité font partie de l'apprentissage non-supervisé.
- Le problème du partitionnement (clustering), dans lequel l'objectif est d'identifier automatiquement des groupes (ou clusters) d'observations partageant des profils communs. On peut aussi parler de classification non-supervisée. Intuitivement, les observations groupées ensemble doivent être davantage similaires entre elles qu'elles le sont d'observations d'un autre groupe. Pour mesurer la similarité entre observations, il faut disposer d'une distance D entre observations.

- La question qui se pose alors est de savoir comment on identifie les groupes à partir des observations du jeu de données. On comprend également qu'il y a une certaine ambiguïté dans l'objectif du partitionnement : « identifier des groupes d'observations partageant un profil commun ». Un problème important est donc de fixer un seuil sur la distance D permettant de décider ce qu'est la limite à « partager un profil commun ». De la même manière, on peut se demander quel est le nombre de groupes à identifier dans l'ensemble des observations.

Apprentissage supervisé

- L'apprentissage supervisé est une méthode d'apprentissage automatique, caractérisée par la création d'un algorithme qui apprend une fonction prédictive. Ceci est possible grâce à un entraînement à partir d'exemples annotés, qui inclut un groupe de variables d'entrée, accompagnées de leurs variables de sortie respectives. Ce processus d'entraînement est répété jusqu'à l'obtention d'une performance satisfaisante. Lors de chaque itération, la machine crée un certain nombre de règles, reliant les variables d'entrée aux variables de sortie. Ce processus permet au modèle d'apprendre à partir des données et d'appliquer les règles afin de prédire, de façon précise, la valeur de sortie lorsqu'une valeur d'entrée est donnée.

- L'apprentissage supervisé peut être divisé en deux sous-catégories : la classification et la régression.

Régression

- Contrairement à la classification qui prédit une catégorie ou une étiquette, les modèles de régression prédisent une valeur de sortie continue, en fonction de la variable indépendante d'entrée. Cette technique est utilisée lorsque la variable de sortie à prédire doit être une valeur continue, soit par exemple pour les prédictions météorologiques ou pour les tendances de marché. Différents modèles de régression existent et diffèrent dépendant de la relation entre le variable dépendante et indépendante considérées, ainsi que du nombre de variables indépendantes utilisées pour le modèle. Parmi les principaux modèles de regression, on trouve la régression linéaire simple et multiple, la régression de Poisson et la régression à vecteurs de support (ou Support Vector Regression, SVR).

- Ainsi, l'analyse de régression est une forme de statistique inférentielle, utilisée pour identifier des tendances au sein des données. Il existe différents algorithmes de régression, partageant le même but qui est de déterminer une valeur de sortie continue en fonction d'une variable d'entrée indépendante.

Un cas d'usage de l'apprentissage supervisé

- Les algorithmes d'apprentissage supervisé prennent le relais dans plusieurs processus du secteur bancaire, qui étaient, par avant, manuels. Par exemple, les data scientists et les ingénieurs chez Bank of America ont créé le Predictive Intelligence Analytics Machine (PRIAM), un système d'IA prédictive qui utilise un réseau d'algorithmes d'apprentissage supervisé. PRIAM permet de comprendre la tendance relationnelle entre les offres des marchés financiers et les investisseurs. Ce système permet également de maximiser la demande des transactions des marchés financiers, en prédisant les meilleurs investisseurs potentiels pour une transactions, dépendant des détails de l'offres, de l'historique des transactions, des échanges et des informations du client.