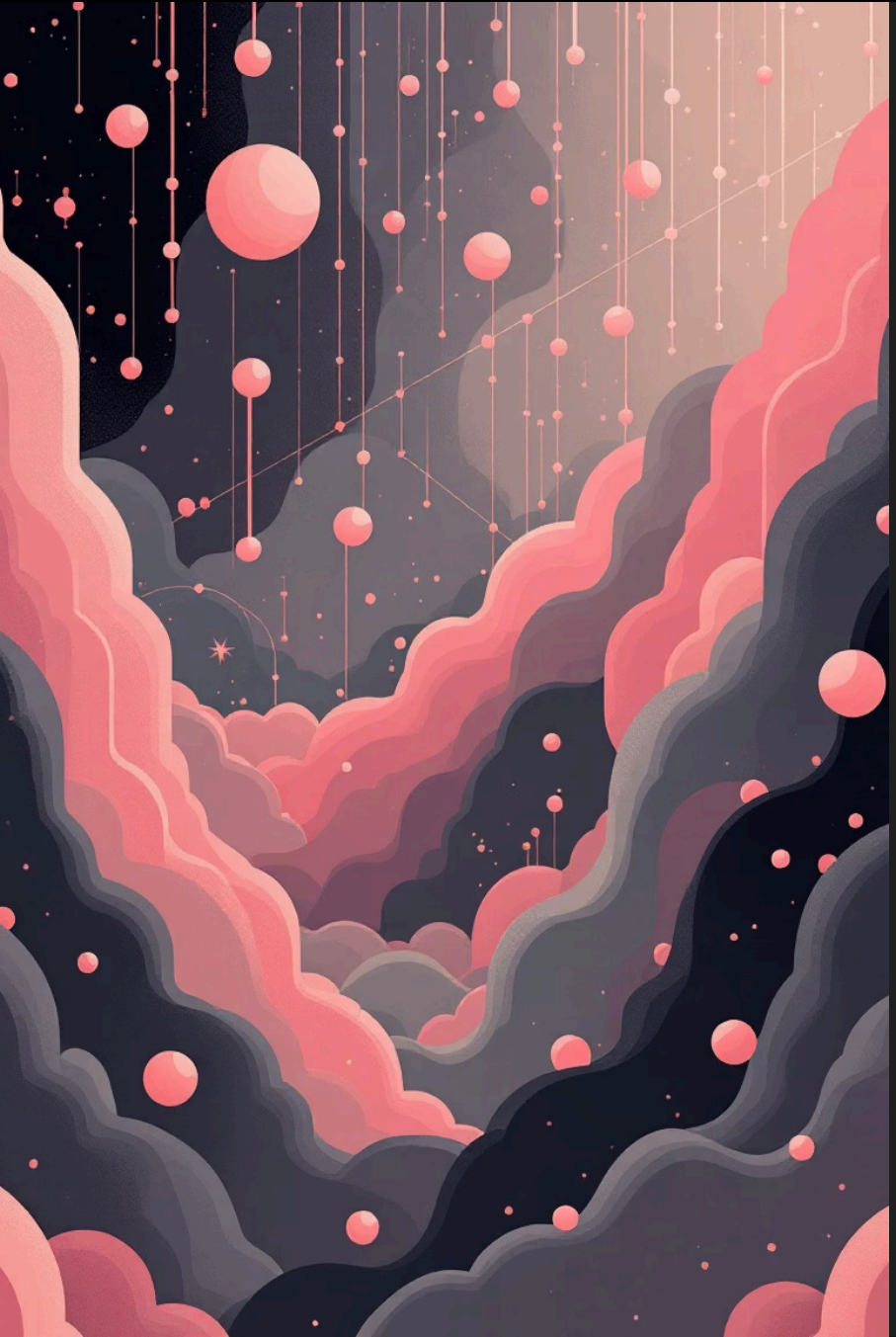


Démystifier l'Apprentissage Automatique : Un Guide pour nos Équipes

Un parcours pédagogique pour comprendre les concepts fondamentaux du Machine Learning et devenir acteur de l'innovation.

Bienvenue dans l'univers du Machine Learning

Cette présentation a été conçue pour vous, équipes produit, ingénierie et commerciales, afin de rendre accessible un domaine qui façonne l'avenir de toute entreprise et de toute industrie.

Les Deux Piliers de l'Apprentissage Automatique

Comprendre la distinction fondamentale entre l'apprentissage supervisé et l'apprentissage non supervisé est la première étape essentielle pour appréhender le domaine du Machine Learning. Ces deux approches répondent à des types de problèmes très différents et utilisent les données de manière distincte.

Apprentissage Supervisé

Apprendre avec des Étiquettes

L'apprentissage supervisé est le type de problème le plus courant. Il se caractérise par l'utilisation d'un jeu de données où nous disposons non seulement de variables d'entrée (appelées features), mais aussi de la variable de sortie que nous cherchons à prédire (appelée cible ou étiquette). Le modèle est "supervisé" car nous lui fournissons des exemples avec les "bonnes réponses" pour qu'il apprenne la relation entre les entrées et la sortie.

Apprentissage Non Supervisé

Découvrir la Structure Cachée

Dans l'apprentissage non supervisé, nous faisons face à un problème où aucune vérité ou étiquette n'est connue à l'avance. L'objectif n'est pas de prédire une valeur spécifique, mais de découvrir des structures, des groupes ou des motifs cachés directement dans les données.

Analogie : Apprentissage Supervisé

C'est comme montrer à un petit enfant des images de chats et de chiens en lui disant "ceci est un chat" et "ceci est un chien". Après avoir vu suffisamment d'exemples étiquetés, vous pouvez lui montrer une nouvelle image et lui demander de quel animal il s'agit.

Exemples concrets :

- Prédire le prix d'une maison en fonction de sa surface, son emplacement et son année de construction
- Catégoriser un objet comme "chat" ou "chien" à partir de ses attributs physiques

Analogie : Apprentissage Non Supervisé

C'est comme donner à un enfant, qui n'a aucune idée de ce que sont les chats et les chiens, une pile d'images d'animaux et lui demander simplement de les regrouper par similitude. L'enfant créera des groupes sans que vous lui ayez donné de nom ou de catégorie au préalable.

Exemple concret :

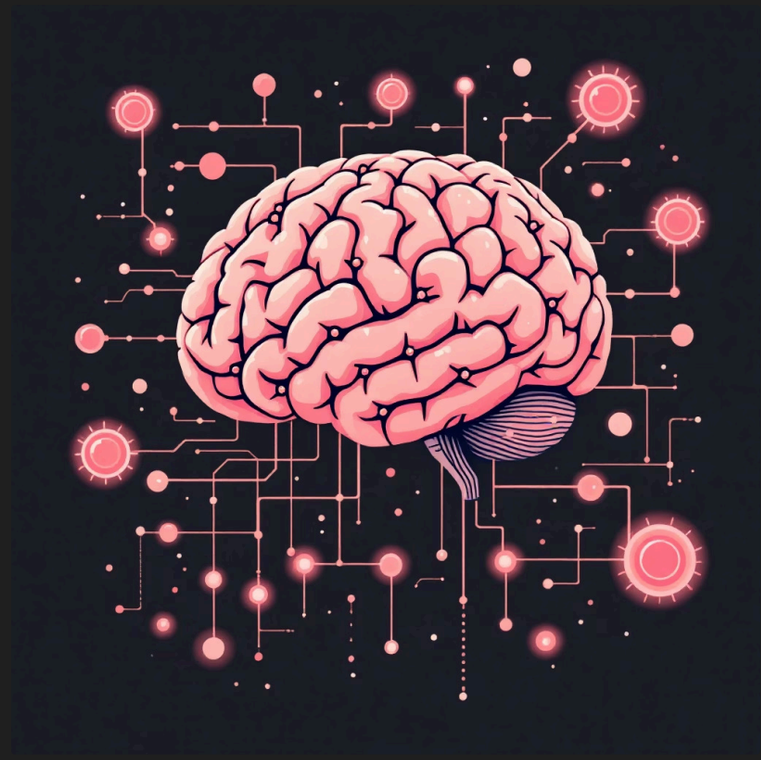
- Trier une grande quantité d'e-mails en trois catégories non spécifiées (clusters) que vous pourrez ensuite inspecter et nommer

Maintenant que cette distinction clé est établie, plongeons dans la branche la plus vaste et la plus importante, l'apprentissage supervisé, pour découvrir la boîte à outils qui permet de résoudre la majorité des problèmes métier.

Qu'est-ce que l'Apprentissage Automatique (Machine Learning) ?

L'apprentissage automatique, ou Machine Learning, est devenu un moteur d'innovation incontournable et une force stratégique dans tous les secteurs. Cependant, la complexité apparente de ses algorithmes peut sembler intimidante. L'objectif de cette présentation est de lever le voile sur ce domaine, de vous fournir une compréhension intuitive des concepts et des algorithmes majeurs, afin que vous ne vous sentiez plus jamais dépassé par le sujet.

Selon la définition formelle, l'apprentissage automatique est « un domaine d'étude de l'intelligence artificielle qui s'intéresse au développement et à l'étude d'algorithmes statistiques pouvant apprendre à partir de données et se généraliser à des données non vues, et ainsi effectuer des tâches sans instructions explicites ». Cette capacité à apprendre de l'expérience, sans être programmé pour chaque scénario, est ce qui le rend si puissant.



Une grande partie des avancées récentes en Intelligence Artificielle est propulsée par les réseaux de neurones. Bien que leur nom puisse paraître complexe, nous verrons que leur fonctionnement est une extension logique et intuitive de principes plus simples que nous aborderons.

Pour bâtir une fondation solide, commençons par explorer les deux grandes familles qui structurent l'ensemble du domaine de l'apprentissage automatique.

Exploration de l'Apprentissage Supervisé : Prédiction et Classification

L'apprentissage supervisé se divise lui-même en deux types de tâches distinctes mais complémentaires : la régression et la classification. Comprendre cette distinction est clé pour choisir le bon outil afin de résoudre un problème métier concret. En termes de stratégie, la régression répond à la question « Combien ? », tandis que la classification répond à la question « Lequel ? » ou « De quel type ? ».

Type de Tâche	Description	Exemple
Régression	L'objectif est de prédire une variable cible numérique et continue. On cherche à modéliser la relation entre les caractéristiques d'entrée et une quantité.	Prédire le prix exact d'une maison (ex: 350 000 €) en fonction de sa superficie.
Classification	L'objectif est d'attribuer une étiquette catégorielle et discrète (une "classe") à une observation. On cherche à séparer les données en groupes définis.	Attribuer l'étiquette "spam" ou "non spam" à un e-mail en fonction de son contenu.

Plongeons maintenant dans les algorithmes les plus fondamentaux qui permettent de réaliser ces tâches.

Régression Linéaire et Régression Logistique

Régression Linéaire : La Base de Tout

Principe Fondamental : C'est la mère de tous les algorithmes d'apprentissage supervisé. Son but est de trouver une relation linéaire (une droite) entre une variable d'entrée et une variable de sortie. Pour ce faire, l'algorithme ajuste une droite qui minimise la distance moyenne entre cette droite et les points de données réels. Cette idée de minimiser l'erreur en ajustant des paramètres est le principe fondamental sur lequel reposent presque tous les algorithmes d'apprentissage supervisé, y compris les réseaux de neurones les plus sophistiqués que nous aborderons plus tard.

Exemple d'Application : Modéliser la relation entre la taille d'une personne et sa pointure. L'algorithme pourrait déterminer que pour chaque augmentation d'une pointure, la taille augmente en moyenne de X centimètres, permettant ainsi de prédire l'un à partir de l'autre.

Régression Logistique : Pour la Classification

Principe Fondamental : Malgré son nom, la régression logistique est un algorithme de classification. C'est une variante de la régression linéaire adaptée aux problèmes où la sortie est une catégorie. Au lieu d'ajuster une droite, elle ajuste une courbe en "S" (la fonction sigmoïde) qui prédit la probabilité qu'une observation appartienne à une classe. Par exemple, le résultat ne sera pas "homme" ou "femme", mais plutôt "probabilité de 80% d'être un homme".

Exemple d'Application : Prédire le genre d'une personne (Homme/Femme) en se basant sur sa taille et son poids.

KNN et Support Vector Machine

K-Nearest Neighbors (KNN)

Le Pouvoir de la Proximité

Principe Fondamental : Cet algorithme est extrêmement simple et intuitif. Il est dit "non-paramétrique" car il n'essaie pas de modéliser une équation mathématique. Pour prédire la valeur d'un nouveau point, il regarde simplement ses 'K' plus proches voisins dans les données d'entraînement. La prédiction est alors la moyenne de ces voisins (pour la régression) ou la classe majoritaire parmi eux (pour la classification).

Mise en Garde : Le choix de 'K' (un "hyperparamètre") est crucial. Un 'K' trop petit rend le modèle très sensible au bruit et risque le sur-apprentissage (overfitting), où il performe parfaitement sur les données connues mais mal sur les nouvelles. Un 'K' trop grand risque le sous-apprentissage (underfitting), où le modèle est trop simple et peu performant.

Support Vector Machine (SVM)

Trouver la Meilleure Frontière

Principe Fondamental : Le SVM est un algorithme de classification qui cherche à dessiner une frontière de décision entre les classes. Son objectif n'est pas seulement de séparer les classes, mais de le faire en maximisant la marge (l'espace) entre la frontière et les points les plus proches de chaque classe. Les points qui définissent cette marge sont appelés "vecteurs de support". Cette approche le rend particulièrement robuste.

Avantage Clé : La puissance des SVM réside dans les "fonctions noyau" (kernel functions). Ces fonctions permettent de transformer les données pour trouver des frontières de décision complexes et non linéaires, en créant implicitement de nouvelles caractéristiques sans effort manuel.

Arbres de Décision et Méthodes d'Ensemble

La Force du Nombre

Principe Fondamental : Un arbre de décision est une série de questions "oui/non" simples qui partitionnent les données de manière hiérarchique. L'objectif est de créer des "feuilles" à la fin de l'arbre qui soient aussi "pures" que possible, c'est-à-dire qu'elles contiennent majoritairement des observations d'une seule et même classe.

Méthodes d'Ensemble

Un seul arbre est souvent peu performant. La véritable puissance vient de la combinaison de nombreux arbres, une approche appelée "méthodes d'ensemble".

Random Forests (Forêts Aléatoires)

Cette technique (Bagging) entraîne de nombreux arbres sur des sous-ensembles de données différents. La classification finale est décidée par un vote majoritaire de tous les arbres.

Boosted Trees (Arbres Boostés)

Cette technique (Boosting) entraîne les modèles en séquence. Chaque nouvel arbre se concentre sur la correction des erreurs commises par le modèle précédent, créant ainsi un modèle final très performant.

Ces algorithmes constituent la boîte à outils essentielle de l'apprentissage supervisé. Il est maintenant temps de découvrir comment les réseaux de neurones, le "roi de l'IA", portent ces concepts à un niveau supérieur.

Le Roi de l'IA : Les Réseaux de Neurones

Les réseaux de neurones sont si puissants car ils résolvent l'un des plus grands défis du Machine Learning : la création de caractéristiques pertinentes (feature engineering). Plutôt que de demander à un expert de définir manuellement des caractéristiques complexes (comme un indice de masse corporelle à partir du poids et de la taille), les réseaux de neurones les conçoivent de manière implicite et automatique.

01

Couche d'Entrée

Vos données brutes, comme les pixels d'une image ou les caractéristiques d'un client.

02

Couches Cachées

Représentent des caractéristiques latentes et inconnues que le réseau apprend par lui-même. Une première couche pourrait identifier des lignes, une autre des formes, puis des objets complexes.

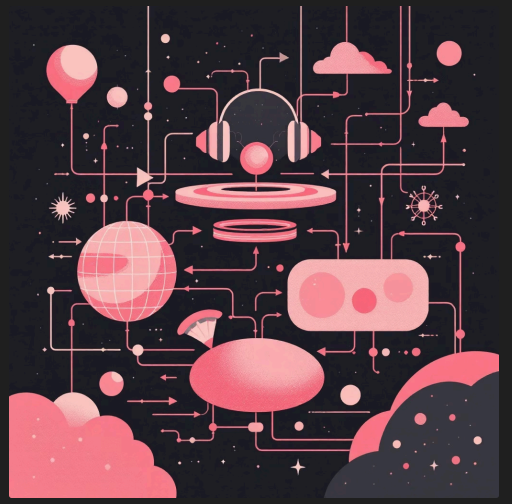
03

Couche de Sortie

La prédiction finale du réseau, qu'il s'agisse d'une catégorie, d'une valeur numérique, ou d'une probabilité.

Pour comprendre leur fonctionnement, revenons à la régression. La forme la plus simple d'un réseau de neurones, appelée "perceptron monocouche", est essentiellement une tâche de régression sur de multiples caractéristiques. C'est en ajoutant des couches supplémentaires entre l'entrée et la sortie que la magie opère.

Le rôle des couches cachées est de représenter des caractéristiques latentes et inconnues que le réseau apprend par lui-même. Par exemple, en analysant des milliers d'images de chiffres manuscrits, une première couche cachée pourrait apprendre à identifier une caractéristique simple comme "présence d'une ligne horizontale", simplement en remarquant que certains pixels adjacents sont souvent allumés ensemble, même si cette caractéristique n'a jamais été définie explicitement par un humain.



Deep Learning (apprentissage profond) désigne simplement l'utilisation de nombreuses couches cachées. Chaque couche successive apprend des caractéristiques de plus en plus complexes.

> En résumé : nous ne savons généralement pas ce que les caractéristiques cachées signifient précisément, mais nous savons qu'en les apprenant, le réseau parvient à réaliser d'excellentes prédictions.

Apprentissage Non Supervisé : Clustering et Réduction de Dimension

Après avoir exploré le monde des prédictions supervisées, tournons-nous vers les situations où aucune vérité n'est connue à l'avance, et où l'objectif est de découvrir une structure cachée dans nos données. L'objectif de l'apprentissage non supervisé est de trouver une structure sous-jacente ou des motifs cachés dans les données lorsque nous ne disposons d'aucune étiquette ou variable cible.

5.1 Clustering : Regrouper les Similitudes

Il est crucial de ne pas confondre le clustering et la classification. La classification est une tâche supervisée : nous connaissons les classes à l'avance et nous avons des données étiquetées pour entraîner le modèle. Le clustering, en revanche, est une tâche non supervisée : nous n'avons pas d'étiquettes et nous cherchons à découvrir des groupes naturels et inconnus dans les données.

5.2 Réduction de Dimension : Simplifier sans Perdre l'Essentiel

L'objectif de la réduction de dimension est de réduire le nombre de caractéristiques (les "dimensions") d'un jeu de données tout en conservant le plus d'informations pertinentes possible. Cette technique permet de rendre les algorithmes de Machine Learning plus efficaces et plus robustes.

Algorithme K-Means

L'algorithme de clustering le plus célèbre est le K-Means. Le "K" dans son nom est un hyperparamètre que l'on doit choisir pour définir le nombre de clusters que l'on souhaite trouver.

Fonctionnement :

1. **Initialisation :** Sélectionner aléatoirement 'K' points qui serviront de centres de clusters initiaux
2. **Assignation :** Assigner chaque point de données au centre de cluster le plus proche
3. **Mise à jour :** Recalculer la position des centres des clusters en faisant la moyenne de tous les points qui leur sont assignés
4. **Répétition :** Répéter les étapes 2 et 3 jusqu'à stabilisation

Analyse en Composantes Principales (PCA)

L'algorithme le plus connu pour la réduction de dimension est l'Analyse en Composantes Principales (ACP ou PCA). Le PCA identifie les directions de plus grande variance dans les données (les "composantes principales"). Ces composantes sont de nouvelles caractéristiques synthétiques qui capturent l'essentiel de l'information.

Exemple : Imaginons que nous analysons des poissons avec de nombreuses caractéristiques. Il est probable que les caractéristiques "hauteur" et "longueur" soient fortement corrélées. Les inclure toutes les deux est redondant. Nous pourrions les combiner en une seule nouvelle caractéristique appelée "forme", réduisant ainsi le nombre de dimensions sans perdre d'informations essentielles.

La réduction de dimension est très utile car elle permet de rendre les algorithmes de Machine Learning plus efficaces et plus robustes en supprimant les dimensions redondantes ou bruyantes qui pourraient nuire à leur performance.