IA & Machine Learning

De la théorie à la pratique

Abdelkarim MOUSSAID

Plan de cours

- Introduction à l’IA et au Machine Learning
- Concepts de base en Machine Learning
- Approche pratique : Premiers pas avec Python
- Perspectives et éthique en IA
- Projet NLP (Natural Language Processing)

Introduction à l’IA et au Machine Learning - Bref historique

- Alan Turing (1912-1954) mathématicien et cryptologue britannique
- Il a joué un rôle très important dans le décryptage des messages de la machine Enigma.
- 1950 : Il a posé les bases de l'apprentissage automatique « L'ordinateur et l'intelligence »
- Test de Turing : la faculté d'une machine à imiter la conversation humaine.
- 1943 : le neurophysiologiste Warren McCulloch et le mathématicien Walter Pitts publient un article décrivant le fonctionnement des neurones en utilisant des circuits électroniques. Ceci est la base de la théorie des réseaux de neurones.

Introduction à l'IA et au Machine Learning - Bref historique

- 1952 : Le terme « Machine Learning » est utilisé pour la première fois par l'informaticien américain Arthur Samuel

- 2012 : Google a développé un réseau de neurones qui a réussi à reconnaître des visages humains ainsi que des chats dans des vidéos YouTube.

- 2014 : le dialogueur Eugene Goostman a réussi a faire croire à 33% des juges humains qu'il parlaient à un enfant de 13 ans.

- 2015 : AlphaGo (google) gagne contre un des meilleurs joueurs au jeu de Go

- 2016 : Une IA nommée LipNet parvient à lire sur les lèvres avec un grand taux de succès

Introduction à l’IA et au Machine Learning - Bref historique

- La popularisation d’internet au début des années 2000 est un élément clé.
- L’utilisation d’internet permet la collecte de beaucoup de données
- Les prouesses technologiques donnent accès à des ordinateurs/serveurs avec beaucoup de puissance de calcul et de stockage.

Introduction à l’IA et au Machine Learning - IA vs ML

IA fait référence à l'utilisation de technologies pour créer des machines et des ordinateurs capables d'imiter des fonctions cognitives associées à l'intelligence humaine

- Visualisation
- Comprendre un langage parlé
- Écrire et répondre
- Analyser des données
- Donner des recommandations

Introduction à l’IA et au Machine Learning - IA vs ML

- ML est un sous-ensemble de l’IA qui permet à un système d’apprendre et de s’améliorer automatiquement.
- ML utilise des algorithmes pour analyser les données, extraire des information pour prendre des décision
- ML apprend de manière continue des données qu’il analyse.
- L'IA est le concept plus large qui permet à une machine ou à un système de détecter, de raisonner, d'agir ou de s'adapter comme un humain
- Le ML est une application d'IA qui permet aux machines d'extraire des connaissances à partir de données et d'en tirer des enseignements de manière autonome.

Introduction à l’IA et au Machine Learning - IA vs ML

Intelligence artificielle

- L'IA permet à une machine de simuler l'intelligence humaine pour résoudre des problèmes.
- L'objectif est de développer un système intelligent capable d'effectuer des tâches complexes.
- Nous créons des systèmes capables de réaliser des tâches complexes comme un humain
- L'IA couvre un large éventail d'applications

Introduction à l’IA et au Machine Learning - IA vs ML

Intelligence artificielle

- L'IA utilise des technologies dans un système de manière à imiter la prise de décision humaine
- L'IA est compatible avec tous les types de données: structurées, semi-structurées et non structurées.
- Les systèmes d'IA s'appuient sur une logique et des arbres de décision pour apprendre, raisonner et se corriger

Introduction à l’IA et au Machine Learning - IA vs ML

Machine learning

- Le ML permet à une machine d'apprendre de manière autonome à partir de données passées
- L'objectif est de créer des machines capables d'exploiter les données pour améliorer la précision du résultat.
- Nous entraînons des machines avec des données à exécuter des tâches spécifiques et à obtenir des résultats précis

Introduction à l’IA et au Machine Learning - IA vs ML

Machine learning

- Le champ d'application des applications de machine learning est limité
- Le ML génère des modèles prédictifs à l'aide d'algorithmes d'auto-apprentissage
- Le ML ne peut utiliser que des données structurées et semi-structurées
- Les systèmes de ML s'appuient sur des modèles statistiques pour apprendre et peuvent corriger automatiquement les nouvelles données

Les données

Tout ce qui peut être stocké numériquement peut être utilisé comme données pour le Machine Learning

Les données - Les sources

- Données en temps réel
- Données sociales
- Données transactionnelles
- Données démographiques
- Données météorologiques

Les données - Les sources

- Données de capteurs
- Données de géolocalisation
- Données de surveillance
- Données de recherche
- Données gouvernementales

Les données - Les méthodes de collecte

- Capteurs et appareils connectés :
- Réseaux sociaux
- Logs de serveur
- Texte et documents
- Données géospatiales
- Systèmes biométriques
- Web Scraping

Les données - La collecte

- API
- Les API permettent aux données de circuler entre différents systèmes
- Les API permettent les automatisations et création de pipeline.
- Il est possible de combiner les capacités de différentes API en une seule application

Les données - Le traitement

- Nettoyage et structuration
- Détection des valeurs aberrantes
- Correction d’erreurs (format, donnée manquante, ...)
- Gestion des variables déséquilibrées
- Extraction de caractéristiques, pattern
- Normalisation des données

Les données

Exemple de collecte et traitement de données

- Liste des villes de France
- Liste des départements
- Liste des régions
- Coordonnées GPS
- Population

- python
- pandas
- seaborn
- jupyter notebook
- …

Concepts de base en Machine Learning

- apprentissage automatique
- apprentissage machine
- apprentissage artificiel
- apprentissage statistique

- approches mathématiques et statistiques
- donner aux ordinateurs la capacité d'« apprendre » à partir de données,
- améliorer leurs performances
- résoudre des tâches

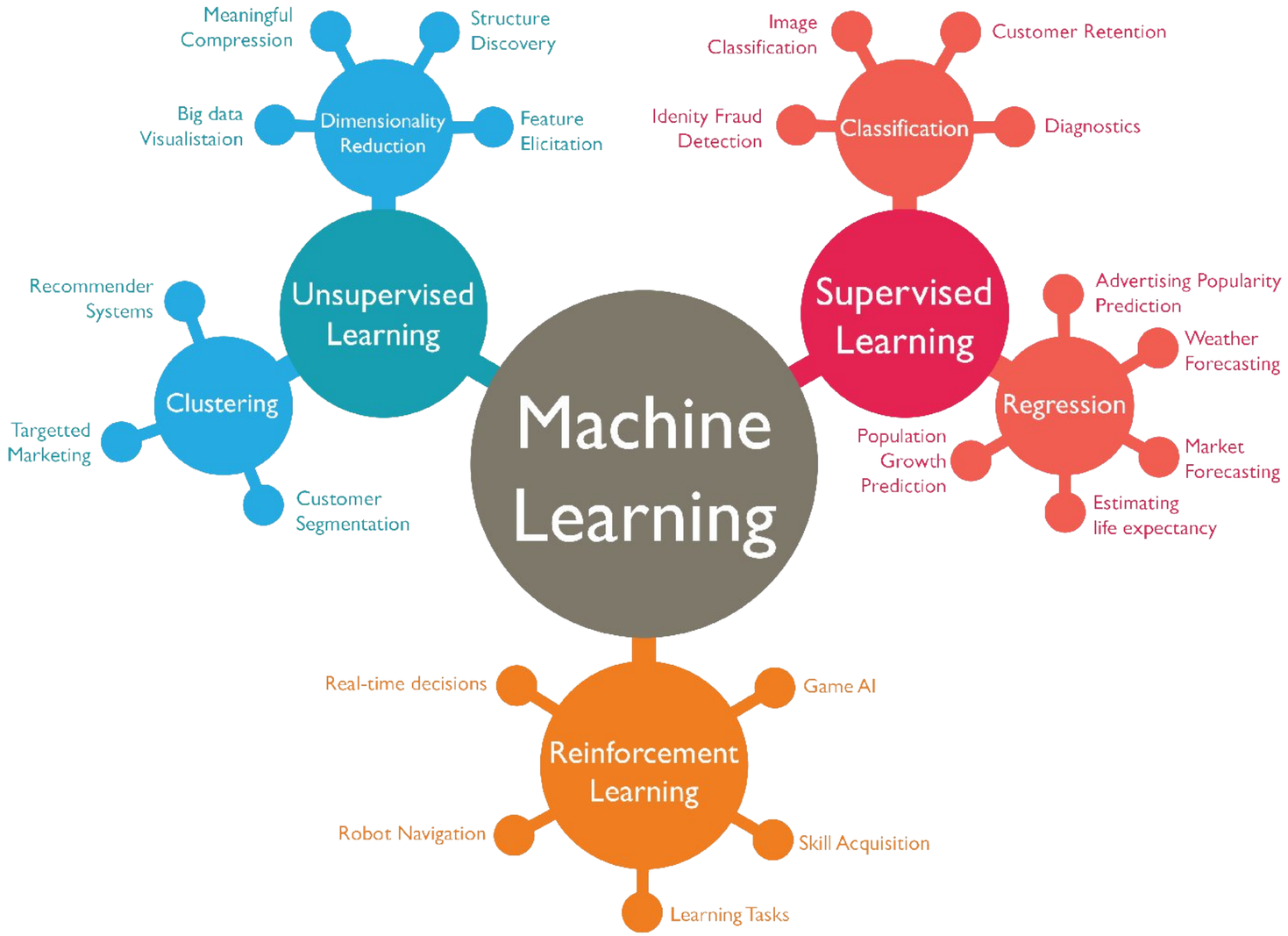
Concepts de base en Machine Learning - 4 étapes

- Sélectionner et à préparer un ensemble de données d’entraînement
- Sélectionner un algorithme : Le type d’algorithme à utiliser dépend du type et du volume de données d’entraînement et du type de problème à résoudre
- Entraînement de l’algorithme
- Utilisation et amélioration du modèle

Concepts de base en Machine Learning - Types d’apprentissage

- Apprentissage supervisé : données étiquetées.
- Apprentissage non-supervisé : données non étiquetées. La machine se contente d’explorer les données à la recherche d’éventuelles patterns.
- Apprentissage semi-supervisé : un mélange de l’apprentissage supervisé et non supervisé.
- Apprentissage par renforcement : l’algorithme apprend de ses erreurs pour atteindre un objectif. Il essayera différentes approches.

Concepts de base en Machine Learning - Types d’apprentissage



Concepts de base en Machine Learning - Types d’apprentissage

Quelques exemples

Apprentissage	Famille d’algorithmes	Algorithme
Supervisé	Régression	Régression linéaire simple et polynomiale
	Régression + Classification	Régression Logistique
		Random Forest
		Réseaux de neurones
		Support Vector Machine (SVM)
Non supervisé	Clustering	K-Means
	Optimisation	Algorithmes génériques
		Optimisation par colonies de fourmis

Concepts de base en Machine Learning - Variables

- Variables quantitatives : valeurs numériques ordonnées
  - Les variables quantitatives continues : Une variable qui permet de mesurer la température ambiante
  - Les variables quantitatives discrètes : prendre que des valeurs numériques définies dans un ensemble fini ou définies dans un ensemble infini, mais dénombrable
- Les opérations mathématiques ont un sens

Concepts de base en Machine Learning - Variables

- Variables qualitatives : portent des informations exprimées généralement sous forme littérale et non pas numérique
  - Les variables nominales : une variable nominale peut prendre des valeurs définies par un ensemble de noms
  - Les variables ordinales : un mix entre les variables quantitatives discrètes et les variables nominales
- Les opérations mathématiques n’ont pas de sens.

Concepts de base en Machine Learning - Régression linéaire

- La régression linéaire simple concerne l'étude de la variation d'une variable à expliquer en fonction d'une et une seule variable explicative.
- ***n*** observations sous la forme de couples $x_i \in \mathbb{R}, y_i \in \mathbb{R}$
- y_i le résultat correspondant à l'observation x_i

$$F(x_i) = ax_i + a_0 \text{ avec } a \in \mathbb{R},\ a_0 \in \mathbb{R}$$

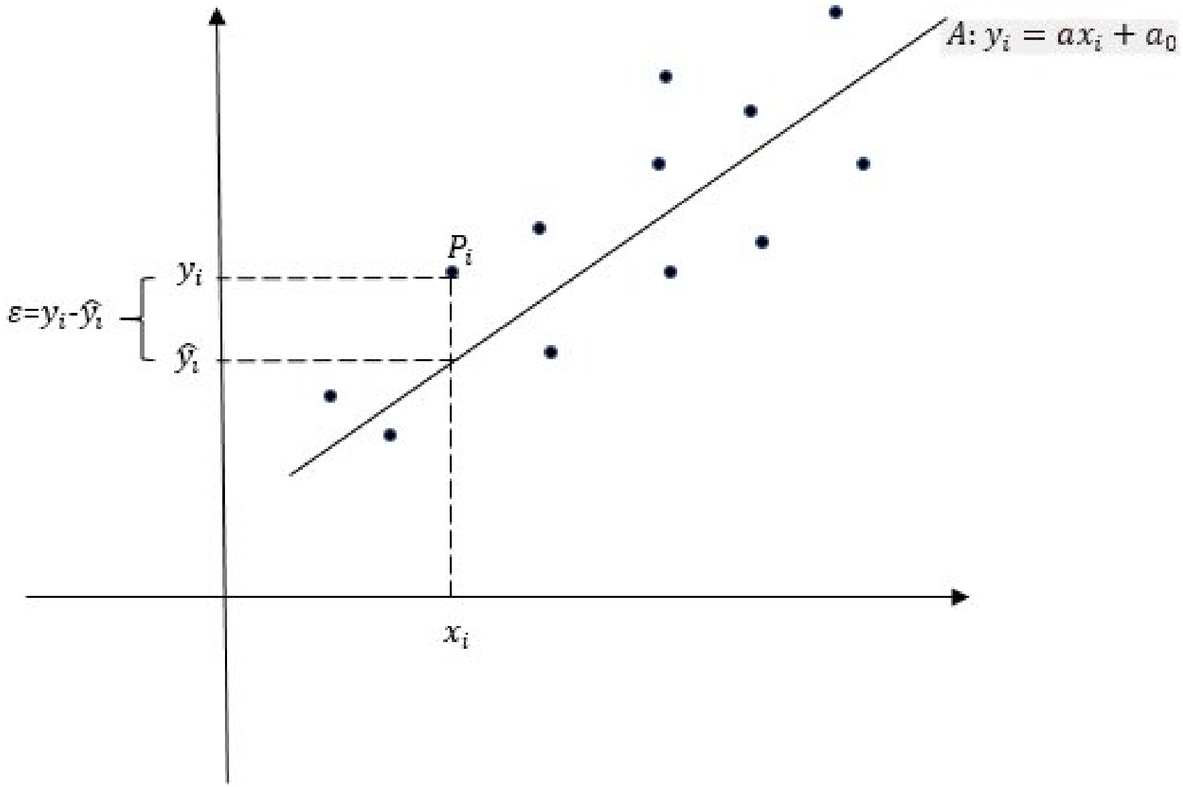
- Le but du jeu est de trouver des bons estimateurs pour les deux paramètres ***a*** et a_0

Concepts de base en Machine Learning - Régression linéaire (géométrique)

Trouver des bons estimateurs de ***a*** et a_0, revient à trouver la droite de régression

$$A: F(x_i) = ax_i + a_0$$

erreur de l'estimation ou le résidu et elle est notée généralement par



Concepts de base en Machine Learning - Régression linéaire (analytique)

- La méthode des moindres carrés permet de rechercher la fonction F, qui minimise la somme des carrés des erreurs

$$E(a, a_0) = \sum_{i=1}^{n} [y_i - F(x_i)]^2$$

- En remplaçant la définition de la fonction F dans l'équation précédente, on obtient :

$$E(a, a_0) = \sum_{i=1}^{n} [y_i - (ax_i + a_0)]^2$$

- Élever au carré les erreurs permet d'éviter que leur somme ne soit pas biaisée par un phénomène de compensation entre les erreurs négatives et les erreurs positives

- Le carré de ces erreurs nous donne des propriétés intéressantes

Concepts de base en Machine Learning - Régression linéaire (analytique)

- La fonction E dépend uniquement des deux paramètres ***a*** et a_0
- Il faut chercher le minimum de la fonction E :
  - Le point qui annule sa dérivée partielle par rapport à a_0

$$\frac{dE}{da_0} = -2 \sum_{i=1}^{n} [y_i - (ax_i + a_0)]$$

  - Le point qui annule sa dérivée partielle par rapport à ***a***

$$\frac{dE}{da} = -2 \sum_{i=1}^{n} [y_i - (ax_i + a_0)]x_i$$

- La fonction E atteint son minimum aux points d'annulation des deux équations

Concepts de base en Machine Learning - Régression linéaire (analytique)

$$\frac{dE}{da_0} = 0 \Leftrightarrow -2 \sum_{i=1}^{n} [y_i - (ax_i + a_0)] = 0$$

$$\Leftrightarrow \sum_{i=1}^{n} [y_i - (ax_i + a_0)] = 0$$

$$\Leftrightarrow \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} (ax_i + a_0) = 0$$

$$\Leftrightarrow \sum_{i=1}^{n} y_i = \sum_{i=1}^{n} (ax_i + a_0)$$

$$\Leftrightarrow \sum_{i=1}^{n} y_i = a \sum_{i=1}^{n} x_i + na_0$$

$$\Leftrightarrow \frac{\sum_{i=1}^{n} y_i}{n} = a \frac{\sum_{i=1}^{n} x_i}{n} + \frac{na_0}{n}$$

$$\frac{\sum_{i=1}^{n} y_i}{n} = \overline{y}$$

$$\frac{\sum_{i=1}^{n} x_i}{n} = \overline{x}$$

$$\overline{y} = a\,\overline{x} + a_0$$

Concepts de base en Machine Learning - Régression linéaire (analytique)

$$\frac{dE}{da} = 0 \Leftrightarrow -2 \sum_{i=1}^{n} [y_i - (ax_i + a_0)]x_i = 0$$

$$\Leftrightarrow \sum_{i=1}^{n} [x_i y_i - (ax_i^2 + a_0 x_i)] = 0$$

$$\Leftrightarrow \sum_{i=1}^{n} x_i y_i = \sum_{i=1}^{n} ax_i^2 + \sum_{i=1}^{n} a_0 x_i \qquad a_0 = \bar{y} - a\,\bar{x}$$

$$\sum_{i=1}^{n} x_i y_i = \sum_{i=1}^{n} ax_i^2 + \sum_{i=1}^{n} (\bar{y} - a\,\bar{x}) x_i$$

$$\Leftrightarrow \sum_{i=1}^{n} x_i y_i = \sum_{i=1}^{n} ax_i^2 + \sum_{i=1}^{n} x_i\,\bar{y} - \sum_{i=1}^{n} ax_i\,\bar{x}$$

$$\Leftrightarrow \sum_{i=1}^{n} x_i y_i = a \sum_{i=1}^{n} x_i^2 + \bar{y} \sum_{i=1}^{n} x_i - a\bar{x} \sum_{i=1}^{n} x_i$$

$$\Leftrightarrow \frac{\sum_{i=1}^{n} x_i y_i}{n} = a \frac{\sum_{i=1}^{n} x_i^2}{n} + \bar{y} \frac{\sum_{i=1}^{n} x_i}{n} - a \frac{\bar{x} \sum_{i=1}^{n} x_i}{n} \qquad \frac{\sum_{i=1}^{n} x_i}{n}$$

$$\frac{\sum_{i=1}^{n} x_i y_i}{n} = a\left(\frac{\sum_{i=1}^{n} x_i^2}{n} - \bar{x}^2\right) + \bar{x}\,\bar{y}$$

$$\Leftrightarrow a\left(\frac{\sum_{i=1}^{n} x_i^2}{n} - \bar{x}^2\right) = \frac{\sum_{i=1}^{n} x_i y_i}{n} - \bar{x}\,\bar{y}$$

$$\Leftrightarrow a = \frac{COV(X,Y)}{VAR(X)}$$

Concepts de base en Machine Learning - Régression linéaire multiple

- La régression linéaire multiple est une généralisation immédiate de la régression linéaire simple
- La fonction F à estimer ne dépend plus d'une seule variable, mais de plusieurs.
- Si nous avons n couple de la forme $X_i = (x_{i,1}, x_{i,2}, ..., x_{i,m}) \in \mathbb{R}^m$, $y_i \in \mathbb{R}$ et y_i le résultat de $X_i = (x_{i,1}, x_{i,2}, ..., x_{i,m})$, alors la fonction à estimer est :

$$F(x_{i,1}, x_{i,2}, ..., x_{i,m}) = a_1 x_{i,1} + a_2 x_{i,2} + ... + a_m x_{i,m} + a_0$$

- L'objectif est de trouver : $A = (a_1, a_2, ..., a_m, a_0)$

Concepts de base en Machine Learning - Régression linéaire multiple

- La méthode des moindres carrés pour la régression multiple
  - On cherche à minimiser

$$E = \sum_{i=1}^{n} [y_i - F(x_{i,1}, x_{i,2}, \dots, x_{i,m})]^2$$

- Les mêmes étapes peuvent être suivies à la différence qu'ici nous allons avoir (m+1) dérivées partielles au lieu de deux

- En annulant et en simplifiant ces m dérivées partielles, nous obtenons

$$\hat{A} = (X^T X)^{-1} X^T Y$$

  - $\hat{A}$ l'estimateur du vecteur $\boldsymbol{A = (a_1, a_2, \dots, a_m, a_0)}$
  - X la matrice de toutes les observations qui est de la forme :

Concepts de base en Machine Learning - Régression linéaire multiple

- En annulant et en simplifiant ces m dérivées partielles, nous obtenons

$$\hat{A} = (X^T X)^{-1} X^T Y$$

  - $\hat{A}$ l'estimateur du vecteur $\boldsymbol{A = (a_1, a_2, \ldots, a_m, a_0)}$
  - X la matrice de toutes les observations

$$X = \begin{pmatrix} x_{1,1}, x_{1,2} & \cdots & x_{1,m} \\ \vdots \quad \vdots & \ddots & \vdots \\ x_{n,1}, x_{n,2} & \cdots & x_{n,m} \end{pmatrix}$$

  - X^T la matrice transposée de la matrice X
  - Y le vecteur des résultats de toutes les observations de la forme

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ . \\ . \\ . \\ y_n \end{pmatrix}$$

Concepts de base en Machine Learning - Régression linéaire multiple

- On suppose que la matrice X^{TX} est réversible.
- L’utilisation de la méthode des moindres carrés peut s’avérer trop consommatrice en temps de calcul quand les matrices sont trop grandes

Concepts de base en Machine Learning - Régression linéaire multiple

- La méthode de la descente de gradient est adapté lorsque la taille des données devient importante

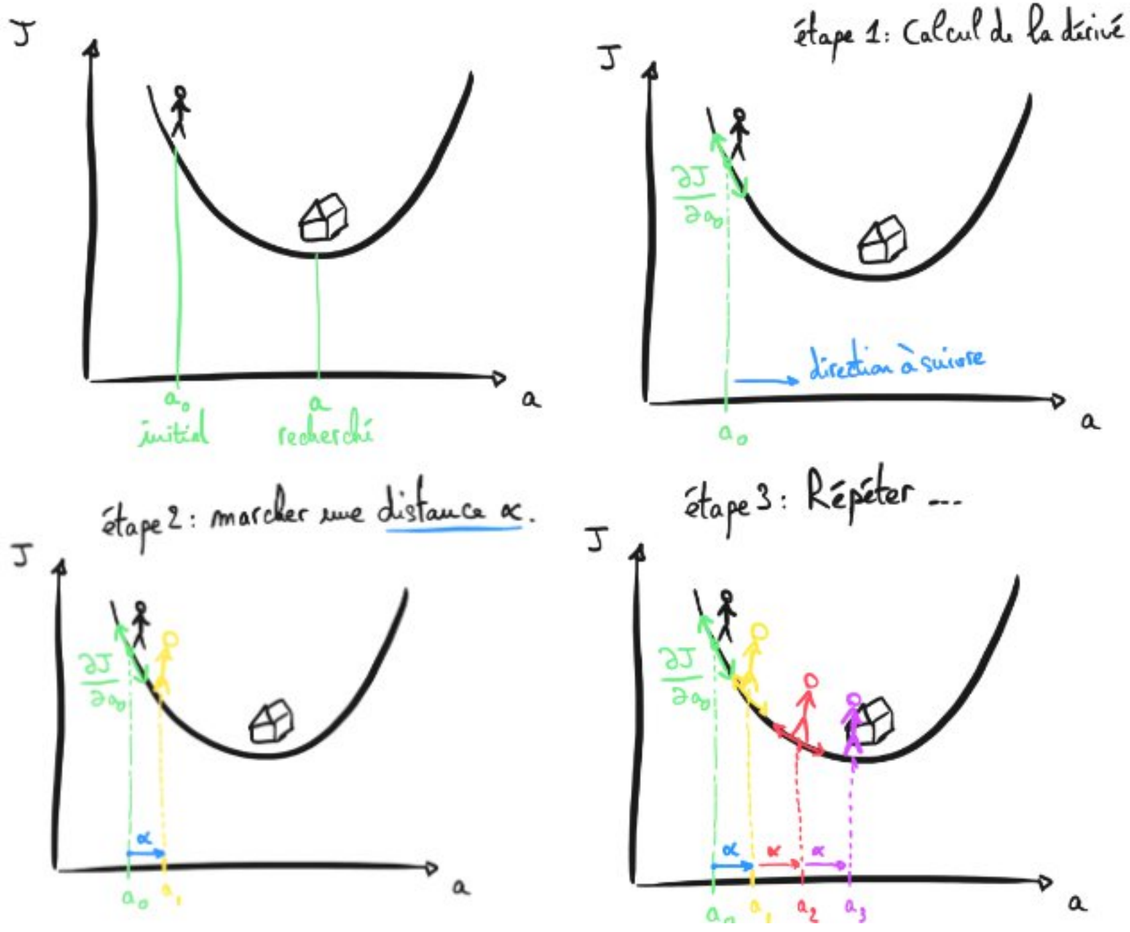
- Pour estimer le vecteur $A = (a_1, a_2, ..., a_m, a_0)$ il faut suivre 3 étapes :
  1. Initialisation aléatoire du vecteur $A = (a_1, a_2, ..., a_m, a_0)$
  2. Pour chaque a_i il faut faire : $a_i = a_i + \alpha \frac{dE}{a_i}$
  3. Si ça ne converge pas, il faut reprendre l'étape 2

- E la fonction d'erreur et α un facteur d'apprentissage

- Plus ce facteur est grand, plus grand est le pas du déplacement de l'algorithme et plus vite la convergence surviendra

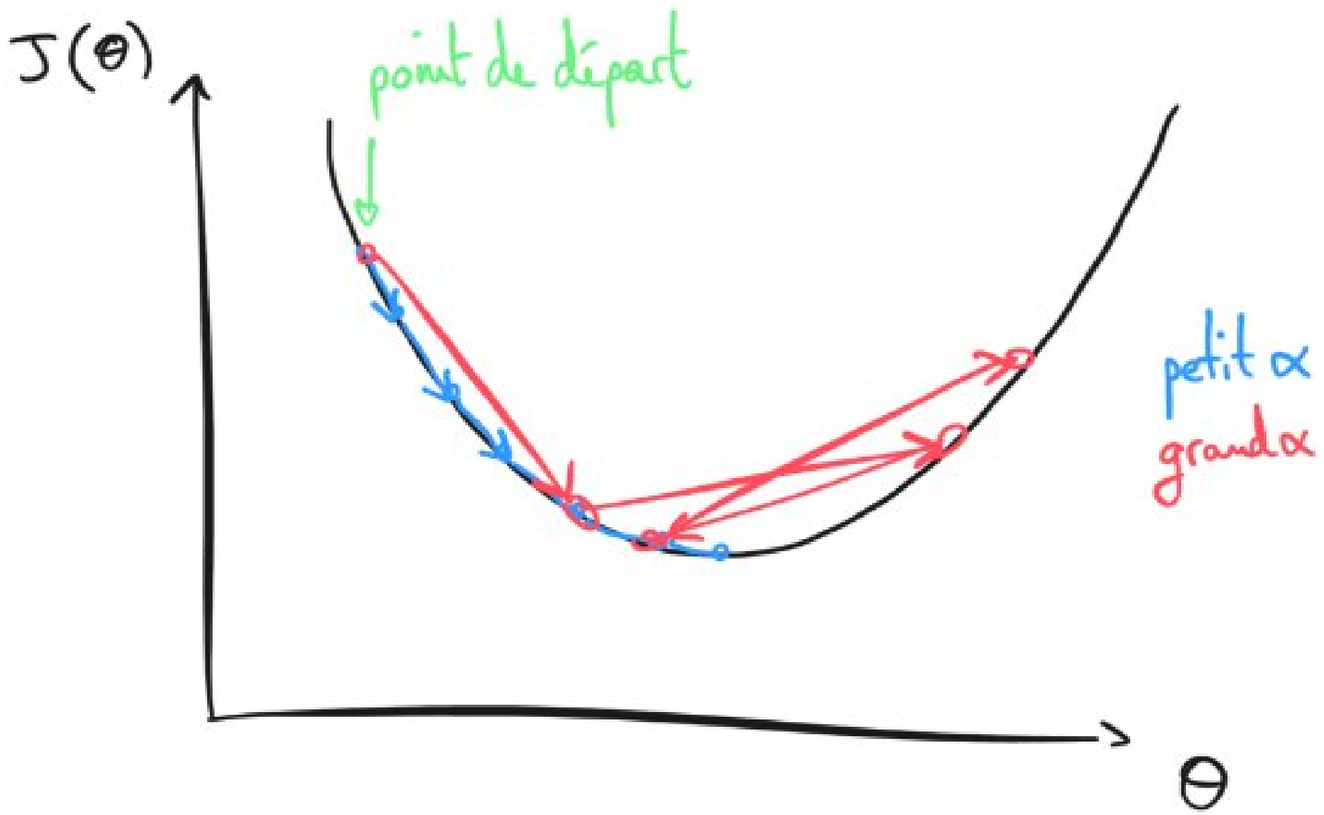
Concepts de base en Machine Learning - Régression linéaire multiple

La méthode de la descente de gradient



Concepts de base en Machine Learning - Régression linéaire multiple

Effet du α un facteur d'apprentissage



Concepts de base en Machine Learning - Régression linéaire multiple

Le problème de surapprentissage

Concepts de base en Machine Learning - Régression linéaire multiple

Exemple d’implémentation d’un algorithme de Machine Learning

Concepts de base en Machine Learning - Classification

- Régression linéaire simple et polynomiale
- Random Forest
- K-Means
- K-nearest neighbors (kNN)
- Réseaux de neurones
- Support Vector Machine (SVM)
- …

Concepts de base en Machine Learning - Classification (Random Forest)

- Les arbres de décision permettent de découper un ensemble d'observations en groupes homogènes en se basant sur des règles appliquées sur les variables descriptives

- CART (1980) parmi les algorithmes les plus connus pour la construction d'arbres de décision.

- Les algorithmes de forêts aléatoires ou Random Forest ont quant à eux été introduits au tout début des années 2000.

- Efficaces, faciles à comprendre et très intuitifs

Concepts de base en Machine Learning - Classification (Random Forest)

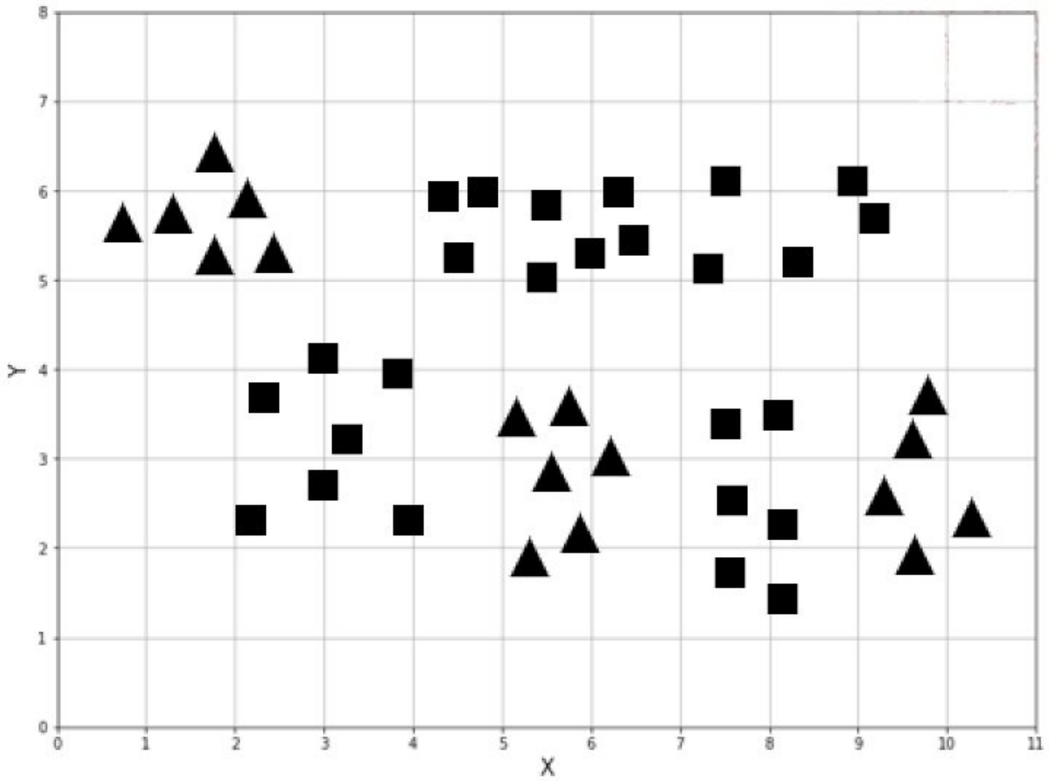
Un arbre de décision est constitué des éléments suivants :

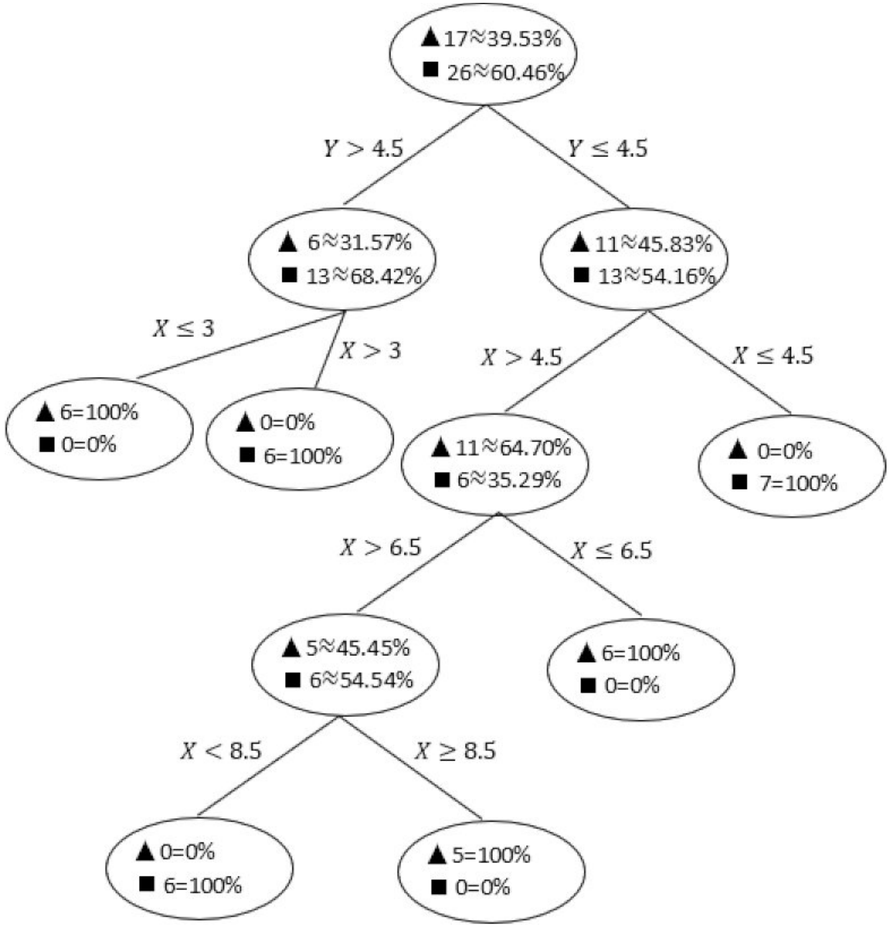
1. Une racine qui représente toutes les observations d’apprentissage.

2. Des sommets pendants communément appelés les feuilles et qui sont situés tout en bas des différentes branches de l’arbre de décision.

3. Des niveaux intermédiaires entre la racine et les feuilles. Chaque niveau intermédiaire correspond à la sélection d’une variable et certaines de ses valeurs.

Concepts de base en Machine Learning - Classification (Random Forest)





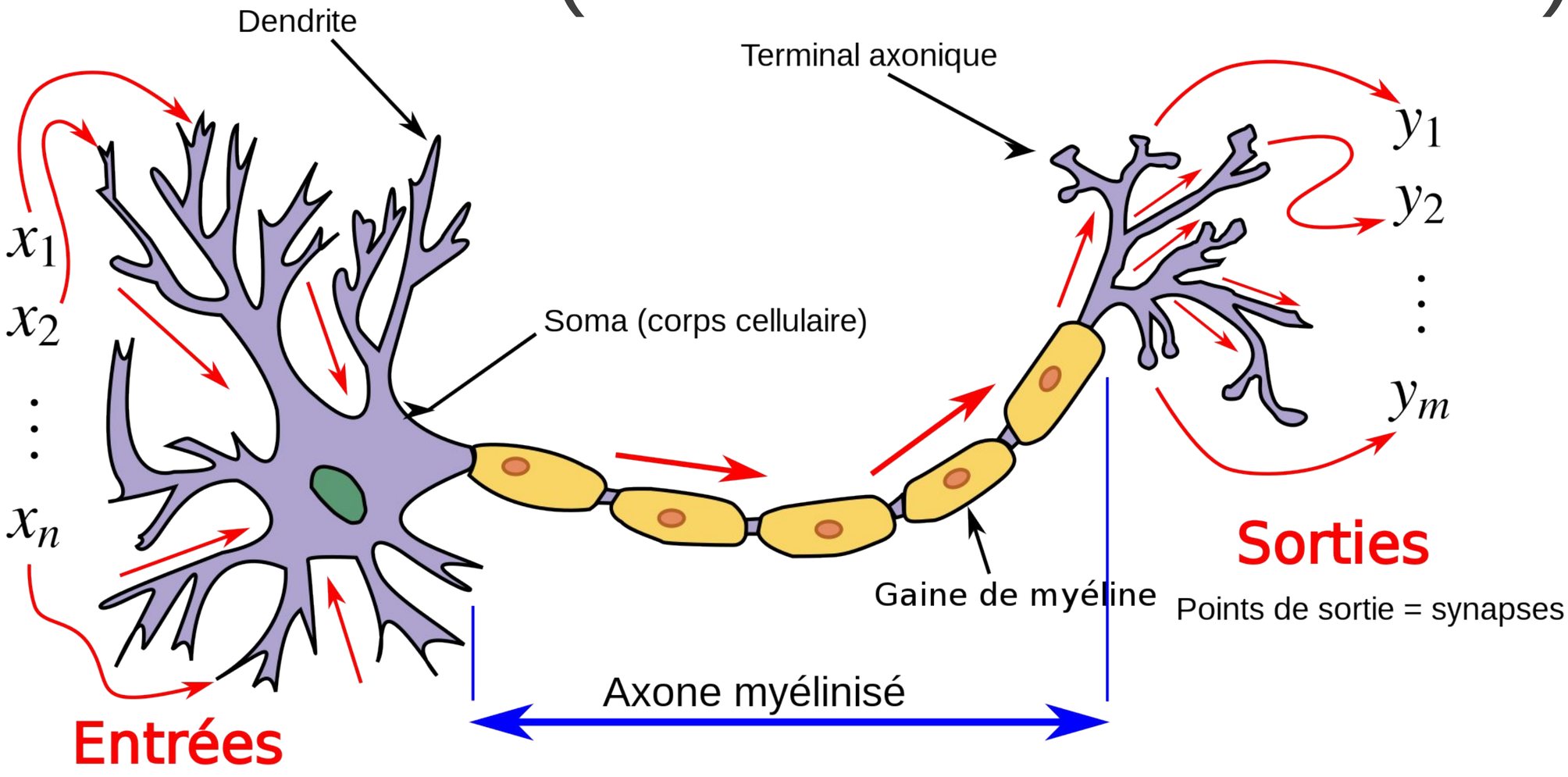
Concepts de base en Machine Learning - Classification (Random Forest)

Exemple d’implémentation d’un Random Forest

Concepts de base en Machine Learning - Classification (Réseaux de neurones)

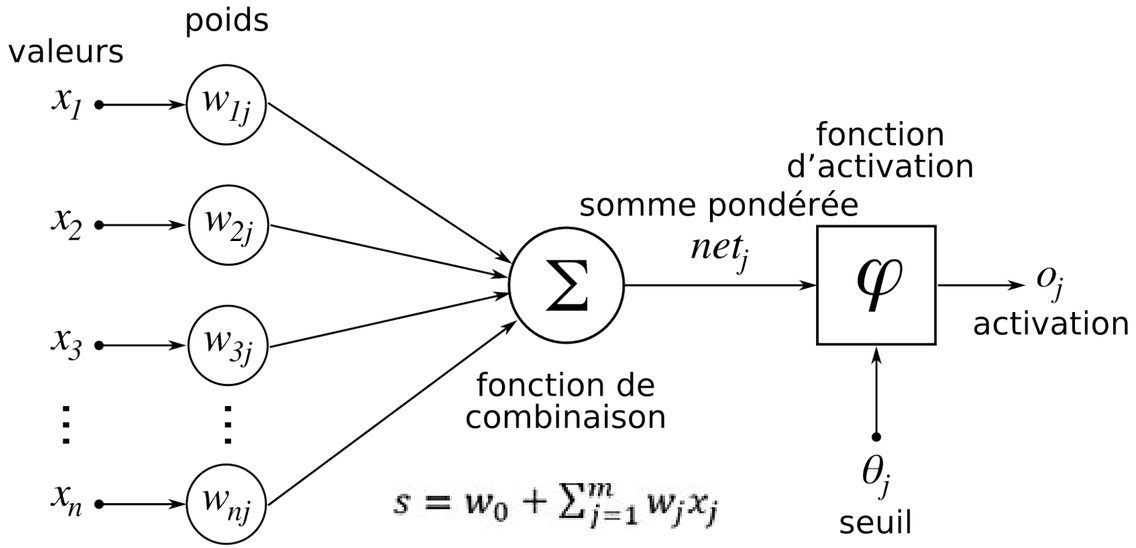
- Les dendrites servent d’interface en entrée pour un neurone. Elles reçoivent l’influx nerveux provenant des autres neurones.
- Le noyau, où ont lieu les réactions électrochimiques.
- L’axone joue le rôle de support de l’influx nerveux électrique qui sera transmis aux terminaisons synaptiques.
- Les synapses transmettent l’influx nerveux électrique provenant de l’axone. Cette transmission est réalisée en libérant des neurotransmetteurs qui seront capturés par les dendrites des autres cellules neuronales.

Concepts de base en Machine Learning - Classification (Réseaux de neurones)



Concepts de base en Machine Learning - Classification (Réseaux de neurones)

- Le neurone artificiel reçoit un ensemble d'entrées et retourne un résultat en sortie
- Le neurone artificiel applique une fonction d'assemblage sur les entrées
- Le neurone applique une fonction d'activation qui donne le résultat final du neurone

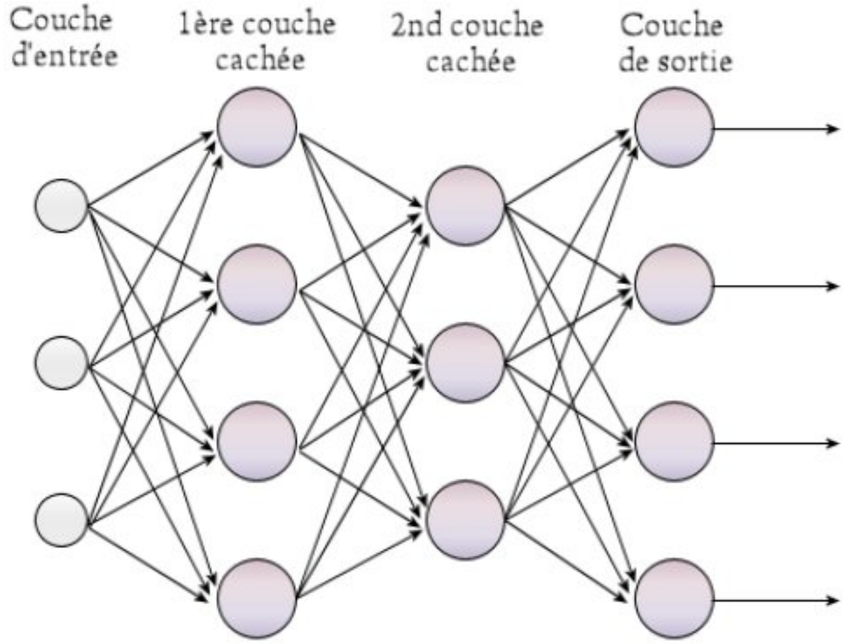


Concepts de base en Machine Learning - Classification (Réseaux de neurones)

- Plusieurs fonctions d'activation existent. Les trois fonctions suivantes font partie des plus utilisées :
  - La fonction sigmoïde $\mathbf{F(s) = \frac{1}{1+e^{s}}}$. Cette fonction transforme le résultat de la fonction d'assemblage en une probabilité.
  - La fonction d'activation à seuil $\mathbf{F(s|\theta)} = \begin{cases} 1\ si\ \ s\ \ \geq \boldsymbol{\theta} \\ 0\ si\ s\ \ \leq\ \boldsymbol{\theta} \end{cases}$
  - La fonction radiale $\mathrm{F(s)}\ = \sqrt{\frac{1}{2\pi}} e^{(\frac{s^2}{2})}$

Concepts de base en Machine Learning - Classification (Réseaux de neurones)

- Un réseau de neurones est constitué de plusieurs couches.
- Chaque couche d’un réseau de neurones est associée à un ensemble de neurones.
- Chaque neurone de la couche est associé à un ensemble de poids
- La couche de sortie fournit la réponse du réseau de neurones



Concepts de base en Machine Learning - Classification (Réseaux de neurones)

Exemple d’implémentation d’un réseaux de neurones

Concepts de base en Machine Learning - Métriques de performance

- La comparaison entre les vrais résultats et les prédictions dépend du type du modèle choisi pour décrire la relation entre X et Y.

- Pour les modèles de régression, les indicateurs utilisés sont :
  - L’erreur absolue moyenne MAE (Mean Absolute Error).
  - L’erreur absolue relative RAE (Relative Absolute Error).
  - La racine carrée de la moyenne du carré des erreurs RMSE (Root Mean Squared Error).
  - Le carré des erreurs relatives RSE (Relative Squared Error).
  - Le coefficient de détermination R2

Concepts de base en Machine Learning - Métriques de performance

- Pour les modèles de classification :
  - La matrice de confusion. À partir de cette matrice de confusion, plusieurs indicateurs de performance peuvent être calculés. Sans être exhaustifs, parmi ces indicateurs les plus utilisés se trouvent l'indicateur de sensibilité, appelé aussi le Recall, l'indicateur de précision, l'indicateur F1 Score.
  - La courbe ROC (Receiver Operating Characteristic).

- Pour les modèles de clustering, les mesures de performances sont essentiellement basées sur des mesures indiquant le degré de rapprochement entre les éléments regroupés ensemble et le degré d'éloignement des éléments répartis dans des groupes différents.

Concepts de base en Machine Learning - Métriques de performance

- L’erreur absolue moyenne MAE (Mean Absolute Error). $\mathrm{MAE} = \frac{1}{n} \sum_{i=1}^{n} | \hat{y}_i - y_i |$
- MAE mesure la moyenne des erreurs en valeur absolue sur les n observations qui forment la base de données d’apprentissage
- Cette mesure est intéressante pour comparer les performances de deux modèles pour le même problème.
- l’algorithme qui obtient le plus petit MAE qui sera considéré comme l’algorithme le plus performant des deux.
- MAE n’est pas adapté pour comparer les performances de deux algorithmes exécutés sur des jeux de données différents.

Concepts de base en Machine Learning - Métriques de performance

- L’erreur absolue relative RAE (Relative Absolute Error).

$$RAE = \frac{1}{n}\sum_{i=1}^{n}\frac{|\hat{y}_i - y_i|}{|\bar{Y} - y_i|}$$

- RAE donne le taux de déviation des prédictions par rapport aux vraies valeurs

- est intéressante pour comparer les performances des modèles, mais malheureusement cette mesure présente certaines limites, à savoir :
  - Elle ne peut pas être utilisée avec des valeurs nulles.
  - Lorsque les écarts entre les valeurs prédites et les valeurs observées deviennent très importants, le taux de déviation peut dépasser la valeur 1.

Concepts de base en Machine Learning - Métriques de performance

- La racine carrée de la moyenne du carré des erreurs

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}$$

- Le RMSE met au carré les différences de prédiction, ce qui met en avant les écarts entre les prédictions et les vrais valeurs.

Concepts de base en Machine Learning - Métriques de performance

- Le carré des erreurs relatives donne le taux de déviation au carré des prédictions par rapport aux vraies valeurs.

$$RSE = \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{n}(\bar{Y} - y_i)^2}$$

- Le coefficient de détermination permet de mesurer à quel point un modèle parvient à expliquer la variance associée à la variable à interpréter à partir des variables explicatives

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{n}(\bar{Y} - y_i)^2}$$

Concepts de base en Machine Learning - Métriques de performance

- Lorsque $\mathbf{R^2 = 1}$, cela signifie que le modèle explique parfaitement l'influence des variables Xi sur la variance de la distribution associée à la variable à expliquer

- Lorsque $\mathbf{R^2 = 0}$, le modèle n'explique pas du tout l'influence des variables Xi sur la variation de la variable à expliquer.

- Lorsque $\mathbf{R^2 < 0}$, cela signifie que les prédictions sont sensiblement loin des vraies valeurs observées et donc le modèle utilisé est de très mauvaise qualité.

Concepts de base en Machine Learning - Métriques de performance

- La matrice de confusion est un outil de mesure de la performance des modèles de classification à 2 classes ou plus.

- Une matrice de confusion n'est pas forcément symétrique et dans la réalité l'est rarement.

- VP (TP) : le nombre des vrais positifs.
- FP (FP) : le nombre des faux positifs.
- FN (FN) : le nombre des faux négatifs.
- VN (TN) : le nombre des vrais négatifs.

Reality

Confusion matrix	Class A	Class B	Class C
Class A	True Positive A : TPA	False Negative B : FNB False Positive A : FPA	False Negative C : FNC False Positive A : FPA
Class B	False Positive B : FPB False Negative A : FNA	True Positive B : TPB	False Negative C : FNB False Positive B : FPB
Class C	False Positive C : FPC False Negative A : FPA	False Positive C : FPC False Negative B : FNB	True Positive C : TPC

Prediction

Concepts de base en Machine Learning - Métriques de performance

- Le taux d'erreur du modèle peut être défini comme suit $\frac{VP+FN}{m_{test}}$
- Le Recall $\frac{VP}{VP+FN}$ correspond au taux des positifs détectés par le modèle
- Le Recall sera égale à 1 si tous les individus qui sont réellement positifs sont classés positifs par le modèle.
- L'importance donnée au Recall et à la précision dépend fortement du problème traité

Concepts de base en Machine Learning - Métriques de performance

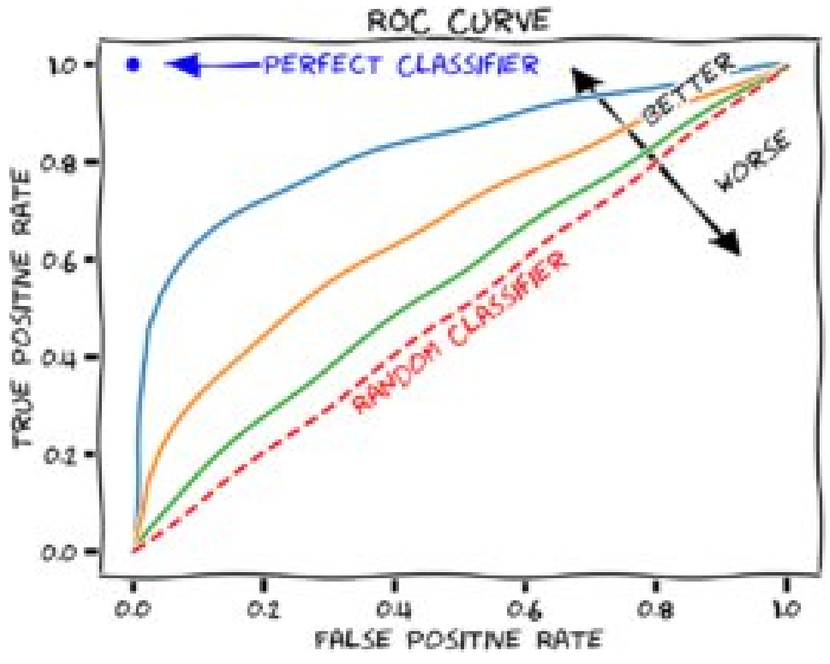
- un modèle qui a plus de précision sera préféré si le fait de considérer un individu positif alors qu'il est négatif entraîne des conséquences graves

- un modèle avec plus de Recall sera préféré si c'est le fait de ne pas détecter un cas positif qui entraîne des conséquences graves.

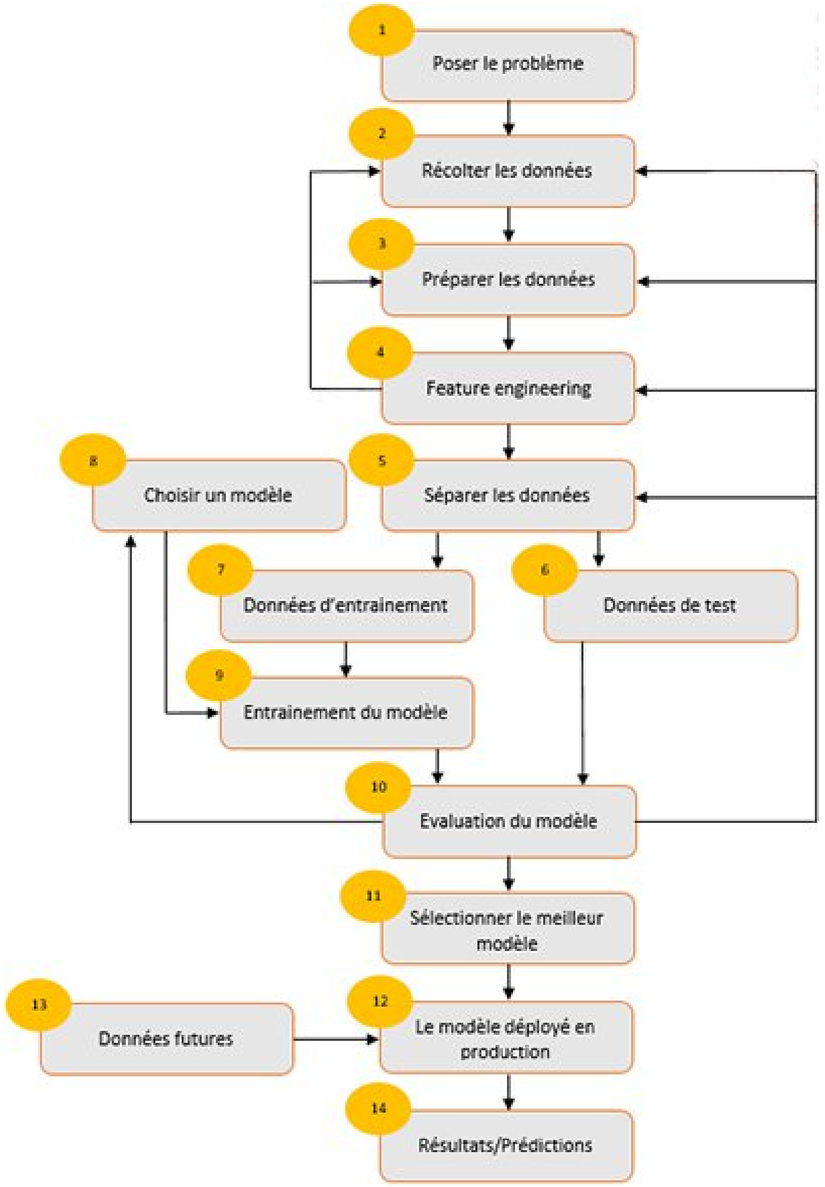
Concepts de base en Machine Learning - Métriques de performance

- La courbe ROC donne les résultats obtenus en fonction de ce seuil de classification.

- La droite qui coupe en deux le plan des abscisses et des ordonnées symbolise un modèle qui fait des prédictions de façon aléatoire



Concepts de base en Machine Learning



Concepts de base en Machine Learning

Exemple d’implémentation d’un réseaux de neurones

Concepts de base en Machine Learning - Le Deep Learning

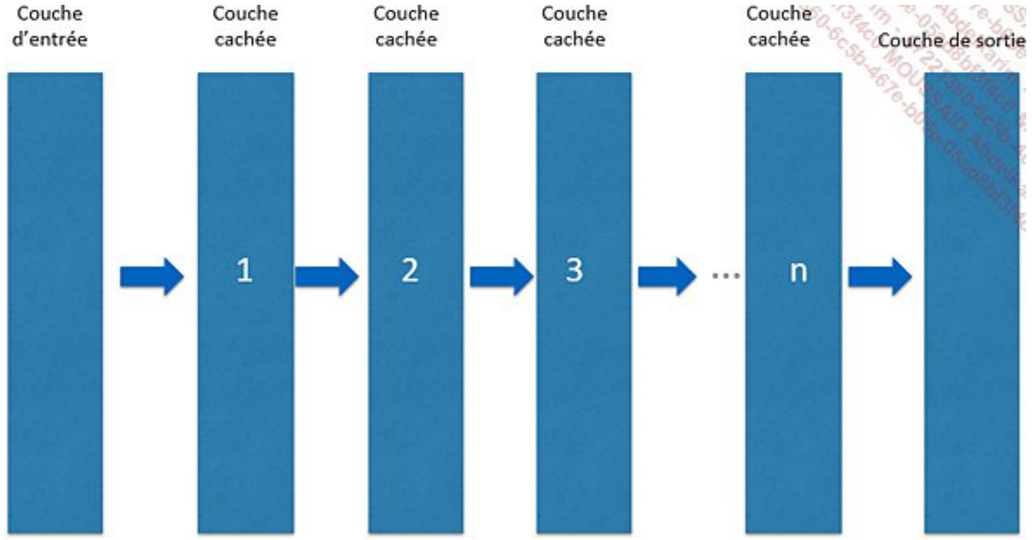
- Traiter le sujet du Deep Learning dans sa globalité avec toutes les versions techniques possibles proposées dans la littérature est quasiment impossible.
- Deep Learning via des exemples en utilisant TensorFlow
- Deep Learning, ou apprentissage profond, il s'agit bien de réseaux de neurones constitués d'une ou plusieurs couches d'entrée, d'une ou plusieurs couches de sortie et de deux ou plusieurs couches cachées.

Concepts de base en Machine Learning - Le Deep Learning

- Chacune des couches cachées d’un Deep Learning sera spécialisée dans la résolution d’une partie du problème à solutionner.
- augmenter le nombre de couches cachées d’un réseau de neurones de manière aléatoire ne fait qu’augmenter la redondance des informations déjà capturées par des couches situées au début de ce réseau
- la multiplicité arbitraire des couches cachées peut détériorer les performances d’un réseau de neurones

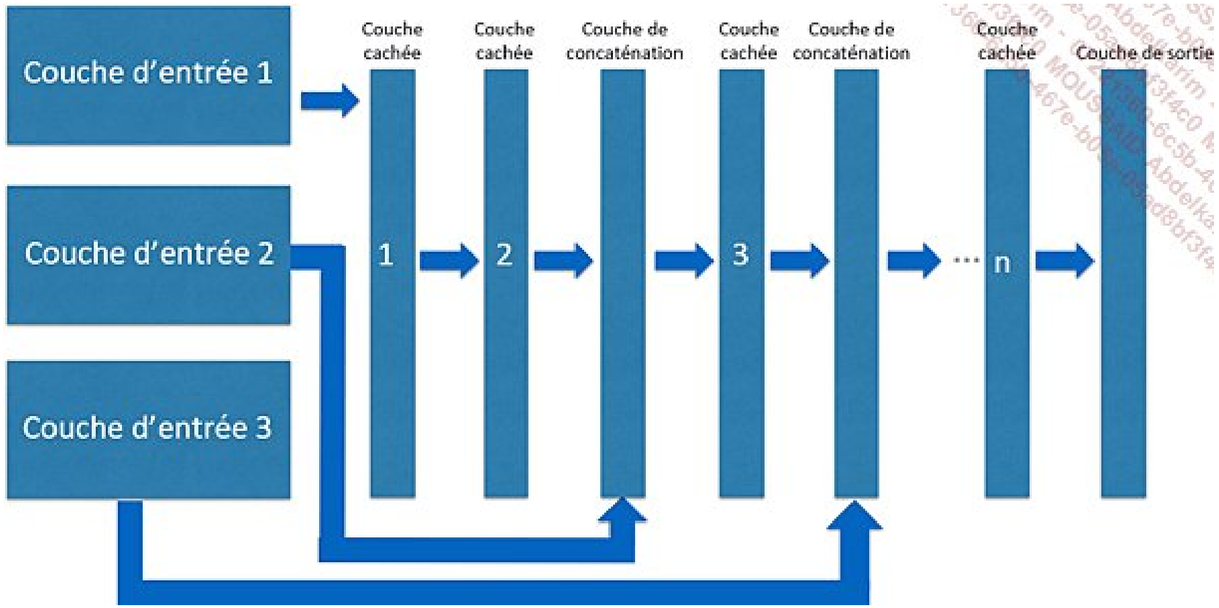
Concepts de base en Machine Learning - Le Deep Learning

Schéma général d’un réseau de neurones dit de Deep Learning



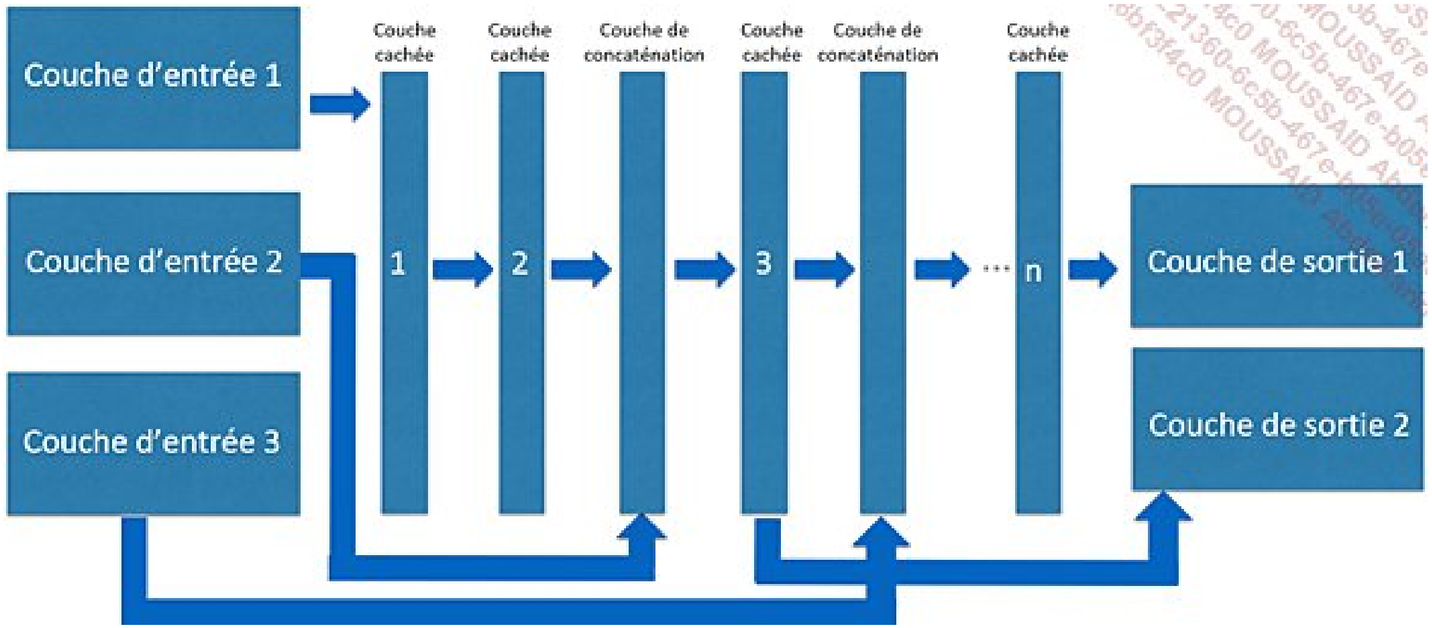
Concepts de base en Machine Learning - Le Deep Learning

Réseau de neurones avec plusieurs couches d’entrée



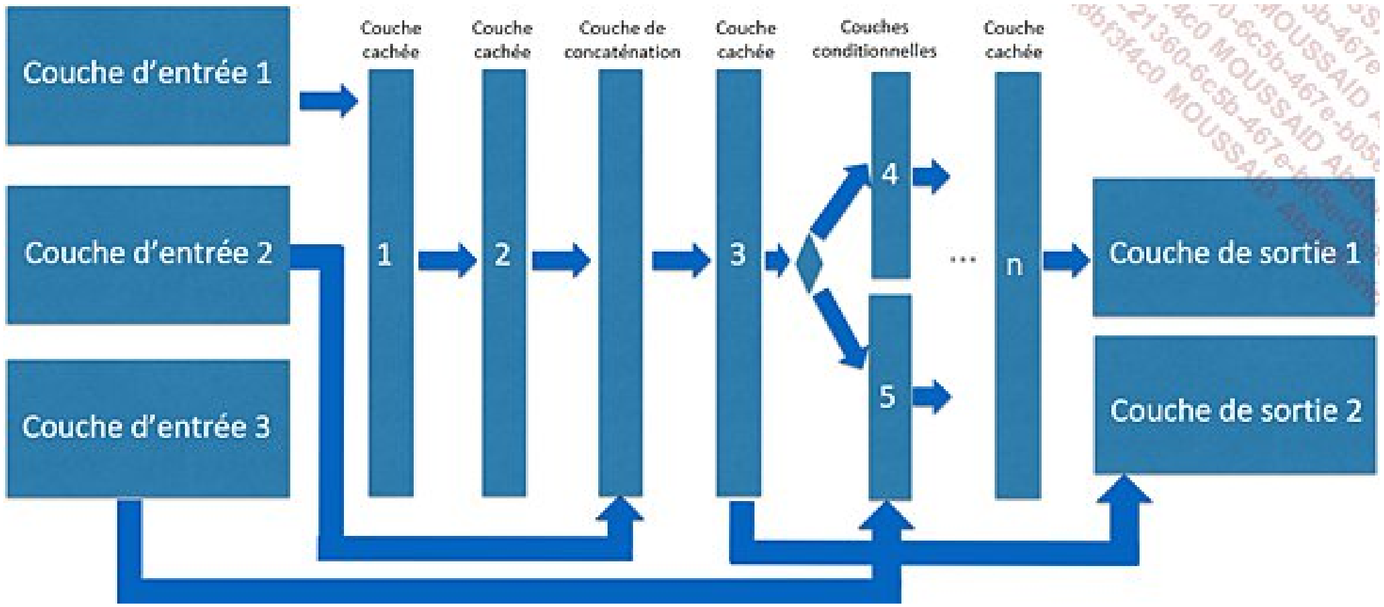
Concepts de base en Machine Learning - Le Deep Learning

Réseau de neurones avec plusieurs couches de sortie



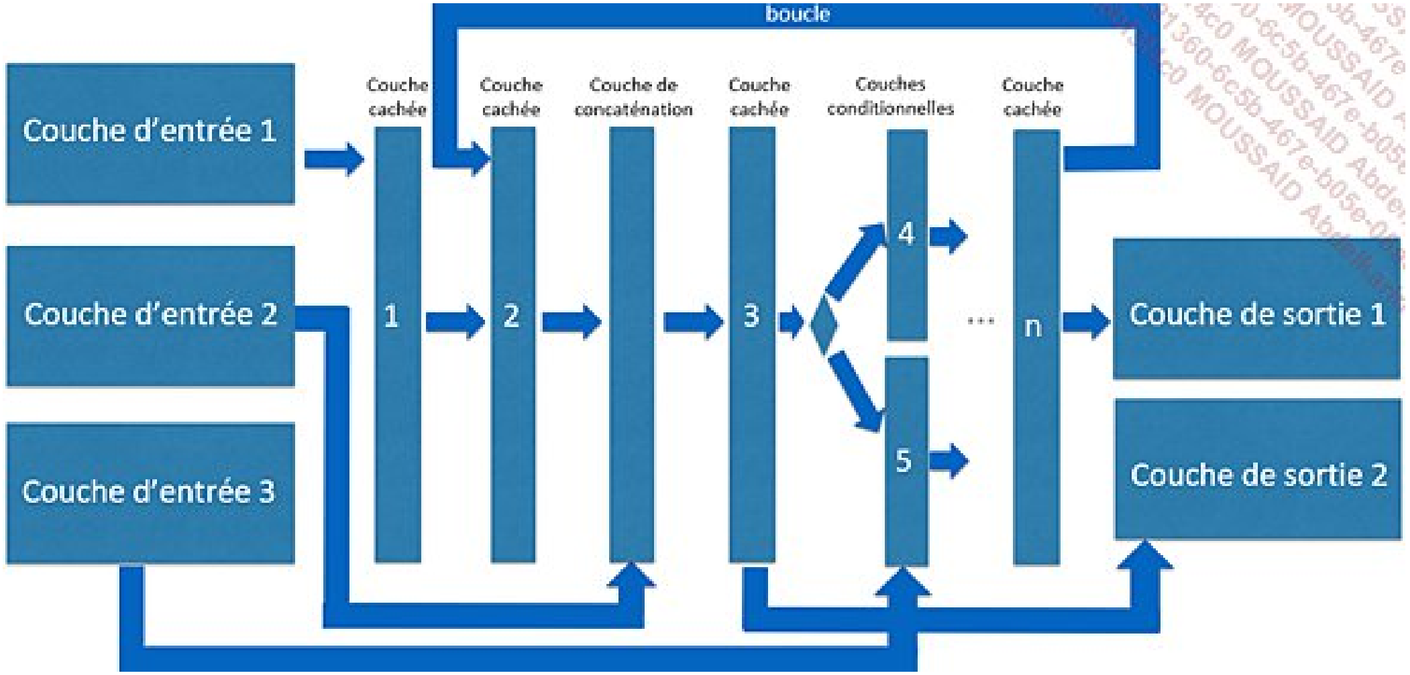
Concepts de base en Machine Learning - Le Deep Learning

Réseau de neurones avec des branchements conditionnels



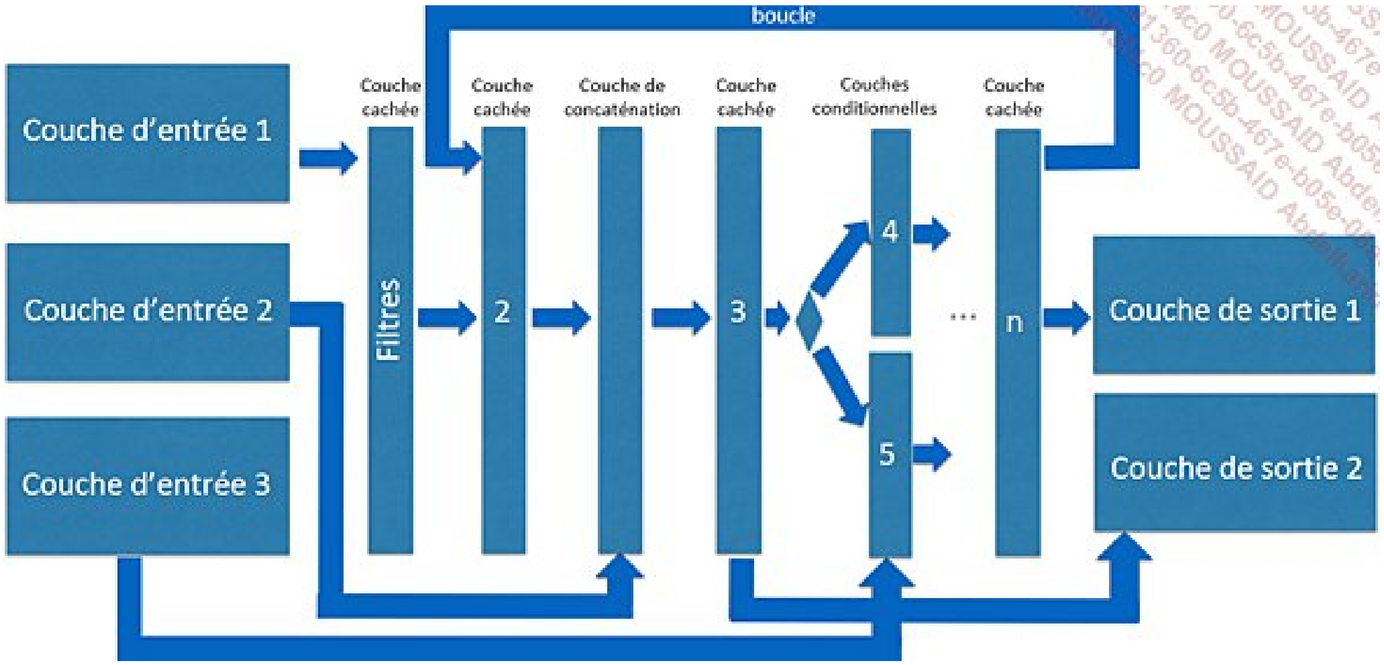
Concepts de base en Machine Learning - Le Deep Learning

Réseau de neurones avec de la récurrence RNN



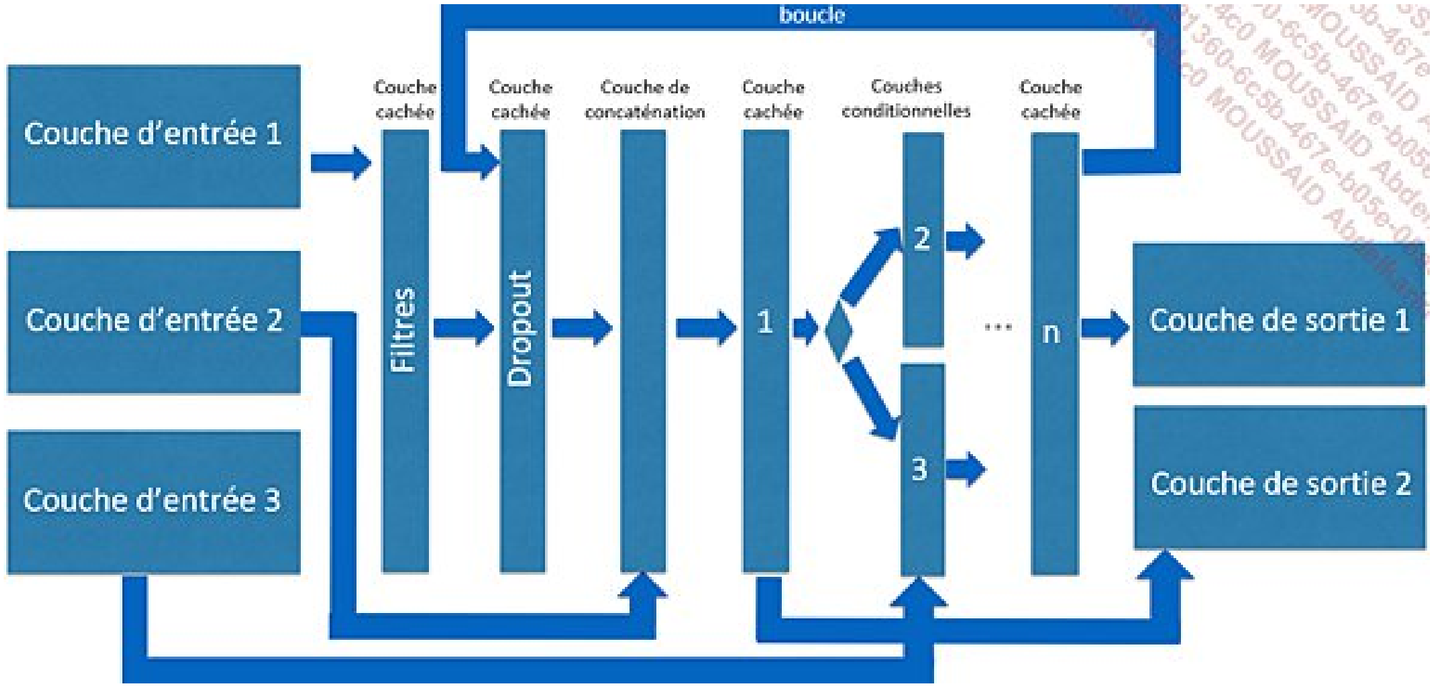
Concepts de base en Machine Learning - Le Deep Learning

Réseau de neurones avec couches de convolution CNN



Concepts de base en Machine Learning - Le Deep Learning

Éviter le surapprentissage avec les couches Dropout

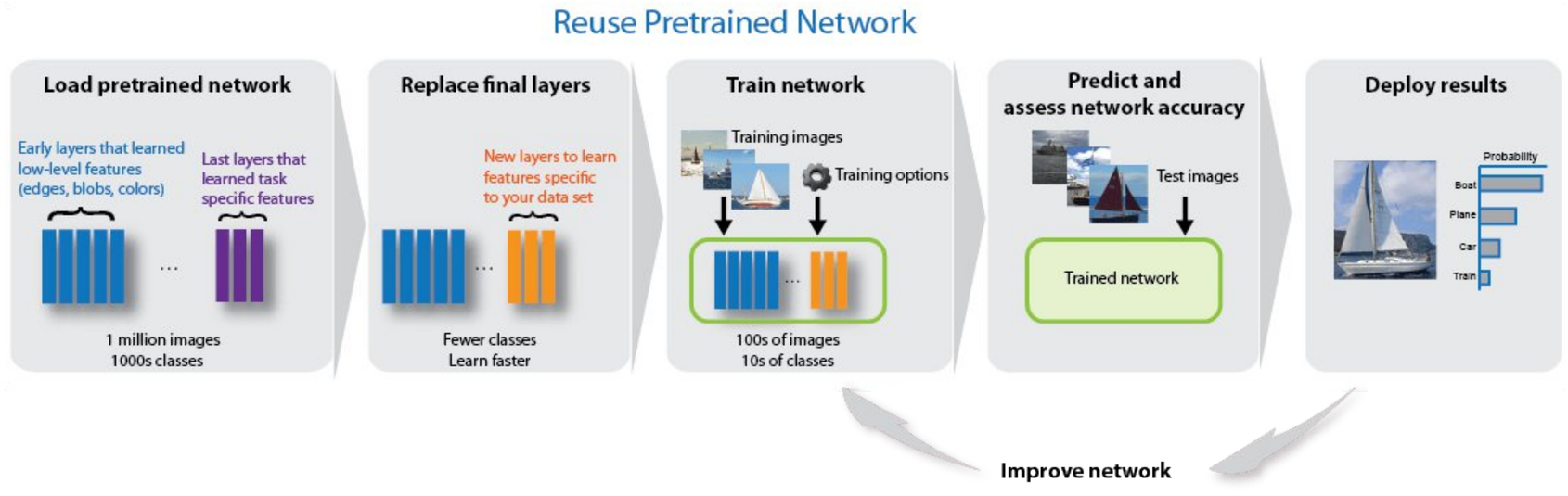


Concepts de base en Machine Learning - Le Transfert Learning

- Le Transfer Learning, ou l'apprentissage par transfert, est une technique très répandue dans le domaine des réseaux de neurones

- Il s'agit de réutiliser une partie d'un réseau de neurones déjà existant et qui est spécialisé dans la résolution d'un problème donné pour résoudre un autre problème

- Le transfert Learning en 5 étapes :



Concepts de base en Machine Learning

Exemple d’implémentation d’un réseaux de neurones

Natural Language Processing (NLP)

- les techniques de nettoyage de documents,
- les concepts de stopwords, de Stemming et de Lemmatization,
- les techniques de vectorisation des données textuelles,
- la vectorisation par comptage d’occurrences de mots,
- la vectorisation avec la méthode TF-IDF,
- la vectorisation avec N-Grams,
- le développement de modèle de classification de documents.

Natural Language Processing (NLP)

- La classification des documents
- L’analyse des sentiments/émotions
- La reconnaissance d’entités nommées
- La génération automatique du texte
- Le Part-Of-Speech (POS) tagging
- Le développement des chatbots

Natural Language Processing (NLP)

Le nettoyage des données textuelles

- La Suppression des stopwords
  - les stopwords sont tous les mots présents dans un document mais qui n'apportent aucune valeur ajoutée à ce document en matière de sémantique
  - les mots moi, tu, seras, je, serai et sur sont généralement considérés comme des stopwords.
  - suppression de tous les signes de ponctuation **!"#$%&\'()*+,-./:;<=>?@[\\]^_`{|}~**

Salut! Fais-moi signe quand tu seras arrivé… je serai déjà sur place!

Salut! Fais signe quand arrivé… déjà place! ==> ***Salut Fais signe quand arrivé déjà place***

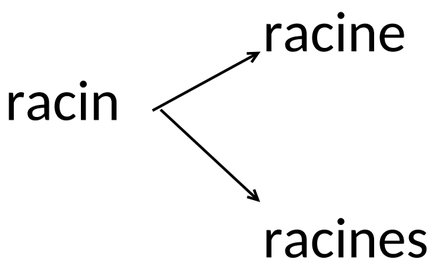
Natural Language Processing (NLP)

Appliquer le Stemming sur un texte

- Stemming permet de réduire la taille d'un document en contractant les mots
- les mots engendrés avec le Stemming peuvent ne pas faire partie des mots du dictionnaire.

racine ==> racin
racines ==> racin

Donc



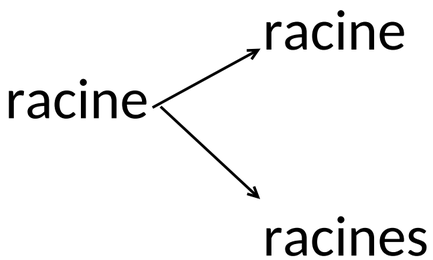
Natural Language Processing (NLP)

Appliquer la Lemmatization sur un texte

- Lemmatization permet de réduire la taille d'un document en contractant les mots
- les mots engendrés avec le Lemmatization font partie des mots du dictionnaire.

racine ==> racine
racines ==> racine

Donc



Natural Language Processing (NLP)

Stemming vs Lemmatization

- Le choix du Stemming ou de la Lemmatization dépend du projet.
- Parfois, cela vaut la peine de tester les deux approches et comparer leurs résultats.
- Le Stemming peut engendrer des mots en dehors du dictionnaire de la langue utilisée ==> Perte du sens
- La Lemmatization engendre des mots du dictionnaire ==> Pas de perte du sens
- La Stemming est beaucoup plus rapide à l'exécution par rapport à la Lemmatization
- Stemming : meanness ==> mean, meaning ==> mean
- Lemmatization : meanness ==> meanness, meaning ==> meaning.

Natural Language Processing (NLP)

La vectorisation par comptage d’occurrences des mots

- Cette méthode permet la transformation des données textuelles en données numériques.
- Consiste à compter le nombre d’occurrence de chaque mot dans la base données

Inconvénients :

- Ne prennent pas en compte le taux de présence de ce mot dans les autres documents
- Ne prend pas en compte le contexte dans lequel les mots sont utilisés
- Venez profiter des soldes ! Nos prix soldés sont tous à moins de 50% ==> Non Spam
- Coucou, le prix dont tu m'as parlé hier ne correspond pas au prix que j'ai trouvé sur leur site ! ==> Spam

Natural Language Processing (NLP)

La vectorisation TF-IDF

- TF-IDF : Term Frequency-Inverse Document Frequency

- La note $W_{i,j}$ correspondant à un mot *i* et un document *j* est donnée par :

$$W_{i,j} = tf_{i,j} * \log\left(\frac{N}{df_i}\right)$$

  - $tf_{i,j}$ est le taux d'apparition du moti dans le document j.
  - N est le nombre de documents total dans la base de données.
  - df_i est le nombre de documents contenant le mot ***i***

- TF-IDF est très couramment utilisée dans le développement des modèles de classification de documents

Doc_id	mot1	mot2	...	Label
1	0.002	0.0000	...	Spam
2	0.0000	0.2500	...	spam
3	0.0000	0.0010	...	ham
4	0.0000	0.0000	...	ham
5	0.2300	0.0000	...	spam
6	0.0000	0.2300	...	ham
7	0.2249	0.0000	...	ham

Natural Language Processing (NLP)

La vectorisation N-Gram

- La méthode N-Gram permet de compter un regroupement de mots et non pas les mots un par un séparément.

- Avec la méthode N-Gram, chaque colonne de la table qui sera construite va correspondre à plusieurs mots en fonction du degré souhaité.

- Si nous utilisons N-Gram avec des constructions de deux mots, alors chaque colonne de la table construite correspond à deux mots retrouvés l'un à la suite de l'autre dans une phrase.

Exemple : Venez profitez des soldes ! Nos prix soldés sont tous à moins de 50%

Natural Language Processing (NLP)

Feature Engineering

- Ajouter de nouvelles variables comme les variables ci-dessous :
  - Taux de caractères de ponctuation.
  - Taille des messages.
  - Taux de lettres en majuscules.
  - Présence ou non de certains mots tels que « lol ».
- La transformation des variables existantes peut s'avérer utile si les valeurs des variables d'origine cachent des informations non exploitables en l'état

Perspectives et éthique en IA

- L'éthique de l'intelligence artificielle est le domaine de l'éthique de la technologie propre aux robots et autres entités artificiellement intelligents.
- L’IA est un outil au service des humains.
- L’IA permet de réaliser des progrès spectaculaires dans de nombreux domaines comme la santé, l’éducation ou l’environnement.
- L’IA peut aussi faciliter le travail de certains acteurs malveillants (les fausses informations, les attaques en ligne, ...).

Perspectives et éthique en IA

Deep fake

https://www.youtube.com/watch?v=S951cdansBI

Scamming

https://www.youtube.com/watch?v=V6_jCGzR020

Perspectives et éthique en IA

Tesla Call

Perspectives et éthique en IA

- Analyse des scanner et détection des cancers
- Analyse des images pour détecter des anomalies
- Détecter un comportement dangereux
- Assistance médical
- Détection de fraude.
- ...

Perspectives et éthique en IA

- Biais de l’IA désigne une situation dans laquelle un système de Machine Learning discrimine un groupe de personnes en particulier.
- Cette discrimination reflète celles que l’on déplore dans notre société
- Technologies de reconnaissance faciale de Microsoft, IBM, et de l’entreprise chinoise Face++.
- Des chercheurs du MIT ont découvert que certaines technologies se révélait plus performante sur des visages à peau blanche que sur des visages à la peau plus sombre
- les trois systèmes se sont avérés bien moins performants sur les visages de femmes que sur les visages

Perspectives et éthique en IA

- Malgré ces biais discriminatoires, les systèmes de reconnaissance faciale sont utilisés.
- Risque important de se tromper lors de l’identification
- Les causes du biais :
  - Déséquilibre des classes dans les datasets
  - La qualité et la véracité des données d’entraînement.
- Exemple : Correctional Offender Management Profiling for Alternative Sanctions (COMPAS)

Perspectives et éthique en IA
- Impact de l’IA sur société

- L’IA n’est pas une technologie autonome, capable de penser par elle-même et de faire preuve d’imagination et de créativité.
- L’IA transforme le marché du travail, certains métiers vont disparaître mais d’autre apparenteront.
- L’IA devient ou deviendra un assistant dans beaucoup de métiers

Perspectives et éthique en IA
- Impact de l’IA sur société

- L’IA n’est pas une technologie autonome, capable de penser par elle-même et de faire preuve d’imagination et de créativité.
- L’IA transforme le marché du travail, certains métiers vont disparaître mais d’autre apparenteront.
- L’IA devient ou deviendra un assistant dans beaucoup de métiers

Projet NLP

Projet implémentation d’un algorithme de Machine Learning pour la détection des sentiments

Merci !