

Rapport - Machine Learning

Romain JAMINET

Louis LAMALLE

Hugo GIANFALDONI

Boyan BECHEV

Romain FOURNIER

7 février 2019

Table des matières

1	Introduction	2
1.1	Qu'est-ce que l'intelligence artificielle	2
1.2	Pourquoi utiliser l'intelligence artificielle	2
1.3	Les différents types d'intelligences artificielles	3
2	Le Machine Learning	3
2.1	Définition	3
2.2	Arbre	3
3	Apprentissage supervisé	3
3.1	Arbre	3
3.2	Définitions	3
3.3	Classification	4
3.3.1	Définition	4
3.3.2	Exemple	4
3.4	Régression	4
3.4.1	Définition	4
3.4.2	Exemple	5
3.5	Algorithmes	5
3.5.1	Simple Linear Regression	6
3.5.2	Logistic Regression	6
3.5.3	Support Vector Machines	6
3.5.4	naïve Bayes	7
3.5.5	Decision Trees	7
3.5.6	K-nearest neighbour	7
3.5.7	Similarity Learning	7
4	Apprentissage non-supervisé	7
4.1	Fonctionnement	7
4.2	Quelques algorithmes	7
4.3	Applications	8
5	Apprentissage par renforcement	9
5.1	Introduction	9
5.2	Réseau de neurones	9
5.2.1	Le neurone	10
5.2.2	Le réseau	10
5.3	Apprentissage	11

1 Introduction

1.1 Qu'est-ce que l'intelligence artificielle

L'intelligence artificielle (IA, ou AI en anglais pour Artificial Intelligence) consiste à mettre en œuvre un certain nombre de techniques visant à permettre aux machines d'imiter une forme d'intelligence réelle. L'IA se retrouve implémentée dans un nombre grandissant de domaines d'application.

La notion voit le jour dans les années 1950 grâce au mathématicien Alan Turing. Dans son livre *Computing Machinery and Intelligence*, ce dernier soulève la question d'apporter aux machines une forme d'intelligence. Il décrit alors un test aujourd'hui connu sous le nom « Test de Turing » dans lequel un sujet interagit à l'aveugle avec un autre humain, puis avec une machine programmée pour formuler des réponses sensées. Si le sujet n'est pas capable de faire la différence, alors la machine a réussi le test et, selon l'auteur, peut véritablement être considérée comme « intelligente ».

De Google à Microsoft en passant par Apple, IBM ou Facebook, toutes les grandes entreprises dans le monde de l'informatique planchent aujourd'hui sur les problématiques de l'intelligence artificielle en tentant de l'appliquer à quelques domaines précis. Chacun a ainsi mis en place des réseaux de neurones artificiels constitués de serveurs et permettant de traiter de lourds calculs au sein de gigantesques bases de données.

[8]Source

1.2 Pourquoi utiliser l'intelligence artificielle

L'intelligence artificielle a connu des hauts et des bas au cours des décennies qui se sont écoulées depuis sa conceptualisation initiale, mais nous constatons enfin de réels progrès et nous sommes en train de transformer notre monde. Les principales raisons à cela sont :

Une puissance de calcul massive est maintenant disponible à bas coût et peut être provisionnée dans le cloud très rapidement. Les améliorations apportées à la conception des GPU (maintenant avec des milliers de cœurs parfaitement adaptés aux charges de travail parallèles) ont multiplié par 50 la vitesse de formation des algorithmes d'apprentissage en profondeur en trois ans.

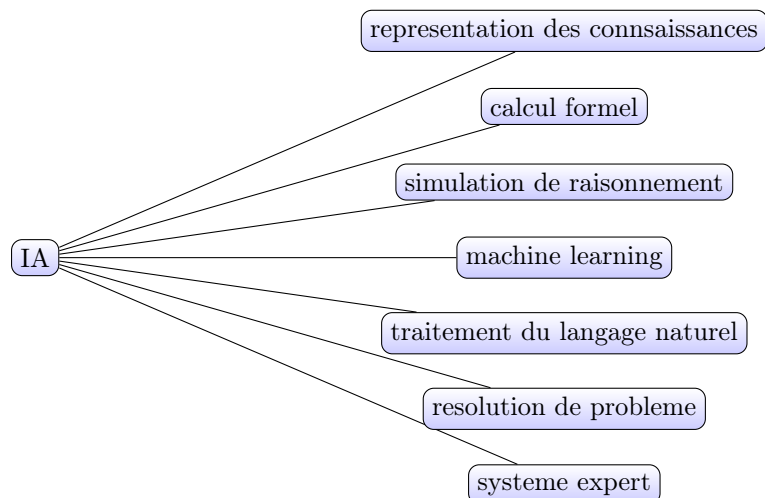
Big Data - La quantité de données créée a explosé, de pair avec une capacité de stockage presque illimitée. Des ensembles de données volumineux et diversifiés fournissent un meilleur matériel de formation pour les algorithmes.

Les algorithmes sont maintenant plus efficaces pour trouver des modèles dans les montagnes de données, et les plates-formes d'intelligence artificielle et d'apprentissage automatique sont recherchées, car Google, IBM et Microsoft facilitent beaucoup le développement d'applications.

L'investissement dans l'IA - en particulier l'apprentissage automatique et l'apprentissage en profondeur - augmente rapidement. Les machines sont déjà aussi performantes, voire meilleures que les humains, dans certaines tâches, par exemple jouer à des jeux tels que jouer aux échecs, reproduire de l'audio, analyser des images et diagnostiquer des maladies.

[1]Source

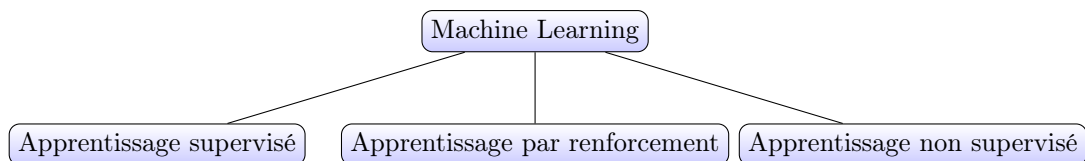
1.3 Les différents types d'intelligences artificielles



2 Le Machine Learning

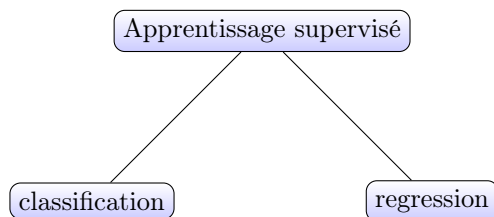
2.1 Définition

2.2 Arbre



3 Apprentissage supervisé

3.1 Arbre



3.2 Définitions

L'apprentissage supervisé consiste à entrainer une machine dans le but de lui faire apprendre une fonction de prediction pour la résolution d'un problème.

Pour se faire , on fournit à la machine une base de donn     tiquet  . L'  tiquette indique    l'algorithme ce qu'il doit avoir en sortie en fonction de la donn  e qui lui      t   donn  e en entr  .

3.3 Classification

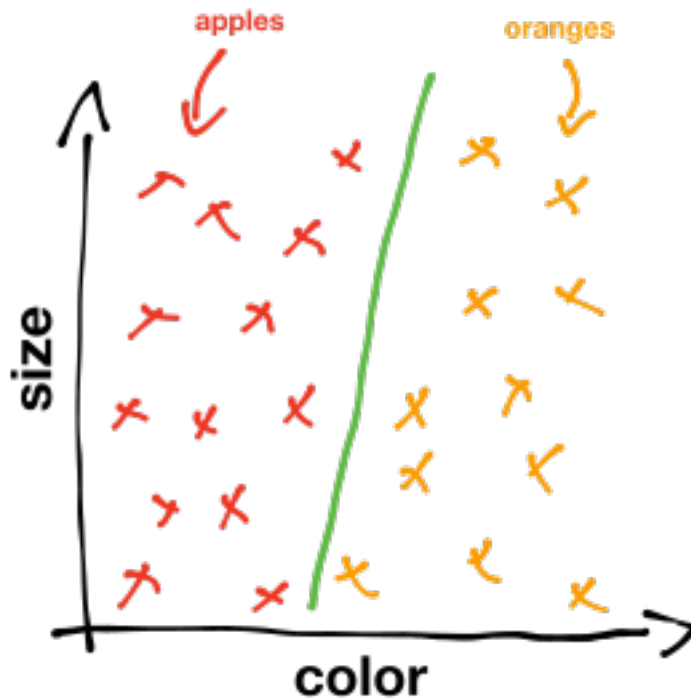
3.3.1 Définition

La classification au sens large correspond à l'organisation d'entité en différentes catégories selon certains critères définis.

En machine learning, on se sert de la classification pour prédire une donnée qualitative d'un objet.

On entraîne la machine avec une base de données étiquetée pour classer les données en différentes "classes". On pourra donc prédire la "future étiquette" d'un objet en fonction des attributs de l'objet.

3.3.2 Exemple



Sur cet exemple, on classe un ensemble de pommes et d'oranges en fonction de leur couleur et de leur taille.

On peut distinguer deux espaces distincts séparés ici par un séparateur. Cela correspond à l'entraînement de l'algorithme.

On peut donc maintenant donner en entrée un fruit (pomme ou orange) en entrée à notre algorithme pour qu'il détermine sa nature (en fonction de sa couleur et sa taille).

En réalité il y a beaucoup plus de 2 attributs à prendre en compte.

Le but final étant de déterminer une fonction pour distinguer les classes en prenant en paramètre les attributs.

[6]Source Image

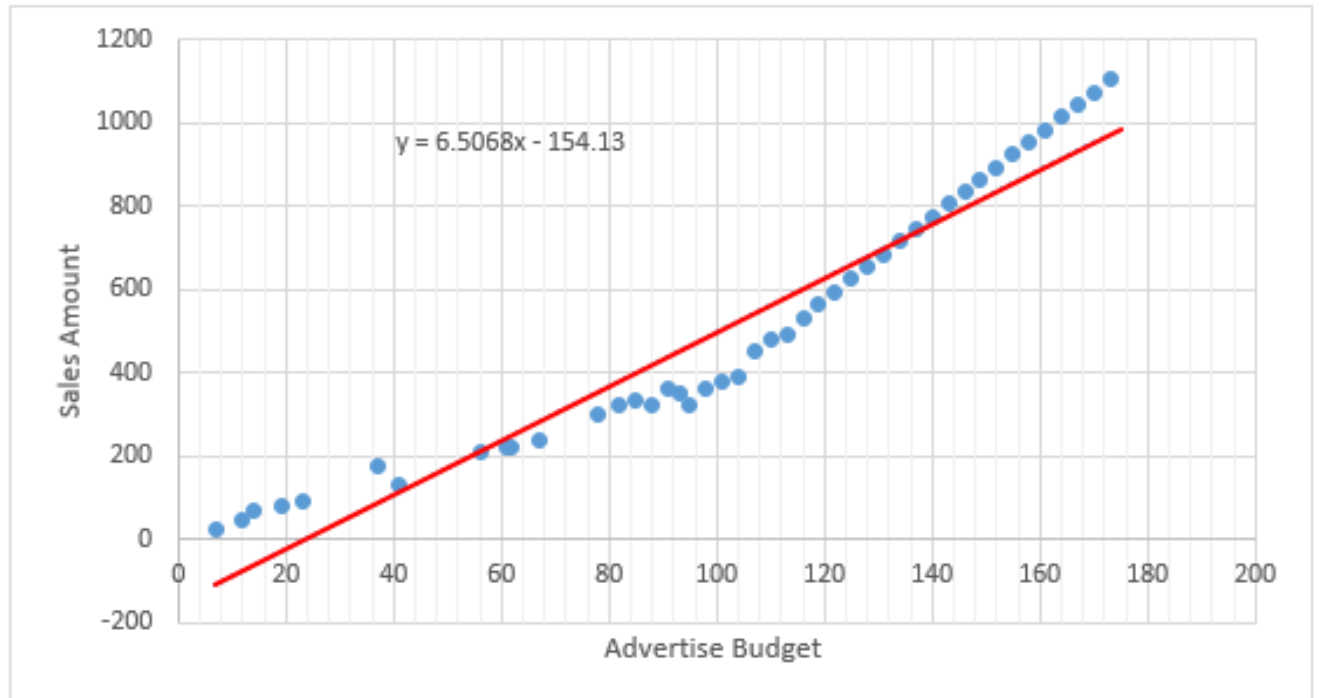
3.4 Régression

3.4.1 Définition

La régression est utilisée en machine learning pour prédire une donnée quantitative d'un objet.

On entraîne l'algorithme pour qu'il estime une fonction qui permette de calculer la donnée cherchée grâce aux attributs de l'objet passé en paramètre.

3.4.2 Exemple



Sur cet exemple , on représente le montant des ventes de certains produits en fonction du budget publicitaire. Grâce à la base de donnée , l’algorithme va approximer une fonction qui permet d’exprimer les ventes en fonction du budget.




On peut donc maintenant prédire les ventes d’un article en fonction de son budget publicitaire.

Le but des algorithmes de régression est donc de déterminer la fonction la plus optimale pour ce type de problème.

[7] Source Image

3.5 Algorithmes

Il existe un nombre d’algorithmes à choisir pour un problème de Machine Learning supervisé. Cependant, il n’existe pas un seul algorithme qui peut traiter tous problèmes.

	TYPE	NAME	DESCRIPTION	ADVANTAGES	DISADVANTAGES
Linear		Linear regression	The "best fit" line through all data points. Predictions are numerical.	Easy to understand -- you clearly see what the biggest drivers of the model are.	X Sometimes too simple to capture complex relationships between variables. X Does poorly with correlated features.
		Logistic regression	The adaptation of linear regression to problems of classification (e.g., yes/no questions, groups, etc.)	Also easy to understand.	X Sometimes too simple to capture complex relationships between variables. X Does poorly with correlated features.
Tree-based		Decision tree	A series of yes/no rules based on the features, forming a tree, to match all possible outcomes of a decision.	Easy to understand.	X Not often used on its own for prediction because it's also often too simple and not powerful enough for complex data.
		Random Forest	Takes advantage of many decision trees, with rules created from subsamples of features. Each tree is weaker than a full decision tree, but by combining them we get better overall performance.	A sort of "wisdom of the crowd". Tends to result in very high quality models. Fast to train.	X Models can get very large. X Not easy to understand predictions.
		Gradient Boosting	Uses even weaker decision trees, that are increasingly focused on "hard" examples.	High-performing.	X A small change in the feature set or training set can create radical changes in the model. X Not easy to understand predictions.
Neural networks		Neural networks	Interconnected «neurons» that pass messages to each other. Deep learning uses several layers of neural networks stacked on top of one another.	Can handle extremely complex tasks - no other algorithm comes close in image recognition.	X Very slow to train, because they often have a very complex architecture. X Almost impossible to understand predictions.

[11]Source

Un grand problème est la balance entre variance et bias. Imaginons qu'on a quelques différentes mais bien structurés datasets. Si on entraîne notre modèle avec une grande bias et une variance basse, notre modèle n'est pas assez "flexible", c.à.d. notre modèle va prédire le correct output, considérant seulement un dataset particulier. En revanche si on rend le modèle "flexible", la grande variance vas nous donner des outputs très différentes que souhaitées, même si le modèle prend en compte beaucoup de datasets.

Un autre problème est la dimension de notre input. Si on ajoutera des nouvelles variables a notre modèle, il va s'embrouiller et on peut anticiper une grande variance dans le résultat.

3.5.1 Simple Linear Regression

Dans le contexte d'une régression simple, on veut dessiner une ligne de la forme $y = ax + b$ à travers une dataset. Dépendamment de la data, on peut choisir une fonction qui nous convient. Par exemple, un cercle, une parabole, etc.

3.5.2 Logistic Regression

La régression logistique est très similaire à la régression linéaire. La fonction principale est la résolution des problèmes de classification. La différence essentielle est qu'on n'utilise pas une ligne pour approximer le résultat. Plutôt on utilise la fonction Sigmoid $S(x) = \frac{1}{1+e^{-x}}$ qui donne un output pour x valeurs entre 0 et 1. Ces valeurs permettent à corriger le modèle plus facilement.

3.5.3 Support Vector Machines

Comment trouver la meilleure ligne quand on parle d'un problème de classification? Une méthode sont des SVM, qui permettent à séparer deux ou plusieurs classes qu'on a défini. L'algorithme cherche la ligne avec la plus

grande marge entre les deux plus proches observations des différentes classes. [11]Source

3.5.4 naive Bayes

Une autre algorithme pour une problème classique - Est-ce que un email est "spam" ou "non spam"? Parce que, on est dans le domaine de learning supervise, un dataset peut donner la probabilité si un email est spam. On peut aussi compter des mots dans l'email spam et déduire quelle est la probabilité d'apparition d'un mot particulier. En utilisant la formule de Bayes

$$P(A | B) = \frac{P(B | A) P(A)}{P(B)}$$

Le modèle peut déterminer la probabilité si un email est spam, sachant que l'email contient le mot particulier. Dans notre exemple $A = \text{'email est spam'}$ et $B = \text{'un mot particulier apparait'}$.

3.5.5 Decision Trees

<https://www.youtube.com/watch?v=tNa99PG8hR8>

3.5.6 K-nearest neighbour

<https://www.youtube.com/watch?v=UqYde-LULfs>

3.5.7 Similarity Learning

<https://www.youtube.com/watch?v=0gI4dqQNNss>

4 Apprentissage non-supervisé

4.1 Fonctionnement

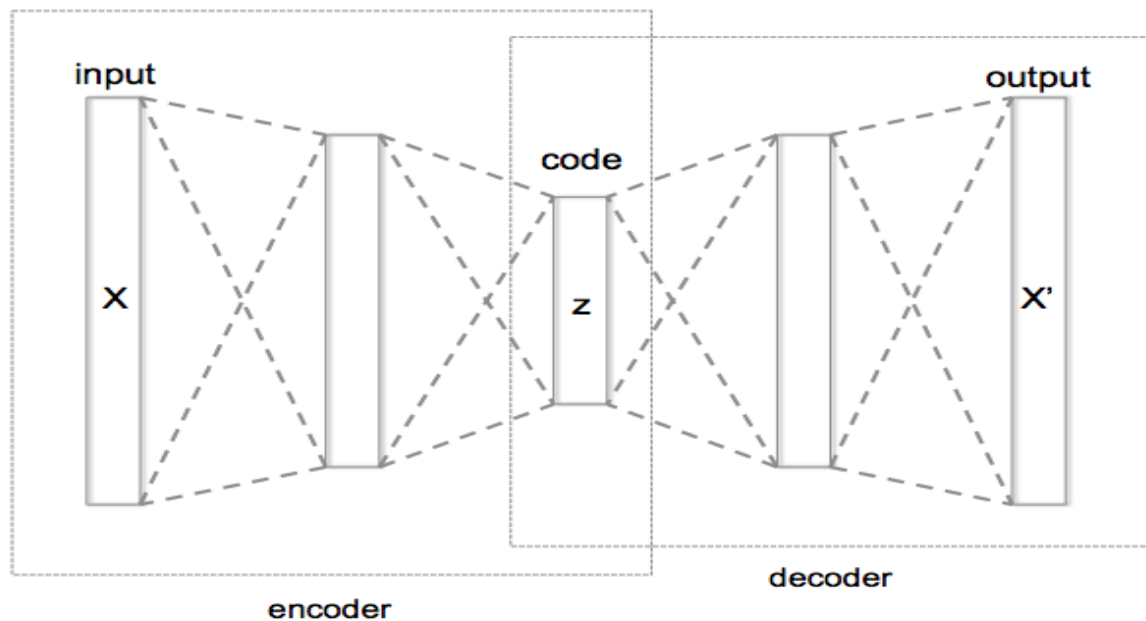
L'apprentissage non-supervisé, au contraire du modèle supervisé, apprend avec des données qui ne sont pas labellisées ni classifiées. Pour apprendre, ce modèle va utiliser les différences entre les différents sets de données et la présence de points commun.

Le nom "non-supervisé" signifie que ce modèle apprend sans superviseur, sans aucune intervention de la part d'un être humain ou d'une autre intelligence artificielle ou programme.

Ce modèle permet d'extraire des classes de données qui sont au début inconnues. Après avoir appliqué ce modèle à un set de donnée, on se retrouve avec un certain nombre de catégorie qui n'était pas connue au départ.

4.2 Quelques algorithmes

Auto-encoder : cet algorithme permet d'apprendre à un réseau de neurones à représenter des données d'une certaine manière qui lui permet de compresser celles-ci.



[3] Source de l'image

Le but de cet algorithme est d'avoir une marge d'erreur la plus petite possible entre x et x' pour ensuite extraire le vecteur z . Chacune des valeurs du vecteur z va correspondre à une caractéristique des données que le réseau a pu extraire.

En faisant, par la suite, varier les valeurs de z , il est possible de générer de nouvelles données qui vont "ressembler" aux données initiales. Chacune des valeurs de z va correspondre à une caractéristique des données.

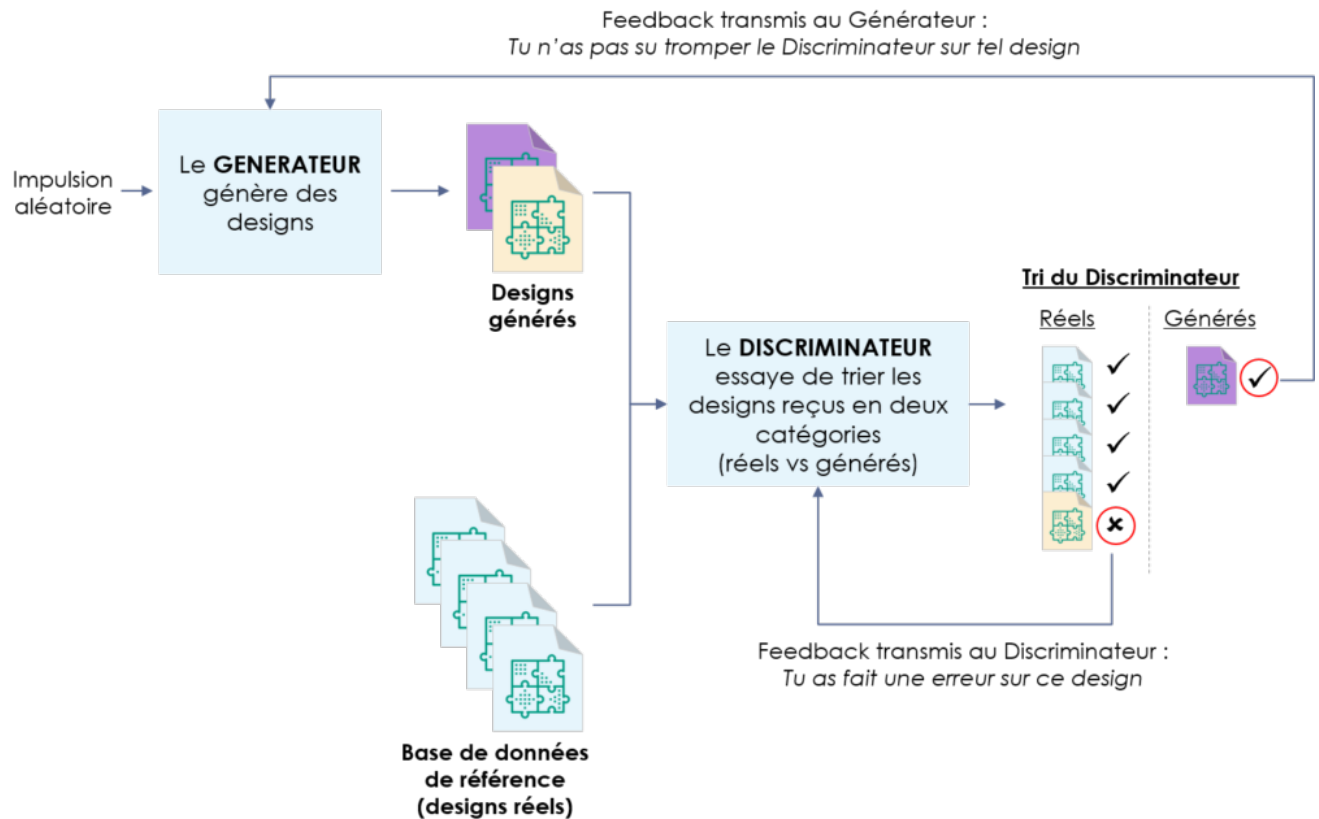
Liste d'autres algorithmes non-supervisés :

- Clustering (hiérarchique, k-moyennes, DBSCAN, OPTICS, mixture models)
- Détection d'une anomalie (LOF)
- Réseaux de neurones (Autoencodeur, DBN, Règle de Hebb, GANs, SOM)
- Approche d'apprentissage de modèles à variables latentes (EM, BSS, Méthode du moment)

4.3 Applications

Les différentes applications de l'apprentissage non-supervisé :

- Génération d'image réaliste :



[9] Source de l'image

— Auto-encodeur : reconnaissance de caractère manuscrits

Un auto-encodeur peut être entraîné pour extraire les caractéristique commune à l'écriture manuscrite pour pouvoir par la suite traiter les données beaucoup plus simplement. [5]

Sources : [10], [4]

5 Apprentissage par renforcement

5.1 Introduction

Le renforcement learning, abrégé "RL", est une branche du domaine du machine learning. Il met en oeuvre un modèle représenté par un "agent" et prédit des sorties en fonction d'entrées. Le RL est utilisé lorsqu'un agent doit évoluer dans un environnement, son apprentissage est dicté par une fonction qui par un système de point peut lui attribuer une récompense lorsqu'il réalise une bonne performance, ou une punition lorsqu'il réalise une mauvaise performance. (fig. 1)

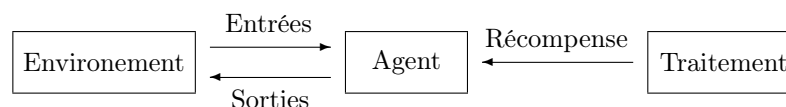


FIGURE 1 – Principe de fonctionnement

5.2 Réseau de neurones

Le réseau de neurone, appelé "Neural network" en anglais, constitue l'agent. C'est une structure de donnée qui s'appuie sur le fonctionnement biologique de neurones. Un réseau de neurones permet de traiter une information de manière probabiliste, et de générer des sorties en fonction des entrées. un réseau neuronal couplé avec les différentes

méthodes d'apprentissage, permet à l'agent d'apprendre à résoudre un problème donné. Le plus souvent, le réseau est traité comme une boîte de pandore et les seules interactions directes avec l'environnement sont les entrées et les sorties. (fig. 2)

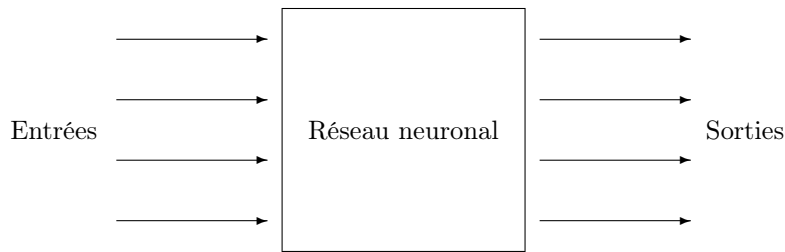


FIGURE 2 – Réseau de neurone

5.2.1 Le neurone

Un neurone est composé de une ou plusieurs entrées et d'une sortie qui correspond à la somme de ses entrées passées par une fonction. Il en existe plusieurs types comme le neurone "simple" ou suivant une table de vérité, mais le plus communément utilisé est le neurone suivant le modèle McCulloch et Pitts dit "MCP" (fig. 3). Ce neurone met à l'échelle ses entrées E_1, E_2, \dots, E_n grâce aux valeurs P_1, P_2, \dots, P_n , en fait la somme et les compare à une valeur T . Mathématiquement, un neurone MCP émet un signal si la formule suivante est vérifiée.

$$E_1P_1 + E_2P_2 + \dots + E_nP_n > T$$

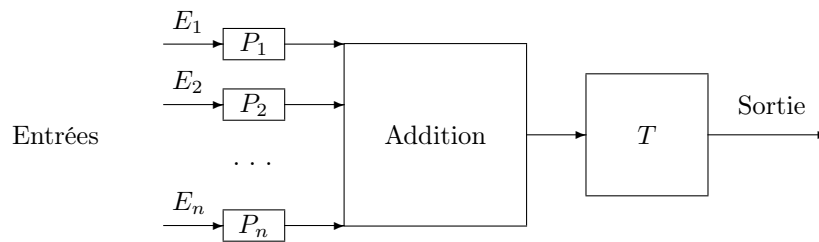


FIGURE 3 – Neurone "MCP"

En modifiant les paramètres P et T des neurones, le comportement du réseau peut être altéré. C'est par ce procédé que l'agent peut apprendre et évoluer.

5.2.2 Le réseau

Un réseau de neurone est une multitude de neurones connectés les uns aux autres. Un réseau possède une couche de neurones d'entrées, une couche masquée représentant la plus grande partie du réseau, et une couche de sortie. Il existe deux grands types de réseau, les réseaux "feedback" acceptant que les neurones puissent former des boucles, et les réseaux "feed-forward" qui arrange les neurones en couches et qui permet aux neurones d'être connectés seulement à une couche supérieure (fig. 4). On notera que la sortie d'un neurone peut être connectée à plusieurs neurones, il s'agit d'une sortie dupliquée et non pas de plusieurs sorties indépendantes.

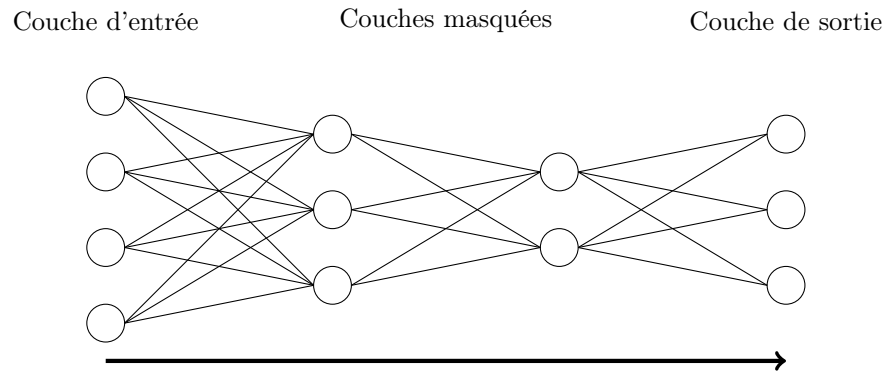


FIGURE 4 – Réseau de neurones "feed-forward"

5.3 Apprentissage

Références

- [1] 10 ways companies use machine learning. URL.
- [2] Apprentissage supervise. URL.
- [3] Autoencoder. URL.
- [4] Autoencoder wikipédia. URL.
- [5] Exemple d'utilisation d'auto-encodeur. URL.
- [6] image classification. URL.
- [7] image regression. URL.
- [8] Intelligence artificielle. URL.
- [9] Réseaux de neurones antagoniste. URL.
- [10] Unsupervised learning. URL.
- [11] MANJUNATH : 15 algos - machine learning. URL.