Cycle 00

Sciences Industrielles de l'Ingénieur

L'IA c'est quoi ?

Cours

Chapitre 1
Introduction

Savoirs et compétences :

- ❒ *Analyser les principes d'intelligence artificielle.*
  - *Régression et classification, apprentissages supervisé et non supervisé.*
  - *Phases d'apprentissage et d'inférence.*
  - *Modèle linéaire monovariable ou multivariable.*
  - *Réseaux de neurones (couches d'entrée, cachées et de sortie, neurones, biais, poids et fonction d'activation).*
- ❒ *Interpréter et vérifier la cohérence des résultats obtenus expérimentalement, analytiquement : Ordre de grandeur. Matrice de confusion (tableau de contingence), sensibilité et spécificité d'un test.*
- ❒ *Résoudre un problème en utilisant une solution d'intelligence artificielle :*
  - *Apprentissage supervisé.*
  - *Choix des données d'apprentissage.*
  - *Mise en œuvre des algorithmes (réseaux de neurones, k plus proches voisins et régression linéaire multiple).*
  - *Phases d'apprentissage et d'inférence.*

Xavier Pessoles & Anthony Meurdefroid
PSI⋆ – MP

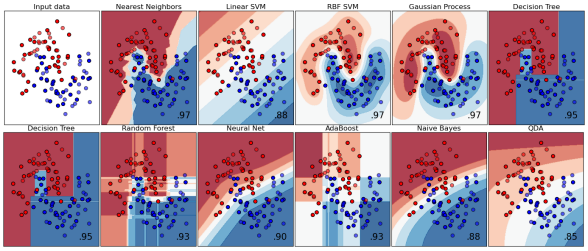
1 L'Intelligence Artificielle, qu'est ce que c'est, à quoi ça sert ?

Définition — Intelligence Artificielle. Wikipedia – L'intelligence artificielle (IA) est « l'ensemble des théories et des techniques mises en œuvre en vue de réaliser des machines capables de simuler l'intelligence ».

En première approche, l'IA peut être assimilée à un ensemble d'algorithmes permettant de réaliser des tris, des classifications.

▪ **Exemple**

Identification d'objets, en temps réel, dans une vidéo.



Comparaison du résultats d'algorithmes de classification permettant de classifier un nuage de points bleus et rouges de transparences différentes.

On remarquera sur l'exemple de gauche que l'algorithme parvient à identifier sur l'image des zones puis à identifier à quoi elles correspondent. On remarquera aussi que la flamme est identifiée comme étant un donut...

`exploring-opencvs-deep-learning-object-detection-library` & `https://scikit-learn.org/`.

▪

Pour utiliser un algorithme d'IA il est nécessaire de disposer de données... en général beaucoup de données.

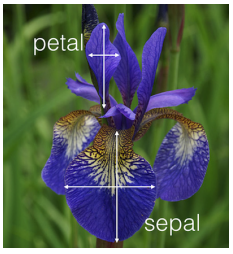
Définition — Données. Les données utilisées par un algorithme d'IA peuvent être sous différentes formes :
- des données quantitatives continues qui peuvent prendre n'importe quelles valeurs dans un ensemble de valeurs (par exemple la température) ;
- des données quantitatives discrètes qui peuvent prendre un nombre limité de valeurs dans un ensemble de valeurs (par exemple nombre de pièces d'un logement) ;
- des données qualitatives (ordinales ou nominales suivant qu'on puissent les classer ou non, par exemple des couleurs, des notes à un test d'opinion ...).

Pour pouvoir traiter ces données, il peut être nécessaire qu'elles soient organisées sous une certaine forme. On peut par exemple identifier :
- les données structurées, dans une base de données par exemple ;
- les données semi-structurées, dans un fichier csv, xml ou json par exemple ;
- les données non structurées comme un image, du texte, ou une vidéo.

Définition — Observations – Caractéristiques. On appelle observation (ou individu ou objet) une « ligne de donnée » qui va être utilisée par un algorithme d'apprentissage.
Une observation est composée de caractéristiques qui varient pour chacun des observations.

▪ **Exemple**

Dans le cas de la base de donnée des Iris, présente par exemple dans `scikit-learn`, les données sont composées de 150 observations, chacune composée de 4 caractéristiques : longueur et largeur du sépale ainsi que longueur et largeur du pétale.



▪

Définition — Big data – Wikipedia. Mégadonnées ou données massives. Le big data désigne les ressources d'informations dont les caractéristiques en termes de volume, de vélocité et de variété imposent l'utilisation de technologies et de méthodes analytiques particulières pour générer de la valeur, et qui dépassent en général les capacités d'une seule et unique machine et nécessitent des traitements parallélisés.

Que fait-on avec ces données ? En général les algorithmes vont chercher un lien entre ces données. Dans le cas où il existe un lien connu par le *Data Scientist* on parle d'apprentissage supervisé.

Si ce lien n'est pas connu, on parle d'apprentissage non supervisé ou de clustering.

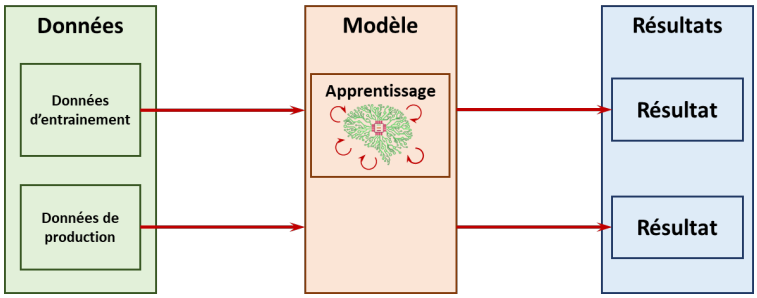
Définition — Apprentissage automatique – Machine learning. L'apprentissage automatique est un champ de l'intelligence artificielle dont l'objectif est d'analyser un grand volume de données afin de déterminer des motifs et de réaliser un modèle prédictif.

L'apprentissage comprend deux phases :

- l'entraînement (ou apprentissage) est une phase d'estimation du modèle à partir de données d'observations ;
- la mise en production du modèle est une phase pendant laquelle de nouvelles données sont traitées dans le but d'obtenir le résultat souhaité.

L'entraînement peut être poursuivi même en phase de production.

▪ Exemple

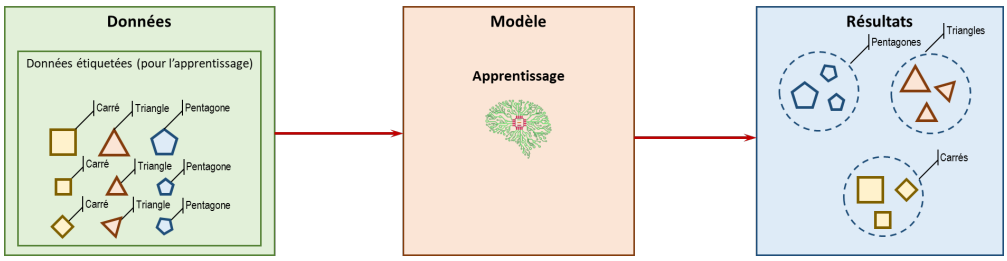


▪

Définition — Apprentissage supervisé – Induction. Tâche d'apprentissage au cours duquel l'algorithme (ou fonction de prédiction) va, à partir d'un ensemble de données **étiquetées**, déterminer un lien entre un ensemble les données et les étiquettes.

Dans un second temps, l'algorithme devra être capable de prédire l'étiquette (sans la connaître) à partir de données originales.

▪ Exemple

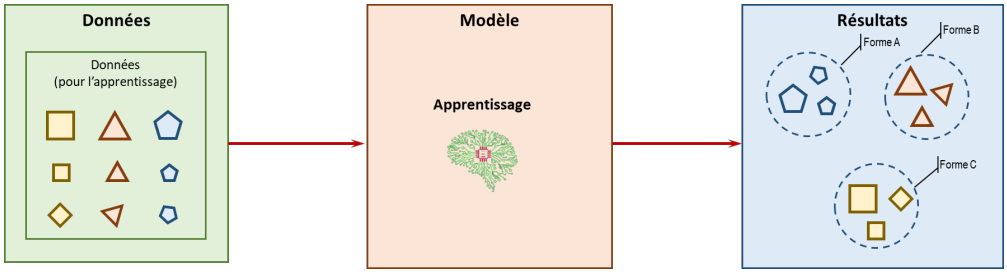


▪

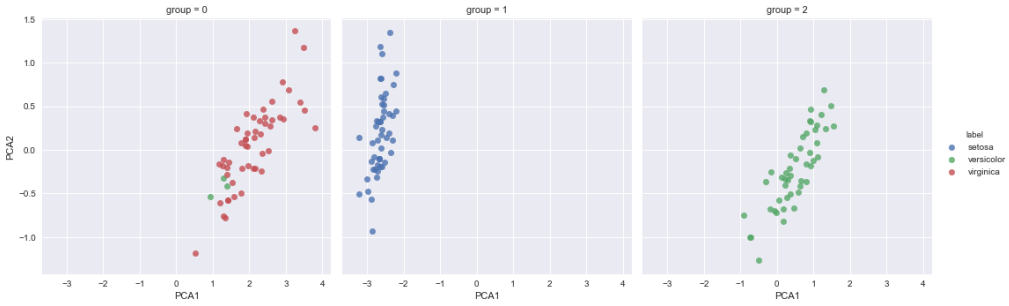
▪ Exemple Soit un ensemble d'images en noir et blanc représentant des chiffres de 0 à 9. L'algorithme doit dans un premier temps *apprendre* le lien entre l'image et le chiffre. Dans un second temps, l'algorithme devra déterminer le chiffre en fonction d'une image seule. ▪

Définition — Apprentissage non supervisé – Clustering. Tâche d'apprentissage au cours duquel l'algorithme (ou fonction de prédiction) va, à partir d'un ensemble de données **non étiquetées**, déterminer un lien entre les données (et les regrouper).

▪ **Exemple**



Classification d'iris en utilisant un algorithme d'apprentissage non supervisé.

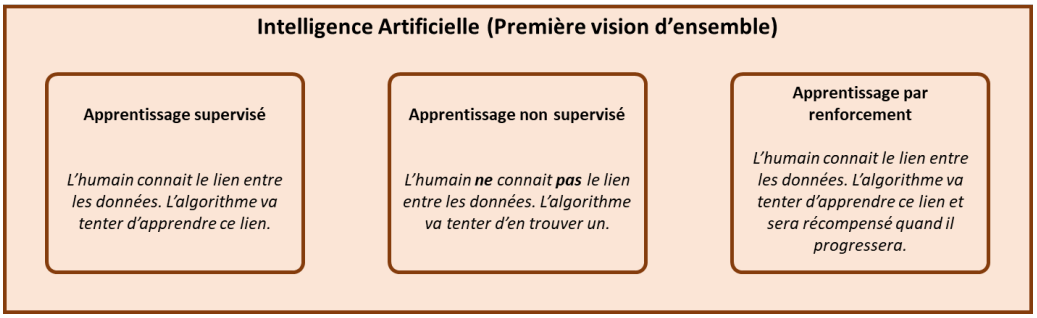


On remarque (lorsque l'image est en couleur) que sur des mesures sur 150 iris, l'algorithme a réussi à retrouver les 3 différentes espèces, sans les connaître, à 3 erreurs près.

▪

Définition — Apprentissage par renforcement. Si au cours de l'apprentissage supervisé, un mécanisme de récompense est mis en œuvre pour améliorer les performances du modèle, on parle d'apprentissage par renforcement.

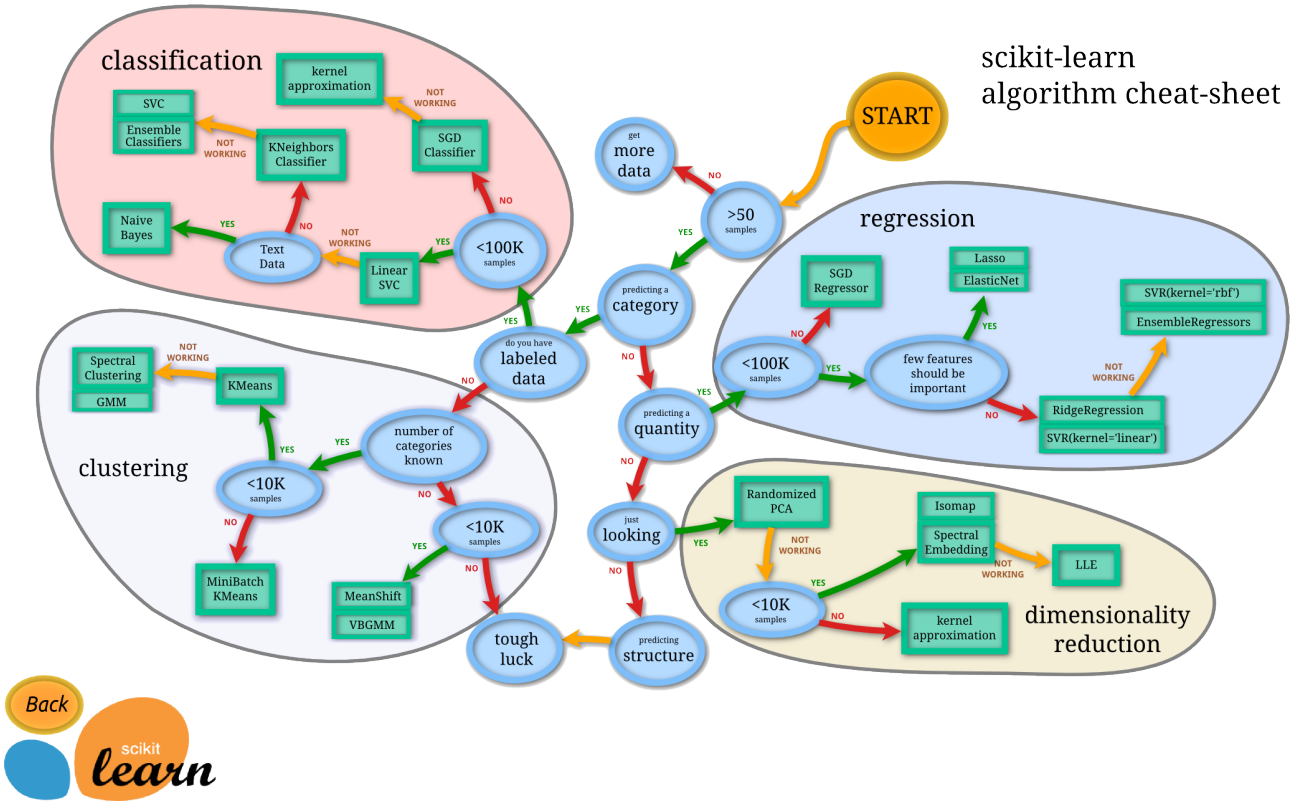
On peut donc faire une synthèse des méthodes d'apprentissage dont nous avons pris connaissance (il en existe d'autres).



Comment situer le deep learning parmi ces méthodes d'apprentissage ? – NDLR (ie XP) pour moi l'utilisation du terme « deep learning » apparaît lorsqu'on commence à utiliser des réseaux de neurones... mais les réseaux de neurones peuvent être utilisés dans chacune des méthodes d'apprentissage sus-citées...

2 Mécanismes d'apprentissages

2.1 Classification des algorithmes d'apprentissage



2.2 Validation du modèle

Une fois un algorithme d'apprentissage choisi, on se pose la question de la validation du modèle. Quels critères et outils vont nous permettre de considérer que notre apprentissage est « bon »?

Lors d'un problème de classification, il est par exemple possible de déterminer les écarts entre les valeurs prédites par l'algorithme et les valeurs cibles.

2.2.1 Critères de validation des problèmes de classification

Définition — Valeur prédictive positive. Valeur prédictive positive (*accuracy classification score*)

$$\text{Précision} = \frac{\text{Nombre de prédictions vraies}}{\text{Nombre de prédictions totales}}$$

```
import sklearn.metrics as skm

# X(numpy.ndarray) : données d'entrées
# Y(numpy.ndarray) : données cibles correspondantes
# Y_pred(numpy.ndarray) : données prédites par l'algorithme de classification

print(skm.accuracy_score(Y, Y_pred))
```

Définition — Matrices de confusion.

2.2.2 Critères de validation des problèmes de régression

2.2.3 Critères de validation des problèmes d'apprentissage non supervisé

2.3 Données et séparation des données/Gestion des données?

Lorsque l'on souhaite disposer d'un modèle défini par un algorithme d'IA, il faut en premier lieu disposer de données.

2.3.1 Types de données – Apprentissage supervisé

En apprentissage supervisé, il est nécessaire de connaître des données d'entrées ainsi que les sorties correspondantes (appelées aussi `cibles` ou `target`).

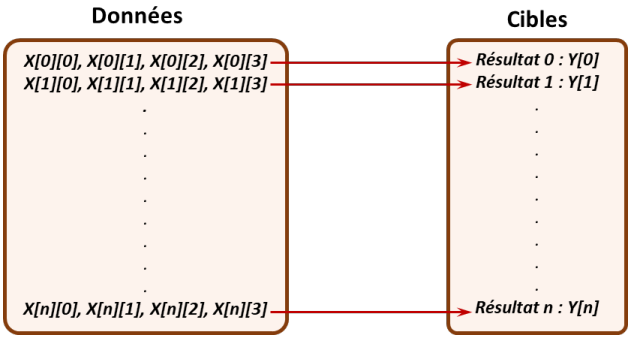
Dans le cadre d'une programmation `Python` avec la bibliothèque `scikit-learn` les données d'entées et de sorties peuvent être implémentées en utilisant des `numpy.ndarray`.

Ainsi, si les données d'entrées, stockées dans la variable `data` sont composées de `n` observations elles mêmes composées de `p` catégories, le shape de `data` sera `(n,p)`.

La donnée cible sera quant à elle un vecteur de `n` valeurs.

▪ **Exemple**

Données d'entrées : matrice `X` de `n` observations avec 4 caractéristiques.
Données de sortie : vecteur `Y` de `n` résultats.



Dans le cadre d'une classification, où `r` résultats sont possibles. On peut attribuer une valeur (entière) entre 1 et `r` à chacun des résultats, notamment si ceux-ci sont des données qualitatives.

Problème de l'identification d'un système : quelles sont les données d'entrées? Quelles sont surtout les données de sorties? Un vecteur aussi?

▪

2.3.2 Normalisation des données

`https://scikit-learn.org/stable/modules/preprocessing.html`

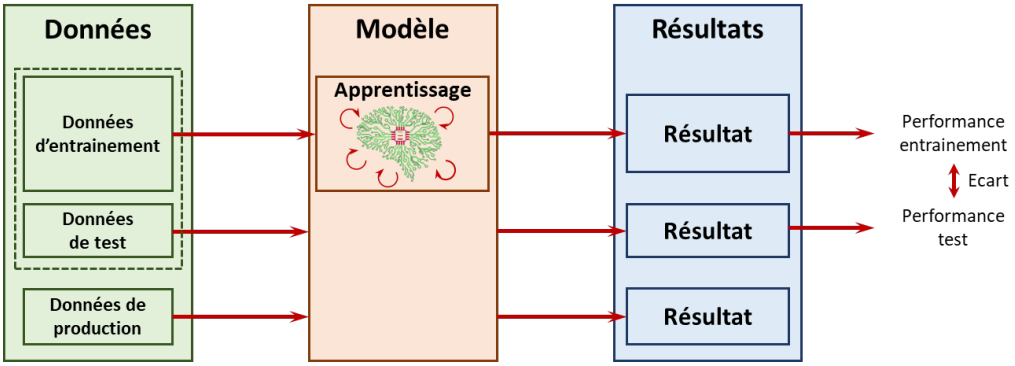
2.3.3 Séparation des données

Lors du démarrage d'un apprentissage supervisé, on dispose d'un jeu de données comprenant les données d'entrée et les cibles correspondantes. Il est d'usage de séparer ces données en deux parties :

- les données d'entraînement permettant... d'entraîner le modèle. Dans la pratique on utilise entre 60 et 80 % des données de base ;
- les données de te test, permettant de valider l'apprentissage (ou le modèle), dans la pratique entre 20 et 40 % des données de base.

On commence donc par réaliser un apprentissage sur les données d'entraînement. Cet entraînement produit des résultats. Connaissant les cibles correspondant aux données d'entraînement, on peut donc en déduire une performance du modèle sur le traitement des données d'entraînement.

On utilise alors le modèle en lui donnant les données de test. De même, on peut donc en déduire une performance du modèle sur le traitement des données de test.



Se pose alors le problème de comment séparer les données. En effet, les données de base pouvant potentiellement être triées (ordonnées par rapport à une des caractéristiques), et sélectionner une partie pourrait créer un biais dans l'apprentissage.

`scikit-learn` permet de séparer aléatoirement des données en fixant le pourcentage de données de validation.

```
from sklearn.model_selection import train_test_split
# X(numpy.ndarray) : données d'entrées
# Y(numpy.ndarray) : données cibles correspondantes
X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size=0.33)
```

On perçoit donc que la qualité de l'apprentissage peut dépendre du choix utilisé lors de la séparation des données. De plus, certains choix de paramètres permettant d'affiner le modèle sont aussi liés aux données d'entraînement sélectionnées. Une des solutions pour résoudre ce problème est d'avoir recours à la validation croisée.

À creuser

`https://scikit-learn.org/stable/modules/cross_validation.html#cross-validation`

3 Algorithmes d'apprentissage

4 Réseaux de neurones

5 Définitions

- Data scientist
- machine learning
- deep learning
- réseaux de neurons
- régression
- data mining
- bigdata
- données continues, données discrètes, données nominales, données ordinales, données (semi-)structurées et non structurées
- algorithmes supervisés, non supervisés
- algorithmes de régression et de classification

Définition — Machine Learning – Apprentissage automatique – Wikipedia. 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.

La première phase de l'apprentissage consiste à estimer un modèle à partir de données, appelées observations, qui sont disponibles et en nombre fini, lors de la phase de conception du système.

La seconde phase correspond à la mise en production : le modèle étant déterminé, de nouvelles données peuvent alors être soumises afin d'obtenir le résultat correspondant à la tâche souhaitée.

Définition — Apprentissage supervisé – Apprentissage non supervisé – Apprentissage par renforcement – Wikipedia. Si les données sont étiquetées (c'est-à-dire que la réponse à la tâche est connue pour ces données), il s'agit d'un apprentissage supervisé. On parle de :

- classification ou de classement si les étiquettes sont discrètes ;
- régression si elles sont continues.

Si le modèle est appris de manière incrémentale en fonction d'une récompense reçue par le programme pour chacune des actions entreprises, on parle d'apprentissage par renforcement.

Dans le cas le plus général, sans étiquette, on cherche à déterminer la structure sous-jacente des données (qui peuvent être une densité de probabilité) et il s'agit alors d'apprentissage non supervisé.

5.1 Quelques définitions

Définition Intelligence Artificielle, première approche

5.2 Le nerf de la guerre, les données

5.3 Méthode de résolution de problèmes d'apprentissage supervisé

1. Choix des données.
2. Normalisation des données.
3. Séparation des données ? (entraînement, test, validation)
4. Choix de la méthode d'entraînement (choix d'un modèle, en fonction du type de données, choix des paramètres du modèle)
5. Entraînement du modèle
6. Test du modèle
7. Observation des métriques et visualisation des résultats

6 Méthode de résolution d'un

Exemples : `https://makina-corpus.com/blog/metier/2017/initiation-au-machine-learning-avec-python`

Références

[1] Éric Biernat et Michel Lutz. *Data science : fondamentaux et études de cas.* Eyrolles.