Université Constantine 2
جامعة قسنطينة 2

Apprentissage machine 1

Chapitre 1 : Introduction

Ouadfel Salima
Faculté NTIC/IFA
salima.ouadfel@univ-constantine2.dz

Université Constantine 2
جامعة قسنطينة 2

Apprentissage machine 1

Chapitre 1 : Introduction

Ouadfel Salima

Faculté NTIC/IFA

salima.ouadfel@univ-constantine2.dz

Etudiants concernés

Faculté/Institut	Département	Niveau	Spécialité
Nouvelles technologies	IFA	Master 1	STIC

Introduction

Semestre *:* 01

UE : APM1

Intitulé de matière: Apprentissage machine 1

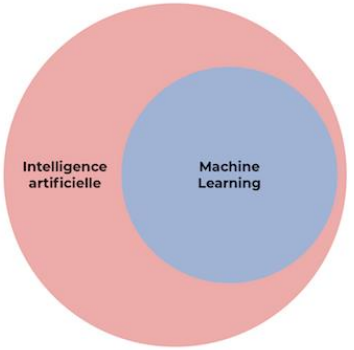
Objectifs

- Acquérir les connaissances de base de l'apprentissage machine
- Maitriser quelques techniques d'apprentissage machine

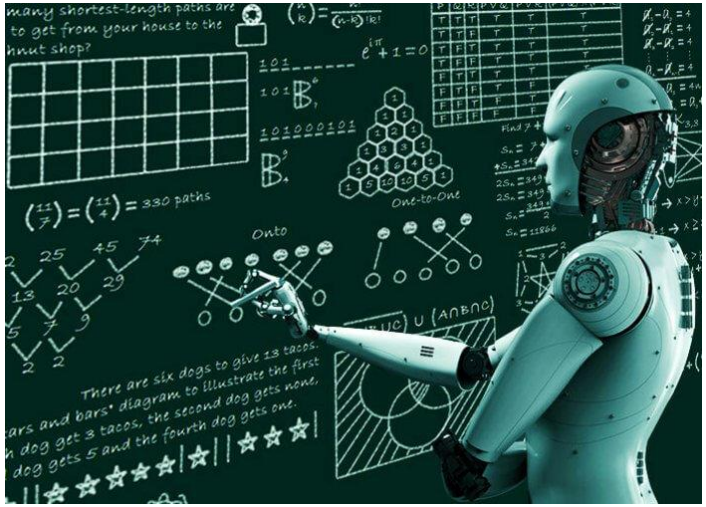
C’est quoi l’apprentissage machine?

Définition 1

- **L’apprentissage machine** est un sous domaine de l’intelligence artificielle.



- **L’intelligence artificielle** (artificial intelligence) est un paradigme qui a pour objectif de doter les machines de capacités des humaines à résoudre des problèmes complexes.



C’est quoi l’apprentissage machine?

Définition2

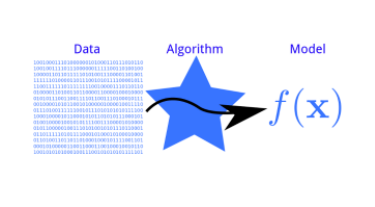
L’apprentissage machine (machine learning) est un ensemble d’algorithmes qui permettent à la **machine d’apprendre à partir des données à résoudre des problèmes sans être explicitement programmée. (Arthur Samuel, 1959)**

Programmation traditionnelle

Pour résoudre un problème, **on doit écrire explicitement (des instructions) un programme** informatique qui décrit toutes les étapes nécessaire à sa résolution. Tous les cas particuliers doivent être considérés,
On présente au programme les données pour qu’il nous donne les sorties attendues,

Machine learning

Pour résoudre un problème, **on entraine un programme informatique sur des données** (des exemples), le programme s’adapte automatiquement à chaque donnée qui représente un cas particulier,
On entraine la machine sur des données pour générer un modèle



C’est quoi l’apprentissage machine?

Définition3

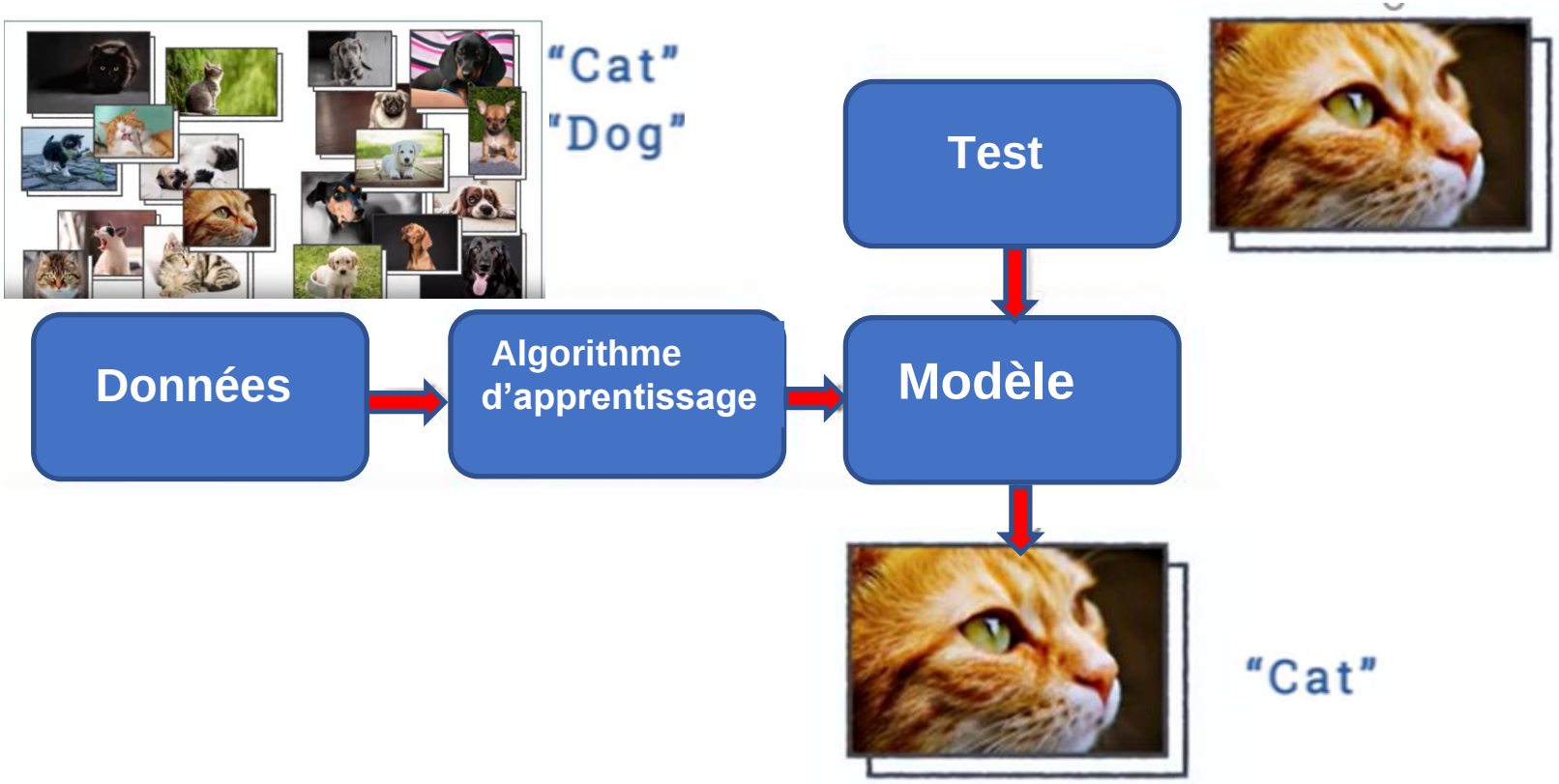
A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with **experience E. (Tom Mitchell (1998))**

On veut utiliser l’apprentissage automatique pour distinguer entre les images de chats et celles des chiens,

- 1- la tache : on veut classer les images chien/chat
- 2- expérience: on présente une à une les images
- 3- performance: le nombre des images de chien et de chat bien classées

A chaque nouvelle expérience, le modèle est ajusté ce qui permet d’améliorer ces performances.

Exemple: classification chat / chien

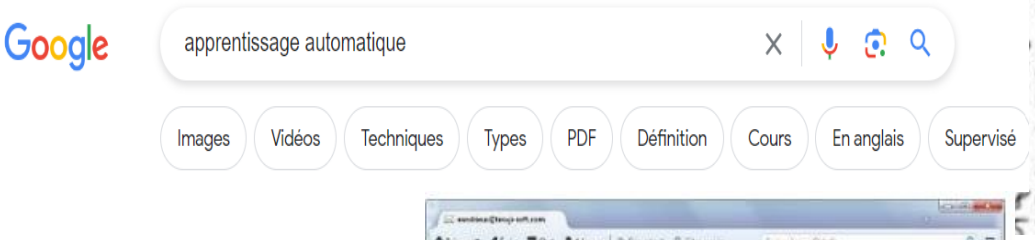


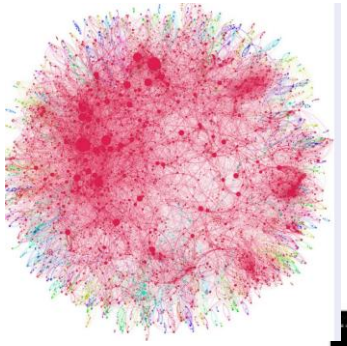
Introduction

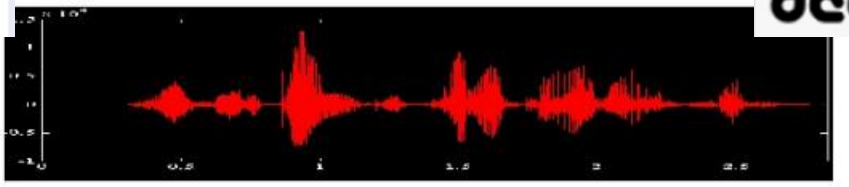
Quand appliquer l'apprentissage machine?

- Pas de modèle analytique disponible
- Beaucoup de cas particuliers pour résoudre un problème donné.
- Disponibilité de données volumineuses et de bonnes qualités

Applications de l'apprentissage machine







Etapes de l'apprentissage machine

1. Définition claire de l'objectif attendu
2. Collecte et représentation les données (Volume et qualité)
3. Prétraitement des données (optionnel)
4. Partage des données en deux sous/ensembles: entrainement et test
5. Choix du modèle d'apprentissage (entrainement)
6. Entrainement le modèle sur les données d'entrainement,
7. Evaluation de la performance du modèle sur les données de test

Collecte et représentation des données

Les données sont un composant crucial à l'apprentissage automatique.

Les données sont aussi appelées instances(samples) ou des exemples

On regroupe les données dans un dataset (ou jeu de données)

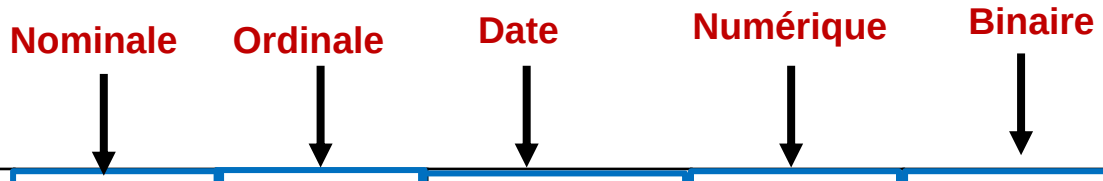
Collecte et représentation des données

Les données peuvent être de deux types: structurées et non structurées.

- **Les données structurées** sont organisées généralement dans des tableaux ou des feuilles de calcul, etc..
- **Les données non structurées** ne présentent pas une structure spécifique: les images, les vidéos, texte etc…

Collecte et représentation des données

Les données structurées sont constituées d'attributs (features) qui peuvent être de types différents: numériques,binaires, date, nominales, ordinales,….

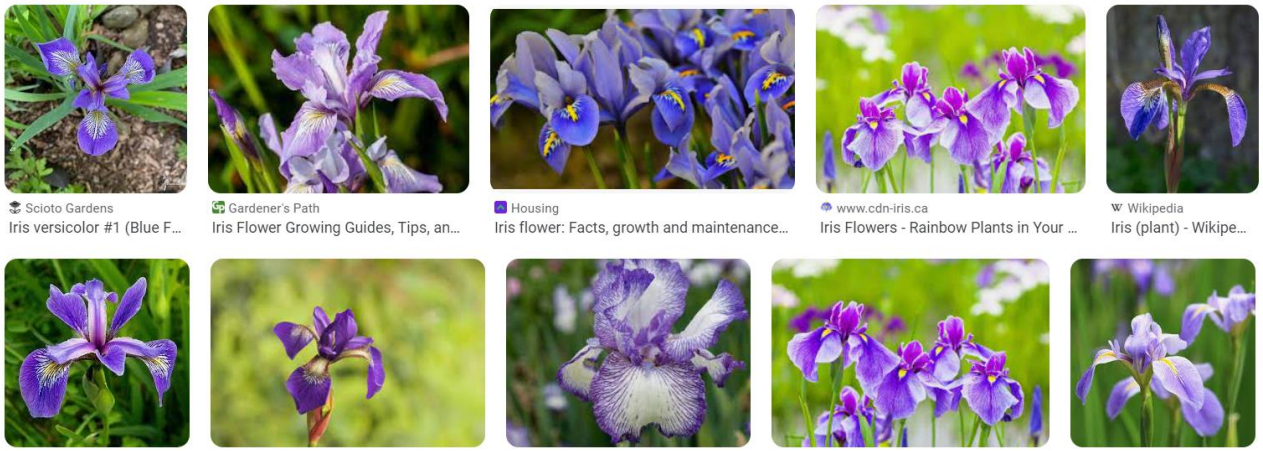


Prénom	Année	Groupe	Date d'inscription	Moyenne	Redoublant
Imane	L1	1	24/08/2023	12.25	0
Mourad	L1	2	24/09/2023	10.45	1
Akram	L1	3	29/08/2023	15.78	0

Collecte et représentation des données

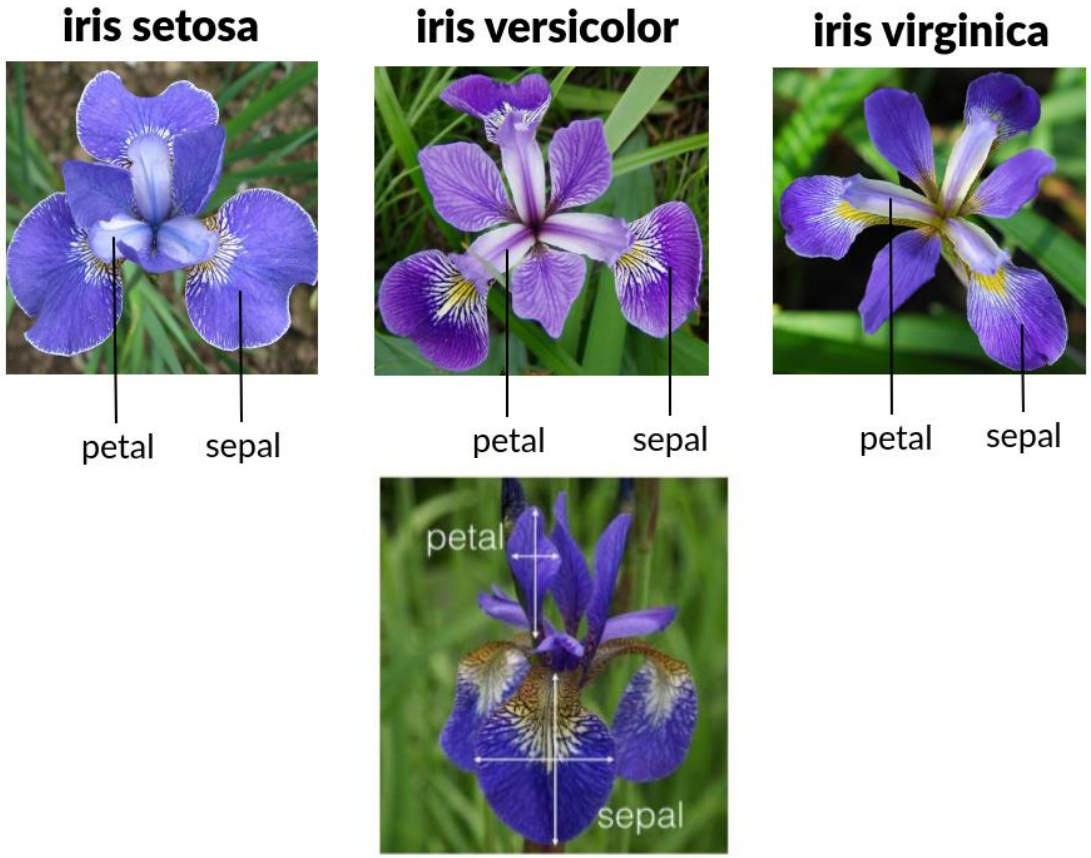
Exemple: classification des fleurs d'iris
en 3 classes: 1-Versicolor
2- Setosa
3- Virgina

1- Collecter des images des fleurs d'iris de chacune des classes



Collecte et représentation des données

2. Chaque image est représentée par des features représentant chacune des classes d'images.



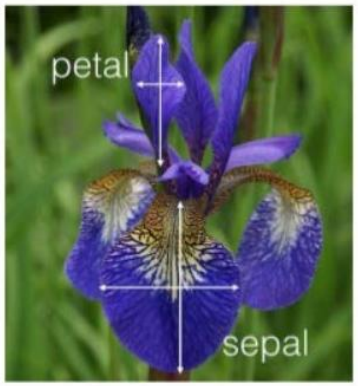
Collecte et Représentation des données

4 Features / 4 caractéristiques

Etiquette / label

Instances / samples

X		SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species
x_0	1	5.1	3.5	1.4	0.2	Iris-setosa
x_1	2	4.9	3.0	1.4	0.2	Iris-setosa
x_2	3		3.2	1.3	0.2	Iris-setosa
	4	4.6	3.1	1.5	0.2	Iris-setosa
	5	5.0	3.6	1.4	0.2	Iris-setosa
	...	...	...	...	...	...
	146	6.7	3.0	5.2	2.3	Iris-virginica
	147	6.3	2.5	5.0	1.9	Iris-virginica
	148	6.5	3.0	5.2	2.0	Iris-virginica
	149	6.2	3.4	5.4	2.3	Iris-virginica
x_{149}	150	5.9	3.0	5.1	1.8	Iris-virginica



$$\text{Donnée } \mathbf{X}_i = \begin{pmatrix} x_{i,1} & \cdots & x_{i,d} \end{pmatrix}^\top$$

i=1..150
d=1..5

$$\mathbf{X} = \begin{pmatrix} x_{1,1} & \cdots & x_{1,d} \\ \vdots & & \vdots \\ x_{n,1} & \cdots & x_{n,d} \end{pmatrix} \in \mathbb{R}^{n \times d}$$

Collecte et représentation des données

Exemple: Prédiction des prix des maisons

1- Collecter des images des maisons

Collecte et représentation des données

2. Extraire les features représentant chacune des maisons

- **Le prix de chaque maison dépend par exemple de :**
  - Superficie de la maison
  - Nombre de chambres dans la maison
  - Nombre de balcons
  - Age de la maison, ….etc

Collecte et représentation des données

3. Représentation des données

4 Features / 4 caractéristiques

Etiquette / label

Instances / samples

X

	A	B	C	D	E
x_0 1	area	bedrooms	balcony	age	price
x_1 2	1200	2	0	2	500000
x_2 3	2300	3	2	5	620000
4	2500	4	2	1	122500
5	3650	5	3	3	6000000
6	1800	3	1	5	2122000
7	3000	3	1	4	120000
8	1222	1	0	2	450000
9	4600	5	3	1	6500000
10	2050	2	2	2	1530000
x_{10} 11	1450	2	2	3	1563330

$$\text{Donnée } \mathbf{X}_i = \begin{pmatrix} x_{i,1} & \cdots & x_{i,d} \end{pmatrix}^\top$$

n=10
d=1..5

$$\mathbf{X} = \begin{pmatrix} x_{1,1} & \cdots & x_{1,d} \\ \vdots & & \vdots \\ x_{n,1} & \cdots & x_{n,d} \end{pmatrix} \in \mathbb{R}^{n \times d}$$

Prétraitement des données

Les algorithmes d'apprentissage machine ont besoin d'un volume important de données pour améliorer leur performance. Si le nombre de données est insuffisant pour lancer l'apprentissage, on utilise principalement deux techniques:

- **Augmentation des données** : générer artificiellement de nouvelles données à partir de données existantes en appliquant certaines transformations telles que des opérations géométriques.
- **Transfert d'apprentissage** : utiliser des algorithmes de l'apprentissage machine pré-entrainés pour résoudre un problème similaire ou connexe avec un réglage fin des paramètres pour former un nouveau modèle qui sera par la suite appliqué sur les nouvelles données .

Prétraitement des données

Nettoyage des données

Dans certains cas, les données collectées dans leur état brut ne peuvent pas être directement utilisées dans les algorithmes de l'apprentissage machine. Il est nécessaire de passer par une étape de nettoyage pour les filtrer. Parmi les opérations de nettoyage, on trouve:

- Traitement des données manquantes
- Gestion des données aberrantes
- Suppression des données en double
- Suppression des données non pertinentes
- Gestion des données déséquilibrées

Prétraitement des données

Transformation des données

Dans certains cas, on doit aussi convertir les données brutes dans un format adapté aux algorithmes d'apprentissage automatique. Par exemple

- Normalisation
- Codage
- Réduction de la dimensionnalité

Partage des données

Afin de préparer les données pour réparation des données pour l'apprentissage automatique, on divise toutes les données collectées (et éventuellement prétraitées) en deux sous-ensembles:

1- sous-ensemble d'entrainement

Il est utilisé pour générer le modèle d'apprentissage automatique. Cet ensemble de données représente généralement 80% du volume total des données.

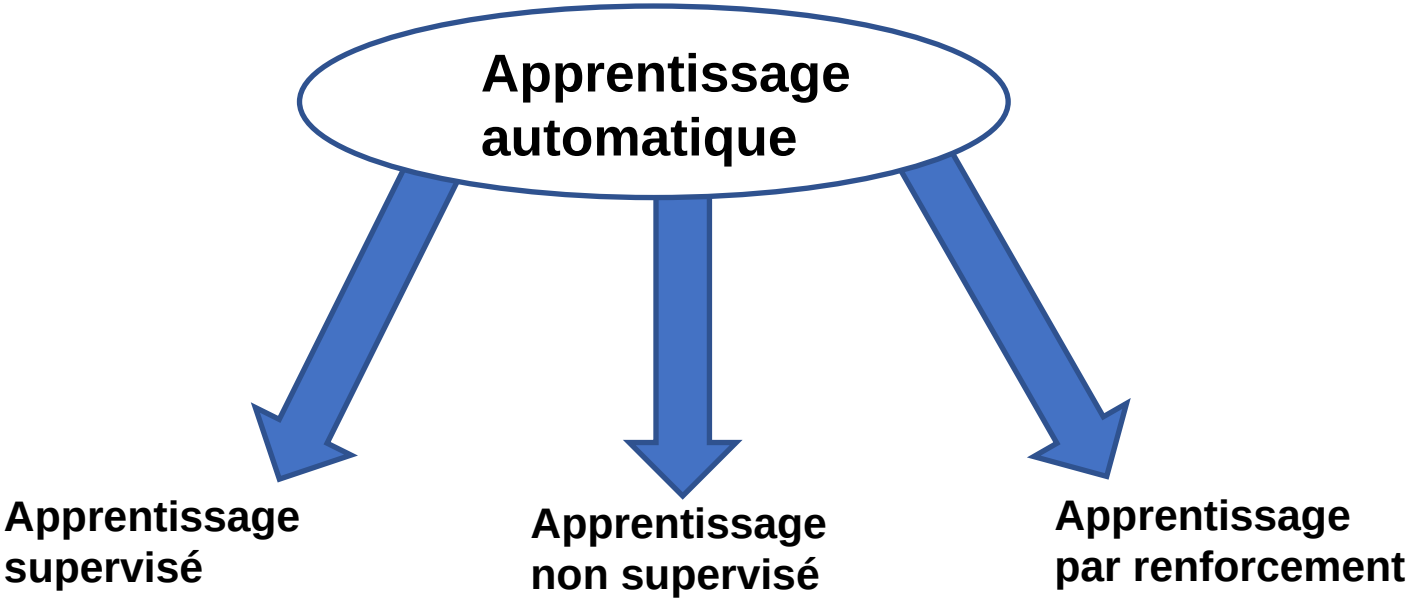
2-sous-ensemble de test

Il est utilisé pour évaluer les performances du modèle obtenu à la suite de l'étape d'entrainement. Son objectif est de tester la performance du modèle sur de nouvelles données qu'il n'a pas vu durant la phase de l'apprentissage.

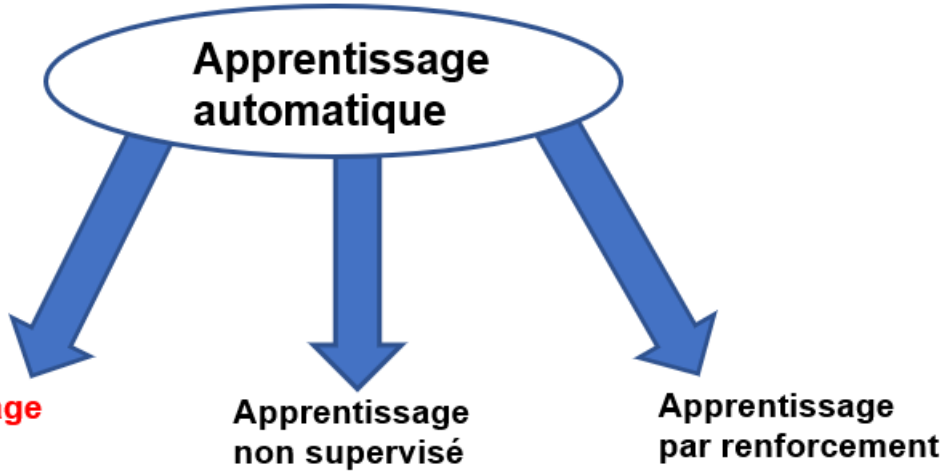
On peut aussi extraire du sous ensemble d'entrainement, un **sous-ensemble de validation** pour corriger les paramètres du modèle durant la phase d'apprentissage.

Choix de l'algorithme d'apprentissage

Types d'apprentissage automatique



Types d'apprentissage automatique

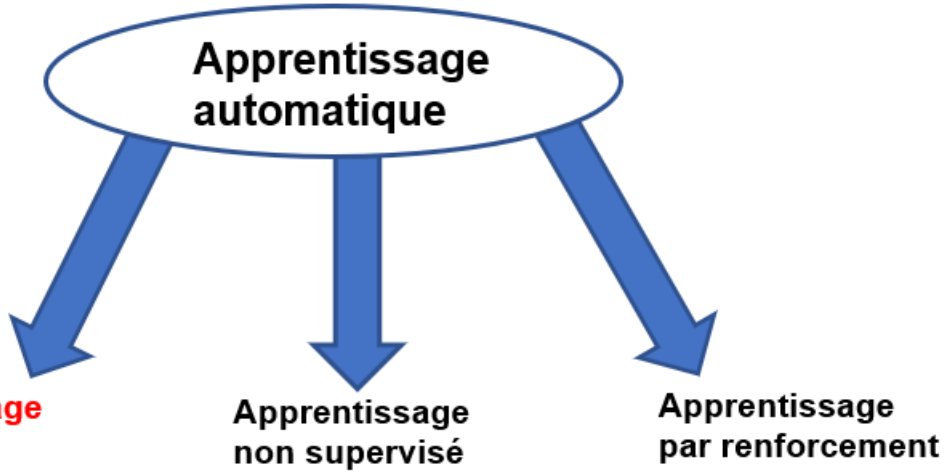


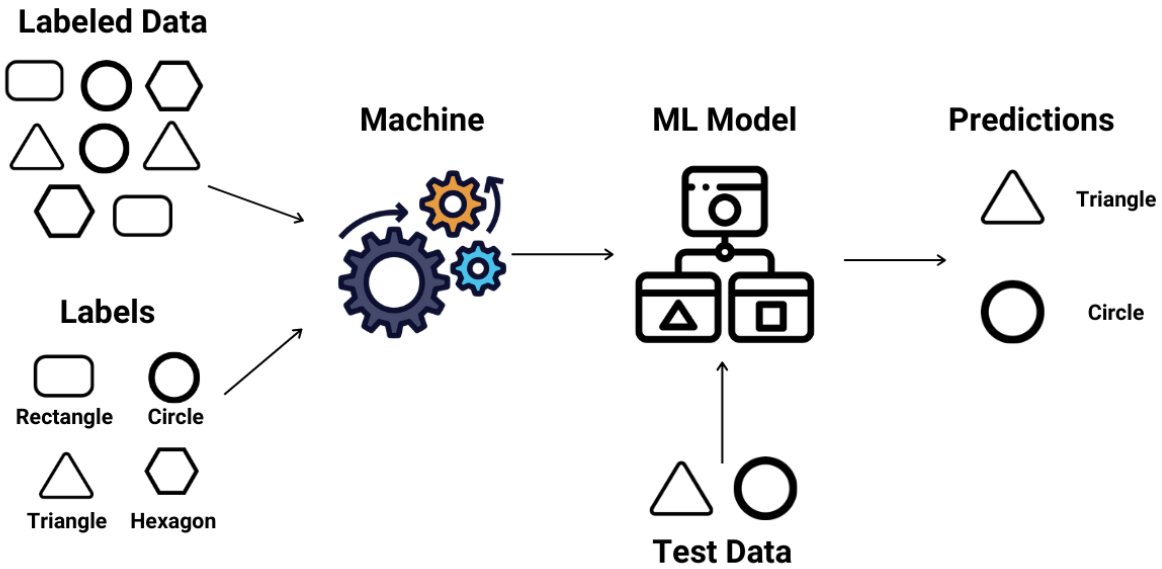
L'apprentissage supervisé (*supervised learning*) consiste à générer **un modèle de prédiction** à partir de données annotées. L'annotation consiste à affecter à chaque donnée une étiquette qui est la réponse à prédire.

En apprentissage supervisé, l'algorithme cherche à minimiser pour chaque donnée, l'erreur entre sa prédiction et la vraie réponse.

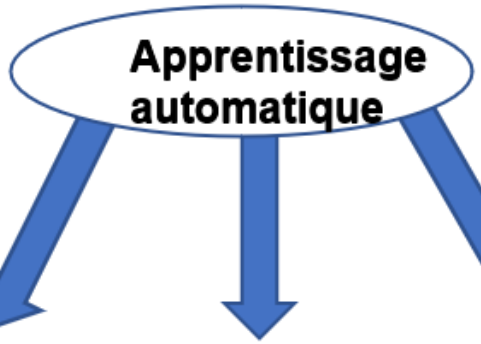
C'est une analyse prédictive

Types d'apprentissage automatique





Types d'apprentissage automatique



Apprentissage supervisé

Apprentissage non supervisé

Apprentissage par renforcement

test

SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm
5.9	3.0	5.1	1.8

SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm
5.1	3.5	1.4	0.2
4.9	3.0	1.4	0.2
4.7	3.2	1.3	0.2
4.6	3.1	1.5	0.2
5.0	3.6	1.4	0.2
...	...	...	...
6.7	3.0	5.2	2.3
6.3	2.5	5.0	1.9
6.5	3.0	5.2	2.0

Données d'apprentissage

Species
Iris-setosa
Iris-setosa
Iris-setosa
Iris-setosa
Iris-setosa
...
Iris-virginica
Iris-virginica
Iris-virginica

Labels

Apprentissage

Modèle

Prédiction

Species
Iris-setosa

label

Types d'apprentissage automatique

Apprentissage automatique

- **Apprentissage supervisé**
- Apprentissage non supervisé
- Apprentissage par renforcement

test

area	bedrooms	balcony	age
4600	5	3	1
2050	2	2	2
1450	2	2	3

area	bedrooms	balcony	age
1200	2	0	2
2300	3	2	5
2500	4	2	1
3650	5	3	3
1800	3	1	5
3000	3	1	4
1222	1	0	2

Données d'apprentissage

price
500000
620000
122500
6000000
2122000
120000
450000

Labels

Apprentissage

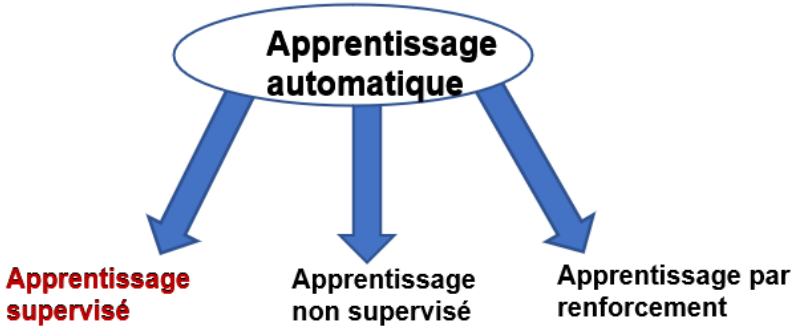
Modèle

Prédiction

price
6500000
1530000
1563330

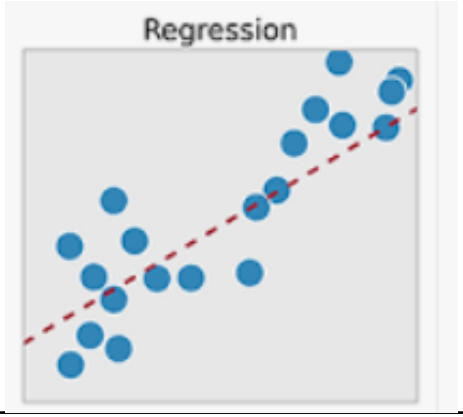
label

Types d'apprentissage automatique

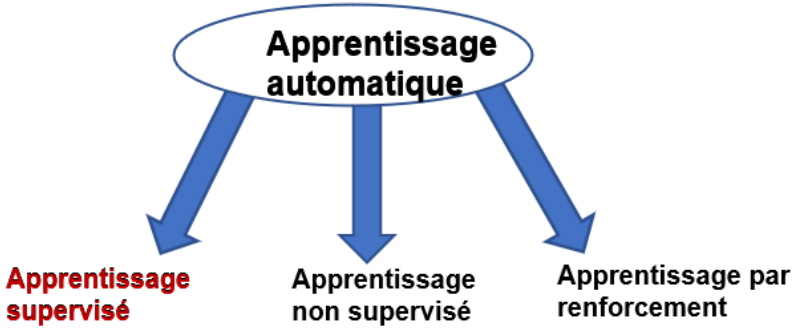


Régression

Problèmes pour lesquels, on cherche à **prédire une valeur quantitative (valeur continue):** prédiction du prix de maisons en fonction de ces caractéristiques , prédiction de la note d'un examen on fonction du nombre d'heures de révision;……

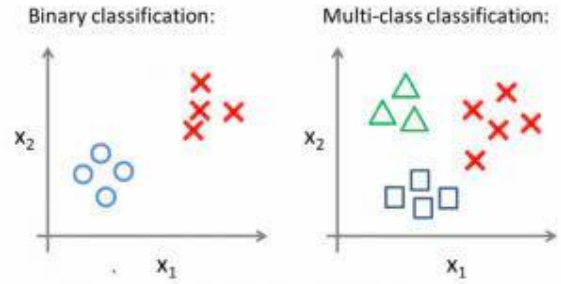


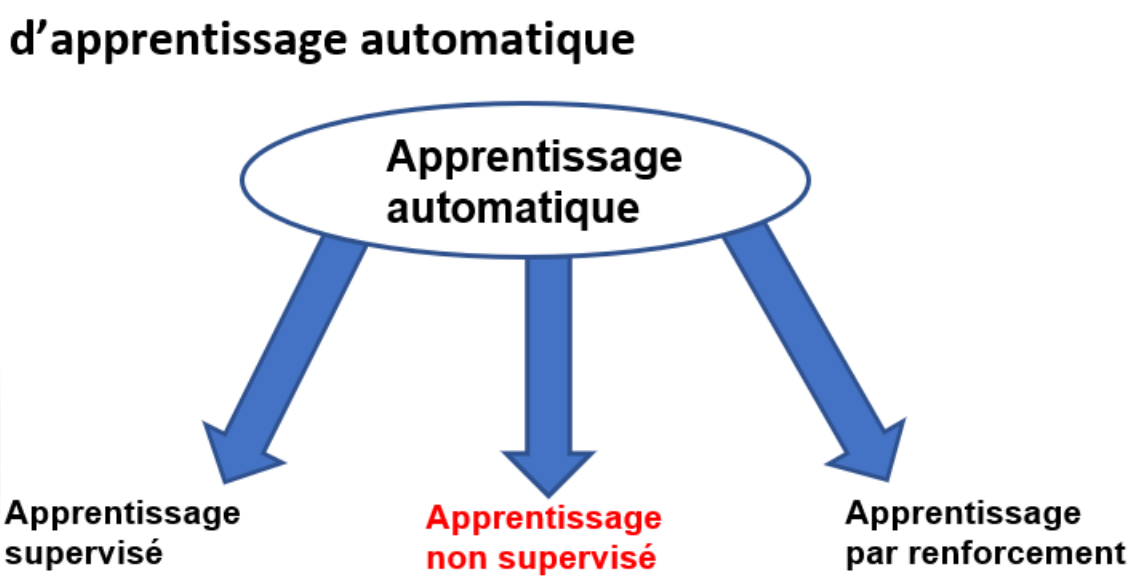
Types d'apprentissage automatique



Classification

Problèmes pour lesquels, on cherche à **prédire une valeur qualitative (valeur discrete):** type d'une fleur, email spam ou non spam, la pluie tombera ou non demain,…….

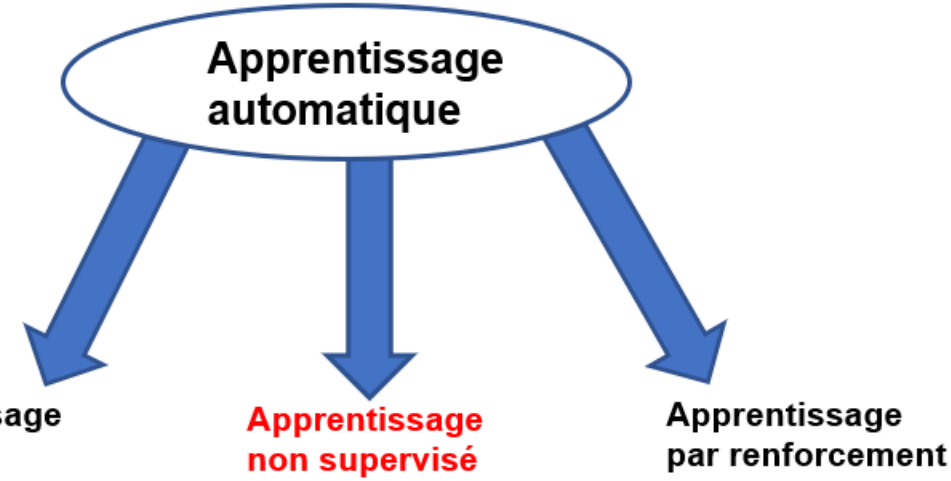


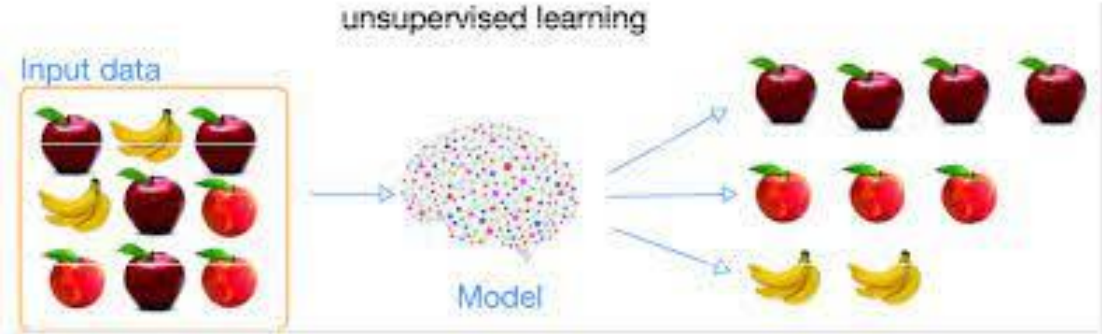


En apprentissage non supervisé (*unsupervised learning*), l'algorithme prend en entrée des données non annotées (sans leur label) et découvre par lui-même des similarités ou des différences entre les données à partir des features qui les décrivent.

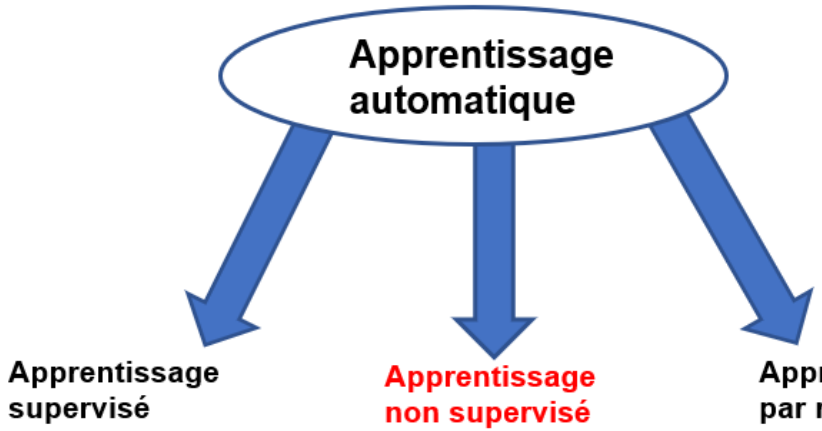
C'est une analyse descriptive

Types d'apprentissage automatique





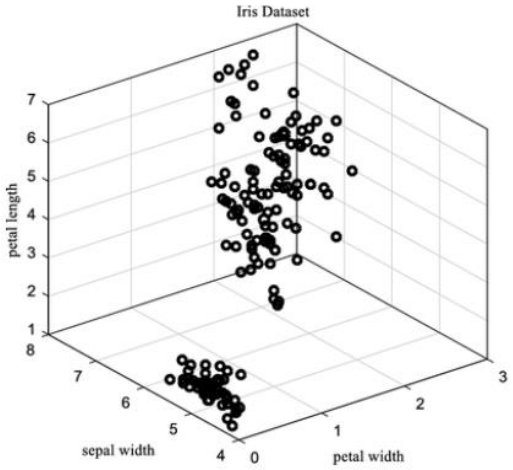
Types d'apprentissage automatique



test

SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm
5.9	3.0	5.1	1.8

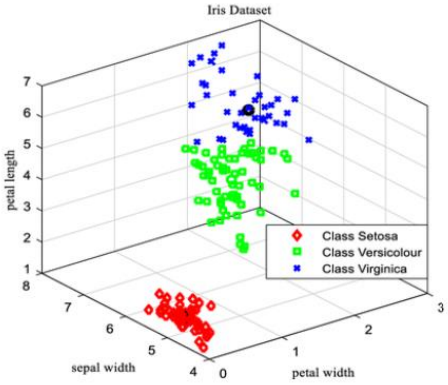
SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm
5.1	3.5	1.4	0.2
4.9	3.0	1.4	0.2
4.7	3.2	1.3	0.2
4.6	3.1	1.5	0.2
5.0	3.6	1.4	0.2
...	...	...	...
6.7	3.0	5.2	2.3
6.3	2.5	5.0	1.9
6.5	3.0	5.2	2.0



Données d'apprentissage

Modèle

Apprentissage

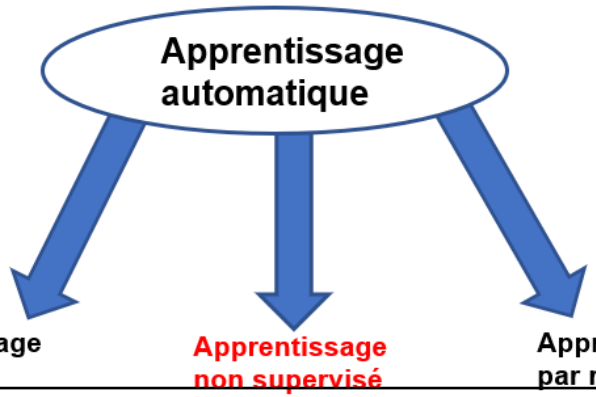


Species
Iris-setosa

label

Introduction

Types d'apprentissage automatique

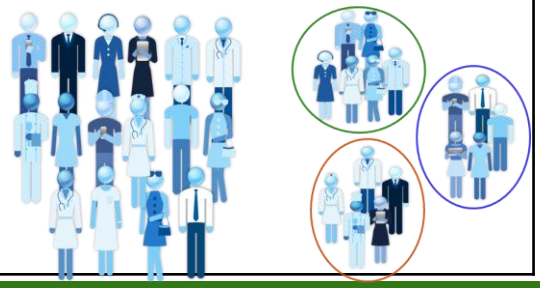


Clustering

Problèmes pour lesquels, on cherche à créer des groupes de données selon leurs caractéristiques communes. En commerce électronique, le clustering consiste à regrouper les clients en des groupes ou clusters selon la similarité de leur comportement d'achat. Un autre exemple est le regroupement des utilisateurs Facebook en des communautés (clusters) selon les profiles des internautes.

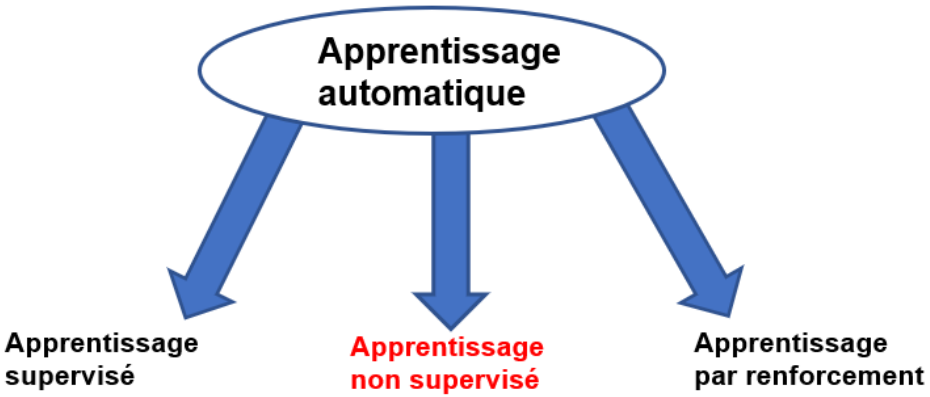
Commerce électronique

Détection de communauté dans Facebook



Introduction

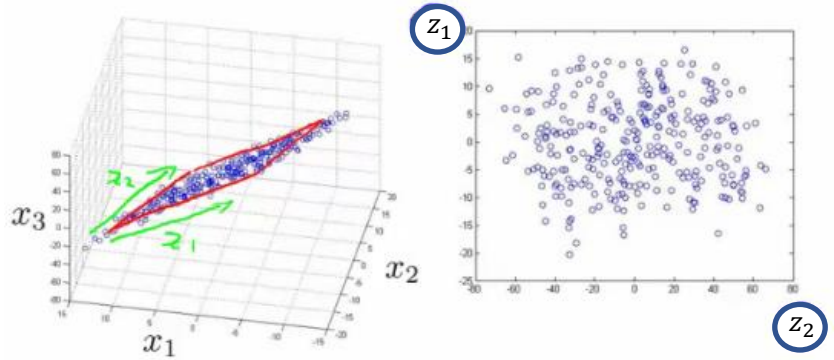
Types d'apprentissage automatique



Réduction de dimension

On cherche à réduire les dimension de l'ensemble de données en appliquant une opération de projection (donc transformation) d'un espace de dimension D dans un espace de dimension d telle que $d \ll D$.

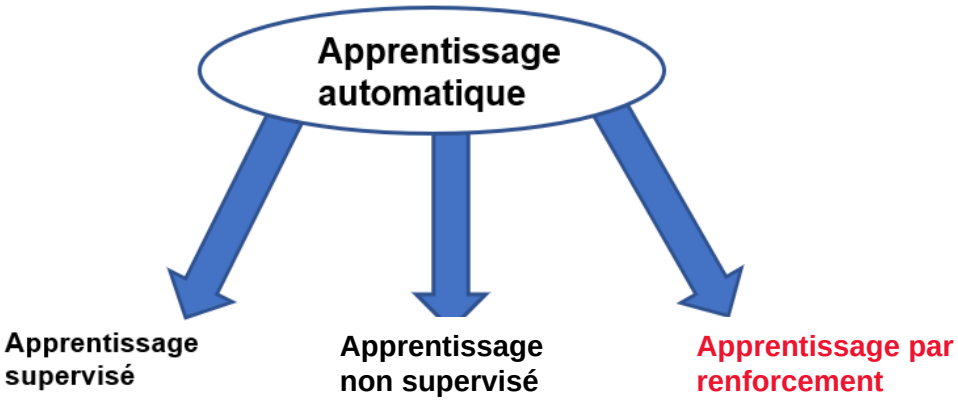
D=3



d=2
Les axes z_1 et z_2 sont différents des axes x_1, x_2 et x_3

Introduction

Types d'apprentissage automatique



En apprentissage par renforcement (*reinforcement learning*) est un type d'apprentissage non supervisé où un agent apprend un comportement à partir d'expériences tout en maximisant des récompenses et minimisant des pénalités .

Evaluation des performances du modèle

Elle dépend du type de prédiction

- Régression

$$MSE = \frac{1}{n}\Sigma\left(y - \hat{y}\right)^2$$

n: nombre de données
y: réponse réelle (annotation)
$\hat{y}$: réponse prédite

- Classification

Matrice de confusion

		Predicted 1	Predicted 0
Actual	1	TP	FN
	0	FP	TN

$$Précision = \frac{TP}{TP + FP}$$

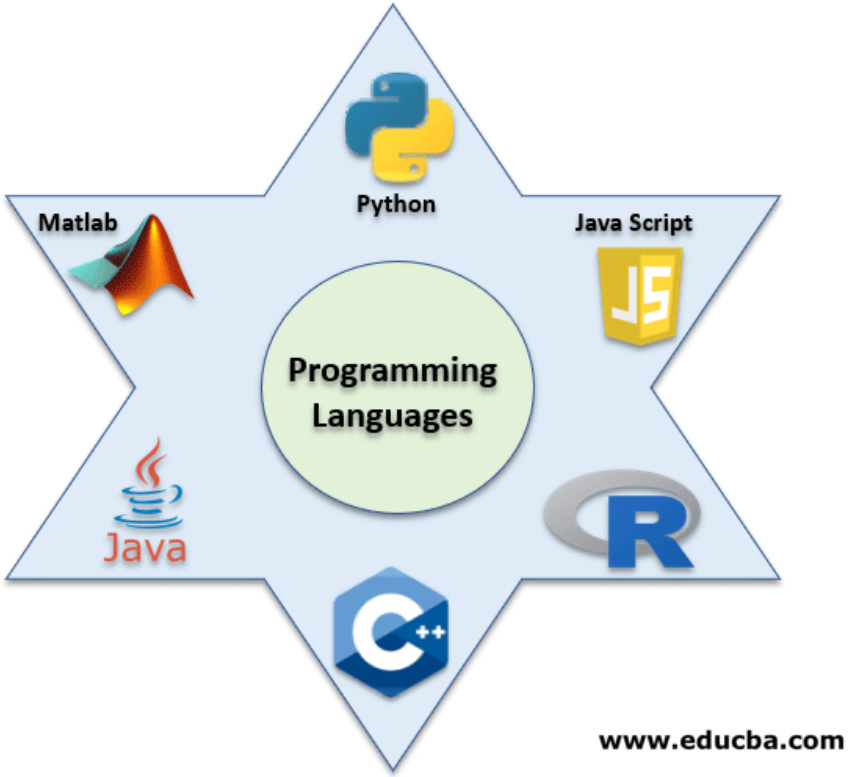
$$Rappel = \frac{TP}{TP + FN}$$

Contenu de la matière

Chapitre 1 Introduction
Chapitre 2 Régression linéaire
Chapitre 3 Arbres de décision
Chapitre 4 Classifieur naïf de Bayes
Chapitre 5 Réseaux de neurones
Chapitre 6 K-means

Mode d'évaluation : Moyenne Matière = Note contrôle*60%+Note de travaux*40%

- **Les outils de l'apprentissage machine**



Introduction

Dans ce cours, nous allons utiliser le langage python avec ces librairies pour découvrir les techniques de l'apprentissage automatique .