

Apprentissage automatique

Chapitre II

Apprentissage automatique

Exemples

	Input / I	Output / O
Filtrage des emails	Un e-mail	Spam / non-spam
Reconnaissance faciale	Une image	Accès / interdiction
Traduction	Phrase en langage A	Phrase en langage B
Reconnaissance parole	Signal audio	Une phrase / mot
Déplacement d'un Robot	Données à partir de capteurs	Mouvement en trajectoire

Types d'apprentissage Automatique

L'apprentissage dépend des données disponibles

- On peut avoir certains données de l'ensemble des entrées et d'autres de l'ensemble des sorties (i,o)

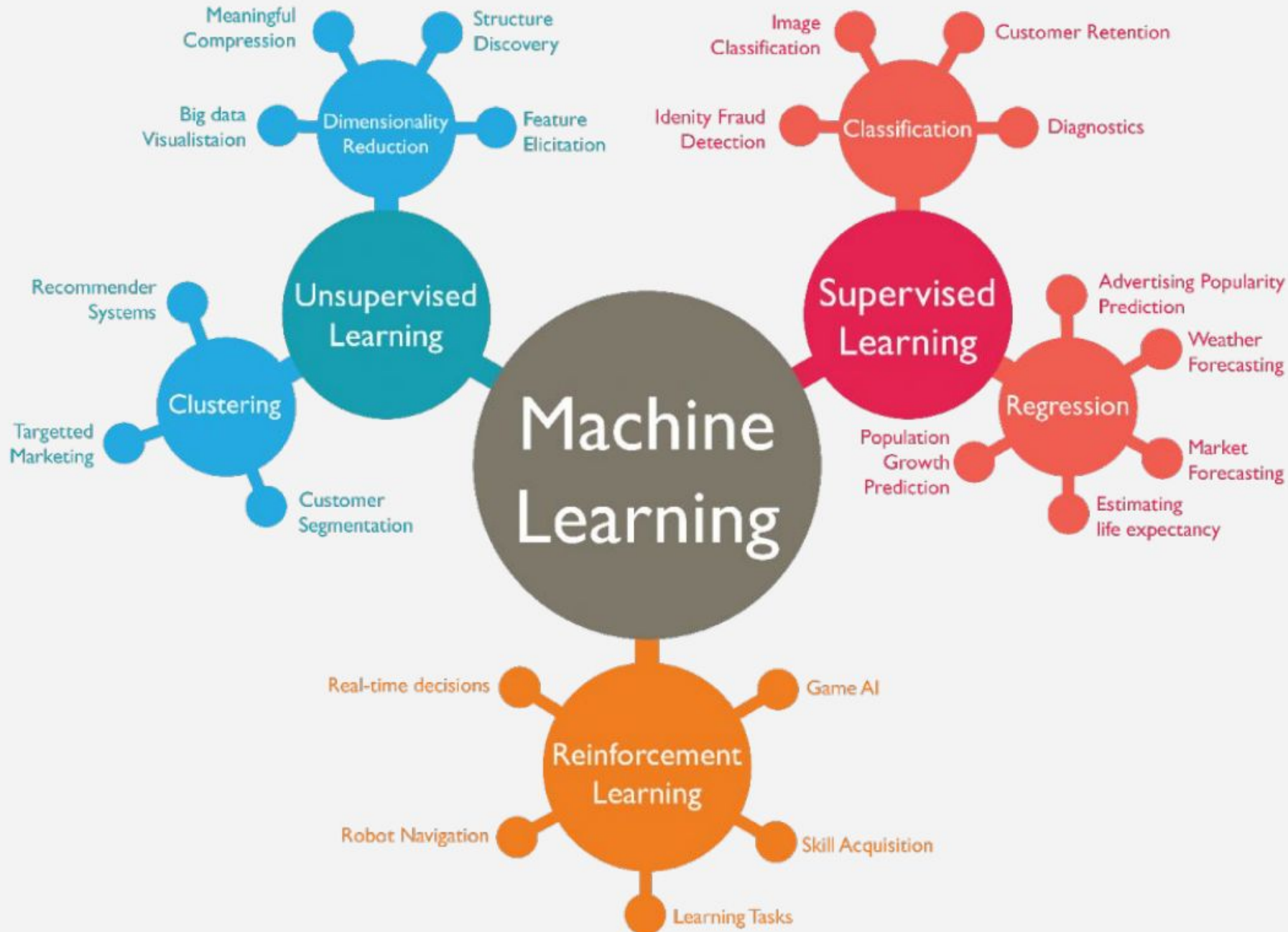
- ☐ Apprentissage supervisé

- On peut avoir seulement des entrées i.

- ☐ Apprentissage non supervisé

- On peut ne pas connaître les sorties "corrects" mais on peut utiliser une certaine mesure qui détermine la qualité de la sortie o étant donnée l'entrée o.

- ☐ Apprentissage par renforcement



L'apprentissage supervisé

- un expert se charge d'étiqueter correctement certaines entrées
- Deux types d'apprentissage supervisé selon le type du résultat obtenu

La classification	La régression
<p>Le résultat obtenu est une valeur discrète</p> <p>La résultat à prédire peut prendre une valeur d'un ensemble fini de valeurs : Classe</p> <p>Par exemple, prédire si un mail est SPAM ou non, le résultat peut prendre deux valeurs possibles : {spam, non spam}</p>	<p>Le résultat obtenu est une valeur continue</p> <p>La résultat à prédire peut prendre n'importe quelle valeur.</p> <p>Par exemple, prédire le prix du véhicule étant données des caractéristiques d'un véhicule</p>

L'apprentissage non-supervisé

- Aucune information n'est disponible sur les sorties o.
- L'algorithme d'apprentissage doit découvrir par lui même La structure des données.

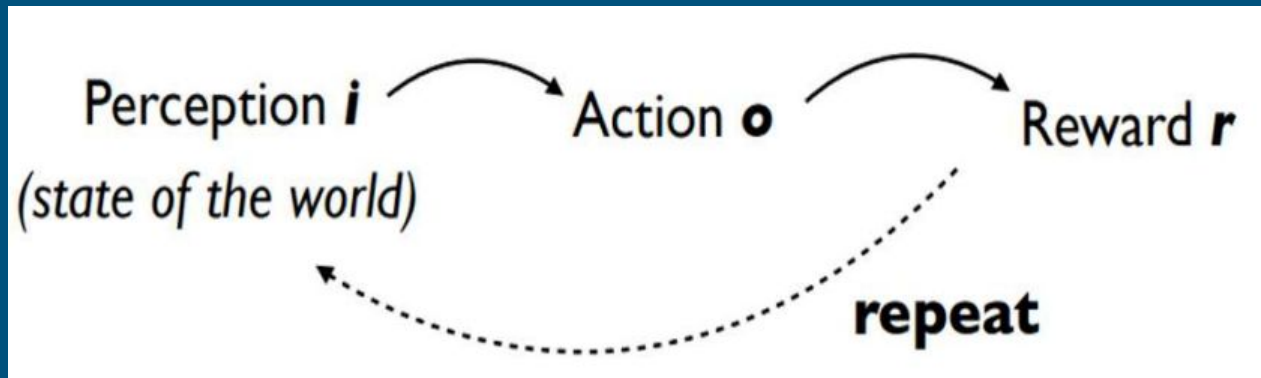
Exemples de tâches associées à l'apprentissage non supervisé

- catégorisation/regroupement/segmentation
Construire des classes automatiquement en fonction des exemples disponibles
- Réduction de dimensions □ Diminuer l'ensemble des attributs
- Règles d'association □ Analyser les relations entre les variables ou détecter des associations

L'apprentissage par renforcement

Nous ne possédons pas la sortie "o" jugée correcte pour un certaine entrée "i".

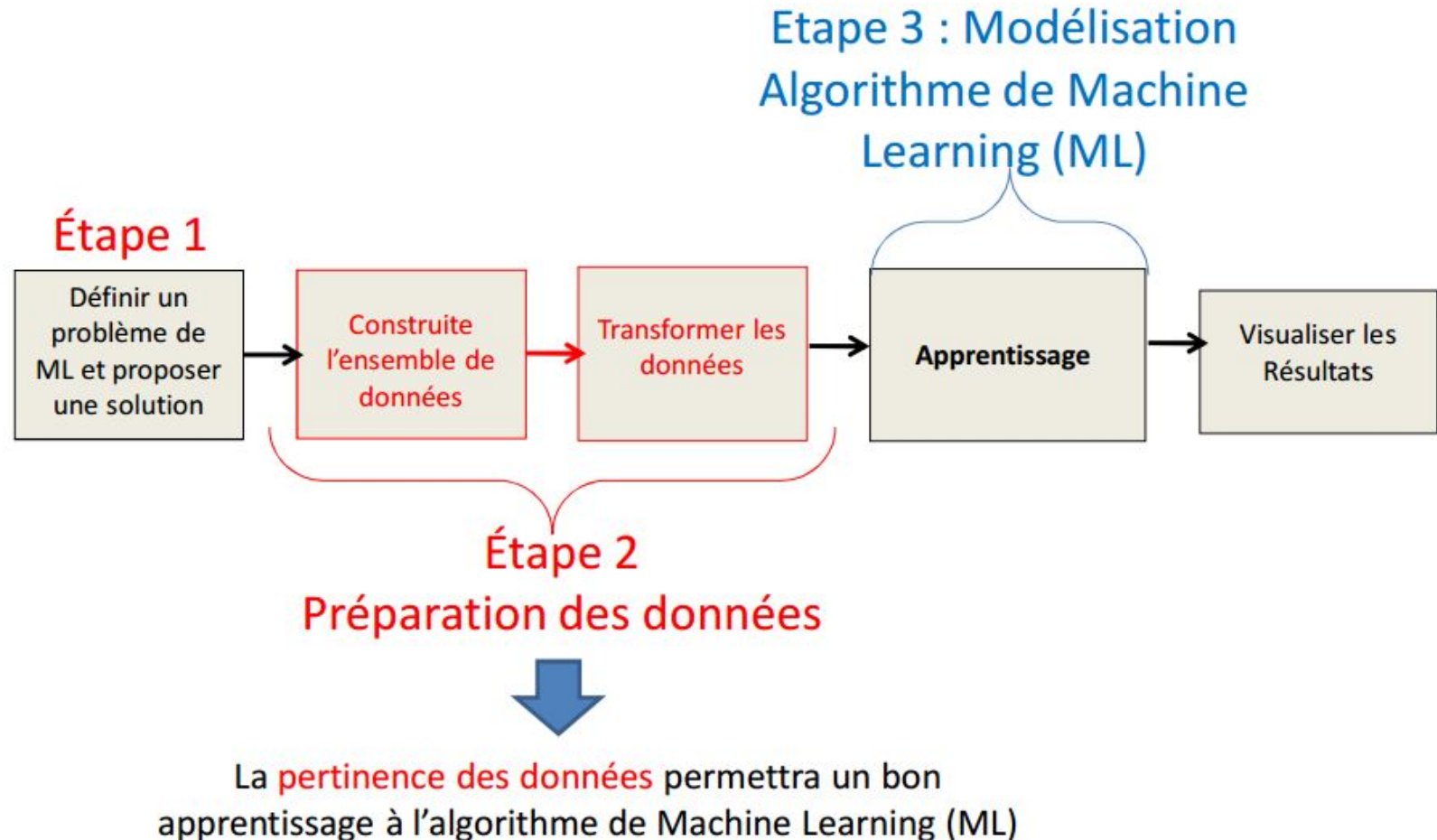
- Il repose sur le principe d'essai/erreur.
- on peut mesurer la qualité d'une sortie "récompense" (reward)



Exemple

- Un robot doit "retourner des crêpes"
 - La récompense peut être
 - » +3 si la crêpe est retournée
 - » -1 si la crêpe reste dans la poêle
 - » -5 si la crêpe tombe
- Le rôle du robot est de maximiser la récompense à travers plusieurs essais

Processus d'apprentissage : du problème vers le modèle



Etape 1 : définir le problème et déterminer le type d'apprentissage

Tâche souhaitée ☐ Problème ML

- Filtrage des e-mails ☐ supervisé
- vision par ordinateur ☐ supervisé + Non supervisé
- Conduite autonome ☐ renforcement
- Chat bots ☐ supervisé + renforcement
- Jouer et gagner des jeux ☐ renforcement
- Diagnostic médical ☐ supervisé
- Recommandations de produits ☐ supervisé + Non supervisé
- Suggestions d'amis/de contenu ☐ supervisé + Non supervisé
- Reconnaissance vocale ☐ supervisé
- Sous-titrage vidéo ☐ supervisé + Non supervisé
- Traduction d'un texte ☐ supervisé

Etape 1 : définir le problème et déterminer le type de traitement

Formuler le problème d'une manière précise et concise







Problème	Type de traitement
d'estimer le prix des appartements.	<input type="checkbox"/> Régression
Dégager les objets dans les images	<input type="checkbox"/> Catégorisation / Classification
programmer le jeu d'échec	<input type="checkbox"/> Apprentissage par renforcement
regrouper ensemble les images similaires dans une base.	<input type="checkbox"/> Catégorisation
détermine l'espèce d'une fleur	<input type="checkbox"/> Classification

Etape 2 : Préparation des données (inputs)

- Les inputs sont aussi appelée observation ou instances
- Les différentes sources : mesures, observations collectée, web scraping, free databases, ...
- Chaque observation admet un certain nombre d'attributs(appelées aussi variables)
- Toutes les observations d'un même ensemble (base) doivent avoir les mêmes attributs (en nombre et en type)
- Les observations peuvent être textuelles, numériques, vidéo, images ou multi-média

Exemple de base de données

Les attributs/variables				
Jour	Ciel	Température	Humidité	Vent
J1	Soleil	Chaud	Elevée	Faible
J2	Soleil	Chaud	Elevée	Fort
J3	Couvert	Chaud	Elevée	Faible
J4	Pluie	Doux	Elevée	Faible
J5	Pluie	Froid	Normale	Faible
J6	Pluie	Froid	Normale	Fort
J7	Couvert	Froid	Normale	Fort
J8	Soleil	Doux	Elevée	Faible
J9	Soleil	Froid	Normale	Faible
J10	Pluie	Doux	Normale	Faible
J11	Soleil	Doux	Normale	Fort
J12	Couvert	Doux	Elevée	Fort
J13	Couvert	Chaud	Normale	Faible
J14	Pluie	Doux	Elevée	Fort

Les attributs/variables			
Customer	Age	Income	No. credit cards
John 	35	35K	3
Rachel 	22	50K	2
Hannah 	63	200K	1
Tom 	59	170K	1
Nellie 	25	40K	4
David 	37	50K	2

Préparation des données : Transformer les données bruts ... Pourquoi!?

- La plupart des algorithmes utilisés en ML nécessite des données sous un format spécifique.
 - Les données peuvent contenir des informations manquantes, redondantes, invalides, erronées et inconsistantes
- Un "Bon" protocole de préparation de données donne de "Bonnes données" qui conduisent à un résultat de prédiction plus fiable

Préparer les données... Pourquoi!?

Exemple :

Les données réelles présentent des problèmes de qualité.

- **incomplétude** : contient des valeurs manquantes ou des données dépourvues d'attributs
- **Bruit** : contient des enregistrements incorrects ou des exceptions
- **incohérence** : contient des enregistrements incohérents

#	Id	Name	Birthday	Gender	IsTeacher	#Students	Country	City
1	111	John	31/12/1990	M	0	0	Ireland	Dublin
2	222	Mery	15/10/1978	F	1	15	Iceland	
3	333	Alice	19/04/2000	F	0	0	Spain	Madrid
4	444	Mark	01/11/1997	M	0	0	France	Paris
5	555	Alex	15/03/2000	A	1	23	Germany	Berlin
6	555	Peter	1983-12-01	M	1	10	Italy	Rome
7	777	Calvin	05/05/1995	M	0	0	Italy	Italy
8	888	Roxane	05/08/1948	F	0	0	Portugal	Lisbon
9	999	Anne	05/09/1992	F	0	5	Switzerland	Geneva
10	101010	Paul	14/11/1992	M	1	26	Ytali	Rome

Valeur
manquante

Erreur valeur

Erreur saisie

Erreur
saisie

Duplication
invalide

Format
incorrecte

dépendance

Exemples de transformation des données brutes (rowdata)

Type de transformation	exemple
La discrétisation des données continues	moyenne <12 □ passable ; moyenne <14 □ assez bien ; moyenne >= 18 □ excellent
modifier le type d'un attributs	date de naissance □ âge date d'abonnement □ durée...
Normalisation = Réduire les attributs à la même échelle	Diviser par la valeur maximal □ avoir des données entre 0 et 1.
Supprimer les données aberrantes	Corriger Erreur de saisi, de mesure
Complétez les données manquantes	Complétez les valeurs des attributs manquants par la moyenne
Suppression	- l'observation incomplète - Attribut avec trop de valeurs manquantes
Réduire la dimension	Supprimer les attributs non pertinents
Etiquetage = label (supervisé)	Ajouter attribut contenant le label

Comment détecter les problèmes liés aux données ?

**Documentation
(lire la
description des
données**

**Exploration des
données**

**Visualisation
des données**

Apprentissage supervisé

1ère étape : phase apprentissage (training)

A partir des Input, on choisit un échantillon dit d'apprentissage dont la sortie est connu. Cet échantillon est utilisé pour estimer une fonction de prédiction = création du modèle. Cette étape peut contenir une phase de validation avec 20% de l'échantillon

2ème étape : phase de test

Utiliser la fonction de prédiction (le modèle) définit dans l'étape précédente afin de déterminer des sorties (output) relatifs aux inputs non utilisées dans la phase apprentissage.

Proportions usuelles : 70% training , 30% test

3ème étape : Évaluation et Recommandations

Calculer le Taux d'erreurs du modèle

La sortie connue d'une entrée de l'ensemble de test est comparée avec le résultat donné par le modèle.

Taux d'erreur = pourcentage de tests incorrectement prédis par le modèle

Types d'apprentissage supervisé

- La régression

- Régression linéaire : Le résultat obtenu est une valeur contenu

- Exemple : prédire le prix du véhicule étant données ses caractéristiques quantitatives

- Régression logistique : le résultat est une probabilité d'appartenir à une ou l'autre des catégories proposées (\approx)

- La classification

- Le résultat obtenu est une valeur discrète d'un ensemble de Classes

- Exemple : prédire si un mail est SPAM ou non, le résultat peut prendre deux valeurs possible : {spam, non spam}

- Exemple : `exempl1_reg_lineaire`

Régression Logistique

- Variable à expliquer Y est qualitative

- 3 types de régression logistique

- binaire $\Rightarrow Y$ binaire (ex : vivant / décès)

- ordinale $\Rightarrow Y$ ordinale (ex : stades de cancer)

- multinomiale $\Rightarrow Y$ qualitative (ex : types de cancer)

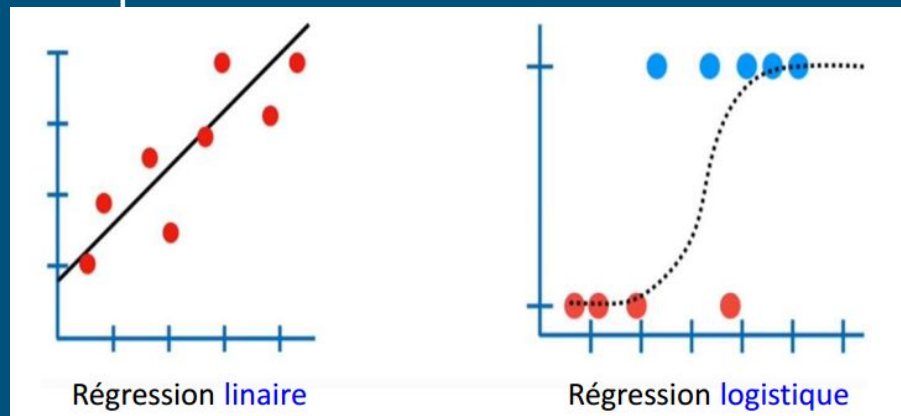
- Variables explicatives $X=[x_1, \dots, x_n]$ sont qualitatives ou quantitatives

Il n'existe pas une méthode analytique pour résoudre ce problème

- Il faut utiliser des méthodes numériques d'approximation du maximum

- Méthode de descente de gradient

- Méthode de Newton-Raphson



Apprentissage supervisé : Classification

Un classificateur est un modèle qui permet de décider de l'appartenance d'une observation donnée à une classe particulière.

□ Deux types de classification

- Classification binaire (SPAM ou non)
- Classification Multi-classes :

exemple : un objet dans une image (chien, cheval, chat)

Les modèles de classification (classificateurs)

K-NN : K Nearst Neighbors

SVM : Support Vector Machine

Arbre de Décision :Random Forests, XGBoost (Gradient Boosting)

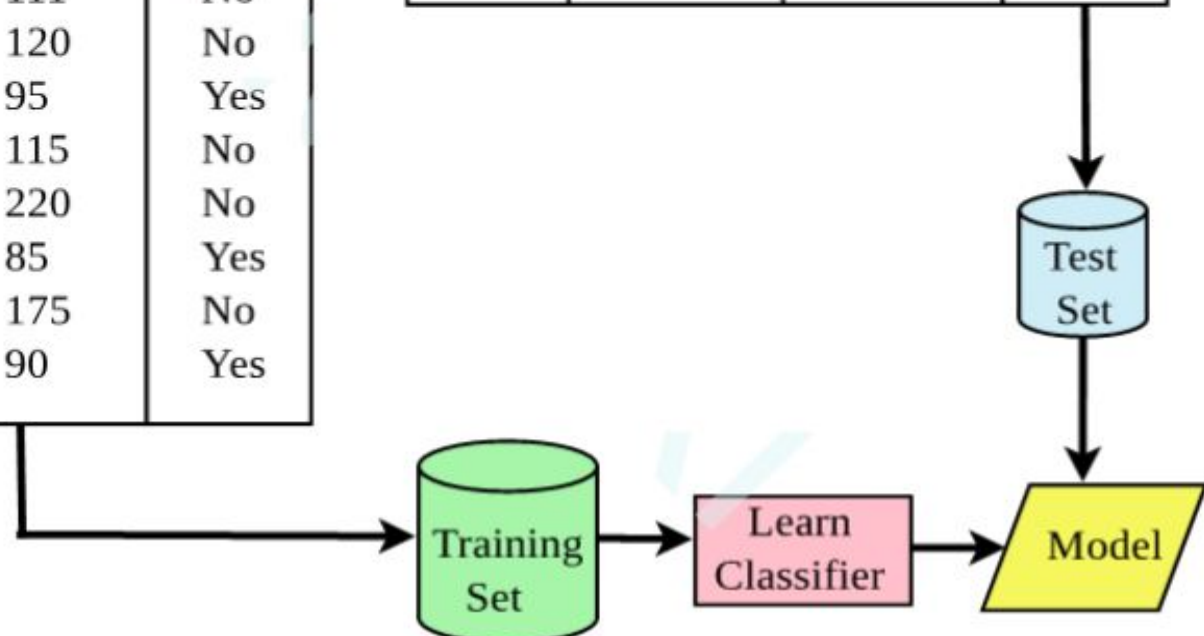
DNN : Deep Neural Networks

CNN : Convolutional Neural Network en combinaison avec un classifie

Elaboration d'un modèle de Classification

Attrib-1	Attrib-2	Attrib-3	Class
Yes	80	125	No
No	95	100	No
No	77	111	No
Yes	101	120	No
No	250	95	Yes
No	1	115	No
Yes	65	220	No
No	140	85	Yes
No	19	175	No
No	200	90	Yes

Attrib-1	Attrib-2	Attrib-3	Class
No	85	175	?
Yes	111	10	?
No	68	130	?
Yes	200	15	?
No	19	215	?



Principe du K-NN

Prédire la classe d'une donnée (observation) O_{new} en s'appuyant sur une base de données étiquetées en 2 étapes :

- calculer toutes les distances entre O_{new} et les données de la base,
- affecter à O_{new} la même classe que celle des k données qui lui sont le plus proches, k étant fixé à l'avance.

Le calcul de la distance est réalisé selon l'une des formules de calcul de distance (euclidienne, manhattan,...)

$$D_e = \sum_{i=1}^n (x_i - y_i)^2$$
$$D_m = \sum_{i=1}^n |x_i - y_i|$$

Les points faibles :

coût élevé en puissance de calcul

surcharge de la mémoire par toutes les données d'entraînement.

□ K-NN convient donc plutôt aux problèmes d'assez petite taille.

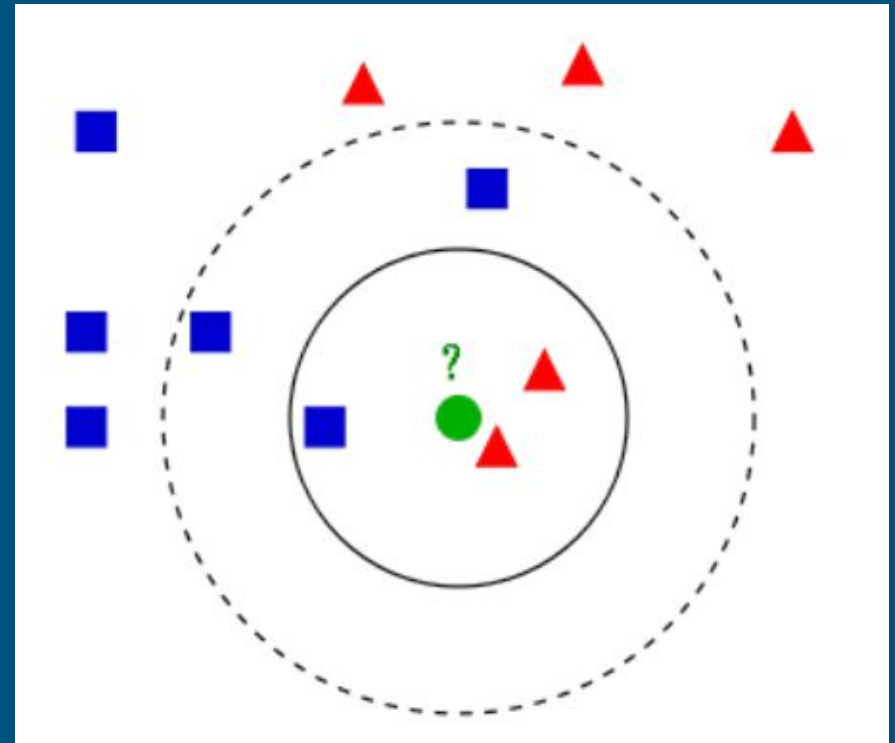
Exercice K-NN

Déterminez la classe du point vert pour :

- $K = 3$



- $K = 5$



Exemple : Modèle KNN

	A	B	C	D	E	F
1	etiquette_fruit	nom_fruit	poids	largeur	hauteur	score-couleur
2		1 pomme	180	8,4	6,8	0,6
3		1 pomme	198	8,2	7,5	0,55
4		1 pomme	174	8	7,7	0,59
5		2 mandarine	85	6	4	0,8
6		2 mandarine	80	5,8	4,6	0,81
7		2 mandarine	83	6,2	4,3	0,77
8		2 mandarine	81	5,9	4,7	0,79
9		1 pomme	152	7,4	7,6	0,69
10		2 mandarine	86	5,3	4,3	0,8
11		1 pomme	174	7,1	7,4	0,81
12		2 mandarine	82	5,7	4	0,77
13		2 mandarine	84	6	4,5	0,79
14		1 pomme	196	7,6	7,9	0,93
15						

- 2 classes : pomme et mandarine
- Attributs pertinents : poids', 'largeur', 'hauteur'

Objectifs : Entraîner et tester un modèle K-NN avec K=3

Onew1 : poids = 20, largeur = 4.3 , hauteur = 5.5

Onew2 : poids = 180, largeur = 8.0 , hauteur =6.8


```

# importer les modules numpy, pandas et sklearn
import numpy as np
import pandas as pd
from sklearn.model_selection import train_test_split
import warnings
warnings.filterwarnings('ignore')
from google.colab import drive
drive.mount('/content/gdrive')
"""# charge dataset + affiche les cinq premieres lignes"""
df = pd.read_excel('/content/gdrive/MyDrive/ML/pratique/fruitDataset.xlsx')
df.head()
"""# créer une correspondance entre la valeur de l'étiquette du fruit et son nom """
nom_fruit_cible = dict( zip (df.etiquette_fruit.unique(), df.nom_fruit.unique()))
print( nom_fruit_cible )
"""# - définir x (attributs) et y (label). + fractionner en train et test (75/25 %)."""
x = df [['poids', 'largeur', 'hauteur']]
y = df['etiquette_fruit']
x_train, x_test, y_train, y_test = train_test_split(x, y, random_state=0)
"""# créer classificateur KNN + train"""
from sklearn.neighbors import KNeighborsClassifier
#instanciation et définition du k
knn = KNeighborsClassifier(n_neighbors = 3)
#training
knn.fit(x_train,y_train)
"""# evaluer le modèle"""
knn.score(x_test,y_test)
"""# prédiction pour une observation """
prediction_fruit = knn.predict([[20,4.3,5.5]])
nom_fruit_cible[prediction_fruit[0]]
prediction_fruit = knn.predict([[180,8.0,6.8]])
nom_fruit_cible[prediction_fruit[0]]

```

Algorithme arbre de décision

- Ensemble de règles de classification basant leur décision sur des tests associés aux attributs, organisés de manière arborescente.
- Motivation : Produire des classifications compréhensibles par l'utilisateur.
- Un arbre est équivalent à un ensemble de règles de décision : un modèle facile à comprendre.
- Principe : Prédire la valeur d'un attribut à partir d'un ensemble de valeurs d'attributs
- Un arbre est composé de :
 - nœuds : classes d'individus de plus en plus fines depuis la racine.
 - arcs : prédicats de partitionnement de la classe source.

Algorithme arbre de

Exemple : Faut-il sortir le chien ?

Attributs

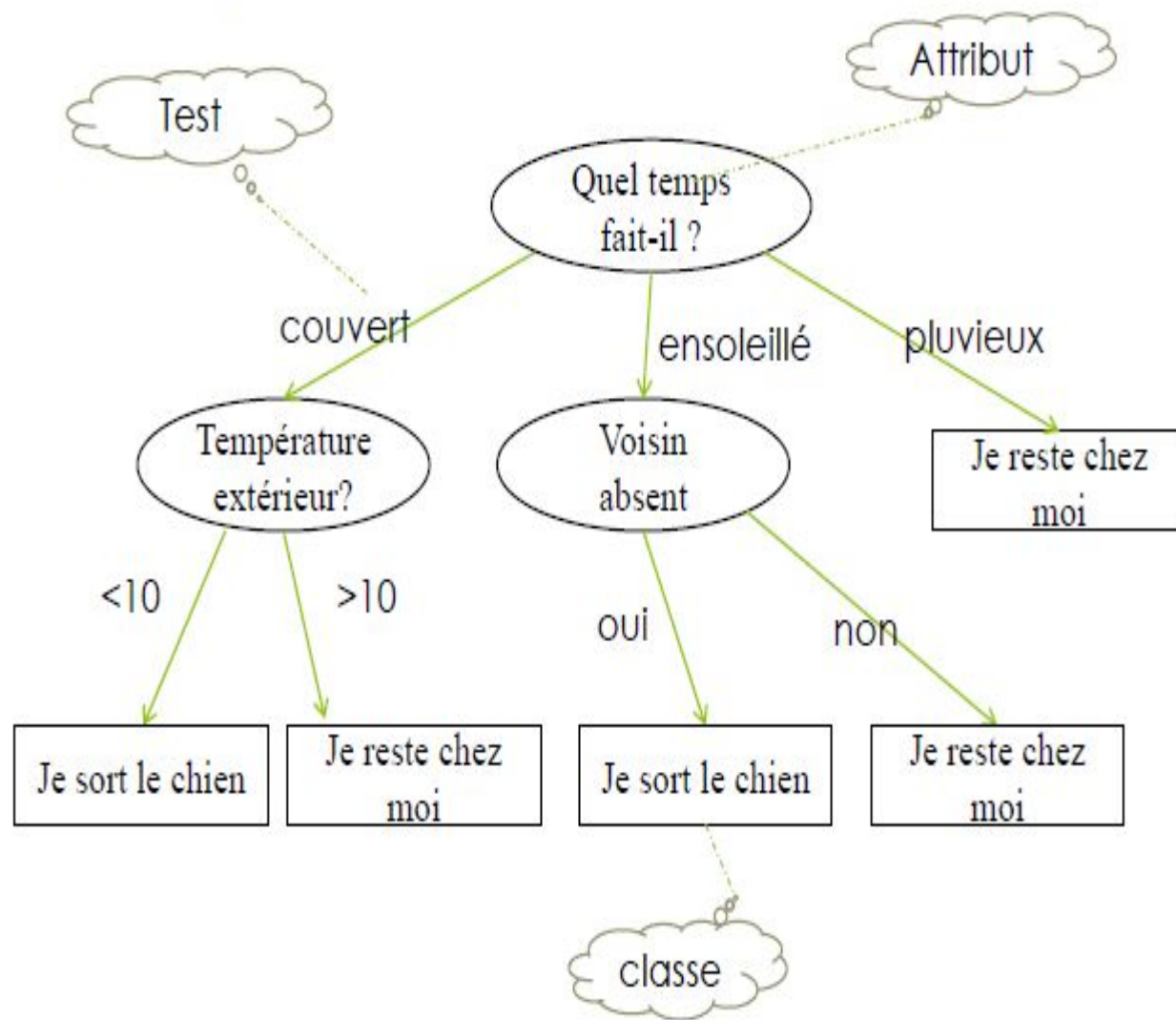
- Quel temps fait-il ? : attribut nominal {couvert , pluvieux , ensoleillé}
- température extérieur : attributs numériques {>10 , <= 10}
- voisin absent : attribut booléen {oui , non}

Décision a prendre (classes)

- C1 : je reste chez moi
- C2 : je sort le chien

Algorithme arbre de

Exemple : Faut-il sortir le chien ?



Arbre de Classification

Permet de prédire l'étiquette de la classe à laquelle une variable cible appartient.



Iris Versicolor

Iris Setosa

Iris Virginica

Arbre de Régression

Permet de prédire une valeur numérique quantitative pour la variable cible.



Fin

Merci Pour Votre Attention