

Introduction au machine learning

Présenté par :
Viktor Iliev et Elian Strozyk

1^{er} février 2024



Qu'est-ce que le machine learning ?

Le machine learning :

- C'est un sous-domaine de l'intelligence artificielle
- Consiste à prédire statistiquement des patterns à partir de données

Les approches du machine learning

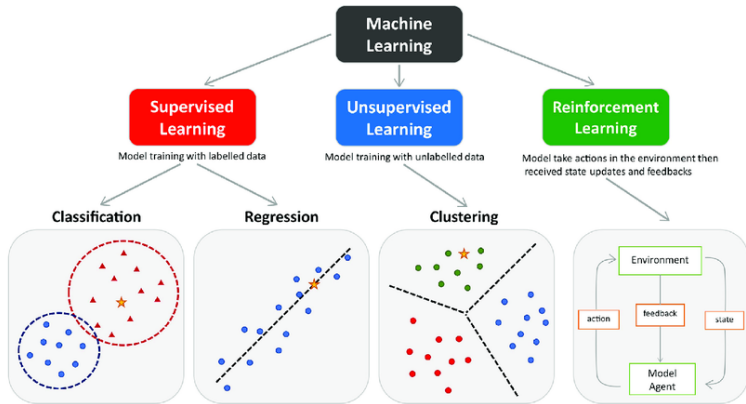


Figure – Différentes approches du machine learning [1]

Exemple simple de classification

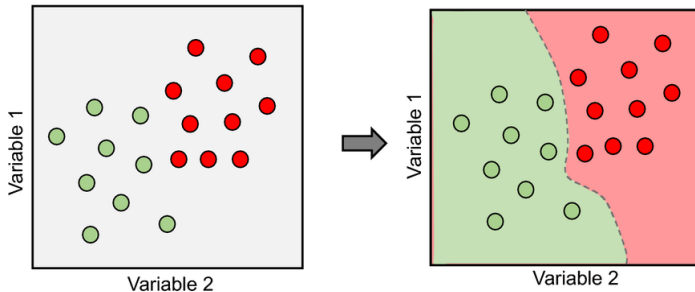
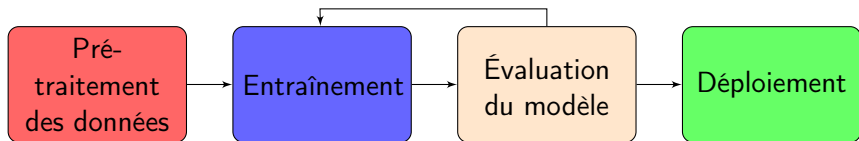


Figure – Prédiction d'une variable catégorielle (couleur) selon 2 variables quantitatives (Variable 1 et Variable 2) [2]

Étapes pour entraîner un modèle



Nettoyage

Extraction
de features

Équilibrage

Découpage

Choix du
modèle

Choix des
hyperpara-
mètres

Matrice de
confusion

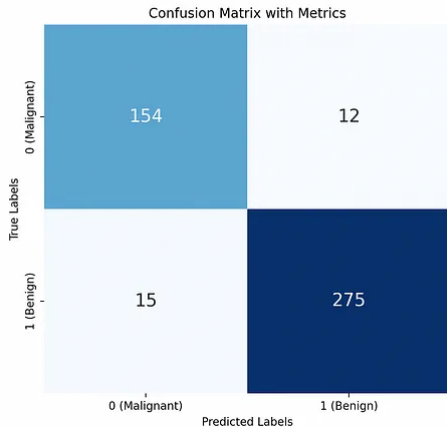
Sauvegarde
du modèle

Matrice de confusion

		PREDICTIVE VALUES	
		POSITIVE (1)	NEGATIVE (0)
ACTUAL VALUES	POSITIVE (1)	TP	FN
	NEGATIVE (0)	FP	TN

Figure – Principe d'une matrice de confusion [3]

Matrice de confusion



Métriques de performance

		PREDICTIVE VALUES	
		POSITIVE (1)	NEGATIVE (0)
ACTUAL VALUES	POSITIVE (1)	TP	FN
	NEGATIVE (0)	FP	TN

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

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

$$F_1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

Ce workshop a pour objectif de :

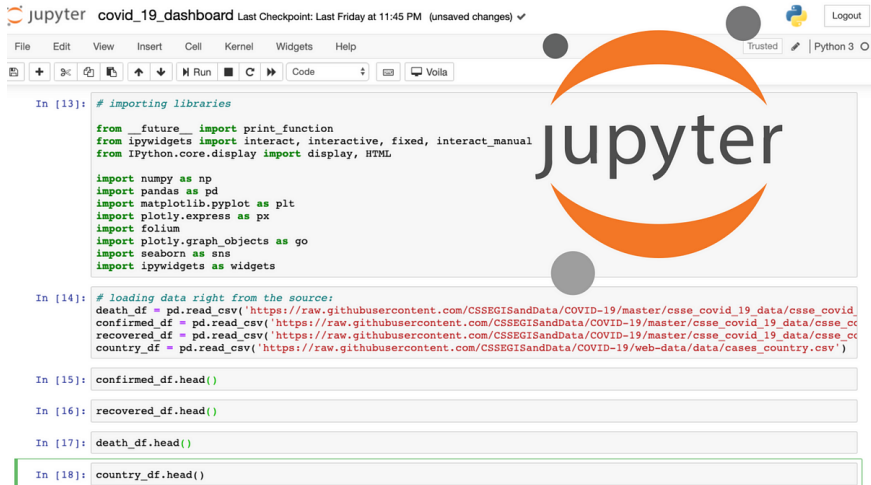
- Présenter des algorithmes d'apprentissage supervisé pour de la classification
- Comprendre les grandes lignes de l'entraînement de modèles de machine learning
- Optimiser l'entraînement de modèles

Outils



On va utiliser :

- Python
- La bibliothèque Scikit-Learn



jupyter covid_19_dashboard Last Checkpoint: Last Friday at 11:45 PM (unsaved changes) ✓

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3 O Logout

```
In [13]: # importing libraries
from __future__ import print_function
from ipywidgets import interact, interactive, fixed, interact_manual
from IPython.core.display import display, HTML

import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import plotly.express as px
import folium
import plotly.graph_objects as go
import seaborn as sns
import ipywidgets as widgets

In [14]: # loading data right from the source:
death_df = pd.read_csv('https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_19_data/confirmed_df.csv')
confirmed_df = pd.read_csv('https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_19_data/recovered_df.csv')
recovered_df = pd.read_csv('https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_19_data/country_df.csv')
country_df = pd.read_csv('https://raw.githubusercontent.com/CSSEGISandData/COVID-19/web-data/data/cases_country.csv')
```

```
In [15]: confirmed_df.head()

In [16]: recovered_df.head()

In [17]: death_df.head()

In [18]: country_df.head()
```

Description générale

Jeu de données modèle de classification sur le cancer du sein

- $N = 569$ instances
- 2 classes : bénigne (benign ou 0) et maligne (malignant ou 1)
- Instances par classe :
 - $N = 357$ pour la classe bénigne
 - $N = 212$ pour la classe maligne
- 30 features (= dimensions)

Features

- radius
- texture
- perimeter
- area
- smoothness
- compactness
- concavity
- concave points
- symmetry
- fractal dimension

Pour chaque feature, on a

- La mesure moyenne (mean)
- L'écart-type des valeurs (error)
- La moyenne des 3 valeurs les plus élevées (worst)

Exemple : mean radius ; radius error ; worst radius

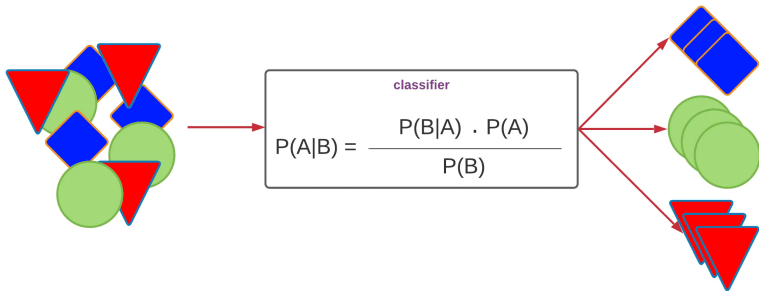


Figure – Classification Naive Bayes : classification selon la classe qui maximise la fonction de probabilité fonctionnelle

Machine à vecteurs de support (SVM)

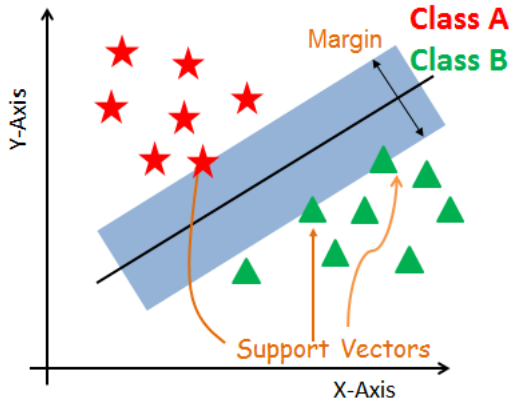


Figure – Exemple d'une vaste marge (hyperplan) discriminant 2 classes [4]

Forêt d'arbres décisionnels

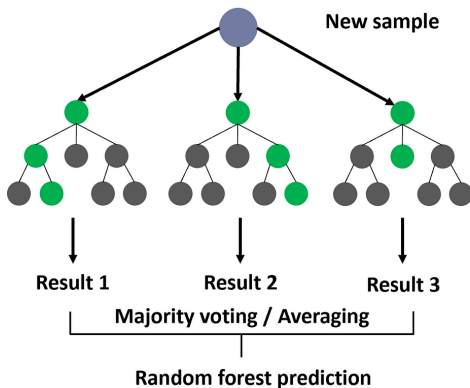


Figure – Exemple d'une forêt d'arbres décisionnels [5]

Extreme Gradient Boosting

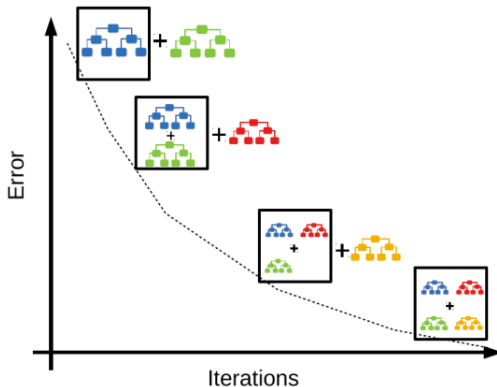


Figure – Exemple de construction d'arbres décisionnels optimisés itérativement [6]

Multi-layer perceptron (MLP)

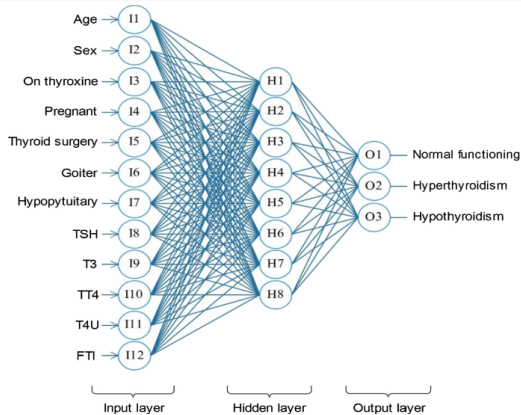


Figure – Exemple des couches d'un MLP pour diagnostiquer une maladie thyroïdienne [7]

Classification selon différents modèles

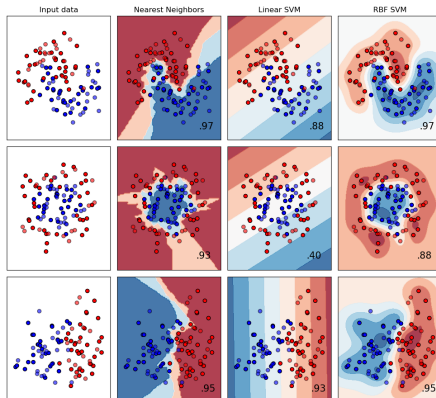


Figure – Exemples de classification pour différentes distributions de données et de modèles [8]

References I



Junjie Peng, Elizabeth Jury, Pierre Dönnès, and Coziana Ciurtin.

Machine learning techniques for personalised medicine approaches in immune-mediated chronic inflammatory diseases : Applications and challenges.

[Frontiers in Pharmacology](#), 12, 09 2021.



Juliano Morimoto and Fleur Ponton.

Virtual reality in biology : could we become virtual naturalists?

[Evolution : Education and Outreach](#), 14, 05 2021.

References II



Romain Cendre.

Classification par méthodes d'apprentissage supervisé et faiblement supervisé d'images multimodales pour l'aide au diagnostic du lentigo malin en dermatologie.

PhD thesis, 02 2021.



Support vector machines with scikit-learn tutorial, 2019.



Random forests, 2023.



Xgboost vs lightgbm : How are they different, 2023.

References III



Mehdi Hosseinzadeh, Omed Hassan Ahmed, Marwan Yassin Ghafour, Fatemeh Safara, Hawkar kamanan hama, Saqib Ali, Bay Vo, and Hsiu-Sen Chiang.

A multiple multilayer perceptron neural network with an adaptive learning algorithm for thyroid disease diagnosis in the internet of medical things.

[The Journal of Supercomputing, 77\(4\) :3616–3637, Apr 2021.](#)



Classifier comparison, 2024.