Machine Learning en Data Science

by REDA OUZIDANE

3 avril 2025

1 Introduction

Le Machine Learning (apprentissage automatique) est une branche de l'intelligence artificielle qui permet aux machines d'apprendre à partir de données et de prendre des décisions ou faire des prédictions sans être explicitement programmées.

2 Types d'apprentissage en Machine Learning

Le Machine Learning se divise principalement en trois types d'apprentissage :

2.1 Apprentissage Supervisé

L'apprentissage supervisé consiste à apprendre une fonction à partir d'un ensemble de données étiquetées (données d'entrée et leurs réponses correspondantes).

2.1.1 Exemples

- **Classification** : Il s'agit de prédire une catégorie ou une classe à partir des données. Par exemple, prédire si un email est un spam ou non.
- **Régression** : Prédiction d'une valeur continue. Par exemple, prédire le prix d'une maison à partir de ses caractéristiques.

2.2 Apprentissage Non Supervisé

Dans l'apprentissage non supervisé, les données ne sont pas étiquetées et l'objectif est de découvrir des structures cachées ou des relations dans les données.

2.2.1 Exemples

- **Clustering (Regroupement)** : Par exemple, segmenter les clients en groupes homogènes selon leurs comportements d'achat.
- **Réduction de dimensionnalité** : Techniques comme le PCA (Principal Component Analysis) utilisées pour réduire le nombre de variables tout en conservant les informations importantes.

2.3 Apprentissage par Renforcement

L'apprentissage par renforcement consiste à entraîner un agent pour qu'il prenne des décisions et apprenne de ses actions dans un environnement donné. L'agent reçoit des récompenses ou des punitions en fonction des actions qu'il prend.

3 Prétraitement des Données

Avant de pouvoir entraîner un modèle, il est souvent nécessaire de prétraiter les données.

3.1 Traitement des Valeurs Manquantes

Les valeurs manquantes peuvent être supprimées ou imputées à l'aide de techniques telles que la moyenne, la médiane, ou KNN Imputer.

3.2 Standardisation et Normalisation

- **Standardisation** : Transformation des données pour qu'elles aient une moyenne de zéro et un écart type de un. La formule est :

$$Z = \frac{X - \mu}{\sigma}$$

- **Normalisation** : Mise à l'échelle des données dans un intervalle défini, souvent [0, 1], à l'aide de la formule :

$$X_{\text{norm}} = \frac{X - X_{\text{min}}}{X_{\text{max}} - X_{\text{min}}}$$

3.3 Encodage des Variables Catégoriques

— **One-Hot Encoding** : Conversion des variables catégorielles en colonnes binaires.
— **Label Encoding** : Attribution d'un entier unique à chaque catégorie.

4 Implémentation en Python

Voici un exemple de code pour implémenter un modèle de régression logistique, qui est un modèle supervisé couramment utilisé pour des tâches de classification.

Listing 1 – Implémentation d'un modèle de régression logistique

```
import numpy as np
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler, OneHotEncoder
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score

# Chargement des donn es
X = pd.read_csv("data.csv")
y = X.pop("target")

# Pr traitement
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)
X_train, X_test, y_train, y_test = train_test_split(X_scaled, y, te

# Mod le
model = LogisticRegression()
model.fit(X_train, y_train)
y_pred = model.predict(X_test)

#  valuation
print("Accuracy:", accuracy_score(y_test, y_pred))
```

4.1 Autres Algorithmes de Machine Learning

Voici quelques autres algorithmes populaires en Machine Learning :

4.1.1 Support Vector Machines (SVM)

Les SVM sont utilisés pour la classification et la régression. Ils cherchent à maximiser la marge entre les classes.

Listing 2 – SVM

```
from sklearn.svm import SVC

# Mod le SVM
svm_model = SVC()
svm_model.fit(X_train, y_train)
y_pred_svm = svm_model.predict(X_test)

#  valuation
print("SVM␣Accuracy:", accuracy_score(y_test, y_pred_svm))
```

4.1.2 K-Nearest Neighbors (KNN)

KNN est un algorithme non paramétrique qui fait des prédictions basées sur les K voisins les plus proches dans les données.

Listing 3 – KNN

```
from sklearn.neighbors import KNeighborsClassifier

# Mod le KNN
knn_model = KNeighborsClassifier(n_neighbors=3)
knn_model.fit(X_train, y_train)
y_pred_knn = knn_model.predict(X_test)

#  valuation
print("KNN␣Accuracy:", accuracy_score(y_test, y_pred_knn))
```

4.1.3 Arbre de Décision (Decision Tree)

Les arbres de décision sont utilisés pour la classification et la régression en divisant les données en fonction des caractéristiques.

Listing 4 – Arbre de Décision

```
from sklearn.tree import DecisionTreeClassifier

# Mod le d'arbre de d cision
```

```
tree_model = DecisionTreeClassifier()
tree_model.fit(X_train, y_train)
y_pred_tree = tree_model.predict(X_test)

#   valuation
print("Decision␣Tree␣Accuracy:", accuracy_score(y_test, y_pred_tree
```

4.1.4 Random Forest

Les forêts aléatoires sont un ensemble d'arbres de décision qui améliorent les performances en combinant plusieurs arbres pour réduire le sur-apprentissage.

Listing 5 – Random Forest

```
from sklearn.ensemble import RandomForestClassifier

# Mod le Random Forest
rf_model = RandomForestClassifier(n_estimators=100)
rf_model.fit(X_train, y_train)
y_pred_rf = rf_model.predict(X_test)

#   valuation
print("Random␣Forest␣Accuracy:", accuracy_score(y_test, y_pred_rf))
```

5 Évaluation des Modèles

L'évaluation des modèles est une étape essentielle pour mesurer la performance d'un modèle de Machine Learning.

- **Accuracy** : Pourcentage de prédictions correctes par rapport au total des prédictions.
- **Métriques de Classification** : Précision, rappel, F1-score, etc.
- **Matrice de Confusion** : Visualisation des erreurs de classification.

6 Conclusion

Le Machine Learning est un domaine puissant de l'intelligence artificielle qui permet de résoudre une large gamme de problèmes. Les algorithmes sont variés et peuvent être appliqués à de nombreuses situations. L'utilisation de bibliothèques Python telles que Scikit-Learn facilite l'implémentation et l'évaluation des modèles.

By REDA OUZIDANE.