Union des Professeurs de classes préparatoires Scientifiques

Compte rendu du stage LIESSE

Machine Learning

EFREI, 11 juillet 2023

Auteur du CR : Damien Riou

/ Contexte

Après un rapide café d'accueil permettant un échange entre les onze stagiaires et l'intervenant, Laurent Cetinsoy, formateur à l'Efrei, les choses sérieuses commencent.

/ Déroulé du stage

Introduction au machine learning

Une introduction au vocabulaire et aux principaux concepts du domaine a été proposée sous la forme d'un cours rapide. Les données ont été qualifiées par Laurent Cetinsoy de nouvel or noir du XXIe siècle, ces données existant déjà dans les silos de données ou étant facilement récupérables dans le monde numérique dans lequel nous vivons. Les données sont multiples, tabulaires, temporelles, géospatiales, sonores, visuelles ou audiovisuelles.

Deux types principaux de machine learning sont mis en avant : le machine learning supervisé (apprentissage sous contrôle, c'est ce qui fonctionne le mieux) ou le machine learning non supervisé (sans les contrôles autre que celui du data scientist, moins fiable). Les modèles présentés sont la régression linéaire, la régression logistique, l'arbre de décision et le réseau de neurones.

Très accessible, cette première présentation a été claire et Laurent Cetinsoy a été très réactif pour répondre à nos questions.

Un exemple a été réalisé très simplement avec teachablemachine.withgoogle.com, site de Google qui donne accès à une interface qu'il est possible de paramétrer sans coder tout en s'appuyant sur du machine learning.

En pratique, travail préparatoire

Laurent Cetinsoy a présenté un certain nombre de bibliothèques liées au machine learning avec Python. Par exemple :

- numpy : vecteurs, matrices, tableaux ;
- pandas : traitement de données tabulaires (csv, svg, sql...), codé en numpy ;
- matplotlib, seaborn, plot.ly : visualisation de données ;
- streamit, gradeio : tableau de bord interactif ;
- scikit-learn, statmodel : machine learning supervisé standard (hors deep learning) ;
- tensorflow.keras : deep learning simple ;
- tensorflow, pytorch : deep learning plus complexe.

Nous avons ensuite suivi un parcours de formation pratique sur numpy et pandas, de manière à pouvoir maîtriser les notions de vectorisation des données, d'index booléens et de masques avec numpy, des dataframes et des séries avec pandas. L'intervenant a insisté sur la concision de l'écriture du code avec ces bibliothèques. L'intervenant a toujours été à nos côtés pour répondre à toutes nos questions, aussi diverses fussent-elles.

Nous avons ensuite déjeuné dans un restaurant à proximité de l'Efrei avec Laurent Cetinsoy, les frais ayant été pris en charge par l'Efrei.

En pratique, introduction au machine learning

À nouveau, une rapide introduction sur le machine learning par apprentissage supervisé a mis l'accent sur les

étapes principales du processus de création d'un modèle, à savoir :

- la compréhension du besoin ;
- la collecte et le nettoyage des données ;
- l'analyse et l'exploration des données ;
- la sélection du modèle à entraîner, la mesure de la performance ;
- la sélection des variables ;
- l'entraînement du modèle sur un jeu de données dédié ;
- l'évaluation sur un autre jeu de données dédié.

Scikit-learn a été la bibliothèque utilisée pour créer un modèle donnant l'estimation du prix d'une maison connaissant un ensemble de paramètres comme le nombre de pièces, la superficie ou l'exposition de la pièce à vivre. Le modèle choisi fut un modèle linéaire avec sklearn.linear_model.LinearRegression, catégorie de modèle semblant adapté au type de problème que nous souhaitions résoudre. Un approfondissement a été proposé sur la prise en compte des paramètres non chiffrés comme l'orientation de la pièce à vivre.

Laurent Cetinsoy a présenté deux fonctions d'erreur utilisables en fonction du type de résultat souhaité : mean square error (MSE) pour le cas général et les problèmes de modélisation, cross entropy pour les problèmes de classification.

Puis une rapide ouverture sur les arbres de décision a été réalisée, ceux-ci pouvant être compris comme un empilement de petits modèles de machine learning.

Une pause nous a permis de découvrir le campus de l'Efrei à Villejuif.

En pratique, introduction aux réseaux de neurones

L'intervenant nous a présenté l'intérêt d'un réseau de neurones, et nous avons pu nous entraîner avec la bibliothèque tensorflow.keras, assez facile à prendre en main pour créer un réseau et l'entraîner.

Enfin, une ouverture a été faite sur la notion d'apprentissage par renforcement avec les réseaux de neurones, ce qui constitue le deep learning.

/ Conclusion et remerciements

Un dernier accent a été mis sur la qualité des données que nous utilisons pour entraîner les données. En effet, des données biaisées donneront un modèle biaisé. La qualification des données est une étape indispensable pour déterminer quelle est la qualité des données sur lesquelles sera entraîné le modèle. Aussi, prédire n'est pas expliquer, et ceci pose des problèmes conceptuels et d'acceptabilité sur les conclusions tirées de modèles obtenus par machine learning.

Pour finir, Laurent Cetinsoy nous a invités si nous le souhaitions à utiliser sa plateforme nowledgeable.com avec nos étudiants, plateforme sur laquelle nous nous sommes entraînés toute la journée.

Nous tenons à remercier Laurent Cetinsoy pour sa présentation très claire et pédagogique d'un domaine où nous étions tous grands débutants.