Consignes de projet AI28 – Machine Learning

3 juin 2024

1 Groupes

(Groupe 1 - ADJ) ARMANI Paul-Antoine, DUJARDIN Cyprien et JOURDAIN Léo
(Groupe 2 - BBL) BENYAGOUB Imene, BOUZAR Massil et LAKRAMTI Saad
(Groupe 3 - CFG) FAVRELLE Paul-Henri, CAO Anton et GORET Axel
(Groupe 4 - CFM) CHERGUI Rayan, FOLLIOT Yohan et MARTIN Victor
(Groupe 5 - CGM) CHEMANACK Thierry, GLANDIER Quentin et MOREIRA Melissa
(Groupe 6 - DES) DAI Abdessalam, ELLOUMI Omar et SIMO TAMO Emmanuel D.
(Groupe 7 - DGK) DELANNOY Jules, GIRAULT Dylan et KAMGA Jessica Kamga
(Groupe 8 - EHM) EL TAYEB EL RAFEI Jana, HASSAN Yousra et MA Haiyang
(Groupe 9 - IL) INCERTI Félix, LUO Eric

2 Projet

Le projet est à effectuer en trinôme. Il consiste à traiter un problème de machine learning avec un jeu de données réelles, dont on vous les fournira. Il faut réaliser une démarche méthodologique de machine learning de bout en bout permettant de répondre au problème posé, à l'aide des algorithmes étudiés dans le cours. Justifier les choix de méthode, de paramètres, etc. ; lier les résultats des éventuelles analyses préalables des algorithmes.

3 Rapport

Le rapport doit inclure une présentation claire et concise du problème et des données, une définition de l'objectif de l'étude, la démarche choisie et les résultats obtenus, ainsi qu'une conclusion, sans dépasser **10 pages de PDF** en utilisant le template LaTeX fourni. D'éventuelles illustrations complémentaires, non essentielles pour la compréhension du travail, pourront être mises dans des annexes (appendices).

Le rapport (obligatoirement en PDF) accompagné du code source (un jupyter notebook (et/ou un script Python) qui a servi à générer les résultats présentés dans le rapport devra être archivé dans un seul fichier .ZIP et déposé sur moodle dans l'activité "Dépôt Projets AI28-P24" avant **Samedi 15/06/2024 à 23h59**.

Chaque figure devra comporter une légende expliquant ce qui est illustré. Attention, une pénalité sera appliquée si une figure n'est pas reproductible.

4 Template LaTeX

Vous allez utiliser le template LaTeX de la conférence *International Conference on Learning Representations 2024*, téléchargeable à l'adresse : https://github.com/ICLR/Master-Template/raw/master/iclr2024.zip. Il est très simple d'éditer ce template : titre, noms de chaque trinôme, sections, l'importation des figures et tableaux, conclusion, etc.. Pour ne pas faire une installation MikeTeX [1] (sous Windows) ou MacTeX [2] (sous MACOS), nous vous conseillons de créer un compte sur Overleaf [3] et charger le template latex ICLR sur Overleaf. Un lien va être généré que vous les partagez, ainsi chacun membre du groupe contribue à la rédaction du rapport.

5 Soutenance

Une séance de soutenance des projets aura lieu le **jeudi 20 juin de 10h15 à 12h45.** Les soutenances seront de durée 15 mm pour chaque trinôme et permettront de vérifier que chaque étudiant.e est bien auteur de son projet.

6 Distribution de projets

Nous vous proposons 8 projets avec des jeux de données réelles répertoriés dans le site web UCI Machine Learning Repository https://archive.ics.uci.edu/ml/index.php. La liste des titres de projets sont mis en ligne dans l'espace moodle. En cliquant sur un titre d'un projet, vous serez amenés à sa page de description ainsi que les données qui viennent avec.

Chaque projet peut être traité **au plus par deux groupes.** Une fois les membres de groupes se mettent d'accord pour le choix d'un projet, vous devez me contacter par mail pour son attribution. Attention ! j'applique la règle "FIFO (First In First Out)", premier arrivé, premier servi.

Un fichier Excel **Groupes et distribution de données** dans la section de Projets est mis à jour régulièrement pour les projets affectés aux groupes. Avant m'envoyer votre mail, vous devez consulter ce fichier pour savoir si le sujet que vous iriez choisir est encore disponible.

7 Points d'évaluation

7.1 Rapport

Le rapport doit contenir les principales étapes pour un projet machine learning de bout-en-bout :

- Description succincte de projet à mettre dans l'abstract (voir template).
- Analyse exploratoire de données.
- Pré-traitement de données.
- Modélisation : un large panel de modèles prédictifs doit figuré dans votre rapport.
- Optimisation : préciser les fonctions objectives utilisées et les algorithmes pour leur optimisation.

1. https://miktex.org
2. https://tug.org/mactex/
3. https://www.overleaf.com/

— Hyper-optimisation des paramètres de modèles (validation croisée, GridSearCV, ...). Nous vous encourageons d'utiliser des packages d'hyper-optimisation comme : `optuna`[4], ou `SHERPA`[5], ou `Hyperopt`[6], ou autre package de votre choix.
— Évaluation des modèles.
— Votre recommandation de modèle(s) pour le problème étudié.

La présentation des résultats et ses interprétations sont des points forts pour l'évaluation.

7.2 Code

Nous vous encourageons à écrire un code "pythonique", c'est à dire un code Python bien conçu et documenté pour faciliter sa compréhension, en particulier vous pouvez coder des petites fonctions avec des pipelines de pre-processing, modélisation, évaluation, etc ... pour autotomiser le plus possible l'exécution de vos programmes.

4. https://optuna.org
5. https://parameter-sherpa.readthedocs.io/en/latest/
6. http://hyperopt.github.io/hyperopt/