Syllabus projet

Année : **2021-2022**

Enseignant(s)	Email(s)
VIDAL Nicolas	nvidal@myges.fr

Projet Annuel 3 Big Data

1 Matières, formations et groupes

Matière liée au projet :

Formations : -

Nombre d'étudiant par groupe : **3 à 4**

Règles de constitution des groupes: **Imposé**

Charge de travail estimée par étudiant : **50,00 h**

2 Sujet(s) du projet

Type de sujet : **Imposé**

3 Détails du projet

Objectif du projet (à la fin du projet les étudiants sauront réaliser un...)

> Implémenter et utiliser des modèles et algorithmes simples relatifs au Machine Learning, combiner le tout dans un cas pratique réel.

Descriptif détaillé

Cadre général

Le résultat du projet annuel sera d'une part un rapport d'un minimum de 20 pages présentant l'étude des performances de tous les algorithmes et modèles sur la problématique choisie. Une version Jupyter interactive sera également à remettre sous format électronique.

Les étudiants seront fortement encouragés à utiliser l'outil Matplotlib pour produire les courbes et graphes retraçant leurs expérimentations ainsi que pour réaliser un document aisément interactif. Une attention toute particulière sera à apporter à l'étude de l'impact des différents paramètres des algorithmes étudiés sur la rapidité de convergence de ceux-ci.

D'autre part, les étudiants devront remettre une application (sous la forme d'un client lourd, d'un site web ou d'une application mobile) interagissant avec une API hébergeant différents modèles pré-entraînés grâce aux datasets appropriés et préalablement constitués.

L'essentiel des démarches des étudiants, des résultats obtenus et de leurs analyses devront être présentés également lors d'une soutenance.

Avant d'appliquer les algorithmes et modèles vus en cours à la problématique choisie, il sera impératif de démontrer la justesse de l'implémentation de ces derniers sur les cas de tests proposés. Ainsi, il est proposé d'implémenter ceux-ci sur des jeux de données classiques (données linéairement séparables, non linéairement séparables, tâches de classification, tâches de régression, etc.) tels que vus en cours de Machine Learning.

Les modèles et algorithmes à appliquer au projet de test sont :
• Modèle linéaire
• Perceptron Multi Couches
• Radial Basis Function Network
• SVM ou proposition du groupe d'étudiants

L'ensemble de ces modèles et algorithmes devront être implémentés en C, C++ ou Rust de manière à pouvoir être aisément utilisé comme une bibliothèque dynamique manipulée depuis des scripts python (optionnellement également en Unity).

Ces mêmes modèles et algorithmes devront être appliqués à la recherche d'un modèle tentant de solutionner une problématique applicative complexe, pour laquelle une implémentation humaine serait difficile (ex. distinction entre deux classes d'images).

Devront être mis en évidence par le biais de courbes les phénomènes vus en cours tels que (liste non exhaustive) :
• Le sous apprentissage
• Le sur apprentissage
• La mise en évidence de biais dans la base d'exemples
• La difficulté de trouver le modèle « suffisamment complexe pour correctement traiter les données d'apprentissage » mais « suffisamment simple pour bien généraliser »
• Etc.

Il est bon de noter que l'implémentation sera nécessaire pour obtenir des résultats, mais qu'il ne s'agit que de la première étape. L'intérêt principal du projet se situant dans la capacité des étudiants à commenter les résultats obtenus et à porter un regard critique sur l'ensemble des outils vus au travers de différents cours et à l'utilité de chacun face à une problématique issue du monde réel.

Les étudiants seront également encouragés à proposer d'autres modèles et algorithmes si ceux-ci s'avèrent pertinents face au choix de leur problématique. De même ils pourront tout à fait faire usage de la bibliothèque fournie de tensorflow/keras à titre de comparaison de leurs implémentations personnelles.

L'utilisation de GIT sera impérative pour assurer un suivi du projet aisé pour l'ensemble des membres du groupe ainsi que de l'encadrant.

L'utilisation d'implémentation externes des algorithmes et modèles vus en cours est proscrite (hormis à titre comparatif).

Le travail de chaque membre du groupe devra être clairement identifié (header de fichier, document de suivi, etc.)

Pour entraîner leur modèle dans le but de l'utiliser au sein d'une application cliente, les étudiants devront se munir d'un dataset approprié.

Plusieurs applications sont proposé es, mais si les é tudiants souhaitent soumettre une autre problé matique et qu'ils disposent d'un moyen de constitution d'un dataset pertinent, il pourra éventuellement être validé par l'enseignant.

Les couples problématiques/datasets envisagés sont les suivants :

Classification :
- Application permettant de catégoriser une image (photo) soit en classe 'Chat' soit en classe 'Chien' soit 'autre'

- Application permettant de différencier une image (screenshot) issue d'un RTS, d'un MOBA ou d'un FPS

- Application permettant de différencier des lettres manuscrites

- Application permettant de différencier une photo d'un match de foot d'un match de rugby

- Application permettant de différencier les drapeaux (photos) issus de différents pays

- Application permettant de différencier le genre d'un film en fonction de son affiche publicitaire

Régression :
- Application permettant de prédire l'âge d'une personne à partir d'une photo de son visage (attention à la constitution du dataset)

- Application permettant de prédire les calories d'un plat/produit en fonction de sa composition

- Application permettant de prédire la qualité d'une position de jeu aux Echecs

- Application permettant de prédire la qualité d'une position de jeu aux Dames

- Application permettant de pré dire la note (imdb) d'un film en fonction de diffé rents paramètres accessibles (attention au texte !)

Ouvrages de référence (livres, articles, revues, sites web...)

- MOOC de ré fé rence pour la partie thé orique des modèles et algorithmes : https://work.caltech.edu/telecourse.html
- https://www.kaggle.com/
- https://datasetsearch.research.google.com/
- https://archive.ics.uci.edu/ml/datasets.php

Outils informatiques à installer

- IntelliJ Pycharm (partie server) + IntelliJ CLion (partie lib c++ ou rust) ou Visual Studio
- Partie cliente au choix :
  * Unity / Rider
  * Front Web quelconque
  * JavaFX / XXX

4 Livrables et étapes de suivi

1	Etape intermédiaire	Problématiques applicatives choisies, repository git créé, pistes de constitution du dataset	**mercredi 30/03/2022 9h45**

2	Etape intermédiaire	Modèle linéaire appliqué aux cas de tests ainsi qu'à un bout du dataset en cours de constitution Transformation non linéaire pour les cas 'KO' PMC appliqué aux cas de tests ainsi qu'à une portion du dataset en cours de constitution • Livrables : o Projet de test opérationnel et sources o Rapport commentant les résultats observés. Début d'implémentation de l'application (tuyauterie)	**vendredi 27/05/2022 8h00**
3	Etape intermédiaire	Radial Basis Function Network et SVM (ou autre) appliqué aux cas de tests • Démonstration de l'implémentation de l'ensemble des algorithmes et modèles de réseaux de neurones étudiés sur les cas de tests ainsi que sur le dataset constitué • Possibilité de sauvegarder/charger des modèles entrainés et de les utiliser grâce au système client/serveur sur de nouvelles données • Livrables : o Projet de démonstration et sources o Première ébauche du rapport intéractif	**mercredi 20/07/2022 23h59**
4	Rendu final	Soutenance publique • Présentation de la démarche scientifique des étudiants pour aborder leurs problématique à l'aide des méthodes et algorithmes vu en cours. Rapide démonstration, analyse et critique des résultats obtenus. • Livrables : o Slides o Projet de démonstration et sources o Document interactif Jupyter o Rapport complet	**vendredi 22/07/2022 10h00**

5 Soutenance

Durée de présentation par groupe : **20 min** Audience : **Publique**

Type de présentation : **Présentation / PowerPoint - Démonstration**

Précisions :