Présentation du projet

Module Mathématiques appliquées aux Data Sciences niv. 2
YNOV - Filière Informatique - Spé DataIA - M1
Année scolaire 2024 / 2025
Intervenant : Nicolas Miotto

Modalités

- Le projet est à rendre sous un format dossier pdf à rendre le 20 / 01 / 2025 lors de la dernière séance du module.
- Il n'y a pas de présentation orale à préparer.
- Le dossier doit comporter 10 pages au maximum.
- Le dossier doit comporter un lien github de vers vos différents scripts python.
- Le dossier, incluant l'ensemble des scripts, sera noté /20 et représente 50% de la moyenne finale du module (les autres 50% pour le partiel écrit aussi noté /20).

Cahier des charges

1. Choisir un dataset :
    - sur le site : https://www.kaggle.com/datasets ;
    - thématique libre ;
    - une colonne contenant la labellisation des données ;
    - régression ou classification acceptée.
2. Présenter le dataset :
    - intérêt personnel concernant la thématique choisie ;
    - contenu ;
    - date de parution ;
    - nombres de lignes de données ;
    - pertinence des différentes dimensions (les colonnes) ;
    - contexte d'utilisation ;
    - exploitation possible des résultats des apprentissages.
3. Nettoyer le dataset :
    - formatage propre dans un DataFrame avec la librairie Pandas (avec indexation et header) ;
    - recherche et suppression des lignes comportant des valeurs NaN et des valeurs extrêmes ;

- ➢ recherche de corrélation entre certaines dimensions (avec la librairie Seaborn) et réduction éventuelle de la dimensionnalité ;
- ➢ normalisation des données ;
- ➢ descriptions des statistiques essentielles (moyennes, écarts-types, médianes, écarts interquartiles).

4. Isoler une ligne (aléatoirement) du dataset :
   - ➢ elle servira à la prédiction finale ;
   - ➢ ne pas l'utiliser lors des entraînements.
5. Concevoir un modèle de Machine Learning :
   - ➢ utilisation de la librairie Scikit-Learn ;
   - ➢ choix et pertinence du modèle algorithmique d'apprentissage supervisé ;
   - ➢ mixage de différents modèles accepté ;
   - ➢ optimisation des paramètres et hyperparamètres du/des modèle(s) ;
   - ➢ entraînement sur le dataset ;
   - ➢ évaluation du modèle (courbes d'erreur et courbe d'apprentissage) ;
   - ➢ compte-rendu exhaustif des savoirs mathématiques utilisés.
6. Concevoir un modèle de Deep Learning :
   - ➢ utilisation des librairies Keras de TensorFlow et/ou Numpy ;
   - ➢ choix et pertinence du nombres de neurones par couches et du nombre de couches cachées ;
   - ➢ optimisation des paramètres et hyperparamètres du modèle ;
   - ➢ entraînement sur le dataset ;
   - ➢ évaluation du modèle (courbes d'erreur et courbe d'apprentissage) ;
   - ➢ compte-rendu exhaustif des savoirs mathématiques utilisés (méthode de descente du gradient, composition de fonctions, activations, rétropropagation matricielle,…).
7. Effectuer une prédiction :
   - ➢ avec la ligne isolée en 4. ;
   - ➢ comparaison de la prédiction avec la labellisation théorique ;
   - ➢ effectuer la prédiction avec le modèle ML et le modèle DL ;
   - ➢ commentaire sur les éventuels écarts de prédiction.

Compétences évaluées (chacune évaluée sur 2 points)

1. Modéliser une problématique de data science en formulations mathématiques.
2. Concevoir et appliquer une solution algorithmique efficace.
3. Appliquer des techniques mathématiques pour analyser et traiter un jeu de données réel.
4. Identifier des modèles algorithmiques (leurs caractéristiques, leurs différences).
5. Utiliser les outils numériques à bon escient pour résoudre un problème.

6. Caractériser un réseau neuronal.
7. Optimiser et normaliser un modèle.
8. Visualiser des courbes de performances.
9. Effectuer une prédiction la plus précise possible.
10. Comprendre les notions mathématiques sous-jacentes à une construction de modèle d'apprentissage.