Sujet du projet du cours “Machine Learning”

Responsable : Julien Ah-Pine

SIGMA Clermont
Mastère Spécialisé Expert en Science des Données 2425

1 Objectif du projet

L'objectif de ce dossier est de vous évaluer sur votre capacité à mener une étude de cas complète sur une tâche d'apprentissage supervisé à l'aide du **langage Python** et dans un contexte de **data science**. Cela implique que votre rendu doit contenir à la fois une analyse descriptive en amont de l'analyse prédictive afin de mettre en perspective les informations contenues dans les *data*. Le protocole expérimental devra ensuite permettre d'argumenter scientifiquement le choix du meilleur modèle de prédiction. Si il y a lieu, en aval de l'analyse prédictive, une étude sur l'importance des variables devra permettre d'extraire des connaissances sur le phénomène à l'étude. Dans ce contexte, vous prendrez garde toutefois, à rester dans votre rôle de *data scientist* dont les modèles prédictifs qu'il pratique s'appuient sur la recherche de relations de corrélation dans des espaces arbitraires et qui en aucun cas ne traduisent des relations de causalité.

2 Organisation et remise du projet

Les projets s'effectuent individuellement. Il est attendu des étudiants qu'ils fournissent :

1. Un fichier `.py` dans lequel se trouveront le code principal et toutes les fonctions implémentées le cas échéant.
2. Un rapport au format `.pdf` présentant l'étude de cas, les analyses et les résultats obtenus avec graphiques à l'appui.

Vous devrez créer une archive .zip contenant ces deux fichiers. Vous nommerez votre archive et les éléments qui y sont contenus de la manière suivante ML_NOM_prenom.(zip|pdf|py). **Vous devrez soumettre votre archive par le biais de la page Moodle du cours ML au plus tard le vendredi 7 février 2025 23h59**.

ATTENTION : vous êtes responsables de votre envoi et donc toute absence de ressource ou tout problème conduisant à l'impossibilité d'accéder correctement à votre travail est de votre responsabilité.

3 Descriptif du travail à réaliser

Le projet consiste à étudier de façon approfondie un jeu de données (qui peut être relatif à l'activité de votre entreprise d'alternance) en mettant en oeuvre et en comparant plusieurs méthodes d'apprentissage supervisé du cours de ML.

La mise en perspective de votre travail devra se faire par une présentation claire du phénomène à l'étude, une description motivée de la tâche à résoudre (notamment dans le contexte de votre activité en entreprise le cas échéant), un exposé des variables cible et descriptives utilisées et la formulation des questions/hypothèses que vous cherchez à analyser.

3.1 Code

La tâche étant supervisée, il s'agit de traiter un problème de prédiction que ce soit dans le cadre de la régression ou de la catégorisation. Toutefois, votre code doit contenir toutes les étapes en amont

et en aval de la modélisation. Vous devrez ainsi implémenter le code permettant de mettre en oeuvre les points suivants :

- La lecture des données (que vous fournirez avec votre dossier en format .csv ou .xls ou .npz).
- Une analyse exploratoire des données utilisant des techniques univariées, bivariées et peremettant d'identifier des données manquantes ou des données aberrantes.
- Le prétraitement des données comprenant le cas échéant, la gestion des données manquantes et des données aberrantes.
- Le protocole expérimental permettant de comparer plusieurs méthodes supervisées et plusieurs combinaisons d'hyperparamètres. L'ensemble des méthodes testées doit contenir a minima :
  - Une méthode linéaire pénalisée par norme ℓ_2, une méthode linéaire pénalisée par norme ℓ_1.
  - Une méthode non paramétrique basée sur les plus proches voisins ou un arbre de décision.
  - Une méthode basée sur des SVM.
  - Une méthode ensembliste basée sur l'agrégation simple par moyenne artithmétique ou vote majoritaire des méthodes individuelles précédentes.
  - Une méthode ensembliste basée sur l'agrégation par stacking des méthodes individuelles précédentes (vous choisirez comme bon vous semble la méthode de la 2ème couche).
  - Une méthode ensembliste des random forest.
  - La méthode ensembliste issue de AdaBoost.
  - La méthode ensembliste issue du Gradient Boosting ou de XGBoost.
- Des graphiques permettant de comparer les résultats de chaque méthode avec les différents paramétres utilisés.

3.2 Rapport

Le rapport devra contenir :

- Une présentation du jeu de données et de la problématique abordée en replaçant ceux-ci dans le contexte de l'entreprise le cas échéant.
- Une présentation succincte des méthodes supervisées testées dans votre rapport.
- Une présentation succincte des librairies Python utilisées pour mettre en oeuvre chaque méthode.
- Une présentation de vos analyses exploratoires de données.
- Une présentation de votre protocole expérimental pour l'analyse prédictive comprenant les différents ensembles de paramètres et d'hyperparamètres utilisés pour chaque méthode.
- Une présentation, analyse et discussion des résultats obtenus pour chaque méthode et vis à vis de chaque ensemble de paramètres utilisés.
- Une présentation, analyse, comparaison et discussion des meilleurs résultats obtenus des différents modèles, conduisant à une conclusion sur le modèle que vous préconisez finalement pour traiter la problématique initiale.
- Le cas échéant, une présentation, analyse de l'importance des variables du modèles sélectionné.