OFFRE DE STAGE MASTER 2
Mathématiques appliquées

Classification de signaux, analyse de similarité et visualisation de données

SUMMIT est une unité de service de Sorbonne Université, créée au 01 janvier 2021, qui a pour mission principale de faciliter les collaborations de recherche avec le monde industriel.

Contexte du projet

Ce stage s'inscrit dans le cadre d'un projet de recherche sur la détection du SARS-CoV-2 dans les eaux usées mené par le réseau Obépine. Différents projets d'intérêt ont récemment émergé et seront proposés au stagiaire.

Le stage pourra faire intervenir une ou plusieurs des thématiques suivantes :

Thème 1 : Data Visualisation

L'objectif du stage est de trouver de nouvelles représentations graphiques permettant au réseau Obépine de mieux communiquer ses résultats auprès du grand public mais aussi des décideurs locaux.

Tâches confiées :

- Effectuer un état de l'art des plate-formes de visualisation des réseaux chargés de la surveillance des eaux usées dans le monde

- Confronter leur approche de présentation des données avec celle du réseau Obépine et du Ministère de la Santé et des Solidarités

- Imaginer de nouvelles représentations graphiques permettant de communiquer les résultats du réseau Obépine au grand public

- Imaginer des outils de visualisation permettant d'aider les décideurs locaux à la prise de décision à partir de ces données

Thème 2 : Apprentissage statistique

Projet 1 : Etiquetage automatique de signaux

Le réseau Obépine génère de manière hebdomadaire environ 200 rapports d'analyses destinés aux ARS et aux décideurs politiques locaux. Ces rapports contiennent :

- Une description exhaustive du contenu et de la méthode de création d'un indicateur des eaux usées
- Différentes représentations graphiques sous forme de séries temporelles de l'indicateur des eaux usées ainsi que des données cliniques (incidence, hospitalisations)
- Une interprétation des séries temporelles des eaux usées, sous forme de commentaire.

Actuellement, les commentaires interprétant les séries temporelles sont générés de façon automatique et décrivent les courbes de la manière suivante :

- La tendance de l'indicateur sur les 7 et 30 derniers jours
- Le niveau de l'indicateur sur les 7 derniers jours

L'automatisation de ces commentaires est néanmoins assez basique. Les indications sur la tendance rentrent dans 3 catégories (hausse, stabilité et baisse) et sont basées sur des seuils numériques arbitraires. Ils ne permettent donc pas de rendre compte de différentes variations successives ayant pu survenir sur la période d'intérêt. Par exemple, un signal qui va connaître une oscillation avec une phase de hausse, de stabilité, puis de baisse sera souvent traduit par un signal stable sur 30 jours, alors que divers régimes se sont succédé. L'idée du projet est donc de pousser un peu plus loin l'approche actuellement déployée en production. Nous imaginons le projet de la manière suivante, mais resterons à l'écoute de nouvelles propositions :

- Étiquetage préalable d'une banque de signaux par des experts métier en concertation avec le ou la stagiaire
- Création de descripteurs adaptés à la résolution du problème à partir des signaux (feature engineering)
- Entraînement et sélection de modèles
- Détermination des seuils de passage d'une classe à une autre par une méthode de classification supervisée et comparaison avec les seuils arbitraires utilisés jusqu'ici

Projet 2 : Analyse de similarité entre des communes à partir de signaux d'eaux usées

A ce jour, le réseau Obépine suit les analyses d'eaux usées d'environ 200 stations d'épuration en France.

L'objectif principal de ce projet est de trouver des descripteurs permettant d'obtenir des mesures de similarité entre ces différentes stations.

Nous imaginons le projet de la manière suivante :

- Récolte des données ouvertes et pertinentes pour la tâche d'analyse de similarité
- Création de descripteurs adaptés à partir de ces données (feature engineering)
- Regroupement des bassins versants ayant des comportements similaires (clustering)

Projet 3 : Inférence des valeurs futures

Les mesures en eaux usées étant soumises à diverses sources de variabilité, l'un des enjeux pour les décideurs est de connaître la probabilité qu'une tendance détectée en semaine S se poursuive effectivement en semaine S+1. En ce sens, un algorithme a déjà été développé et permet d'estimer les probabilités de changement ou de maintien de la tendance sur les 7 derniers jours jusqu'à la prochaine mesure. L'objectif de ce projet consiste à améliorer les performances de ce modèle.

Nous imaginons le projet de la manière suivante :

- Étude bibliographique portant sur les solutions adaptées à un problème de classification en classes multiples
- Feature engineering
- Confrontation des performances à l'algorithme actuellement déployé
- Amélioration de la fenêtre temporelle de prédiction

Profil recherché

Étudiant en Master 2 Mathématiques appliquées/Science des données.

Pour le thème 2, une connaissance des méthodes d'apprentissage supervisé et non-supervisé et des approches de sélection de modèle serait fortement appréciée.

Pour le thème 1, des compétences de programmation en Javascript, D3.js et Python seront requises, ainsi que des connaissances de base en HTML et CSS.

Pour le thème 2, des compétences de programmation en Python seront requises et une bonne connaissance des librairies standards d'analyse de données et d'apprentissage statistique (numpy, pandas, scikit-learn, etc) sera appréciée.

Mots-clés : science des données, apprentissage statistique, data visualisation, mathématiques appliquées

Durée : 6 mois à partir de début avril 2022.

Lieu : SUMMIT – tour 33-34

Rémunération : gratification de stage

Merci d'adresser votre candidature (CV et lettre de motivation) à :

valerie.neyrolles@sorbonne-universite.fr

nora.aissiouene@sorbonne-universite.fr

nicolas.cluzel@sorbonne-universite.fr