Sujet de postdoc - Financement de 12 mois
Institut de Mathématiques de Bordeaux
LyRE, centre de recherche de SUEZ - Bordeaux

Algorithmes stochastiques du second ordre pour l'apprentissage statistique et applications à la recherche de fuites dans des réseaux de distribution d'eau potable.

Encadrants (IMB) : Jérémie Bigot & Vincent Couallier
jeremie.bigot@u-bordeaux.fr; vincent.couallier@u-bordeaux.fr

Contexte du sujet de postdoc

Ce projet de recherche est motivé par une collaboration entre l'Institut de Mathématiques de Bordeaux et le LyRE, centre de recherche de SUEZ à Bordeaux dont l'objectif est d'innover dans la gestion quantitative mais aussi qualitative de l'eau pour limiter l'impact des métropoles sur leur environnement.

L'un des axes de recherche du pôle Data du LyRE est centré sur le développement d'un module de localisation de la recherche de fuites dans les réseaux d'eau potable. L'objectif du module localisation de la recherche de fuite est de cibler les ouvrages (canalisation et/ou ouvrage) susceptibles d'être fuyards, permettant ainsi une meilleure organisation de la recherche de fuite, et une meilleure protection de la ressource en économisant l'eau grâce à des durées d'écoulement des fuites plus courtes. Ce module de localisation doit permettre dans un premier temps d'identifier les secteurs les plus à risque de fuite et dans un second temps quels sont les ouvrages responsables. Il existe des solutions au sein de SUEZ pour la gestion patrimoniale des ouvrages d'eau potable, cependant ces solutions sont préférées pour la localisation du renouvellement de ces ouvrages. Elles ne sont pas adéquates pour orienter la recherche de fuite, la raison principale étant qu'elles ne prennent pas en compte des variables ayant une dynamicité suffisante pour permettre d'identifier sur un pas de temps suffisamment fin les ouvrages à prioriser pour la recherche de fuite. Le module localisation doit ainsi permettre de pallier ces limites en utilisant des données dynamiques dans un modèle de type machine learning, sur un pas de temps adéquat pour la recherche de fuite.

La construction de ce modèle de localisation est basé sur des données d'étude qui proviennent de différentes sources : Sytème d'Information Géographique (SIG), données opérationnelles, informations récoltées auprès des experts métier, open data, etc... En particulier, les données sont des mesures d'environ une centaine de caractéristiques statiques et dynamiques, relevées par ouvrage, par semaine, sur une période de plusieurs années sur le réseau métropolitain d'une grande agglomération. Les données proposées sont majoritairement des séries temporelles, appelées mesures dynamiques (capteurs acoustiques, température de l'eau, météo, interventions), par opposition aux propriétés intrinsèques des ouvrages, appelées mesures statiques (caractéristiques patrimoniales, environnementales). Chaque ouvrage est identifié par un numéro et une date de relevé des mesures. Pour chaque ouvrage, une étiquette indique si une fuite existait à cette date pour cet ouvrage. Toutefois, il est à noter que la problématique est plus délicate que celle du cadre usuel de classification supervisé. En effet, il existe une imperfection dans la labellisation de ces données. Il peut parfois y avoir des fuites qui coulent plusieurs semaines ou des mois avant d'être identifiées. Dans ce cas, l'étiquette sera qu'il n'y a "pas de fuite" à tort. Pour y palier, il a été défini des zones "grises" avant la detection de fuites, et ces périodes pourraient exclue de l'entrainement d'un modèle d'apprentissage statistique. Toutefois,

il pourrait être très interessant d'envisager une meilleure manière de prendre en compte cette incertitude.

Un tel jeu de données permet ainsi d'envisager le développement d'algorithmes d'apprentissage supervisé en grande dimension afin d'estimer le risque de casse sur les canalisations et les ouvrages par l'estimation du risque de casse sur un pas de temps hebdomadaire

Principaux objectifs du postdoc

Les objectifs de ce postdoc sont d'étudier, pour le développement d'un module de localisation de recherche de fuites, l'intérêt de méthodes d'apprentissage statistique basés sur des algorithmes stochastiques du second ordre du type Newton [2, 3] dans des modèles de régression logistique [1] ou des modèles plus complexes de machine learning du type réseaux de neurones profonds. Les principales difficultés liées à la base de données qui sera considérée sont d'une part la grande dimensionnalité du nombre de mesures ainsi que le faible taux de mesures liées à des fuites. Il s'agit donc d'un problème de classification supervisée avec un fort déséquilibre entre les 2 classes. Il conviendra donc d'adapter les méthodes d'apprentissage statistique à ce contexte, et d'étudier également la potentielle influence de la taille de modèle de réseaux de neurones dans des problèmes de classification très déséquilibrée.

Le postdoc est de nature à la fois théorique, numérique et appliqué. Les principales notions abordées feront appel à des outils de statistique mathématique et computationnelle. Il nécessite un doctorat de mathématiques appliquées en statistique, ainsi que la maîtrise du langage de programmation Python pour la science des données.

La durée de financement du postdoc est de 12 mois, et la date de début est prévue entre novembre 2023 et janvier 2024. La candidature (incluant CV + lettre de motivation + lettre de recommandation) est à envoyer à

`jeremie.bigot@u-bordeaux.fr` et `vincent.couallier@u-bordeaux.fr`

References

[1] Bach, F. Adaptivity of averaged stochastic gradient descent to local strong convexity for logistic regression. *Journal of Machine Learning Research 15* (2014), 595–627.

[2] Bercu, B., Bigot, J., Gadat, S., and Siviero, E. A stochastic Gauss-Newton algorithm for regularized semi-discrete optimal transport. *Information and Inference: A Journal of the IMA* (05 2022). iaac014.

[3] Bercu, B., Godichon, A., and Portier, B. An efficient stochastic newton algorithm for parameter estimation in logistic regressions. *SIAM Journal on Control and Optimization 58*, 1 (2020), 348–367.