Stage de Recherche en Data Privacy

Présentation de Sarus

Les technologies développées par Sarus permettent d'entraîner des modèles d'IA sans accéder à la donnée d'origine. En implémentant une protection mathématiques des données personnelles (dite de confidentialité différentielle ou *differential privacy*), Sarus accélère l'innovation sur données sensibles et ouvre de nouvelles formes de collaboration sur la donnée. La technologie s'insère directement dans l'infrastructure de données du client et permet aux analystes et data scientists de travailler avec leurs outils standards.
Sarus est une startup parisienne financée par des VC français, elle est membre de Y Combinator et est active en santé, finance et énergie.

Projet de stage

La génération de données synthétiques sous contrainte de confidentialité différentielle est au cœur de l'offre Sarus. Le projet consistera à creuser les aspects théoriques du problème et à développer des approches concrètes inspirées des travaux de recherche dans le produit Sarus. Les questions théoriques posées par la génération de données synthétiques privées sont nombreuses, et incluent notamment:

- Génération de données synthétiques en faible dimension : peut-on modéliser toutes les marginales de rang faible à un coût contrôlé en privacy en utilisant des algorithmes de Machine Learning (ML) classiques (plutôt que DP-SGD) ?
- Génération de données synthétiques en grande dimension : Peut-on améliorer DP-SGD (une revue détaillée de la littérature sera nécessaire) ? Peut-on consommer moins de privacy lorsque l'on passe à la limite d'un réseau de neurones infiniment large ?

Une partie du stage consistera à donner des éléments de réponse à ces questions.
Une autre partie consistera à développer un module permettant d'encoder des variables réelles en vecteurs utilisables dans notre modèle de données synthétiques (real embedding) et à décoder un vecteur de sortie de notre modèle en distribution réelle (real distribution embedding) duquel on peut échantillonner des valeurs ou à partir de laquelle on peut calculer une log-vraisemblanc.

Cadre du stage

Le stage se déroulera au sein de l'équipe Sarus R&D sous la responsabilité de Nicolas Grislain (CSO), en collaboration avec Olivier Cappé (directeur de recherche CNRS au DI ENS à Paris).
L'objectif de ce stage est d'entamer un travail de recherche, se poursuivant en thèse CIFRE dans l'équipe R&D de Sarus et encadrée par Olivier Cappé à l'ENS.

Votre profil

Issu d'une grande école d'ingénieur ou autre institution de premier plan prouvant votre excellence en mathématiques appliquées et machine learning, vous cherchez un stage de M2 recherche dans une startup en vue de continuer en thèse.

Vous disposez de solides en connaissances en:

- théorie des probabilités,
- modélisation statistique,
- optimisation,
- machine learning,
- algorithmique, structures de données

Vous avez de l'expérience en programmation, de préférence en python et java, et une connaissance de base du développement sous Unix.

Mais surtout vous faites preuve de :

- Motivation et envie d'apprendre (ouverture à la critique constructive, au feedback)
- Ouverture aux autres et sens de l'entraide
- Intérêt pour les produits et plus généralement la mission de Sarus

Processus de candidature et recrutement

Pour candidater, merci d'envoyer votre CV à careers@sarus.tech avec un court (3 lignes max) email d'introduction et de motivation.

- Discussion téléphonique pour se connaitre
- Entretien CSO, tests techniques et rencontre avec toute l'équipe

Informations complémentaires

- Type de contrat : Stage (3 à 6 mois)
- Date de début : selon disponibilité du candidat
- Lieu : Paris centre
- Niveau d'études : Master 2
- Rémunération : à discuter mais aligné avec le marché