Offre n°2024-07919

PhD Position F/M AI model audit

Le descriptif de l'offre ci-dessous est en Anglais

Type de contrat : CDD

Niveau de diplôme exigé : Bac + 5 ou équivalent

Fonction : Doctorant

Niveau d'expérience souhaité : Jeune diplômé

A propos du centre ou de la direction fonctionnelle

The Inria Saclay-Île-de-France Research Centre was established in 2008. It has developed as part of the Saclay site in partnership with Paris-Saclay University and with the Institut Polytechnique de Paris since 2021.

The centre has 39 project teams , 27 of which operate jointly with Paris-Saclay University and the Institut Polytechnique de Paris. Its activities occupy over 600 scientists and research and innovation support staff, including 54 different nationalities.

Contexte et atouts du poste

Contex: Rich and complex AI systems are increasingly used across multiple actors of society, from large language models to diagnostic models.
Stakeholders generally ask for these systems to be audited, with initiatives such as the AI safety institute and many questions from
regulatory bodies. However, there is a mismatch between the probabilistic objects of modern AI and the desired safety warrants: certitudes at the level of individual cases. From an engineering standpoint, in many complex settings, answers are best expressed with uncertainty quantification. From an audit standpoint, this quantification needs to be evaluated, both in its own right and insofar as it is linked to decision-making. This control of uncertainty is recognized as one of the main challenges of machine learning in high-stakes applications such as healthcare [11].

One challenge lies in the fact that individual probabilities are never directly observed; instead, only discrete labels are available. The
machine-learning literature has predominantly focused on the concept of "calibration error" [5], which controls the error rate given a
confidence score (i.e. a probabilistic output). A calibration error of zero implies that a predictor is neither over-confident nor under-
confident. However, this measure being an average control applied across all individuals, it does not preclude the possibility of systematic over-confidence for some individuals and under-confidence for others. Likewise, conformal prediction methods come with certain guaranties on uncertainty, yet the strong results are marginal [9]. "Proper scoring rules" give finer a characterization: these functions fully control errors on individual probabilities via observed samples [4]. Yet, their value does not relate simply to an error rate that can be understood in application terms. Our recent research has delved into the decomposition of these into calibration, grouping, and irreducible errors. We have introduced an estimator for the grouping term, thereby completing existing estimators of calibration error. This approach allows for a comprehensive characterization of errors in probabilistic predictions [6]. Using these tools on large language models reveals cultural biases where the models' uncertainty is more erroneous for answers about east-Asians than north Americans; biases that can then be partly corrected [2].

Environment: This PhD will take place at Inria Saclay, in the Soda team. Soda is a team of 25 people with 4 PIs
doing computational and statistical research, both fundamental and applied, to harness large databases on health and society. Soda also develops core software tools such as scikit-learn. For the application partners of soda and their corresponding application contexts, model validation and auditing are crucial, and the team has developed expertise on related topics.
The two advisors have developed and applied crucial prior work for auditing predictors.
Weekly meetings will be organized with the candidate in Inria.

We will also collaborate with stakeholders outside of mathematics and computer science, to make sure that we bring answers that are
relevant to the broader society and that we run empirical studies close to important application senarios.
For this, we will work with the workforce of the PEPR Santé Numérique that considers evaluation of

medical devices, some of them embarking AI.

References:

[1] Solon Barocas, Moritz Hardt, and Arvind Narayanan. Fairness and machine learning: Limitations and opportunities. MIT Press, 2023.

[2] Lihu Chen, Alexandre Perez-Lebel, Fabian M Suchanek, and Gaël Varoquaux. Reconfidencing LLMs from the grouping loss
perspective. arXiv preprint arXiv:2402.04957, 2024.

[3] Jérôme Dockès, Gaël Varoquaux, and Jean-Baptiste Poline. Preventing dataset shift from breaking machine-learning biomarkers.
GigaScience, 10(9):giab055, 2021.

[4] Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. Journal of the American statistical
Association, 102(477):359–378, 2007.

[5] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In International conference on machine learning, pages 1321–1330. PMLR, 2017.

[6] Alexandre Perez-Lebel, Marine Le Morvan, and Gaël Varoquaux. Beyond calibration: estimating the grouping loss of modern neural
networks. ICLR, 2023.

[7] Alexandre Perez-Lebel, Gaël Varoquaux, Marine Le Morvan, Julie Josse, and Jean-Baptiste Poline. Benchmarking missing-values
approaches for predictive models on health databases. GigaScience, 11, 2022.

[8] Ievgen Redko, Emilie Morvant, Amaury Habrard, Marc Sebban, and Younès Bennani. Asurvey on domain adaptation theory: Learning bounds and theoretical guarantees. arXiv preprint arXiv:2004.11829, 2020.

[9] Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. Journal of Machine Learning Research, 9(3), 2008.

[10] Ryan J Tibshirani, Rina Foygel Barber, Emmanuel Candes, and Aaditya Ramdas. Conformal prediction under covariate shift.
Advances in neural information processing systems, 32, 2019.

[11] Ben Van Calster, David J McLernon, Maarten Van Smeden, Laure Wynants, Ewout W Steyerberg, Topic Group 'Evaluating diagnostic tests, and prediction models' of the STRATOS initiative Patrick Bossuyt Gary S. Collins Petra Macaskill David J. McLernon Karel GM Moons Ewout W. Steyerberg Ben Van Calster Maarten van Smeden Andrew J. Vickers. Calibration: the achilles heel of predictive analytics. BMC medicine, 17:1–7, 2019.

Mission confiée

The goal of this PhD is to provide mathematical results and tools to guide auditing such systems: How to best interrogate a system? Given a set of results, inputs and outputs of the system, what can we extrapolate on other inputs?

The work will strive to move away from population level, enabling answers as tight as possible at the individual level. We will consider
statistical results on estimating expected errors and bounds for a black-box AI system given observations of its inputs and outputs. We will also consider procedures to best audit a system. To give a meaningful quantification of the impact of errors, we will consider bridging to decision theory and notions of utility. Finally, we will loop back with stakeholders to define utilities and understand better the non-formal aspects of audits.

Principales activités

* Finding and reading related bibliographic material

* Formalizing the problem

* Deriving mathematical results and related algorithms

* Numerical implementation and experimentation

* Writing corresponding publications

Compétences

* Good written and spoken English proficiency

* Proficiency in Python, numpy, and scikit-learn

* Knowledge of mathematics, statistics, and computer-science fundamentals of machine learning

* Good statistical background, with machine learning knowledge.

* Curious mindset.

Avantages

- Subsidized meals
- Partial reimbursement of public transport costs
- Leave: 7 weeks of annual leave + 10 extra days off due to RTT (statutory reduction in working hours) + possibility of exceptional leave (sick children, moving home, etc.)
- Possibility of teleworking and flexible organization of working hours
- Professional equipment available (videoconferencing, loan of computer equipment, etc.)
- Social, cultural and sports events and activities
- Access to vocational training
- Social security coverage

Rémunération

Minimum remuneration: 2,100 € gross/month

Order of August 29, 2016 setting the remuneration of contract doctoral students

Informations générales

- **Thème/Domaine :** Optimisation, apprentissage et méthodes statistiques
  Statistiques (Big data) (BAP E)
- **Ville :** Palaiseau
- **Centre Inria :** Centre Inria de Saclay
- **Date de prise de fonction souhaitée :** 2024-09-01
- **Durée de contrat :** 3 ans
- **Date limite pour postuler :** 2024-08-31

Contacts

- **Équipe Inria :** SODA
- **Directeur de thèse :**
  Varoquaux Gael / Gael.Varoquaux@inria.fr

A propos d'Inria

Inria est l'institut national de recherche dédié aux sciences et technologies du numérique. Il emploie 2600 personnes. Ses 215 équipes-projets agiles, en général communes avec des partenaires académiques, impliquent plus de 3900 scientifiques pour relever les défis du numérique, souvent à l'interface d'autres disciplines. L'institut fait appel à de nombreux talents dans plus d'une quarantaine de métiers différents. 900 personnels d'appui à la recherche et à l'innovation contribuent à faire émerger et grandir des projets scientifiques ou entrepreneuriaux qui impactent le monde. Inria travaille avec de nombreuses entreprises et a accompagné la création de plus de 200 start-up. L'institut s'efforce ainsi de répondre aux enjeux de la transformation numérique de la science, de la société et de l'économie.

Attention: Les candidatures doivent être déposées en ligne sur le site Inria. Le traitement des candidatures adressées par d'autres canaux n'est pas garanti.

Consignes pour postuler

Sécurité défense :
Ce poste est susceptible d'être affecté dans une zone à régime restrictif (ZRR), telle que définie dans le décret n°2011-1425 relatif à la protection du potentiel scientifique et technique de la nation (PPST). L'autorisation d'accès à une zone est délivrée par le chef d'établissement, après avis ministériel favorable, tel que défini dans l'arrêté du 03 juillet 2012, relatif à la PPST. Un avis ministériel défavorable pour un poste affecté dans une ZRR aurait pour conséquence l'annulation du recrutement.

Politique de recrutement :
Dans le cadre de sa politique diversité, tous les postes Inria sont accessibles aux personnes en situation de handicap.