

Stage : Machine Learning Engineer (H/F) Apprentissage adaptatif et *stream mining*

Qui sommes-nous ?

Nous pourrions commencer en nous présentant comme l'un des leaders sur le marché international du conseil en data, avec un chiffre d'affaires défiant ceux des G.A.F.A. Mais non. Nous préférons mettre en avant notre cadre de travail, nos réussites et surtout nos consultants. Allez venez, c'est par là...

Quantmetry c'est une centaine de personnes qui travaille de concert pour accompagner nos clients dans leurs réflexions, P.O.C., industrialisation autour de la Data, l'Intelligence Artificielle, le Machine Learning et la Big Data. Nous sommes ce qu'on peut appeler communément un cabinet de conseil pure player en Data.

N'étant pas sectorisés, nous intervenons sur différents sujets (voici une petite liste non-exhaustive) :

- La maintenance prédictive,
- La segmentation clients, le scoring,
- La détection de fraude, de corruption, de blanchiment d'argent,
- Le développement des architectures de plateformes Data,
- L'analyse de textes et d'images dans le cadre de la lutte contre le cancer du sein.

Dans le cadre de notre incessante croissance, nous recherchons des personnes motivées pour nous accompagner et ainsi travailler sur des projets de recherche.

Préalable

Quantmetry propose ci-dessous le volet R&D d'une offre de stage. Tout stagiaire entrant chez Quantmetry, outre le travail de R&D qui lui est proposé et pour lequel il sera encadré, aura aussi pour objectif de participer à certaines missions de conseil chez des clients variés, lui permettant d'aborder le monde du consulting.

Contexte

En présence de données fortement volatiles, les modèles standards de machine learning voient leurs performances décroître dans le temps, et devenir obsolètes. C'est le cas par exemple des algorithmes de détection de fraude, qui doivent se battre contre un adversaire aux stratégies changeantes, les algorithmes d'enchères publicitaires, qui doivent surfer sur les vagues de promotions et les changements de consommation, ou encore les algorithmes de prévision de consommation électrique, soumis aux aléas de la vie et situations familiales et professionnelles. La dérive des signaux contenus dans les données est appelée dérive conceptuelle (concept drift). Il existe une famille d'algorithmes spécialement dédié à l'apprentissage sur des données évolutives à forte volumétrie (stream mining), les algorithmes adaptatifs.

L'essentiel du savoir-faire et de la recherche en apprentissage adaptatif est centralisé sur trois outils :

- Scikit-multiflow, écrit en python, développé par Telecom ParisTech,
- Massive Online Analysis (MOA), écrit en Java, développé par l'université de Waikato (Nouvelle-Zélande)
- Apache Scalable Massive Online Analysis (SAMOA), écrit en Java et incubé par la fondation Apache, développé par Yahoo Labs (Barcelone).

SAMOA vise la scalabilité des algorithmes. MOA concentre l'état de l'art en la matière et est entièrement compatible avec SAMOA. Scikit-multiflow est une version allégée de MOA écrite en python et avec la même syntaxe que scikit-learn, dans une volonté de diffusion élargie.

Les objectifs de ce stage sont :

- Prendre en main les 3 outils et les appliquer à des jeux de données simples en accès libre. On évaluera notamment la capacité de SAMOA à s'intégrer dans un environnement Big Data.
- Comparer leurs fonctionnalités, évaluer leur complétude vis-à-vis de l'état de l'art et lister leurs avantages/inconvénients
- Construire un guide méthodologique et un tutoriel d'apprentissage adaptatif avec les outils jugés les plus pertinents
- Animer une formation en interne sur l'apprentissage adaptatif