

Stage : Data Scientist (H/F) Adversarial Machine Learning - GANs

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

Quantmetry se positionne depuis plusieurs années comme pionnier sur des sujets d'Intelligence Artificielle nécessitant une forte composante R&D. Afin de satisfaire au mieux les besoins de ses clients, Quantmetry se structure autour de pôles d'expertises.

Au cours de cette R&D, les méthodologies de modèles génératifs basés sur le Deep Learning ont été investigués, et plus particulièrement les **Generative Adversarial Networks (GANs)**.

Les GANs sont un cadre d'estimation de modèle génératif. Il s'agit d'une façon d'entraîner des réseaux de neurones à représenter implicitement la distribution qui a permis de générer un certain jeu de données. On peut alors générer de nouveaux échantillons, plus ou moins réalistes selon le niveau de qualité du modèle. Plus précisément, deux réseaux sont entraînés simultanément en concurrence : un générateur G (responsable de la génération de données, par exemples des images) et un discriminateur D (dans le cas général, il s'agit d'un classifieur estimant la probabilité qu'une donnée provienne des données d'origine).

Appliqués à des données images, mais aussi tabulaires, de nombreux usages sont à portée de main, et la recherche académique est prolifique.

TÂCHE PROPOSÉE

En suite de premiers travaux sur la génération de **données tabulaires** numériques continues, discrètes et catégorielles, ayant pour objectif de générer des données synthétiques suffisamment qualitatives pour remplacer des données réelles lors de l'entraînement de modèles de Machine Learning de type classification, le stage proposé vise à capitaliser sur ces développements afin de les généraliser ces travaux, avec un objectif d'amélioration de la robustesse des modèles ML.

L'objectif sera donc d'investiguer l'**adversarial machine learning**, ou comment **générer de nouvelles données améliorant la robustesse d'un modèle**.

Les GANs utilisés pourront être par exemple WassersteinGAN avec Gradient Penalty, Cramer GAN, éventuellement MedGan ou encore couplés à un autoencodeur.

Les méthodes hors GAN doivent également être benchmarkées.

RÉSULTATS ATTENDUS

Les résultats attendus sont le code source générique documenté accompagné d'une documentation scientifique et technique.

Une ou plusieurs restitutions sous forme de présentation orale à des publics experts ou novices sont également à prévoir.

COMPÉTENCES SOUHAITÉES

Savoir-être

- Autonome
- Travail en équipe
- Proactif
- Esprit critique

Data Science

- Machine Learning
- Deep Learning
- Statistiques

Engineering

- Python
- TensorFlow, Keras
- Bonnes pratiques de développement logiciel