Quantmetry
Building AI with pionneers

Stage : Data Scientist (H/F)
Adaptation de domaine et transport optimal

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.

Quantmetry
Building AI with pionneers

Contexte

Le machine learning est bâti sur une hypothèse rarement remise en cause, à savoir : les données sont indépendantes et identiquement distribuées. C'est l'hypothèse implicite et sous-jacente à toutes les techniques de séparation train/test et de validation croisée. Toutefois en pratique, cette hypothèse est rarement vérifiée. En effet, les données sont des signaux envoyés par un monde en constante évolution. A ce titre, les données n'échappent pas à la règle : elles évoluent dans le temps également, plus ou moins rapidement. C'est un vrai problème pour la mise en production de modèles prédictifs, car le jeu d'entraînement est souvent éloigné dans le passé, rendant le modèle obsolète pour prédire aujourd'hui. L'adaptation de domaine est la science qui cherche à résoudre ce problèmen: sachant un jeu de données labellisées (X, Y) et un jeu de prédiction non-labellisé Z, comment transformer X pour qu'il ressemble le plus possible à Z et corriger l'obsolescence du modèle ?

Plusieurs outils implémentent des solutions d'adaptation de domaine, et plus largement de transport optimal. Certaines sont disponibles en accès libre, en particulier :

- libTLDA, développé par l'université de Delft (Pays-Bas)
- salad, développé par l'université de Tubingen (Allemagne)
- POT, développé par l'université Nice Sophia-Antipolis et l'Université Bretagne-Sud

La promesse de ces outils est de fournir une interface type sckit-learn et de démocratiser l'usage de l'adaptation de domaine.

Les objectifs de ce stage sont :

- Prendre en main les 3 outils et les appliquer à des jeux de données simples en accès libre
- 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'adaptation de domaine avec les outils jugés les plus pertinents
- Animer une formation en interne sur l'adaptation de domaine