Sujet de Stage de Master 2/ de fin d'étude d'école d'ingénieur
Durée 6 mois

Apprentissage automatique pour la prévision et *feature engineering* basé sur la sémantique des données

Mots-clés

Apprentissage automatique, régression, prévision, machine learning.

Le contexte

De plus en plus de sociétés utilisent les solutions de Talend pour leurs traitements de données, toujours plus nombreuses. La valorisation et un traitement complet des données est devenu un enjeu majeur pour les entreprises d'aujourd'hui. Talend recherche un/une stagiaire en apprentissage statistique pour développer son offre et ses services en matière de prévision et de traitements avancés de la donnée.

Sujet

L'objectif du stage est de commencer le développement d'une chaine automatique de construction de modèles de prévision. En se basant à la fois sur des bibliothèques open source d'apprentissage (ex : scikit-learn) et sur des outils développés par Talend (ex : librairie de reconnaissance de types sémantiques, fonction de pré-processing de données développées dans les librairies de data quality ou l'outil Talend Data Preparation), l'objectif du stagiaire sera de développer une première approche de construction de modèle de régression pour la prévision. On se restreindra volontairement à un cadre de régression temporelle pour commencer. Une attention particulière sera portée à la généricité des résultats, l'approche proposée devant se généraliser au mieux à la diversité des cas d'usages de nos clients. Un autre point d'attention sera porté à la pédagogie autour des modèles. On s'intéressera en particulier aux travaux récents portant sur l'explication de modèles et l'interprétation de ceux-ci. Enfin, un des challenges sera la construction astucieuse de nouvelles « *features* ». La construction sans a priori de « *features* » informatives à partir de données arbitraires est un problème compliqué. Talend a développé des librairies de reconnaissances de types sémantiques pour identifier la nature des données et leur donner un sens. L'enjeu sera de tirer parti de ces fonctionnalités et éventuellement de les enrichir afin de faciliter la construction de nouvelles *features* intéressantes pour nos clients.

Proposition de déroulé de stage

- Etude biblio (papiers académiques, librairies de *feature engineering*, etc..), familiarisation avec l'écosystème Talend et de ses outils.
- Construction d'un protocole d'apprentissage de modèle. Identification de jeux de données type. Elaboration de la stratégie d'apprentissage des modèles, du processus de validation, de la construction des *features* conditionnellement aux types de données.
- Application sur les jeux de données types (ex : anciens challenges data…).

- Rédaction du mémoire de stage.

En fonction de l'avancée et de la pertinence des travaux, les résultats pourront donner lieu à des présentations ou une éventuelle publication scientifique.

Profil

Vous avez suivi une formation d'ingénieur spécialisé en statistiques et/ou apprentissage automatique ou équivalent (grande école d'ingénieur avec majeur en apprentissage statistique, master 2 statistiques et informatiques, etc). Vous êtes à l'aise avec les notions de régression avancée, avec les méthodes ensemblistes. Vous êtes à l'aise avec la lecture d'articles scientifiques.
Vous avez une réelle appétence pour la programmation et maitrisez les outils et les langages de la science des données (R et/ou Python), utilisation de notebooks ou d'IDE dédié, utilisation de git.
La réalisation de projets de data science ou la participation à des challenges (Kaggle, Challenge Data Ens, KDD) est un plus.
La notion de qualité est importante pour vous.
Votre rigueur et votre sens de l'organisation vous permettent de respecter les délais impartis. Vous êtes autonome et curieux mais savez également travailler en équipe.
Anglais courant : lu, écrit, parlé.

Talend

Les solutions d'intégration de Talend aident les entreprises à tirer le meilleur parti de leurs données. A travers le support natif des plates-formes modernes de Big Data, Talend réduit la complexité de l'intégration, tout en permettant aux départements informatiques de répondre plus rapidement aux besoins métiers, le tout pour un coût prévisible. Reposant sur des technologies open source, les solutions hautement évolutives de Talend répondent à tous les besoins d'intégration, actuels et émergents. Plus de 4000 entreprises du globe s'appuient sur les solutions et services de Talend. Basée à Suresnes (France) et à Redwood City (Californie), la société est implantée en Amérique du Nord, en Europe et en Asie, et s'appuie sur un réseau mondial de partenaires. Pour plus d'informations : www.talend.com sur le Web et @Talend sur Twitter.

Lieu

Le poste sera basé sur l'ile de Nantes, à Nantes (44) et rattaché au Lab, à la R&D.

Contact

Raphaël Nedellec rnedellec@talend.com, Sebastiao Correia scorreia@talend.com

Quelques liens

Librairies de feature engineering : TPOT http://epistasislab.github.io/tpot/, auto-sklearn https://automl.github.io/auto-sklearn/master/, etc
Talend types sémantiques :
https://help.talend.com/reader/t2uv4oMU8x6C8T4ptuLorA/kSH9NUxqve7mZNqHEDpSKw
LIME (Local Interpretable Model-Agnostic Explanations)
https://www.oreilly.com/learning/introduction-to-local-interpretable-model-agnostic-explanations-lime