Chaire Data Science and Artificial Intelligence for Digitalized Industry and Services

Offre de stage

Sujet

apprentissage non-supervisé (clustering,ranking,seriation), apprentissage actif, robustesse

Possibilité de poursuivre sur une thèse

OUI / NON ?

A propos de la chaire

La Chaire Data Science and Artificial Intelligence for Digitalized Industry and Services, portée par Florence d'Alché-Buc, enseignante-chercheuse dans le département Image, Données, Signal de Télécom Paris, la chaire DSAIDIS réunit cinq partenaires industriels : Airbus Defence & Space, Engie, Idemia, Safran et Valeo. Son objectif général est de développer, en liaison étroite avec les partenaires, une formation et une recherche de niveau international.

Ses quatre principaux axes de recherche sont :

1. Analyse, prévision de séries temporelles (Predictive Analytics on Time Series) ;

2. Exploitation de données hétérogènes, massives et partiellement étiquetées (Exploiting Large Scale and Heterogeneous, Partially Labelled Data) ;

3. Apprentissage pour une prise de décision robuste et fiable (Learning for Trusted and Robust Decision) ;

4. Apprentissage dans un environnement dynamique (Learning through Interactions with a Changing Environment).

Description du stage

Encadrement

Yann Issartel, et Stephan Clémençon

Lieu et dates du stage

Adresse : Télécom Paris, 19 Place Marguerite Perey, 91120 Palaiseau

Date de début du stage : en 2024

Équipe(s) d'accueil de la thèse

Département IDS, équipe Signal, Statistique et Apprentissage (S2A)

Mots clés

Apprentissage non-supervisé, Clustering, Ranking, Seriation, Apprentissage actif (on-line), Robustesse, Données contaminées

Sujet détaillé

Le but de ce projet est d'apporter des résultats théoriques pour des problèmes majeurs en

apprentissage supervisé, tels que, le clustering, le ranking ou encore la sériation. Un point central consistera à caractériser les vitesses optimales d'apprentissage (au sens minimax) pour des modèles généraux. Les résultats existants souffrent de plusieurs limitations, notamment :

(i) l'hypothèse d'observation en `batch' : le statisticien observe toutes les données avant la construction de son algorithme d'apprentissage. Cette hypothèse ne rend pas compte de la nature `online' de nombreuses applications, où les données sont collectées à la volée, et les décisions sont faites en parallèle, de façon séquentielle.

(ii) l'hypothèse de données propres, alors que les données réelles sont souvent contaminées pour diverses raisons.

(iii) l'hypothèse de données homogènes, qui peuvent être classées sous l'un des onglets `ranking/sériation/clustering'. En pratique, les données sont hétérogènes, comme une mixtures de ces données canoniques.

En fonction de ses intérêts, le/la candidat(e) travaillera sur des sujets en lien avec le projet.

Profil du candidat

Etudiant(e) titulaire d'un master 2 recherche

- Apprentissage statistique
- Bonnes bases en mathématique

Candidatures

A envoyer à yann.issartel@telecom-paris.fr:

- court CV
- brève lettre de motivation
- Relevés de notes des années précédentes / OU / liste de deux personnes qui peuvent être contactées pour des lettre de recommandation