Stage de recherche/développement (L3)

Représentation barycentrique pour l'apprentissage non supervisé

Contexte :

Dans de nombreuses applications, les observations ne sont pas naturellement représentées sous forme d'un nombre fixé de valeurs numériques, i.e., sous forme de vecteurs. Les données réelles peuvent en effet être de taille variable, être décrites par des variables qui ne sont pas directement comparables, ne pas être numériques, etc. On peut évoquer par exemple les données textuelles ou les données symboliques (intervalles, distributions, etc.). Or, beaucoup de méthodes d'analyse de données ont été construites pour des données représentées dans un espace vectoriel. Pour être appliquées à des données non vectorielles, les méthodes en question doivent être modifiées et adaptées. Une approche particulièrement fructueuse consiste à s'appuyer sur la définition de mesures de (dis)similarités entre données complexes. L'avantage évident de cette stratégie est de séparer la construction d'algorithmes d'analyse et le choix de la représentation des données. Cela permet de proposer une implémentation unique d'un algorithme d'analyse qui pourra être utilisée avec toute sorte de données, à condition de pouvoir calculer une (dis)similarité entre les observations. L'algorithme et son implémentation deviennent alors universels.

Cependant, ce type de méthodes s'adapte difficilement aux données massives, malgré une augmentation très rapide de la taille des jeux de données due aux nouvelles technologies. Le problème vient de l'impossibilité de placer l'ensemble de ces données en mémoire vive. Les données doivent être traitées une par une ou par « paquets ». Or, dans les méthodes à base de similarité, les données (ou leurs représentants) sont décrites par leurs distances à toutes les autres données, ce qui est impossible si on ne conserve pas en mémoire l'ensemble des données. Il est donc nécessaire de proposer un de nouveaux algorithmes d'analyse de données à base de similarité qui soient adaptés à ce type de problème.

L'objectif de ce stage sera de développer une représentation des données basées sur le système de coordonnées barycentrique. L'objectif sera de transformer les mesures de similarité entre les données et un ensemble de représentants en des coordonnées vectorielles dans espace de représentation formellement défini, ce qui permettra en particulier de construire des barycentres de données et de calculer des similarités entre ces barycentres et les données, ouvrant la voie à des algorithmes à base de prototypes pour données non vectorielles (K-means, SOM, etc.).

Objectifs du stage :

- Se familiariser avec la notion de coordonnées barycentriques et avec les méthodes de clustering à base de prototypes.
- Programmer une fonction qui calcule les coordonnées barycentriques des données à partir d'une matrice de similarité.
- Tester la méthode sur différents jeux de données.

Compétences souhaitées :

- Intérêt pour l'analyse de donnée et l'intelligence artificielle
- Développement en Python

Durée du stage : 2 mois
Lieu du stage : LIPN, UMR 7030, Sorbonne Paris Cité, Université Paris 13, Villetaneuse

Contact : Basarab Matei (matei@lipn.univ-paris13.fr)