Machine Learning pour données médicales hétérogènes de grande dimension

Stage de M2

L'enjeu en médecine personnalisée est d'utiliser les données du patient pour permettre d'adapter le traitement. Les analyses médicales récentes produisent un grand nombre de données différentes : données cliniques, données métagénomiques, données métabolomiques ou encore protéomiques ou sous formes de questionnaires. Le challenge est de traiter des données et de grande dimension et d'origine hétérogène, on parle de *data integration*.

Pour l'étude qui nous concerne, nous nous intéressons à l'obésité et principalement à une technique qui permet la réduction de poids : la chirurgie bariatrique. L'objectif d'une analyse approfondie de grandes bases de données sur des patients obèses est de mieux comprendre la maladie ainsi que les particularités de chaque personne afin de déterminer le meilleur type de chirurgie et évidemment si celle-ci a des chances de fonctionner.

En ce qui concerne les problématiques du stage, nous nous focaliserons sur le métabolisme de la flore bactérienne qui est essentielle à la santé des personnes et qui est fortement impliquée dans l'obésité. Les données se présentent sous forme de graphes et il s'agit de développer des méthodes qui s'appuient sur des modèle à blocs stochastiques [1]. Cette méthode permet de faire du clustering des nœuds d'un graphe, en regroupant les nœuds qui ont le même type d'interaction avec leur environnement. Pour les calculs on utilise en général un algorithme de type EM variationnel. L'objectif ultime serait d'exploiter ce modèle et la stratification des patients obtenue pour faire des prédictions pour des futures nouveaux patients.

Plus précisément, les étapes de l'étude sont les suivantes

- **Définition d'un modèle** de type à blocs stochastiques pour des données de plusieurs sources différentes, avec éventuellement une prise en compte des données manquantes dans le modèle.
- **Implémentation d'un algorithme** pour ajuster le modèle sur les données, en veillant particulièrement à une programmation efficace afin d'éviter de longs temps de calcul dûs à la taille des données.
- **Application sur des données médicales et interprétation**. Les résultats obtenus par l'algorithme apportent-ils une meilleure compréhension de la maladie ? Sont-ils exploitable pour faire des prédictions pour des futures patients ?

Références

[1] C. Matias & S. Robin, Modeling heterogeneity in random graphs through latent space models: a selective review. Esaim Proc. & Surveys, 47: 55-74, 2014.

Encadrement

Par Hédi Soula, Nataliya Sokolovska et Tabea Rebafka.

Profil recherché

- Etudiant de M2 ou de dernière année d'école d'ingénieur en machine learning, statistique, bioinformatique.
- Bonnes connaissances de programmation en Python ou R.
- Eventuellement intéressé par une poursuite en thèse.

Informations pratiques

- Durée de 4 à 6 mois.
- Localisation : Hôpital La Pitié – Salpétrière, 91 Bld de l'Hopital, 75013 Paris.
- Indemnité de stage d'environ 500 € par mois.
- Contact : tabea.rebafka@upmc.fr