Offre de stage de Master 2
Étude et implémentation d'un modèle de classification multi-partitions en grande dimension

Durée : 4 à 5 mois à compter de début avril
Profils du candidat : Master 2 ou équivalent disposant de compétences en classification supervisée, non supervisée, optimisation et programmation R.
Contact : Vincent Vandewalle (vincent.vandewalle@inria.fr), maître de conférence à l'Université Lille et membre de l'équipe Modal d'Inria Lille Nord-Europe.
Lieu du stage : Équipe Modal de l'Inria Lille Nord-Europe, 40 Avenue du Halley, 59650 Villeneuve-d'Ascq, France.
Rémunération : Environ 500 euros par mois.
Date limite de candidature : vendredi 2 février 2018.

1 Présentation du stage

En classification non supervisée on suppose souvent l'existence d'une seule variable de classe sensée expliquer l'hétérogénéité de l'ensemble des données. Cependant quand les variables proviennent de diverses sources cette hypothèse est souvent irréaliste.

Dans ce stage on propose d'étudier une extension du modèle de classification multi-partitions développé dans [1]. Ce modèle suppose l'existence de plusieurs variables de classe, chacune d'entre-elle expliquant l'hétérogénéité d'une partie des variables observées. Afin d'adapter l'approche précédente à des problèmes de très grande dimension, nous proposons de développer dans ce stage un algorithme de classification multi-partitions de type $k-means$ en reprenant des idées développées par [3] dans le cadre de la sélection de variables. Il s'agit essentiellement de substituer au critère probabiliste optimisé (typiquement une vraisemblance pénalisée) une critère géométrique avec une pénalité de type LASSO. L'avantage est alors de permettre au praticien de considérer le problème de la classification multi-partitions à travers une extension l'algorithme classique des $k-means$, et de pouvoir considérer de très gros volumes de données.

Le stage proposé consistera dans un premier temps à transposer le modèle développé par [1] au cadre du $k-means$ multi-parititions en s'appuyant sur les idées [3]. Dans un second temps, il consistera à proposer une stratégie d'optimisation du critère défini, notamment en s'appuyant sur des approches de type LASSO [2]. Dans une troisième temps, il consistera à implémenter et à tester sous R la stratégie proposée sur données réelles et simulées. Enfin, en fonction de son avancement le stage pourra donner lieu à la création d'un package R.

Références

[1] Matthieu Marbac and Vincent Vandewalle. A tractable multi-partitions clustering. *arXiv preprint arXiv :1801.07063*, 2018.

[2] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[3] Daniela M Witten and Robert Tibshirani. A framework for feature selection in clustering. *Journal of the American Statistical Association*, 105(490) :713–726, 2010.