Proposition de sujet de stage

Niveau : 3ème année ingénieur ou Master 2

Titre : « Classification supervisée par arbres de décision obtenus par des programmes linéaires en nombres entiers »

Sujet du stage :

L'objectif de la **classification supervisée** consiste à déterminer la classe d'appartenance de vecteurs de données x. Pour cela on se base sur un jeu de données d'entraînement $X = \{(x_1, y_1), \ldots, (x_p, y_p)\}$ dans lequel à chaque donnée x_i est associée une classe d'appartenance y_i. On souhaite déterminer une fonction f permettant de **prédire la classe $f(x)$ d'une donnée x** tout en minimisant les erreurs de prédiction (par exemple en minimisant : $\Sigma_i |f(x_i) - y_i|$). De nombreuses méthodes telles que les réseaux de neurones ou les SVM ont été introduites pour la résolution de ce type de problème. Dans ce stage nous nous intéressons aux **arbres de décision**. Le principe est de calculer des règles de séparation d'un père à ces deux fils ($a_n x \leq b_n$ et $a_n x \geq b_n$) afin construire un arbre dont le simple parcours permettra de classifier des données.

Dans [1,2] une approche de construction d'un arbre de classification est proposée par la résolution d'un Programme linéaires en variables mixtes. Sa solution optimale fournit un **arbre de classification optimal** pour des règles de branchements qui sont décrites par des fonction linéaire. En d'autres termes, la solution fournit les vecteurs de coefficients a_n et les seconds membres b_n des équations de branchements $a_n x \leq b_n$ et $a_n x \geq b_n$ de chaque nœud n.

Le travail à effectuer dans ce stage est à la fois théorique et expérimental. Le travail théorique portera sur l'**extension de la modélisation** proposée dans [1,2] à des règles de branchements décrites par des fonctions séparables au lieu de linéaires. La principale difficulté apparaissant en considérant cette modélisation plus fine est le passage à l'échelle. Le travail expérimental consistera à **implanter les méthodes de résolution** proposées et à comparer leurs résultats. On utilisera pour ce faire les logiciels standards de programmation mathématique.

Mots clés : Classification supervisée, arbres de décisions, Optimisation linéaire en variables mixtes,
Connaissances requises :
Cours de programmation mathématique.
Connaissance d'un langage quelconque de programmation

Encadrants :
Zacharie Ales, Maitre de conférences, CEDRIC-ENSTA, zacharie.ales@ensta-paris.fr
Amélie Lambert, Maître de Conférences, CEDRIC-Cnam, amelie.lambert@cnam.fr
Lieu : CEDRIC-Cnam (Paris) ou ENSTA (Palaiseau).
Durée : 6 mois
Poursuite en thèse : envisageable

Réferences :
[1] Bertsimas, D., Dunn, J. *Optimal classification trees*. Mach Learn 106**,** 1039–1082 (2017). https://doi.org/10.1007/s10994-017-5633-9
[2] Dunn, J. *Optimal trees for prediction and prescription*, Phd Thesis, (2018). http://hdl.handle.net/1721.1/119280
[3] Gambella, C., Ghaddar, B. Naoum-Sawaya, J. *Optimization problems for machine learning: A survey.* European Journal of Operational Research (2020) https://doi.org/10.1016/j.ejor.2020.08.045