Interprétabilité et forêts aléatoires

Sébastien Da Veiga

Pour des raisons de compréhension mais désormais aussi de certification, la thématique de l'interprétabilité des modèles d'apprentissage se pose de plus en plus à l'heure actuelle. En particulier parmi les modèles d'apprentissage ayant de très bonnes performances en pratique, les forêts aléatoires sont souvent considérées comme des boites noires non interprétables, car elles combinent des milliers d'opérations pour construire une prédiction. Mais si intrinsèquement elles ne sont pas interprétables, elles sont connues également pour leur compatibilité naturelle avec des approches dites post-hoc (post-traitement), et plus spécifiquement les mesures d'importances de variable.

Dans cet exposé, nous proposerons dans un premier temps la construction d'un modèle d'apprentissage interprétable à base de règles, dont la construction se base sur les forêts aléatoires et leur capacité à identifier des motifs fréquents et robustes dans les données. Ce modèle hérite de la précision des forêts aléatoires, construit des prédictions de manière simple mais aussi et surtout est stable par rapport à des perturbations sur les données.

De manière complémentaire, nous nous intéressons aussi à une des méthodes de mesure d'importance des forêts les plus utilisées, le MDA (Mean Decrease Accuracy). Pour la première fois nous démontrons sa convergence vers une quantité théorique, dont l'analyse fine montre que contrairement à son utilisation pratique fréquente, cette mesure d'importance est trompeuse dès que le modèle n'est pas additif ou que les variables explicatives ne sont pas indépendantes. Pour corriger cet effet délétère, nous proposons une modification de la mesure en nous basant sur des projections et non des permutations, ce qui permet de démontrer que la version corrigée tend bien vers une mesure d'importance.