CONFÉRENCE MAGISTRALE (MASTERCLASS)

APPRENTISSAGE STATISTIQUE PAR MÉTHODES ENSEMBLISTES À BASE D'ARBRES : INTRODUCTION, ALGORITHMES, INTERPRÉTABILITÉ ET CAS D'USAGE DE LA STATISTIQUE PUBLIQUE

Sébastien Da Veiga (*)

(*) Centre de recherche en économie et statistique (CREST) et Ecole nationale de la statistique et de l'analyse de l'information (ENSAI) - Enseignant-chercheur en statistiques

---

Pour lever les limitations des arbres de décision en apprentissage statistique, en particulier leur instabilité par rapport à des données aberrantes ou à une modification de l'échantillon d'entraînement, il est désormais usuel de considérer des algorithmes qui agrègent une grande collection d'arbres différents pour construire un modèle prédictif : on parle alors de méthode ensembliste à base d'arbres. Dans cette catégorie, on retrouve deux classes d'algorithmes particulièrement employés à l'heure actuelle sur des données tabulaires, les forêts aléatoires et le *gradient boosting*. Dans cet exposé, après avoir rappelé les fondements théoriques de l'apprentissage supervisé et le principe de construction d'un arbre de décision, nous justifierons la construction de modèles ensemblistes à la fois d'un point de vue théorique, mais aussi pratique, en se concentrant sur les forêts aléatoires (et variantes) et le *boosting* d'arbres (et le *gradient boosting*).

Même si ces dix dernières années ces algorithmes ont montré d'excellentes performances sur de très nombreuses applications pratiques, ils sont souvent caractérisés comme des "boîtes noires", au mécanisme interne obscur et complexe à la différence d'un arbre de décision ou d'une régression linéaire, car une prédiction est le résultat de centaines ou de milliers d'opérations sur les covariables. La quête d'interprétabilité pour éclairer ou expliquer leurs prédictions fait l'objet de récentes recherches académiques, que nous introduirons et synthétiserons lors de cet exposé.

Enfin, nous discuterons du potentiel fort de ces méthodes sur une variété de cas d'usage du Service statistique public, qui seront mis en lumière lors de la conférence par Mélina Hillion et Olivier Meslin, du SSP Lab de l'Insee.

Pré requis

Bases en probabilités et statistique, notions d'optimisation.

Durée : 3 heures.