Big data analytics: trees & aggregation methods (bagging, random Forests & boosting)

Nom : **TOKPAVI** Prénom : **Sessi** Année : **M2** Semestre : **9**

Nature : **CM** Volume horaire : **12 H** ECTS / Coef : **2**

Prérequis	- Des connaissances théoriques niveau Master 1 en Statistiques et Econométrie: notions de variables aléatoires, de spécification, d'estimation et de prévision dans le cadre de la régression (variable cible quantitative) et de classification (variable cible qualitative-logit/probit). - Une bonne maîtrise du logiciel SAS (SAS-base/stat).

Résumé	Ce cours a pour objet l'étude d'un algorithme ou modèle d'apprentissage supervisé connu sous le nom d'arbres de décision. Le principe général consiste à partitionner l'univers des individus de l'échantillon d'apprentissage en groupes d'individus homogènes du point de vue de la variable cible. La partition est hiérarchique en tenant compte de la capacité prédictive relative de chacun des prédicteurs. Les arbres de décision présentent de nombreux atouts : assez performants pour modéliser de manière non-paramétrique des relations non-linéaires ; adaptés à des bases de données volumineuses, dans la dimension individuelle mais également celle des prédicteurs ; autorisent des prédicteurs quantitatifs et qualitatifs ; gèrent de manière élégante les données manquantes ; fournissent une hiérarchie dans l'importance des prédicteurs. Ils ont cependant des pouvoirs prédictifs limités. Les méthodes d'agrégation, telles que les forêts aléatoires ou Random Forest (Breiman, 2001) et les méthodes de Boosting (Freund et Schapire, 1996) permettent de pallier à cette insuffisance, en combinant plusieurs arbres de décision. Les forêts aléatoires sont l'application aux arbres de décision du raffinement d'une méthode d'agrégation plus générale connue sous le nom de Bagging ou Bootstrap Aggregation (Breiman, 1996).

Objectifs	A l'issue de ce cours, les étudiants doivent maîtriser : - les arbres de décisions en tant que méthode d'apprentissage supervisé, aussi bien pour la régression que pour la classification. - les méthodes d'agrégation des arbres de décisions en tant qu'apprentis faibles (Bagging, Random Forest, Boosting) pour l'augmentation du pouvoir prédictif. - la mise en œuvre dans la pratique sous le logiciel SAS des méthodes ci-dessus mentionnées.

Bibliographie	- Breiman, L. (1996), "Bagging predictors," Machine Learning, 26, 123-140. - Breiman, L. (2001), "Random Forest," Machine Learning, 45, 5-32. - Freund, Y. et R. E. Schapire (1996), "Experiments with a new boosting algorithm," Proceedings of the Thirteenth International Conference in Machine Learning, 148–156. - Hastie, T., Tibshirani, R. et J. H. Friedman (2009), The Elements of Statistical Learning, Data Mining, Inference, and Prediction, Springer Series in Statistics, 2nd Edition. - James, G., Witten, D., Hastie, T. et R. Tibshirani (2016), An Introduction to Statistical Learning, with Applications in R, Springer Series in Statistics. 6th Edition.

PLAN

- Introduction
- Arbres de décision & l'algorithme CART
    - Implication de la division binaire
    - Choix de la variable de césure : cas de la classification
    - Choix de la variable de césure : cas de la régression
    - Hauteur optimale de l'arbre : la question de l'élagage
    - Les règles de prédiction
- Les méthodes d'agrégation
    - Arbres de décision et Bagging : cas de la régression
    - Arbres de décision et Bagging : cas de la classification
    - Les forêts aléatoires
    - Boosting: présentation générale
    - AdaBoost
    - Gradient Boosting
    - Gradient Boosting & descente de gradient
    - Généralisation du Gradient Boosting
- Applications sous SAS