Big data analytics: penalized regressions (Lasso, Adaptative Lasso, Elastic-Net)

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 la classification (variable cible qualitative-logit/probit). - Une bonne maîtrise du logiciel SAS (SAS-base/stat).

Résumé	Le cours aborde les méthodes de pénalisation pour régression et classification. Ces méthodes sont appropriées au cas où l'on dispose d'un nombre important de variables (parfois supérieur au nombre d'exemples ou d'individus), où la méthode des moindres carrés ordinaires (MCO), lorsque faisable, conduit à des estimateurs et des prédictions très instables. Dans un tel cas, une sélection des variables les plus pertinentes est nécessaire via les méthodes ou techniques dites de « régularisation » ou de « pénalisation ». L'objectif est d'arbitrer entre biais et variance avec des estimateurs et des prédictions qui sont biaisés mais plus stables que ceux issus de la méthode des MCO. Les méthodes abordées dans le cours sont successivement, le Lasso (Tibshirani, 1996), l'Elastic-net (Zou et Hastie, 2005), et l'Adaptive Lasso (Zou, 2006). Les deux dernières méthodes palliant à certains défauts de la méthode séminale du Lasso. Bien qu'elle ne procède pas à de la sélection de variables, la régression dite Ridge (Hoerl et Kennard, 1978) est également abordée, car elle correspond aussi à de la pénalisation et est très utile pour résoudre les problèmes de multi-colinéarité.

Objectifs	- A l'issue de ce cours, les étudiants doivent maîtriser les méthodes modernes de pénalisation pour la régression et la classification. Ces méthodes qui permettent d'éviter le sur-ajustement en présence d'un nombre important de variables incluent entre autres la régression ridge, Lasso, Elastic Net, Adaptive Lasso. - La mise en œuvre dans la pratique, sous le logiciel SAS, des méthodes ci-dessus mentionnées est également abordée.

Bibliographie	- 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. - Hoerl, A. E. et R. Kennard (1978), "Ridge regression: Biased estimation for nonorthogonal problems," Technometrics, 12, 55-57. - 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. - Tibshirani, R. (1996), "Regression shrinkage and selection via the lasso," Journal of the Royal Statistical Society, Series B, 58(1), 267-288. - Zou, H. et T. Hastie (2005), "Regularization and variable selection via the elastic net", Journal of the Royal Statistical Society, Series B, 67, 301-320. - Zou, H. (2006), "The Adaptive Lasso and Its Oracle Properties," Journal of the American Statistical Association, 101, 476, 1418-1429.

PLAN

- Introduction
    - La révolution Big Data & ses enjeux
    - Le cadre global des solutions analytiques
    - Apprentissage supervisé : principe et quelques résultats théoriques
    - Apprentissage supervisé : flexibilité et arbitrage biais-variance
- Au-delà des MCO
- La régression Ridge
    - Estimateur et propriétés
    - Validation croisée et choix du paramètre de régularisation
    - Multicollinéarité et régression Ridge
- La régression Lasso
    - Motivations et critère d'estimation
    - Algorithmes d'estimation
    - Quelques propriétés de l'estimateur Lasso
- Lasso : extensions
    - La régression Elastic-Net
    - La régression Adaptive Lasso
- Cas de la classification
- Séparateurs à vaste marge
    - Présentation
    - Propriétés théoriques
    - Estimation
- Applications SAS