FORMATION CONTINUE

FSMP Fondation Sciences Mathématiques de Paris

cnrs formation entreprises

Apprentissage statistique sur données complexes: grande dimension, masses de données, Big Data

Objectifs et contenu

L'augmentation continue des capacités de mesure font qu'il est de plus en plus nécessaire de disposer d'outils statistiques capables de résumer et d'extraire l'information contenue dans les données. Cependant, la nature même des données modernes (grande dimension, masse de données) n'autorise pas l'utilisation de la plupart des méthodes statistiques classiques (tests, régression, classification). En effet, ces méthodes ne sont pas adaptées à ces conditions spécifiques d'application et souffrent en particulier du fléau de la dimension. Cette formation présentera les outils statistiques récents permettant d'analyser les données modernes ainsi que leur mise en œuvre dans des situations réalistes à l'aide du logiciel R.

Les thèmes suivants seront abordés :
- régression en grande dimension : régularisation, méthodes sparses, cas où *n* est plus petit que *p*
- clustering de masse de données : méthodes online, modèle de mélange
- classification des données de grande dimension : méthodes de sous-espaces, sélection de variables, visualisation

La mise en oeuvre des méthodes se fera à l'aide de paquets (gratuits) pour le logiciel R. L'accent sera mis sur l'interprétation des résultats et les méthodes seront mises en œuvre sur des données réelles provenant de domaines d'application variés.

Intervenant(s)

Pr. Charles Bouveyron (MAP5, Université Paris-Descartes).
Page personnelle : http://w3.mi.parisdescartes.fr/~cbouveyr/
Page du laboratoire : http://www.math-info.univ-paris5.fr/map5/

Public visé

Bac + 4/5

Prérequis

Connaissances de bases en statistique descriptive (histogramme, boxplot, corrélation, ...) et en statistique inférentielle (modèle linéaire, maximum de vraisemblance, ...). La connaissance du logiciel R n'est pas requise, une initiation sera proposée.

Lieu

Institut Henri Poincaré (IHP) rue Pierre et Marie Curie, 75005 Paris

Durée et emploi du temps

2,5 jours.

	Jour 1	Jour 2	Jour 3
Matin	1. Tour de table et présentation des domaines d'application 2. La problématique des « Big Data » et ses méthodes 3. Rappels de statistique descriptive et multivariée 4. Introduction au logiciel R	1. Classification supervisée (LDA, FDA, régression logistique) 2. Travaux dirigés avec le logiciel R	1. Applications des méthodes enseignées aux problématiques des stagiaires 2. Conclusion
Après-midi	1. Classification non supervisée (k-means, CAH, modèle de mélange, grande dimension) 2. Visualisation des données de grande dimension (ACP, ADF) 3. Travaux dirigés avec le logiciel R	1. Régression en grande dimension (régression ridge, lasso) 2. Travaux dirigés avec le logiciel R	

Infrastructure requise

Les logiciels open source R (www.r-project.org) et Rstudio (www.rstudio.com) devront être installés avant la formation. Les apprenants devront disposer de leur ordinateur personnel.