Modélisation statistique en présence de données manquantes

Projet M2 MIGS – 2016-17

Il n'est pas rare en pratique que certaines variables ne soient pas renseignées pour tous les individus, en particulier dans les grandes bases de données. Les approches classiques de construction d'estimateurs telles que les moindres carrés ou le maximum de vraisemblance ne peuvent alors plus s'appliquer directement. Par ailleurs supprimer les individus manquants peut conduire à des estimations fortement biaisées lorsque le mécanisme de non réponse n'est pas indépendant des caractéristiques des individus.

Ce projet vise à comprendre et explorer quelques pistes classiques qui ont été proposées par les statisticiens pour aborder ces questions. Il pourra porter sur l'étude des approches basées sur l'algorithme EM qui a été développé pour traiter les questions d'optimisation de la vraisemblance en présence de données observées partiellement. Cette approche s'applique pour la régression et plus généralement pour les modèles basés sur la famille exponentielle (la régression logistique pour la classification supervisée ou encore le modèle log-linéaire pour les modèles de comptage avec des lois de Poisson).

Des comparaisons sur données réelles et/ou simulées (la non réponse sera alors contrôlée et les vraies valeurs des paramètres connues) seront mises en œuvre.

Projet encadré par H. Cardot.

Bibliographie

Roderick J. A. Little (1992). Regression With Missing X's : A Review. *Journal of the American Statistical Association*, Vol. 87, pp. 1227- 1237.

McLachlan, G-J. and Krishnan, T. (2008). *The EM algorithm and Extensions*. Wiley Series in Probability and Statistics, 2nd edition.