Scoring sur données mixtes : Guide pour l'utilisation de MixtComp

MixtComp est un logiciel permettant de classifier des données mixtes (quantitatives, qualitatives, comptages, etc.) ; il est développé par l'INRIA et fait partie d'un ensemble appelé MASSICCC (Massive Clustering with Cloud Computing) comportant trois applications (MixtComp, MixMod, BlockCluster) dédiées à la classification. MixtComp s'utilise en ligne à l'adresse : `https://massiccc.lille.inria.fr/`.

Un exemple d'utilisation

Les données. On considère les données `loan` accessibles à l'adresse `http://alexandrelourme.free.fr/scoring_data_loan`. Il s'agit de soixante clients bancaires décrits par quatre variables quantitatives : le salaire, le solde, l'âge, le nombre d'enfants et deux variables qualitatives : le type de véhicule utilisé et le lieu de résidence.

L'objectif. On souhaite estimer la classe d'emprunteur (1/2) à laquelle appartiennent les dix derniers clients en apprenant une règle de classement sur les cinquante premiers clients.

Les étapes à suivre.

(*i*) Créer deux fichiers au format csv. Le premier fichier `loantrain.csv` contient les cinquante premiers clients de `loan` décrits par sept variables : les six descripteurs et la classe. Le second fichier `loan.test` est composé des dix derniers clients de `loan` décrits par les six descripteurs.

(*ii*) Télécharger le fichier `loantrain.csv` sous MASSICCC en spécifiant `classe` pour Index Column (la classe à prédire) et en vérifiant Data Type pour chaque variable.

(*iii*) Télécharger le fichier `loantest.csv` sous MASSICCC en ne spécifiant rien pour Index Column et en vérifiant Data Type pour chaque variable.

(*iv*) Créer un travail qui s'appellera `loanlearn` en spécifiant `classe` pour Labels Column et 2 pour le nombre de groupes.

(*v*) Créer un travail qui s'appellera `loanpredict` en spécifiant `loantest.csv` pour Data File, `Predict` pour option et `loanlearn` pour classifieur.

(*vi*) Télécharger le résultat de `loanpredict` sous forme d'un fichier `num1.rdata` dans un dossier `myfolder`.

(*vii*) Importer les résultats sous R : `load('myfoler/num1.rdata')`.

(*viii*) Les scores (probabilités conditionnelles) et les classes estimées s'obtiennent ainsi : `output$variable$data$z_class`.

(*ix*) Pour connaitre la valeur des statistiques estimées : (a) importer le résultat de `loanlearn` sous forme d'un fichier `num0.rdata` dans le dossier `myfolder` (b) importer les résultats sous R par : `load('myfoler/num0.rdata')` (c) la valeur de BIC (par exemple) s'obtient par : `output$mixture$BIC`.

A. Lourme, Faculté d'économie, gestion & AES, Université de Bordeaux `http://alexandrelourme.free.fr`