Méthodes statistiques pour le Big Data

Composante
Institut universitaire de technologie de Poitiers-Châtellerault-Niort

Présentation

Description

L'objectif de cette ressource est de comprendre la difficulté du passage aux grandes dimensions avec, notamment, la problématique des 6 V (Volume, Variété, Vélocité, Véracité, Valeurs et Visualisation). En particulier, deux difficultés de la (très) grande dimension sont que les données peuvent être réparties sur plusieurs serveurs et comporter énormément de données manquantes.

Contenus :

- Définition les problématiques du Big Data par les 6V
- Limite des outils classiques en statistique qui ne fonctionnent plus (tests toujours significatifs, problématique des tests multiples, intervalles de confiance tout petits…).
- Introduction à la difficulté d'avoir plus de variables que d'individus (par exemple en abordant la pénalité LASSO)
- Sensibilisation au fait que les données sont sur plusieurs serveurs (reprendre les statistiques descriptives de bases et voir le passage provenant de différents échantillons). Réflexion en même temps sur la parallélisation.
- Complétion des valeurs manquantes : différence entre une valeur manquante par censure (qui ne sera pas traitée) et aléatoire puis réflexion sur la complétion (comprendre que remplacer par la moyenne biaise les éventuelles corrélations ; on peut, pour cela, s'appuyer sur le package missMDA même s'il ne passe pas bien à la grande dimension).

L'utilisation d'un logiciel d'analyse statistique pour la mise en œuvre des méthodes est indispensable.

La diversité et les volumes de données nécessitent d'adapter les traitements et les méthodes d'analyse pour en tirer des informations exploitables et prendre des décisions. Les méthodes classiques ne sont plus adaptés à ces problématiques et les étudiants doivent aborder de nouveaux outils.

Heures d'enseignement

CM	CM	3h
TD	TD	18h