Fouille de données

Composante
École Nationale Supérieure des Ingénieurs en Arts Chimiques

Volume horaire
25,33h

En bref

- **Code:** LP1A203W
- **Ouvert aux étudiants en échange:** Oui

Présentation

Objectifs

Être capable de prendre en compte, d'explorer et d'analyser un ensemble de données multidimensionnelles.

Maîtriser les statistiques descriptives multidimensionnelles.

Utiliser et valider des outils de statistiques prédictifs (régressions multilinéaires, arbre de régression,..) sous R et/ou Matlab

Description

Ce module est une sensibilisation aux concepts et applications de la fouille de données (ou Data Mining en anglais), qui constitue un des piliers du "Big Data".

Il reprend les principes de base des calculs et analyses statistiques de l'ingénieur via le logiciel R, et propose une ouverture aux méthodes d'analyse de grands jeux de données (analyse en composantes principales, classification hiérarchique, etc...).

Ce module comporte deux parties "assez distinctes" :

- La première partie porte sur le calcul statistique, avec une approche numérique, mis en œuvre au moyen du langage R. Cette partie du cours aboutit à l'écriture en binôme d'un code de calcul statistique programmé en R.
- La seconde porte sur l'analyse de type "data mining" (fouille de données en français), en groupe de 3-4 élèves, d'un jeu de données particulier. Ce dernier travail, réalisé très largement en autonomie, nécessite de formuler les objectifs de l'analyse des données

en question, d'identifier une ou plusieurs techniques d'analyse statistique appropriées, d'étudier le principe de ces techniques et de les mettre en œuvre pour répondre aux objectifs formulés.

Pré-requis obligatoires

Connaissances élémentaires en probabilités et statistique : variables aléatoires, indépendances, distribution, tests,...

Contrôle des connaissances

L'évaluation consiste en :

- une remise d'un code R sur la partie calcul statistique (**comptant pour 40% de la note de l'enseignement**)

- une remise d'un rapport et une présentation orale pour chaque groupe , devant l'ensemble de la classe, d'autre part (**comptant pour 60% de la note de l'enseignement**). Compte-tenu du fait que les groupes ne mettent pas nécessairement en œuvre les mêmes méthodes d'analyse statistique, ce format d'évaluation permet à chacun de se sensibiliser à l'application d'un nombre significatif de méthodes d'analyse de données.

Syllabus

Introduction aux Statistiques Appliquées

Introduction générale, incertitude et prise de décision

Signification et visualisation de l'incertitude

Rappels sur les variables aléatoires et les distributions de probabilités

Calcul statistique et programmation en langage R

Estimateurs, barres d'erreur et tests statistiques

Estimateur du maximum de vraisemblance (notions)

Propagation de la variance et liaison variance/incertitude

Calcul d'incertitude par simulation Monte-Carlo (Génération de variables aléatoires)

Corrélation et calcul d'incertitude (Génération de variables aléatoires corrélées) - Propagation de distributions et analyse de sensibilité

Mise en pratique par l'utilisation du langage R pour le calcul statistique

Quelques méthodes d'analyse de données

Analyse de variance (ANOVA)

Analyse en composantes principales (ACP)

Arbres de décision

Classification Ascendante Hiérarchique (CAH)

Méthode des k-moyennes (k-means)