Journées de Statistique Mathématique et Data Science (JSMDS 2021)
Hammamet, Tunisie, 25-28 novembre 2021

Introduction à la data science

François Husson[1],

Résumé :

Nous présenterons dans un premier temps différents domaines et problématiques qui mettent en jeu des données de plus en plus volumineuses, hétérogènes, complexes, etc. La science des données est une discipline qui s'attache à visualiser, analyser et donner de la valeur à ces données. Les différents exposé présenteront des méthodes permettant de visualiser les données et de prédire une variable quantitative ou qualitative en fonction de variables quantitatives ou qualitatives.

La méthode d'analyse en composante principale permet de décrire un tableau croisant des individus statistiques en ligne et des variables quantitative en colonnes, et permet de résumer et synthétiser l'information contenue dans un tel tableau avec es graphiques simples. La classification permet de construire un arbre hiérarchique permettant de visualiser les distance entre individus ou groupes d'individus, et de constituer des classes d'individus.

Les arbres de régression et de classification sont des méthodes dont le principe est simples à comprendre et dont les résultats sont faciles à interpréter pour prédire une variable Y en fonction de variables explicatives. Plusieurs arbres peuvent être construits et la synthèse des résultats de plusieurs arbres améliore la qualité de prédiction : on parle alors de forât aléatoire quand on agrège les résultats de plusieurs arbres.

Mots clés : Analyse en composantes principales; classification; arbre de classification et de régression; forêts aléatoires

Références

1. Hastie T., Tibshirani R, Friedman J. The elements of statistical learning. Springer. `https://web.stanford.edu/~hastie/ElemStatLearn/`

[1]l'institut Agro

2. Husson F., Lê S., Pagès J. Analyse de données avec R. Presses Universitaires de Rennes, 2008.