Discipline(s) : Infomatique et télécommunications

MAKING DATA SPEAK: ADVANCED PROBABILISTIC DATA ANALYSIS AND MODELING

Nature	UE

CONTACTS

Guillaume Gravier (enseignant)

OBJECTIFS

Les données, quelle que soit leur nature, n'ont que peu de valeur s'il n'est pas possible d'en extraire des informations pour mieux synthétiser, comprendre et prédire. Les outils statistiques d'analyse de données et les modèles probabilistes sont très largement utilisés pour ce faire. L'objectif de ce cours est d'acquérir les techniques de bases de l'analyse de données et de la modélisation probabiliste et d'étudier leurs applications possibles à différents types de données. Le cours s'articule naturellement autour de deux grandes parties qui correspondent aux deux étapes d'une démarche de modélisation : comprendre les données puis les modéliser. Après un bref rappel des notions importantes des probabilités, on s'intéressera tout d'abord aux techniques de la statistique exploratoire, permettant ainsi à l'analyste de ce faire une idée de ses données, d'en comprendre la structure et les aspects importants. En particulier, les techniques d'analyse factorielle, de clustering, et les tests d'hypothèses seront au coeur de cette première partie. L'analyse des données débouche naturellement sur la modélisation.
La seconde partie du cours pose tout d'abord les bases théoriques de la statistique inférentielle (théorie de la décision, de l'estimation) avant de détailler les principales familles de modèles utilisées en pratique pour la prédiction et la classification. Nous y étudions en particulier les modèles de mélange et l'algorithme EM, les modèles de Markov cachés (HMM) à temps discret, les réseaux bayésiens et les modèles à maximum d'entropie.
Le cours est avant tout théorique et vise à dresser un panorama des différentes approches, l'objectif étant de donner aux étudiants la capacité à mener une démarche complète d'analyse de données, à choisir les techniques appropriées à un problème et à les mettre en oeuvre. Le cours ne se limite pas à un type de données particulier et sera illustré par des exemples issus de domaines variés comme l'économie, le traitement des contenus multimédias, la bioinformatique, le diagnostic, etc.

MOTS-CLÉS

Analyse de données, analyse factorielle, ANOVA, clustering, test d'hypothèse, théorie de la décision et de l'estimation, processus markoviens, algorithme EM, réseaux bayésiens, champs aléatoires

PRÉREQUIS

Bases de probabilités

CONTENU

Partie 1 - Statistique exploratoire

Rappel de probabilité et statistique : variable, densité, lois usuelles, corrélation, CCA, etc.
Analyse factorielle : analyse en composante principale, analyse discriminante, analyse factorielle
Analyse de variance Clustering : k-moyenne, clustering hiérarchique, méthodes spectrales, méthodes par densité
Tests d'hypothèse : anatomie d'un test, test classiques, rapport de vraisemblance

Partie 2 - Statistique inférentielle

Introduction rapide à la théorie de la décision et à la théorie de l'estimation
Modèle de mélange et algorithme EM
Modèle de Markov, modèle de Markov cachés, algorithme de Viterbi
Réseaux bayésiens, inférence, propagation des croyances
Modèles par maximum d'entropie, champs aléatoires conditionnels

COMPÉTENCES ACQUISES

Savoir : Méthodologie de l'analyse des données ; Outils de description des données ; Inférence statistique ; Apprentissage probabiliste ; Modèles de Markov ; Réseaux bayésiens ;
Savoir-faire : Décrire statistiquement des ensembles de données ; Modéliser des problèmes et estimer les paramètres inconnus ; Évaluer des résultats

APPARTIENT À

Master 2 informatique parcours Science Informatique

Mise à jour le 17 juillet 2017

CONTACT(S)

Département Informatique et télécommunications
École normale supérieure de RennesCampus de Ker LannAvenue Robert Schuman
35170 BRUZ
Tél. : 02 99 05 52 43
E-mail
Site Internet