IF.1105 / IF.1205 – Sciences et traitement des données

INFORMATIONS GÉNÉRALES

Titre du module : Science et traitement des données
Identifiant du module : IF.1105 / IF.1205
Responsable du module : Patricia CONDE-CESPEDES / Hélène URIEN
ECTS : 3
Quantité de travail moyenne par élève : 75h dont 48h encadrées
Travail en équipe : un projet en data science à faire en groupe de 2 ou 3 personnes.
Mots clés : probabilités, statistiques et science de données

PRÉSENTATION

De nos jours nous pouvons facilement avoir accès à une énorme quantité de données. La science des données est un domaine d'étude de l'intelligence artificielle qui combine des outils issus de l'informatique, des probabilités et des statistiques pour extraire des informations significatives à partir de données brutes. En effet, les Probabilités et les Statistiques constituent une clé de voûte pour construire des modèles en Data Science. La théorie des probabilités est une branche des mathématiques qui étudie le degré d'incertitude dans un processus aléatoire décrit par des variables aléatoires. Alors que la statistique consiste à utiliser l'échantillonnage de données, principalement dans deux buts principaux : décrire certains phénomènes (statistiques descriptives) et inférer des propriétés sur la distribution de probabilité des variables aléatoires décrivant la population de l'échantillon (inférence statistique). La plupart des méthodes statistiques dépendent de la théorie des probabilités. En science des données, les probabilités et les statistiques sont principalement utilisées pour l'estimation et la prédiction d'un phénomène aléatoire.

OBJECTIFS PÉDAGOGIQUES

Les connaissances et compétences développées dans ce module relèvent du domaine des probabilités et statistiques, dans des contextes d'utilisation relevant de l'analyse de données, du traitement du signal, et de l'apprentissage de la méthode scientifique. L'essentiel des exemples d'application seront contextualisés.

Prérequis

- Notions de probabilités, notions d'algèbre linéaire

Contenu/programme

- Probabilités
  - Notion d'évènement et de probabilité
  - Probabilités conditionnelles & indépendance
  - Variable aléatoire réelle
  - Valeurs typiques d'une variable aléatoire réelle
  - Fonction caractéristique d'une variable aléatoire réelle
  - Transformation d'une variable aléatoire réelle
  - Variables aléatoires réelles bidimensionnelles
  - Esperance, fonction caractéristique et moments pour 2 variables aléatoires
  - Notion de convergence, loi des grands nombres et théorème de la limite centrale
- Statistiques

    - La statistique descriptive
    - Théorie statistique de l'estimation
    - Test d'hypothèses
- Science de données
    - Introduction à la science de données
    - Rappels d'algèbre linéaire pour la science de données
    - Régression linéaire simple et multiple
    - Analyse en composantes principales et applications

Outils utilisés

- R (pour la partie Statistiques)
- Python (pour la partie science des données)

MODALITÉS PÉDAGOGIQUES

Méthodes d'apprentissage

Ce module repose sur une modalité d'approche par problèmes, par le recours systématique à des problèmes contextualisés. Chaque composante du cours théorique est suivie/accompagnée des travaux dirigés et des travaux pratiques sur machine avec le logiciel R et Python (pour science de données).

Déroulement du module (Heures de face-à-face pédagogique) :

- Cours (12 séances de 1h30 et 1 séance sur machine de 3h)
- TD (12 séances de 2h)
- TD sur machine (1 séances de 3 heures)

Modalités d'évaluation

- 1 examen de probabilités vers le milieu du semestre.
- 1 examen de statistiques vers la fin du semestre.
- 1 projet en science de données à faire en binôme ou à 3.
- La participation en classe est prise en compte pour des points supplémentaires.

Langue de travail

- La langue de travail principale est le français, mais certaines ressources bibliographiques peuvent être en anglais.
- Les supports de cours de science de données sont en anglais.

BIBLIOGRAPHIE, WEBOGRAPHIE, AUTRES SOURCES

- Gilbert Saporta (2011) Probabilités, analyse des données et statistique. 3ème édition.
- MIT-OPEN-Courseware: « Probabilities and applied statistics »
- Poly pour les deux parties du cours avec quelques références utiles.