Data Mining

Infos pratiques

> **ECTS** : 3.0

> **Nombre d'heures** : 24.0

> **Niveau d'étude** : BAC +5

> **Période de l'année** : Enseignement neuvième semestre

> **Méthodes d'enseignement** : En présence

> **Forme d'enseignement** : Cours magistral et Travaux dirigés

> **Ouvert aux étudiants en échange** : Oui

> **Composante** : Sciences économiques, gestion, mathématiques et informatique

Objectifs

L'objectif de ce cours/TP est d'introduire les principaux éléments de Data Mining et d'analyse de données, ainsi que des concepts et outils fondamentaux de la modélisation. L'accent est mis sur la pratique plus que sur la théorie. L'objectif est que les étudiants soient capables d'effectuer une analyse descriptive des données, d'en extraire des questions d'intérêts, les formaliser à l'aide de modèles/outils statistiques classiques adaptées (comme la régression) et les interpréter. Le logiciel **R** et l'environnement **RStudio** seront utilisés.

Approche pédagogique et plan de cours.

Séance # 1 : Qu'est-ce que la modélisation de données ? Les différents types de données.

* Identification de questions concrètes à travers quelques exemples de la vie réelle (en informatique, finance,....,etc)
* Introduction au logiciel R et à l'environnement RStudio.
* Rapports automatisés : générer des documents word, pdf et html de manière automatique avec R Markdown.
* Lecture des données de différents formats et graphiques simples.

Séance # 2 : Manipulation, nettoyage, visualisation et analyse de données

* Construction de sous échantillons, manipulation de tableaux, résumés, stat
* Manipulation des données et analyse descriptive simple avec les packages dplyr et tidyverse du logiciel R.
* Visualisation des données: graphiques univariés et bivariés avec le package ggplot2 de R.
* Comparaison de groupes visuellement avec des graphiques pertinents

Séance # 3 : Analyse multivariée et clustering

* ACP
* Clustering (CAH, Kmeans)

Séance # 4 : Text mining.

* Extraction de texte
* Comptage de mot (TF-IDF).
* Nuage de mots

Séance # 5 : Data Camp :
mise en application des outils et méthodes présentés en cours sur un jeu de données réelles. Ce travail sera réalisé en trinôme et un compte-rendu de 5 pages maximum sera demandé (avec R Markdown).

Séance # 6 : Test d'hypothèses.

* Présentation des tests d'hypothèses.
* Qu'est-ce que la p-valeur ? (Illustration via des simulations sur R)
* Techniques de ré-échantillonnage

Séance # 7 : Régression multiple

* Présentation du modèle et prédiction.
* Etude de la qualité prédictive du modèle – estimation de l'erreur de prédiction à l'aide de techniques de ré-échantillonnage (package Caret de R)
* Limites du modèle

Séance # 8 : Choix de modèles

* Compromis Biais-Variance et critères pénalisés.
* **Régularisation : Ridge, Lasso**

Évaluation

Session 1 : Évaluation continue (cf. règle par défaut de la section « Modalités spécifiques » des M3C spécifiques)

Session 2 : Règle par défaut décrite dans la section « Modalités de contrôle et examens / Modalités spécifiques »

Pré-requis nécessaires

Ce cours/TP peut être suivi par des étudiants n'ayant qu'une connaissance basique des statistiques (au moins les concepts de statistique descriptive : population, échantillon, proportions, moyennes et variances et représentations graphiques de type diagramme en bâtons, histogrammes et quelques distributions connues en statistique) grâce à des synthèses présentées en début de séances.

Compétences visées

* Acquérir le vocabulaire et les concepts fondamentaux de Data Mining et d'analyse des données. Etre capable d'identifier des questions concrètes.
* Conduire une analyse statistique pour répondre à la question posée : modéliser, analyser et interpréter les résultats fournis par le logiciel R.
* Réaliser des rapports automatisés avec R markdown
* Apprendre le logiciel R et ses récentes extensions (dplyr, tidyverse, ggplot2)
* Être capable de mener un mini projet de Data Mining, analyse de données et choisir/appliquer le type d'algorithme de résolution adapté.

Bibliographie

* Benjamin S. Baumer et al. Modern Data Science with R
* François Husson et al. R pour la statistique et la science des données
* Pierre-André Cornillon, Eric Matzner-Lober. Régression (Théorie et applications).
* Philippe Besse. Statistique et Big Data Mining
* https://www.math.univ-toulouse.fr/~besse/enseignement.html