ANALYSE DE DONNÉES

ECTS
4 crédits

Composante
UFR de mathématiques et informatique (UFR27)

Volume horaire
42h

Période de l'année
Printemps

plugin.odf:CONTENT_PROGRAM_TAB01_TITLE

Description

Objectifs:

- L'objectif du cours est d'introduire les fondements et principales questions et approches qui interviennent de façon incontournable dans tous les cours ultérieurs du M2 MMMEF, M2MO ou du M2 TIDE où l'Apprentissage Statistique (machine learning), la Data Science et l'intelligence artificielle apparaissent en bonne place.

- Les séances alterneront entre cours magistraux et séances de travaux dirigés selon l'avancement et en fonction des besoins.

-Les supports de cours (slides) seront en anglais. Cette langue est en effet, qu'on le veuille ou non, un outil indispensable pour acquérir et transmettre des savoirs/informations dans le monde professionnel. Sa maîtrise (au moins pour le vocabulaire technique) est donc devenue indispensable à quiconque envisage une carrière dans le domaine de la Data Science ou de l'Intelligence Artificielle.

Prérequis**:**

-avoir suivi des cours en théorie des probabilités

-avoir suivi des cours en algèbre linéaire

Contenu du cours:

1. Linear regression:

rappels d'algèbre linéaire, décomposition SVD, description du modèle, estimation des paramètres par minimisation du risque empirique/maximisation de la vraisemblance, l'hypothèse gaussienne.

2. Model selection:
   Sur la base du modèle de régression linéaire, on envisage plusieurs modèles candidats.
   On introduit l'idée de quantification de la performance et justifie l'utilisation de critères pénalisés de type AIC. La procédure de validation-croisée sera également détaillée.
3. Classification:
   Différents types de modèles seront envisagés tels que la régression logistique, l'analyse linéaire discriminante, le classifieur des k plus proches voisins.
4. Clustering:
   Nous envisagerons différentes stratégies pour définir des mesures de similarité et étudierons deux principales approches pour le clustering : les K -means et le clustering hiérarchique ascendant.
5. Data Visualization:
   Afin de pouvoir visualiser les résultats de l'analyse, nous envisagerons deux principales techniques de réduction de dimension telles que l'analyse en composantes principales et l'analyse canonique des corrélations. La question de l'estimation de densité par estimateurs histogramme ou à noyau sera détaillée.

Références:

-Hastie, Tibshirani, Friedman. The elements of statistical learning: Data Mining, Inference and Prediction. Springer series in Statistics.