

Parcours Données et décision



Crédits ECTS
4 crédits



Période de
l'année
Automne

En bref

> **Langue de cours:** Français

Présentation

Prérequis

- Connaissances en optimisation, probabilités et statistique (équivalent des contenus de MAT-1A et de l'Approfondissement MIE M1)
- Connaissances en algorithmique et en programmation Python (équivalent du contenu de INFO-1A)

Objectifs d'apprentissage

- Connaître les principes de base de gestion des données dans les projets de data science.
- Savoir acquérir, agréger et manipuler des données.
- Savoir modéliser les problèmes de régression et de classification standards et savoir utiliser un langage informatique approprié pour implémenter les solutions des problèmes.
- Savoir utiliser les données pour prendre des décisions.
- Comprendre l'importance de la gouvernance et la qualité des données

Description du programme

This course unit consists of three courses Statistical learning, Python for data science and Data-driven decision making, of 24 hours each, and is complemented by the second part of the data project (9 hours course and 12 hours project) devoted to data issues.

Statistical learning

1. Introduction
 - i. Classical problems: regression, classification

- ii. Supervised, unsupervised and semi-supervised learning
- iii. Curse of dimensionality
- 2. Regression
 - i. Multiple linear regression, OLS method
 - ii. Shrinkage-type methods (LASSO, Ridge)
 - iii. k-nearest neighbors
- 3. Classification
 - i. Logistic regression
 - ii. k-nearest neighbors
 - iii. SVM
 - iv. Rosenblatt perceptron and neuronal networks

Python for data science

- 1. Dataframe: data exploration and data description
 - i. Spotting patterns using factor
 - ii. Principal Component Analysis
 - iii. Correspondence analysis
- 2. Prediction using trend analysis
 - i. Linear regression
 - ii. Logistic regression
- 3. Data classification
 - i. Classification using partitions
 - ii. Hierarchical methods

Data-driven decision making

- 1. What is data?
- 2. How do we take decision?
- 3. Data governance and data quality
- 4. How to develop data-based decision making?
- 5. Data platform and data architecture

Data science projects: data issues

- 1. Starting a data science project
- 2. The constraints of data science projects
- 3. Finding data
- 4. Acquiring information
- 5. Playing with data

Compétences et connaissances scientifiques et techniques visées dans la discipline

- Savoir manipuler des données pour commencer un projet en data science
- Savoir appliquer les méthodes classiques pour la classification supervisée et non-supervisée et savoir comparer plusieurs modèles.

- Savoir appliquer les méthodes standards de régression (Moindres Carrés Ordinaires) et des méthodes avancées pour sélectionner les variables et tenir compte de la malédiction de la dimension (Ridge, LASSO, Elastic Net)
- Savoir appliquer des méthodes de réduction de dimensionnalité et de description des données telles que l'ACP et l'Analyse des correspondances.
- Etre capable de construire des indicateurs de performance pour un modèle appliqué à des données.
- Comprendre et mesurer la valeur des données.

Modalité de contrôle des connaissances

- Tests et projets (Statistical learning): 30%
- Projet (Python for data science): 35%
- Projet de groupe et présentation (Data-driven decision making): 35%

Bibliographie

Statistical Learning

- James G., Witten D., Hastie T. and al. (2013). An introduction to statistical learning: with applications in R. New York: Springer
- Hastie T., Tibshirani R. and Friedman J. (2013). The elements of statistical learning: data mining, inference, and prediction. New York: Springer.
- Cornillon P-A., Matzner-Løber E. et al. (2010). Régression avec R. Paris: Springer.

Python for data science

- Jannach, D., Zanker, M., Felfernig, A. and Friedrich, G. (2010). Recommender Systems: An Introduction. Cambridge.

Data science projects

- Zeng, A and Casari, A. Feature Engineering for Machine Learning. O'Reilly Media.
- Müller, A. and Guido, S. Introduction to Machine Learning with Python. O'Reilly Media.

Equipe pédagogique

- Statistical learning: Christophe Pouet (Centrale Marseille)
- Python for data science: Emmanuel Daucé (Centrale Marseille)
- Data-driven decision making: Mickaël Chalamel (Yves Saint-Laurent), Franck Chevalier (EY)
- Data project: Maximilien Défourné (Mantiks)

Objectif de Développement Durable



Partenariats pour la
réalisation des objectifs

Total des heures

CM	Cours Magistral	100h 81h
PJ		19h

Infos pratiques

Nom responsable UE

Responsable pédagogique

Renaud Bourles

✉ renaud.bourles@centrale-med.fr