MASTER FOR SMART DATA SCIENCE

COURSE DESCRIPTIONS

Note: In order to ensure that the curriculum is adapted to the needs of the current job market and its students, ENSAI reserves the right to modify the proposed curriculum and the following descriptions at any time during the academic year.

BEFORE SEMESTER 1

Before the main courses start, some preliminary modules are organized. The list of these courses is subject to change from one year to another. There are no ECTS credits granted for these preliminary modules but they are mandatory. The preliminary courses allow students to complete the prerequisites for the MSc. Tentative list of preliminary courses for 2023/2024:

- GNU Linux & Shell Scripting (12h)
- Multivariate Data Exploration (12h)
- SQL (6 hrs)
- Statistical languages: R, Python (18h)
- Topics in probability: Markov Chains (12h)

SEMESTER 1

Machine Learning for Data Science

(Lectures and Tutorials: 30hrs)

This course focuses on supervised learning methods for regression and classification. Starting from elementary algorithms such as ordinary least squares, we will cover regularization methods (crucial in large scale learning), nonparametric decision rules such as *support vector machine*, the *nearest neighbor* algorithm and *CART*. Finally, bagging and boosting techniques will be discussed while presenting random forest, and the XGboost algorithm.
The focus will be on methodological and algorithmic aspects, while trying to give an idea of the underlying theoretical foundations. Practical sessions will give students the opportunity to apply the methods on real data sets using either R or Python. The course will alternate between lectures and practical lab sessions.

Deep Learning

(Lectures and Tutorials: 15hrs)

This course is devoted to neural network (NN) architectures and their extension known as deep learning. Beforehand, the stochastic gradient descent algorithm and the back-propagation - its application to feedforward neural networks - are introduced to be further used as the learning basis. This is followed by the study of most spread NN architectures for regression and classification. Among those, convolutional neural networks (CNN) are investigated in detail and other structures like Recurrent Neural Networks, Restricted Boltzmann machines (RBM), and the contrastive divergence algorithm (CD-k) are examined. Furthermore, practical aspects will be addressed about the usage of Deep Learning to resolve typical problems like pattern recognition or object/detection tracking. Presented material shall be motivated by the theoretical background together with real data illustrations. There will be specific labs for each topic held in R & Python.

Dimension Reduction & Matrix Completion

(Lectures and Tutorials: 18hrs)

In modern datasets, many variables are collected and, to ensure good statistical performance, one needs to circumvent the so-called "curse of dimensionality" by applying dimension reduction techniques. The key notion to clarify the performance of dimension reduction is sparsity, understood in a broad sense meaning that the phenomenon under investigation has a low-dimensional intrinsic structure. Sparsity is also at the core of compressive sensing for data acquisition. The simplest notion of sparsity is developed for vectors, where it provides an opening to high-dimensional linear regression (LASSO) and non-linear regression, such as for instance generalized high-dimensional linear models, using regularization techniques. Such methods can be extended to deal with the estimation of low-rank matrices, that arise for instance in recommender systems under the problem of matrix completion. Sparsity is also helpful in the context of highly non-linear machine learning algorithms, such as clustering.
While clearly stating the mathematical foundations of dimension reduction, this course will focus on methodological and algorithmic aspects of these techniques

Machine Learning for Time Series

(Lectures and Tutorials: 18hrs)

When learning from structured data such as time series data, special attention has to be paid to the models used. Indeed, designing machine learning models requires thinking of the invariants to be learned, and either encoding them in the model or designing the model so that it is able to discover such invariants and encode them. In this course, we will cover the use of alignment-based methods in traditional machine learning models. Dedicated neural network architectures will also be tackled. All these models will be illustrated on real datasets. After this course, the student will be able to choose an adequate machine learning model and apply it for a given time series task.

High-Dimensional Time Series

(Lectures and Tutorials: 24hrs)

In this course, we will introduce the primary tools for analyzing time series data. We will begin by presenting the key concepts for dealing with univariate time series, such as trend, seasonality, and stationary processes. Subsequently, we will delve into the main models and inference methods for multivariate linear time series. In the latter part of the course, we will explore the scenario of multiple time series with a substantial number of components. To address high-dimensional parameter spaces, we will introduce the LASSO penalty and its variants, along with dimension reduction techniques using factor models. Towards the end of the course, we will provide an introduction to neural networks and to clustering or classification problems in the context of time series analysis. Real-world data examples and the software R will be used to illustrate all the methods.

Functional Data Analysis
(Lectures and Tutorials: 24hrs)

This course aims to provide an introduction to functional data analysis (FDA). The fundamental statistical tools for modeling and analyzing such data will be explored. This course introduces ideas and methodology in FDA as well as the use of software. Students will learn the idea of different methods and the related theory, and also the numerical and estimation routines to perform functional data analysis. Students will also have an opportunity to learn how to apply FDA to a wide array of application areas. The course will contain several examples where FDA techniques have clear advantage over classical multivariate techniques. Some recent development in FDA will also be discussed.

Machine Learning for Natural Language Processing (NLP)
(Lectures and Tutorials: 18hrs)

The course will introduce the main notions of NLP and detail machine learning based approaches to modern NLP, going through the following: word representation, text classification, word tagging, language modeling, transformers and large language models, text generation.
The courses will tightly mix lectures and guided hands-on practice that will be complemented by small personal projects pursuing the guided hands-on practice sessions.

Data Visualization
(Lectures and Tutorials: 15hrs)

Data visualization is a fundamental ingredient of data science as it "forces us to notice what we never ex-pected to see" in a given dataset.
Dataviz is also a tool for communication and, as such, is a visual language
All along the courses, we will focus on methods and strategies to represent datasets, using dynamic and interactive tools.

Parallel Computing with R & Python
(Lectures and Tutorials: 18hrs)

First, an introduction of code profiling is proposed (micro and macro profiling, memory monitoring). Then, the two standard methods for CPU parallel computations are presented (forking and socket).
In the R section, we will go through the basic tools in parallel programming, how to detect bottlenecks in their code, and how to perform simulations using parallelization.
With Python, we will cover basic ideas and common patterns in parallel computing, including embarrassingly parallel map, unstructured asynchronous submit, and large collections.

IT Tools 1: Hadoop & Cloud Computing
(Lectures and Tutorials: 18hrs)

The goal of this course is to give a brief introduction to Cloud Computing: definitions, types of cloud (IaaS/PaaS/Saas, public/private/hybrid), challenges, applications, main cloud players (Amazon, Microsoft Azure, Google etc.), and cloud enabling technologies (virtualization). Then we will explore data processing models and tools used to handle Big Data in clouds such as MapReduce and Hadoop. An overview on Big Data including definitions, the source of Big Data, and the main challenges introduced by Big Data, will be presented. After that, we will discuss distributed file systems. We will then present the MapReduce programming model as an important programming model for Big Data processing in the Cloud. Hadoop ecosystem and some of major Hadoop features will then be discussed.

IT Tools 2: NoSQL & Big Data Processing with Spark
(Lectures and Tutorials: 24hrs)

One of the main goals of this module is to understand the fundamentals of NoSQL databases and the features and specific challenges NoSQL databases are addressing compared to classic SQL databases. Evaluate and select appropriate NoSQL technologies for particular situations. Gain hands-on experience in deploying and using NoSQL databases, such as MongoDB or Neo4j.
Another goal of this module is to understand the stakes of distributed computing through the Apache Spark architecture. Discover how to use Apache Spark, platforms & tools available. Practice PySpark coding to learn Apache Spark features, from data management to machine learning.

Smart Data Project or Research Project
(Lectures and Tutorials: 24hrs)

Smart Data Project:
The main part of courses focuses on studying several facets of statistics, mathematics and computer sciences, according to the Big/Smart Data paradigm. One of the main objectives of this project is to apply this new knowledge learned among the 1st semester into a unique application. This project puts into practice theoretical methods studied in different courses and starts with project management.

The learning objective is not limited to putting the theory learned in other courses into practice, but aims to raise awareness of other aspects linked to project management among students, such as communication (between students and also with the client that proposed the project).

This project should provide additional support, be carried out by an expert of the field, according to the needs of students. The expert is expected to provide

- Supervising at start for requirement
  - Distant supervising on technical queries
  - Technical supervising during implementation phase
  - Help for defense preparation

Research Project:
Depending on the profile of the student, a research project can be proposed as an alternative to the Smart Data project. The aim of this project will be an initiation to a modern research topic in the field of statistics or machine learning. Such a project will be considered as a priority for students interested in pursuing a PhD.

Topics, Case Studies, Conferences
(24hrs)

Several conferences held by specialists or researchers from the academic, industrial, or business world will be organized. “Smart Data” is becoming a major issue in modern society. The purpose of these conferences is to provide an up-to-date review of the ongoing data revolution, on the stakes for analyzing the information in a smart way, on presenting recent case studies, and on providing complementary perspectives (economic, business, management) for Smart Data students.

French Summer Program
(August - Duration: 4 weeks)

Non-French speakers arrive 1 month early to France for intensive French language and culture courses, while being hosted with a French family. While the Smart Data Science program is taught in English, this allows students to acquire vital skills for daily life and cultural integration.

Courses for Non-French Speakers: Written and/or Oral French Language Courses
(Duration: 2 or 4 hours/week over the 1st semester)

Designed specifically for foreign students, these weekly evening courses give students practical written and/or oral French skills, necessary for everyday life in France.

SEMESTER 2

End-of-Studies Internship
(Duration: 4 to 6 months from the end of February)

This final phase of the Master for Smart Data Science program involves a four to six-month paid internship, which can take place either in France or abroad, in either the professional world or academic/research laboratories. This experience should allow the student to apply the data science and computer science theory and methods that they have learned during the first semester of coursework. The internship should allow students to meet at least two objectives: a technical and a professional one. The student must write a master’s thesis and defend it in front of a jury in September.