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. The preliminary courses only allow students to complete the prerequisites for the MSc. Depending on their background, students will be asked to take some or all of these courses.
Tentative list of preliminary courses for 2022/2023:

- Statistical languages: R, Python (18h)
- Multivariate Data Exploration (12h)
- Markov Chains (12h)
- GNU Linux & Shell Scripting (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 deep learning. Beforehand, the stochastic gradient descent algorithm and the back-propagation - its application to feedforward neural networks - are introduced in this context. 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 use 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, mainly held in Python with TensorFlow/Pytorch illustrations.

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 focus on time series and will dig into two main ways of encoding / learning these invariants. First, we will cover the design of alignment-based metrics that tackle the problem of (temporal) localization invariance. Standard similarity measures will be introduced and their use at the core of machine learning models will be discussed. Second, we will discuss standard neural network architectures and the kind of invariants they encode.

High-Dimensional Time Series
(Lectures and Tutorials: 24hrs)

Vector-valued time series are ubiquitous in many application domains. With common models for inference and forecast multivariate time series, practitioners often face a high dimensional problem due to the large number of parameters involved in the dynamics. In the first part of the course, it is shown that regularization techniques originally introduced for linear regression models, can be particularly useful for fitting vector autoregressive models to the data. An alternative route for reducing complexity is provided by dimension reduction approaches. This is the spirit of factor models, where one assumes a small number of unobserved variables summarize the information contained in large time-series vectors. The second part of the course presents some common factor models.

Functional Data Analysis
(Lectures and Tutorials: 18hrs)

Functional data analysis (FDA) deals with data that are in the form of functions, images, or more general objects. The functional datum is such an object, or a vector of such objects. Each object is measured continuously or at several discrete points in a domain. This course introduces ideas and methodology in FDA as well as the use of software. Students will learn 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 demonstrate applications where FDA techniques have clear advantage over classical multivariate techniques. Some recent developments in FDA will also be discussed.

Graphical Models and Latent Structures
(Lectures and Tutorials: 24hrs)

The course will focus on probabilistic graphical models, which give compact and analytically useful representations of joint distributions over a large number of variables, using graphs. Each graph represents a family of distributions – the nodes of the graph represent random variables, the edges encode independence assumptions. First we will introduce the basics of probabilistic graphical models and will study both directed and undirected graphical models. We will study their mathematical properties, algorithms for their implementation, and applications to real problems. The course will then provide a comprehensive survey of state-of-the-art methods for statistical learning and inference in graphical models. In particular, we will discuss EM and latent variable models, approximate inference, variational inducing, sampling techniques and sequential Monte Carlo methods for static and dynamic random graphs. Finally, the problem of causality will be introduced.

Data Visualization
(Lectures and Tutorials: 15hrs)

Data visualization is a fundamental ingredient of data science as it "forces us to notice what we never expected to see." In this course, we show through examples and case studies that graphical methods are powerful tools for revealing the structure of the data, patterns and (ir)regularities, groups, trends, outliers... Dataviz is relevant for data analysis, when the analyst wants to study data, but also, as any statistics, to question the data. Additionally, it is a tool for communication and, as such, a visual language with a theory of the functions of signs and symbols used to encode the visual information. All along the course, we will focus on methods, tools and strategies to represent simple and then complex or high-dimensional datasets, highlighting the growing development of dynamic and interactive tools.

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

The content of the course is devoted to the implementation of calculation on different central processing units (CPU) and the use of servers.
First, code and memory profiling methods are introduced, and then different ways to improve the code on a single CPU are presented.
Second, the forking and sockets methods which allow for parallelization of independent chunks of codes in R are described.
Finally, parallelizing MPI use shall be demonstrated in R with the Rmpi, extending it to packages and Scientific Python library in Python.
(As the tools that are in the scope of the course evolve rapidly, the professor reserves the right to adapt the content to the latest developments.)

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

The goal of this module is to introduce Cloud Computing: definitions, types of clouds (IaaS/PaaS/Saas, public/private/hybrid), challenges, applications, main cloud players (Amazon, Microsoft Azure, Google etc.), and cloud enabling technologies (virtualization). Then data processing models and tools used to handle Big Data in clouds such as MapReduce, Hadoop, and Spark will be explored. An overview on Big Data including definitions, the source of Big Data, and the main challenges introduced by Big Data, will be presented. The MapReduce programming model will be presented as an important programming model for Big Data processing in the Cloud. Hadoop ecosystem and some of Hadoop's major 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 address, compared to classic SQL databases. An introduction to deploying and using NoSQL databases, such as MongoDB, or CouchDB will be provided.
Another goal of this module is to understand key concepts and get practice of distributed data processing frameworks such as Apache Spark. All steps of a typical data science project using large volumes of data will be covered: accessing data sources, preparing and processing data, storing them, but also using distributed machine learning libraries such as Apache Spark MLlib and H20 to train and fire models. Emphasis will on practice & hands-on sessions.

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

Smart Data Project:
The program courses focus mainly on studying several aspects of Statistics, Machine Learning, and Computer Science, within the framework of the Big Data paradigm. One of the main objectives of this project is to apply all the new knowledge learned among the 1st semester to a unique application. The project is supervised by specialists or researchers from academic, industrial, or business fields. The Smart Data project puts into practice theoretical methods studied in different courses and

starts with project management. The learning objective is not limited simply to putting theory learned in courses into practice. It also 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).

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.