Course Catalog
Master for Smart Data Science

ACADEMIC YEAR 2022 / 2023

ENSAI
École nationale
de la statistique
et de l'analyse
de l'information

List of courses

General Presentation and Objectives

The world is producing previously unimaginable amounts of data every second. This data could help to understand and improve our society, to predict and prevent, to combat diseases and generally improve life. Extracting valuable information and creating knowledge from the massive and heterogeneous data require skills in statistical modelling, machine learning algorithms, as well as computer science. The synergy of these academic fields, oriented towards their application, is the guiding idea of the Master for Smart Data Science at ENSAI.

ENSAI is part of the network of prestigious higher-education establishments in France known as Grandes Ecoles, or specialized graduate schools. ENSAI trains its students to become qualified, high-level specialists in information processing and analysis.

The graduates of this Master will be capable of creating and implementing methodologies and algorithms for analyzing large flows of data arriving from different sources, of using statistical tools and machine learning algorithms to identify correlations, effects, patterns and trends in data, and of formalizing predictions. As such, they will be qualified for data scientist and artificial intelligence jobs in industry, marketing, banking and insurance, media, or further pursuing a PhD.

This Master's program is composed of 1 semester of coursework at ENSAI, followed by a four to six-month paid internship in France or abroad within the professional world, academia, or research laboratories.

Since this program welcomes students with varying academic levels and skills in Computer Science, Applied Mathematics and Statistics, preliminary coursework is put in place to bring all students to the same scientific level in these fields, with respect to their existing training, knowledge, and skills.

For the 2022-2023 academic year, most of the lectures are intended to take place at ENSAI, in live. Some of the lectures are scheduled online using Microsoft Teams. Depending on the pandemic evolution, more courses could move online.

Curriculum – Program Overview and Credits

	Semester Hours	ECTS Credits	Total in the block
UE–MSD01 – Machine Learning			
Machine Learning for Data Science	30	3.5	
Deep Learning	15	1.5	7
Dimension Reduction & Matrix Completion	18	2	
UE–MSD02 – Models for Dependent Data			
Machine Learning for Time Series	18	2	5
High-Dimensional Time Series	24	3	
UE–MSD03 – Statistics for New Data			
Functional Data Analysis	18	2	5
Graphical Models & Latent Structures	24	3	
UE–MSD04 – Advanced Tools for Data Analysis & Computing			
Data Visualization	15	1	3
Parallel Computing with R & Python	18	2	
UE–MSD05 – IT Tools			
IT Tools 1 (Hadoop & Cloud Computing)	18	2	5
IT Tools 2 (NoSQL, Big Data Processing with Spark)	24	3	
UE–MSD06 – Case Studies and Project			
Smart Data Project / or Research Project	24	2.5	5
Topics & Case Studies, Conferences / or Research Project	24	2.5	
TOTAL Semester 1	270 H	30 credits	
UE-MSD07- Internship End-of-Studies Internship	(4 to 6 months)		30
TOTAL Semester 2		30 credits	
TOTAL Academic Year	270 H	60 credits	

Prior to the start of the first semester, the students will be given the opportunity to attend courses designed to reinforce different topics in Computer Science, Statistics, and Mathematics. The tentative list of these courses for September 2022 is the following:

GNU Linux & Shell Scripting	12 h
Statistical Languages – R, Python	18 h
Multivariate Data Exploration	12 h
Markov Chains	12 h

List of Professors and Lecturers

Code	Topic	Professor/Lecturer
Preliminary 1	GNU Linux & Shell Scripting	Guillaume GRABE
Preliminary 2	Statistical Language: R	Matthieu MARBAC-LOURDELLE
Preliminary 3	Statistical Language: Python	Pierre NAVARO
Preliminary 4	Multivariate Data Exploration	Cesar SANCHEZ SELLERO
Preliminary 5	Markov Chains	Adrien SAUMARD
MSD 01-1	Machine Learning for Data Science	François PORTIER
MSD 01-2	Deep Learning	Pavlo MOZHAROVSKYI
MSD 01-3	Dimension Reduction & Matrix Completion	Adrien SAUMARD
MSD 02-1	Machine Learning for Time Series	Romain TAVENARD
MSD 02-2	High-dimensional Time Series	Valentin PATILEA Jad BEYHUM
MSD 03-1	Functional Data Analysis	Eftychia SOLEA
MSD 03-2	Graphical Models & Latent Structures	Eftychia SOLEA
MSD 04-1	Data Visualization	Laurent ROUVIERE
MSD 04-2	Parallel Computing with R & Python	Matthieu MARBAC-LOURDELLE Pierre NAVARO
MSD 05-1	IT Tools 1 (Hadoop & Cloud Computing)	Shadi IBRAHIM
MSD 05-2	IT Tools 2 (NoSQL, Big Data Processing with Spark)	Nikolaos PARLAVANTZAS Hervé MIGNOT
MSD 06-1	Smart Data Project / or Research Project	Industrial/lab partners
MSD 06-2	Topics, Case Studies, Conferences / or Research Project	
	Reinforcement Learning	Pascal BIANCHI
	Some Recent Advances for Big Data Processing in the Cloud	Shadi IBRAHIM
	Gans	Ugo TANIELIAN
	Case Studies in Smart Data	Thomas ZAMOJSKI
MSD 07-1	End-of-Studies Internship	

Preliminary Courses

Preliminary 1 – MSD - Before the start of the 1st Semester

GNU Linux & Shell Scripting

Professor	:	Guillaume GRABE (PAYFIT)
ECTS Credits	:	0 (preliminary course)
Estimated personal workload (beyond lecture and tutorial time)	:	5 to 7 hrs
Lectures and Tutorials	:	12 hrs (ENSAI) including 2 to 3 hrs of independent work
Teaching language	:	English
Software	:	Linux + Shell (installed during the lecture)
Course materials	:	A computer (lent by ENSAI)
Prerequisites	:	A computer + internet connection + VirtualBox

Learning Objectives

This class teaches students the concepts that they should understand before they start working with GNU/Linux. During this course, students will configure a distribution on their computer and learn how to interact with the shell, from basic tasks (navigation, file edition, network configuration) to more advanced operations with shell scripting.
GNU/Linux is essential in particular when using and developing Big Data technologies.

Main Subjects covered

1. GNU/Linux
- Introduction to GNU/Linux
- Installing a distribution
- The shell
- Users, groups, permissions
- Packages management
- Network management

2. Shell scripting
- Shell scripting principles
- Variables in the shell, operations on variables
- Conditional expressions, basic statements, functions
- Regular expressions

References

1. https://wiki-dev.bash-hackers.org/
2. http://tldp.org/index.html
3. B. FOX and C. RAMAY, Bash Reference Manual, Free Software Foundation

Preliminary 2 – MSD - Before the start of the 1st Semester

Statistical Language: R

Professor	:	Matthieu MARBAC-LOURDELLE (ENSAI)
ECTS Credits	:	0 (preliminary course)
Estimated personal workload (beyond lecture and tutorial time)	:	9 hrs
Lectures and Tutorials	:	9 hrs (ENSAI)
Teaching language	:	English
Software	:	R
Course materials	:	Slides and tutorials on Moodle
Prerequisites	:	A laptop with R and RStudio installed

Learning Objectives

At the end of the lectures, the student will know the basic concepts of R programming.

Main Subjects covered

This course is organized in three parts:

- Introduction to the data analysis with R
- Presentation of the programming elements
- Performing a simulation with R

References

1. WINSTON CHANG, R Graphics Cookbook, O'Reilly, 2013.
2. CORNILLON P-A et al., R for Statistics, Chapman & Hall, 2012.
3. COTTON R, Learning R, O'Reilly, 2013.
4. GROLEMUND G., Hands-On Programming with R, O'Reilly, 2014.

Preliminary 3 – MSD - Before the start of the 1st Semester

Statistical Language: Python

Professor	:	Pierre NAVARO (Université Rennes 1)
ECTS Credits	:	0 (preliminary course)
Estimated personal workload (beyond lecture and tutorial time)	:	1 hour
Lectures and Tutorials	:	9 hrs (ENSAI) including 1 h of independent work on a small project to implement the linear regression model using a Python class
Teaching language	:	English
Software	:	miniconda (https://docs.conda.io/en/latest/miniconda.html)
Course materials	:	https://github.com/pnavaro/python-notebooks
Prerequisites	:	Experience in programming with another language

Learning Objectives

Python is a programming language used for many different applications. In this practical course, students will start from the very beginning, with basic arithmetic and variables, and learn how to handle data structures, such as Python lists, Numpy arrays. Students will learn about Python functions, control flow and data visualizations with Matplolib.
At the end of the lecture, the students are expected to know how to code with Python.

Main Subjects covered

- Setting up your Python environment
- Write functions using control flow tools and manage files input and output
- Introduction to object orienting programming.
- Jupyter Notebook
- NumPy
- Matplotlib
- Implementation of a simple regression model

References

1. Python documentation http://docs.python.org/
2. LUTZ M., ASCHER D., Learning Python, O'Reilly
3. LANGTANGEN H.P, Python Scripting for Computational Science, Springer
4. Python Data Science Handbook https://jakevdp.github.io/PythonDataScienceHandbook/
5. How to Think Like a Computer Scientist: Learning with Python
6. http://interactivepython.org/runestone/static/thinkcspy/

Preliminary 4 – MSD - Before the start of the 1st Semester

Multivariate Data Exploration

Professor	:	Cesar SANCHEZ SELLERO (Universidad de Santiago de Compostela)
ECTS Credits	:	0 (preliminary course)
Estimated personal workload (beyond lecture and tutorial time)	:	12 hrs
Lectures and Tutorials	:	12 hrs (online)
Teaching language	:	English
Software	:	R
Course materials	:	Slides on Moodle and scripts on R
Prerequisites	:	Basic knowledge of Statistics (notions of estimation, confidence intervals and hypothesis testing) and basic Algebra (vectors, matrices, scalar product, norms…)

Learning Objectives

This course provides an introduction to the main exploratory methods used to analyze multivariate data and to summarize their main characteristics. The concepts will be illustrated by applications using R packages. The contents are structured in the following chapters.
At the end of the lectures, the students are expected to know how to analyse and represent multivariate data and how to make groups in data.

Main Subjects covered

- Principal Components Analysis: Algebraic derivation of the principal components of a random vector. Geometric properties of the principal components as a least squares approximation and comparison with regression. Rescaling principal components. Choosing the number of components. Interpreting the components. Simultaneous representation of individuals and variables: the biplot.
- Correspondence Analysis: Contingency tables. Chi-Squared statistic as a measure of the variability between conditional distributions. Decomposing the variability. Simultaneous representation of rows and columns in a contingency table.
- Hierarchical Clustering: Distances, similarities and hierarchical clustering. Agglomerative and divisive methods. Single, complete or average linkage methods. Ward's method. Representation of hierarchical clustering: the dendrogram.
- Non-hierarchical Clustering: K-means method. Clustering mixtures of Gaussian distributions.

References

1. EVERITT, B.S, An R and S-Plus companion to multivariate analysis, Springer, 2005.
2. EVERITT, B.S, Dunn, G. Applied multivariate data analysis. Hodder Education, 2001.
3. HUSSON, F., Le, S., PAGES, J. Exploratory multivariate analysis by example using R. CRC Press, 2011.
4. JOHNSON, R.A., WICHERN, D.W, Applied multivariate statistical analysis, Pearson Education, 2007.

Preliminary 5 – MSD - Before the start of the 1st Semester

Markov Chains

Professor	:	Adrien SAUMARD (ENSAI)
ECTS Credits	:	0 (preliminary course)
Estimated personal workload (beyond lecture and tutorial time)	:	12 to 24 hrs depending on the student's knowledge about Markov Chains
Lectures and Tutorials	:	12 hrs (ENSAI)
Teaching language	:	English
Software	:	N/A
Course materials	:	Blackboard
Prerequisites	:	Basic probability notions

Learning Objectives

Markov chains are a central family of random processes that naturally arises in various fields of application through modelisation. Markov chains allow also to describe a great variety of (stochastic) optimization techniques. It is thus very important to recall the basic notions related to Markov chains and to their long-time behavior, which is the primary goal of this course.
At the end of the lecture, the students are expected to be able to:

- Identify a Markov chain in a modelisation context and prove that a stochastic process is a Markov chain.
- Analyze the static structure of a Markov chain (i.e. establish the associated transition graph, identify the structure in communication classes, show the recurrence or transience of the states of the chain, calculate the periodicities of the classes).
- Describe the limit behaviour of an ergodic chain (i.e., by calculating the stationary, possibly reversible law; cite and apply the limit theorems of the course).

Main Subjects covered

- Basic definition, discrete state space
- Chapman-Kolmogorov equation and Markov properties.
- States classification, periodicity, recurrence and transcience.
- Stationary law and limit theorem (long time behavior)

References

1. NORRIS J.R., Markov Chains, Cambridge Series in Statistical and Probabilistic Mathematics, 1997.
2. GRIMMETT G.R. & STIRZAKER D.R., Probability and Random Processes, Oxford Sciences Publications, 1992 (2nd edition).
3. PARDOUX E., Processus de Markov et applications: Algorithmes, réseaux, génome et finance. Dunod, 2007.

First Semester

1st Semester

TEACHING UNIT MSD-01 : MACHINE LEARNING

Supervisor	: François PORTIER (ENSAI)
ECTS Credits	: 7
Estimated personal workload (beyond lecture and tutorial time)	: 50 to 60 hrs
Lectures and Tutorials	: 63 hrs

Learning Objectives of the Teaching Unit

Introduce fundamental and modern machine learning approaches and provide computing tools for effective implementation. Topics in model/feature selection and regularization methods, regression trees, aggregation methods and support vector machine (more generally RKHS regression), as well as neural network methods and algorithm will be presented. Dimension reduction techniques are also presented. The students are expected to know the main up to date algorithms and to be able to implement them.

Description

The Machine Learning unit includes 3 courses:

1. Machine Learning for Data Science
2. Deep Learning
3. Dimension reduction & Matrix completion

Machine Learning for Data Science is a general and introductory course on Machine Learning. It will cover most of the techniques used in Machine Learning. Deep Learning is a more specific course on the use of Neural networks in Machine Learning. They have become one of the leading class of algorithms — due notably to their success in image processing. The last course is about Dimension reduction in Machine Learning. It covers several sets of techniques that are essential to treat large scale data.

Acquired Skills

Knowledge of a large panel of algorithms, use of modern machine learning approaches for complex data problems, implementation of algorithms using packages and notebooks.

Pre-requisites

Regression models, notions of probability theory, linear algebra and geometry, algorithm complexity.

UE-MSD01 – Machine Learning – MSD 01.1 - 1st Semester

Machine Learning for Data Science

Professor	:	François PORTIER (ENSAI)
ECTS Credits	:	3.5
Estimated personal workload (beyond lecture and tutorial time)	:	5 to 10 hrs
Lectures and Tutorials	:	30 hrs (ENSAI) including 3 hrs of independent work
Teaching language	:	English
Software	:	Python and R
Course materials	:	Slides and lecture notes
Prerequisites	:	Linear algebra, probability, optimization

Learning Objectives

Upon completing this course, students should be able to:

- select the appropriate methods;
- implement these statistical methods;
- compare leading procedures based on statistical arguments;
- assess the prediction performance of a learning algorithm;
- apply these key insights into class activities using statistical software.

Main Subjects covered

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 XGboost algorithm.

We shall focus on methodological and algorithmic aspects, while trying to give an idea of the underlying theoretical foundations. Practical sessions will give 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.

Evaluation

Final exam and computer class

References

1. HASTIE T., TIBSHIRANI R., FRIEDMAN J.H., The elements of statistical learning: data mining, inference and prediction; 2009
2. JAMES G., WITTEN D., HASTIE T., & TIBSHIRANI R., An Introduction to Statistical Learning. New York: Springer. R; 2013.

UE-MSD01 – Machine Learning – MSD 01.2 - 1st Semester

Deep Learning

Professor	:	Pavlo MOZHAROVSKYI (Telecom ParisTech)
ECTS Credits	:	1.5
Estimated personal workload (beyond lecture and tutorial time)	:	12 hrs
Lectures and Tutorials	:	15 hrs (ENSAI) including 1,5 hrs of independent work
Teaching language	:	English
Software	:	R, Python
Course materials	:	Slides, lab subjects and codes for practical sessions
Prerequisites	:	Regression analysis, gradient descent, (matrix) algebra, R, Python (basics).

Learning Objectives

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 Restricted Boltzmann machines (RBM) and the contrastive divergence algorithm (CD-k) are examined. Further 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 and Python.

Main Subjects covered

- Introduction to deep learning.
- Neural network architectures.
- Stochastic gradient descent and the back-propagation algorithm.
- Neural networks for regression and classification.
- Convolutional neural networks, Restricted Boltzman Machines.
- Applications: Pattern recognition, object detection

Evaluation

Written Exam + Lab

References

1. GOODFELLOW, I., BENGIO, Y., COURVILLE, A. Deep Learning. MIT Press. 2016.
2. HASTIE, T., TIBSHIRANI, R., FRIEDMAN, J. The Elements of Statistical Learning. Springer-Verlag. 2009.
3. HAYKIN, S.O. Neural Networks and Learning Machines. Pearson. 2008.
4. VAPNIK, V.N. Statistical Learning Theory. Wiley-Blackwell. 1998.

UE-MSD01 – Machine Learning – MSD 01.3 - 1st Semester

Dimension Reduction & Matrix Completion

Professor	:	Adrien SAUMARD (ENSAI)
ECTS Credits	:	2
Estimated personal workload (beyond lecture and tutorial time)	:	36 hrs
Lectures and Tutorials	:	18 hrs (ENSAI)
Teaching language	:	English
Software	:	Python
Course materials	:	Slides
Prerequisites	:	Basic statistics, linear algebra and probability.

Learning Objectives

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.

- Understand the curse of dimensionality and the notion of sparsity.
- Know the definition of the Lasso and its main variants, as well as its main algorithmic implementations.
- Understand the tuning of the Lasso and know the main techniques.
- Know how to regularize a high-dimensional generalized linear model.
- Understand the matrix completion problem and the collaborative filtering approach.
- Know how to use the SVD and solve a low-rank matrix estimation problem.

Main Subjects covered

- High-dimensional linear regression.
- High-dimensional generalized linear models.
- Low-rank matrix estimation.

Evaluation

Mean {1 project, 1 oral examination}

References

1. HASTIE T., TIBSHIRANI R., WAINWRIGHT M., Statistical Learning with Sparsity, The Lasso and generalizations, CRC Press, 2015.
2. BÜHLMANN P., VAN DE GEER S., Statistics for high-dimensional data, Springer, 2011.
3. WAINWRIGHT M., High-dimensional statistics, A non-asymptotic viewpoint, Cambridge Series in Statistical and Probabilistic Mathematics, 2019.
4. GIRAUD C., Introduction to High-dimensional Statistics, CRC Press, 2nd Edition, 2014
5. FOUCART S., RAUHUT H., A Mathematical Introduction to Compressive Sensing, Springer, 2013.

1st Semester

TEACHING UNIT MSD-02 : MODELS FOR DEPENDENT DATA

Supervisor	: François PORTIER (ENSAI)
ECTS Credits	: 5
Estimated personal workload (beyond lecture and tutorial time) /	: 60 to 70 hrs
Lectures and Tutorials	: 42 hrs

Learning Objectives of the Teaching Unit

In many modern applications, temporal dependency needs to be considered to build reliable models and to reach a fine prediction accuracy. Typical examples include weather forecasting or predicting the price of a given financial derivative. The first part of the unit presents methods, algorithms to handle time-series data. The second part of the unit is interested in modeling multivariate (potentially high-dimensional) time-series. The aim is to account for the possible interaction between different time series.

Description

The Models for dependent data unit includes 2 courses:

1. Machine Learning for Time Series
2. High-Dimensional Time Series

Acquired Skills

Build statistical model taking into account time-dependency in data, estimate the model using dependent data

Pre-requisites

Probability theory (covariance matrix, correlation, Gaussian vector), linear algebra, basic machine learning algorithms.

UE-MSD02 – Models for Dependent Data – MSD 02.1 - 1st Semester

Machine Learning for Time Series

Professor	:	Romain TAVENARD (Université Rennes 2)
ECTS Credits	:	2
Estimated personal workload (beyond lecture and tutorial time)	:	10 hrs
Lectures and Tutorials	:	18 hrs (ENSAI) including 1.5 h of independent work
Teaching language	:	English
Software	:	Python
Course materials	:	https://rtavenar.github.io/ml4ts_ensai/contents/foreword.html
Prerequisites	:	Basics of neural networks

Learning Objectives

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.

Main Subjects covered

- Shift invariance in time series
- Alignment-based methods for time series
- Recurrent neural networks
- Convolutional models for time series

Evaluation

Written exam + report on a real-data analysis.
The report will be initiated during class hours, and more specifically during the 1.5 hours marked as "independent work".

References

1. GOODFELLOW, I., BENGIO, Y., COURVILLE, A. (2016). Deep learning. MIT Press, 2016.
2. Lecture notes for the course are made available at: https://rtavenar.github.io/ml4ts_ensai/contents/foreword.html

UE-MSD02 – Models for Dependent Data – MSD 02.2 - 1st Semester

High-Dimensional Time Series

Professors	:	Valentin PATILEA (ENSAI) Jad BEYHUM (ENSAI)
ECTS Credits	:	3
Estimated personal work-load (beyond lecture and tutorial time)	:	60 hrs
Lectures and Tutorials	:	24 hrs (ENSAI)
Teaching language	:	English
Software	:	R
Course materials	:	Slides, codes, articles, book chapters
Prerequisites	:	Standard background in probability theory. Gaussian vectors. Matrix Algebra. Notions of univariate time series: autocorrelation function, ARMA processes, least squares method. Principal component analysis. Basic regularization and classification methods.

Learning Objectives

Time Series analysis accounts for the fact that data points (numbers or vectors) observed over time have a specific structure (such as autocorrelation or seasonal variation). Time series models aim revealing such structures, making inference, building forecasts, ... The standard concepts and models for univariate and multivariate time series will be reviewed.

Time series are called high-dimensional when the number of variables is large relative to the number of dates of observations. We will present two-dimension reduction approaches for high-dimensional time series. The first approach is based on regularization, such as l1-norm penalization, under sparsity or low-rank assumptions. The second approach is based on the approximate factor model. Motivated by the increasing amount of available information, such models are a versatile approach to summarize information contained in large vectors of data. Estimation of the factors by principal components analysis will be discussed. Different estimators of the number of factors will be introduced. We will show how to use the estimated factors for forecasting.

After this lecture, the students will know and will be able to apply the main diagnosis tools and models to analyze and forecast 1- , multi- or high-dimensional time series.

Main Subjects covered

- Univariate Time series: stationarity, autocorrelations, basic models
- Vector autoregressive models. Stationarity and statistical inference of VAR models.
- High-dimensional VAR models: regularized estimation under sparsity or low-rank assumptions.
- Factor models, estimation by principal components analysis, factor-augmented regression

Evaluation

Written exam + report on a real-data analysis

References

1. AHN, S, HORENSTEIN, A. Eigenvalue ratio test for the number of factors. Econometrica 81:1203–27. 2013.
2. BAI, J., and PENG W. Econometric analysis of large factor models. Annual Review of Economics 8:53-80. 2016.
3. BAI, J. Inferential theory for factor models of large dimensions. Econometrica 71:135–72. 2003.
4. BAI, J., NG S. Confidence intervals for diffusion index forecasts and inference for factor-augmented regressions. Econometrica 74:1133–50. 2006.

5. BASU, S. and MICHAILIDIS, G. Regularized estimation in sparse high-dimensional time series models. Annals of Statistics 43:1535–1567. 2015.
6. BASU, S, LI, X., and MICHAILIDIS, G. Low rank and structured modeling of high-dimensional vector autoregressions. IEEE Transactions on Signal Processing 67:1207– 1222. 2019.
7. LUTKEPOHL, H. New introduction to multiple time series analysis. Springer. 2005.
8. STOCK, J. H., and WATSON, M.W. Dynamic Factor Models. In The Oxford Handbook of Economic Forecasting. Oxford University Press. 2011.
9. TSAY, R.S. Multivariate time series analysis: with R and financial applications. Wiley. 2014.

1st Semester

TEACHING UNIT MSD-03 : STATISTICS FOR NEW DATA

Supervisor	: François PORTIER (ENSAI)
ECTS Credits	: 5
Estimated personal workload (beyond lecture and tutorial time)	: 63 hrs
Lectures and Tutorials	: 42 hrs

Learning Objectives of the Teaching Unit

Modern applications produce data under various complex forms. Many existing data are composed of curves, images or are graph-structured, for which standard regression technique or the time-series paradigm is not applicable or relevant. This unit presents, on one hand, the functional data analysis (FDA) approaches and, on the other hand, graphical models to handle such kind of particular data. This unit will hence explore alternative approaches to the classical ones as developed in the Machine Learning and time-series unit.

Description

There are two courses:

1. Functional data analysis
2. Graphical models and latent structures

Acquired Skills

Model complex data using advanced methods.

Pre-requisites

Analysis and linear algebra, Probability theory

UE-MSD03 – Statistics for New Data – MSD 03.1 - 1st Semester

Functional Data Analysis

Professor	:	Eftychia SOLEA (ENSAI)
ECTS Credits	:	2
Estimated personal workload (beyond lecture and tutorial time)	:	1.5 h personal workload per 1h lecture
Lectures and Tutorials	:	18 hrs (ENSAI) including 1.5 hrs of independent work
Teaching language	:	English
Software	:	R
Course materials	:	Lecture notes and textbook (see below)
Prerequisites	:	Statistical inference and methods, Multivariate statistical analysis

Learning Objectives

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 demonstrate applications where FDA techniques have clear advantage over classical multivariate techniques. Some recent development in FDA will also be discussed.

Main Subjects covered

Chapter 1. Introduction.
Chapter 2. Representing functional data and exploratory data analysis. Including: basic expansions, FPCA, derivatives, penalised smoothing, registration, fda package.
Chapter 3. Basic elements of Hilbert space theory and random functions.
Chapter 4. Estimation and inference from a random sample. Including, estimation of functional principal component analysis (FPCA). Inference about the mean function.
Chapter 5. Functional Linear regression models. Including: Functional linear regression models with scalar and functional response variable (function-on-scalar, scalar-on-function and function-on-function models).
Chapter 6. Functional generalised linear models
Chapter 7. Analysis of functional time series and the ftsa package.

Evaluation criteria

The final grade will be determined by three criteria: Homework (20%), small project (30%) and Final exam (50%)

References

1. RAMSAY, J.O. and SILVERMAN, B. W. Functional Data Analysis. Springer. 2005.
2. RAMSAY, J.O., HOOKER, G. and GRAVES, S. Functional Data Analysis in R and Matlab. Springer. 2009.
3. SHI, J. Q. and CHOI, T. Gaussian Process Regression Analysis for Functional Data. Chapman & Hall/CRC Press. 2011.
4. HORMANN, S. and KIDZINSKI, L., HALLIN, M. Dynamic Functional Principal Components. JRSSB, Vol. 77, No. 2, pp. 319-348. arXiv 1210.7192v5. 2015.
5. SHANG, H. L. ftsa: An R package for analysing functional time series. The R journal, 64-72. 2013.
6. HORVATH, L. and KOKOSZKA, P. Inference for Functional Data with Applications. Springer Series in Statistics, Volume XIV. 2012.
7. KOKOSZKA, P and REIMHER, M. Introduction to Functional Data Analysis. Chapman & Hall/CRC, Texts in Statistical Science. 2017.

UE-MSD03 – Statistical for New Data – MSD 03.2 - 1st Semester

Graphical Models & Latent Structures

Professor	:	Eftychia SOLEA (ENSAI)
ECTS Credits	:	3
Estimated personal workload (beyond lecture and tutorial time)	:	1.5 h personal workload per 1h lecture
Lectures and Tutorials	:	24 hrs (ENSAI) including 1.5 h of independent work
Teaching language	:	English
Software	:	R
Course materials	:	Lecture notes, textbook (see list of references below)
Prerequisites	:	Basic knowledge in probability theory, mathematics & programming recommended

Learning Objectives

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 conditional independence. 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 learning the graphs, and applications to real problems. Then, the course will 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.

Main Subjects covered

1. Introduction.
2. Conditional distributions and conditional independence. Conditional independence models. Directed and undirected graphical models. Markov properties.
3. Structure learning: Gaussian graphical models and Ising graphical models
4. Basic inference in graphical models (variable elimination etc)
5. Parameter estimation: MLE, Bayesian, Exponential family distributions, learning with latent variables, EM algorithm and latent variable models.
6. Approximate inference: variational techniques and sampling techniques.
   Directed graphs: causal inference (if time allows)

Evaluation

Grades will be based on components on homework (45%) and final exam (55%)

References

1. WAINWRIGHT, M.J., JORDAN, M.I. "Graphical models, exponential families, and variational inference." Foundations and Trends® in Machine Learning, Vol. 1, No 1-2: 1-305, 2008.
2. HJSGAARD, S., EDWARDS, D., LAURITZEN, S. "Graphical Models with R. Springer, New York. 2012.
3. BISHOP, C. "Introduction to graphical modelling". 2nd edn. Springer, New York. 2000.
4. LAURITZEN, S.L. "Graphical models". Clarendon Press, Oxford. 1996.
5. KOLACZYK, E.D., GABOR, C. "Statistical analysis of network data with R". New York: Springer, 2014.
6. KOLACZYK, E.D. "Statistical Analysis of Network Data: Methods and Models". Springer. 2009.
7. NEWMAN, M. "Networks: An Introduction". Published to Oxford Scholarship Online. 2010.

1st Semester

TEACHING UNIT MSD-04 : ADVANCED TOOLS FOR DATA ANALYSIS & COMPUTING

Supervisor	: François PORTIER (ENSAI)
ECTS Credits	: 3
Estimated personal workload (beyond lecture and tutorial time)	: 20 to 30 hrs
Lectures and Tutorials	: 33 hrs

Learning Objectives of the Teaching Unit

This teaching unit develops two important topics that lies at the heart of any data analysis: data visualization and parallel computing. Those 2 topics are often left aside in most Machine learning or statistical course (which most often are interested in modeling and predicting). They here have their own places. Data visualization is a set of technique allowing to summarize visually some piece of information contained in the data but also to allow determining some patterns in the data. Parallel computing consists in sending what needs to be computed to different machine in order to reduce computing time. This is necessary in most large scale learning problems.

Description

There are two courses:

1. Data visualization
2. Parallel computing with R & Python

Acquired Skills

Algorithm complexity

Pre-requisites

Basics on R an Python. A minimal knowledge of the basic tools used in data science, as well as in statistics is required such as: PCA, classification algorithms.

UE-MSD04 – Advanced Tools for Data Analysis & Computing – MSD 04.1 - 1st Semester

Data Visualization

Professor	:	Laurent ROUVIERE (Université Rennes 2)
ECTS Credits	:	1
Estimated personal workload (beyond lecture and tutorial time)	:	10 to 15 hrs
Lectures and Tutorials	:	15 hrs (ENSAI)
Teaching language	:	English
Software	:	R
Course materials	:	
Prerequisites	:	During this course, we will manipulate basic notions used in data science. A minimal knowledge of the basic tools used in data science, as well as in statistics is required such as: PCA, classification algorithms. Basics on R are also necessary.

Learning Objectives

Data visualization is a fundamental ingredient of data science as it "forces us to notice what we never expected to see" in a given dataset. In this course, we show through examples and case studies that graphical methods are powerful tools for revealing not only the structure of the data, but also patterns and (ir)regularities, groups, trends, outliers…
Dataviz is relevant both for data analysis, when the analyst wants to study data and, as any statistics, to question the data. It is also a tool for communication and, as such, is a visual language with a theory of the functions of signs and symbols used to encode the visual information. All along the course, we'll 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.

Main Subjects covered

- Data visualization for data sciences
- Classics in Data visualization
- Grammar of graphics with ggplot2
- Mapping with sf and leaflet
- Interactive and dynamic visualization

Evaluation

The evaluation consists on a data visualization project. The students will have to:
- deploy a shiny web application and to publish it on the web.
- write a markdown report to present the application
They will work in groups with two members.

References

1. BERTIN, J. 1983. Semiology of Graphics, translation from Sémilogie graphique . 1967.
2. TUFTE, E. R The Visual Display of Quantitative Information. 2 ed. Graphics Press. 2001.
3. http://ggplot2.org
4. https://ggplot2-book.org
5. https://statnmap.com/fr/2018-07-14-initiation-a-la-cartographie-avec-sf-et-compagnie/
6. https://rstudio.github.io/leaflet/
7. *https://rmarkdown.rstudio.com/flexdashboard*

UE-MSD04 – Advanced Tools for Data Analysis & Computing – MSD 04.2 - 1st Semester

Parallel Computing with R & Python

Professors	:	Matthieu MARBAC-LOURDELLE (ENSAI) - lectures on "R" Pierre NAVARO (Université Rennes 1) - lecture on "Python"
ECTS Credits	:	2
Estimated personal workload (beyond lecture and tutorial time)	:	12 to 15 hrs
Lectures and Tutorials	:	18 hrs (ENSAI)
Teaching language	:	English
Software	:	R and Python
Course materials	:	Material on Moodle for R and https://github.com/pnavaro/big-data for Python
Prerequisites	:	Knowledge of R and Python

Learning Objectives

- Detecting the slow parts of a script by using graphical tools for code profiling. Students will be able to detect the parts of a script where the code should be improved and where the memory allocations should be reduced.
- Improving the code performances using CPU parallel computation. Students will be able to use both of the forking and socket methods of parallel computation.

Main Subjects covered

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).
With Python, we will cover basic ideas and common patterns in parallel computing, including embarrassingly parallel map, unstructured asynchronous submit, and large collections.

Evaluation

Lab 2 hrs (the report is written at home)

References

1. https://www.r-project.org (R-packages Rmpi, RHadoop assembly, gpuR).
2. https://wiki.python.org/moin/ParallelProcessing (ScientificPython library).
3. https://computing.llnl.gov/tutorials/mpi/
4. DEAN, J., GHEMAWAT, S. MapReduce: simplified data processing on large clusters. Proceedings of OSDI'04. 2004.
5. https://www.khronos.org/opencl/

1st Semester

TEACHING UNIT MSD-05 : IT TOOLS

Supervisor	:	François PORTIER (ENSAI)
ECTS Credits	:	5
Estimated personal workload (beyond lecture and tutorial time)	:	35 to 40 hrs
Lectures and Tutorials	:	42 hrs

Learning Objectives of the Teaching Unit

In modern applications, the collected data is often associated with a large dimension (big data's one v is volume) and needs to be treated in a small amount of time (big data's other v is velocity). In such context, IT tools have become of prime importance in data science. This unit presents a panorama of modern computer/cloud tools for processing massive amounts of complex data.

Description

Courses of Linux, NoSQL, Hadoop and Spark are proposed.

Acquired Skills

Using the most recent computer/cloud computing tools (Hadoop and Spark) for data processing.

Pre-requisites

Basics in programming and databases: Java, Python, R, Linux, SQL.

UE-MSD05 – IT Tools – MSD 05.1 - 1st Semester

IT Tools 1 (Hadoop & Cloud Computing)

Professor	:	Shadi IBRAHIM (INRIA – Rennes)
ECTS Credits	:	2
Estimated personal workload (beyond lecture and tutorial time)	:	9 to 15 hrs
Lectures and Tutorials	:	18 hrs (ENSAI) including 1,5 hrs of independent work
Teaching language	:	English
Software	:	Hadoop, Virtual Machine Mangers (e.g., Virtual Box, VMware-Player, VMware Fusion, etc)
Course materials		All course materials presentations, tutorials and hand-ons, libraries and codes will be available online on the course website in pdf and zip format.
Prerequisites		Familiar with Linux command-line Familiar with Java/Python

Learning Objectives

At the end of the lectures, the student will realize the potential of Big Data and will know the main tools to process this tsunami of data at large-scale. In particular, the students will understand the main features of MapReduce programming model and its open-source implementation Hadoop, and will be able to use Hadoop and test it using different configurations.
Data volumes are ever growing, for a large application spectrum going from traditional database applications, scientific simulations to emerging applications including Web 2.0 and online social networks. To cope with this added weight of Big Data, we have recently witnessed a paradigm shift in computing infrastructure through Cloud Computing and in the way data is processed through the MapReduce model. First promoted by Google, MapReduce has become, due to the popularity of its open-source implementation Hadoop, the de facto programming paradigm for Big Data processing in large-scale infrastructures. On the other hand, cloud computing is continuing to act as a prominent infrastructure for Big Data applications.

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. 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.

Main Subjects covered

Throughout the course we will cover the following topics:

- Cloud Computing: definitions, types, Challenges, enabling technologies, and examples (2.25 hrs)
- Big Data: definitions, the source of Big Data, challenges (1.5 hrs)
- Google Distributed File System (1.5 hrs)
- The MapReduce programming model (1.5 hrs)
- Hadoop Ecosystem (2.25 hrs)
- Practical sessions on Hadoop (7 hrs)
    - ü How to use Virtual Machines and Public Cloud Platforms
    - ü Starting with Hadoop
    - ü Configuring HDFS
    - ü Configuring and Optimising Hadoop
    - ü Writing MapReduce applications

Independent work (tentative): Students will be divided into groups where each group will do a 15 - 20 min presentation on one of the main subjects or a life demonstration on one of the practical sessions (2 hrs)

Evaluation

Written exam

References

1. JIN Hai, IBRAHIM Shadi, BELL Tim, GAO Wei, HUANG Dachuan, WU Song. Cloud Types and Services. Book Chapter in the Handbook of Cloud Computing, Springer Press, 26 Sep 2010.

2. JIN Hai, IBRAHIM Shadi, BELL Tim, LI QI, HAIJUN Cao, WU Song, XUANHUA Shi. Tools and technologies for building the Clouds. Book Chapter in Cloud Computing: Principles Systems and Applications, Springer Press, 2 Aug 2010.

3. ARMBRUST Michael, FOX Armando, GRIFFITH Rean,. JOSEPH Anthony D, KATZ Randy, KONWINSKI Andy, LEE Gunho, PATTERSON David, RABKIN Ariel, STOICA Ion, and ZAHARIA Matei. 2010. A view of cloud computing. Commun. ACM 53, 4 - April 2010.

4. GHEMAWAT Sanjay, GOBIOFF Howard, and LEUNG Shun-Tak. The Google file system. In SOSP '03.

5. DEAN Jeffrey, GHEMAWAT Sanjay, OSDI, MapReduce: Simplified Data Processing on Large Clusters. 2004.

6. JIN Hai, IBRAHIM Shadi, LI QI, HAIJUN Cao, WU Song, XUANHUA Shi. The MapReduce Programming Model and Implementations. Book Chapter in Cloud Computing: Principles and Paradigms.

7. VAVILAPALLI Vinod Kumar, MURTHY Arun C., DOUGLAS Chris, AGARWAL Sharad, KONAR Mahadev, EVANS Robert, GRAVES Thomas, LOWE Jason, SHAH Hitesh, SETH Siddharth, SAHA Bikas, CURINO Carlo, O'MALLEY Owen, RADIA Sanjay, REED Benjamin, and BALDESCHWIELER Eric. Apache Hadoop YARN: yet another resource negotiator. In SOCC '13.

UE-MSD05 – IT Tools – MSD 05.2 - 1st Semester

IT Tools 2 (NoSQL, Big Data Processing with Spark)

Professors	:	Nikolaos PARLAVANTZAS (IRISA Rennes) - NoSQL Hervé MIGNOT (Equancy) – Big Data Processing with Spark
ECTS Credits	:	3
Lectures and Tutorials (total)	:	24 hrs (9+15)
Teaching language	:	English

NoSQL

Professor	:	Nikolaos PARLAVANTZAS (IRISA Rennes) - NoSQL
Estimated personal workload (beyond lecture and tutorial time)	:	10 hrs
Lectures and Tutorials	:	9 hrs (ENSAI) including 1,5 h of independent work
Software	:	Redis, Elasticsearch, Cassandra, Neo4j
Course materials	:	Slides and lab subjects on Moodle
Prerequisites	:	Basic knowledge of SQL, databases, and computer systems

Learning Objectives

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.

Main Subjects covered

- NoSQL origins (history & players)
- NoSQL / SQL comparison
- Key concepts of NoSQL databases:
  - ü Data models
  - ü Distribution models
  - ü Query languages
  - ü Consistency
- NoSQL database types
- NoSQL database technologies & comparisons (MongoDB, Cassandra, Neo4j, Redis, ElasticSearch…)
- Neo4j introduction + lab
- ElasticSearch introduction + lab

Evaluation

Questionnaire

References

Many online resources are available

Big Data Processing with Spark

Professor	:	Hervé MIGNOT (Equancy)
Estimated personal workload (beyond lecture and tutorial time)	:	14 hrs
Lectures and Tutorials	:	15 hrs (ENSAI)
Software	:	Spark
Course materials	:	
Prerequisites	:	Computer systems and architecture basic knowledge, Python & SQL language practice

Learning Objectives

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.

Main Subjects covered

- Distributed computing introduction
- Apache Spark origins & history, links to Apache Hadoop
- Apache architecture and main concepts:
  - Apache Spark "modules"
  - Architecture: driver & executors
  - Transformations vs. actions
  - Lazy evaluation
  - Data structures: RDD, dataframes & datasets
- Using Apache Spark:
  - ü Create sessions and connect to clusters
  - ü Use data management functions
  - ü Leverage SQL with Spark SQL
  - ü Train & test machine learning models
- Use Spark Web UI

Evaluation

Questionnaire and Project

References

1. Apache Spark online documentation: https://spark.apache.org/docs/latest/
2. KARAU H., WARREN R. High Performance Spark (2017). Note: old but with details about Spark internals.

1st Semester

TEACHING UNIT MSD-06 : CASE STUDIES & PROJECT

Supervisor	: François PORTIER (ENSAI)
ECTS Credits	: 5
Estimated personal workload (beyond lecture and tutorial time)	:
Lectures and Tutorials	: 48 hrs

Learning Objectives of the Teaching Unit

This teaching unit has been organized to offer the students the occasion to work on new subjects using the knowledge they acquired during the first semester. This is an important step toward completing the master as the students should demonstrate their ability to draw some links between the previous teaching units in order to discover new topics. The first part consists in working on a project as a team (two or three by team, supervised by a field expert) and the second part is divided into multiple seminar sessions (each is dedicated to a recent data science topic).

Description

There are two phases in this teaching unit:

1. The project
2. The seminars

Acquired Skills

Knowledge on some specific hot-topics in data science and the ability to work as a team on a (research-like) project

Pre-requisites

All previous courses given in the Master.

UE-MSD06-Case Studies and Project – MSD 06.1 - 1st Semester

Smart Data Project or Research Project

Supervisors : Several industrial or lab partners
ECTS Credits : 2.5

Learning Objectives

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

Main Subjects covered

The topic of the Smart Data project could be related to any type of application requiring advanced data science tools.

Evaluation

The evaluation is two-fold:

1 - a report written by all students of each project team, eventually supervised by the external organism.
2 - a project defense in front of a jury

UE-MSD06-Case Studies and Project – MSD 06.2 - 1st Semester

Topics, Case Studies, Conferences /or Research Project

Professors		Pascal BIANCHI (Telecom Paris Tech)
		Shadi IBRAHIM (INRIA - Rennes)
		Ugo TANIELIAN (Criteo)
	:	Thomas ZAMOJSKI (DATASTORM)
ECTS Credits	:	2.5
Lectures and Tutorials (total)	:	24 hrs (Ensai)
Teaching language	:	English

Reinforcement Learning

Professor	:	Pascal BIANCHI (Telecom Paris Tech)
Estimated personal workload (beyond lecture and tutorial time)	:	1,5 hrs
Lectures and Tutorials	:	6 hrs
Course materials	:	Python notebook
Prerequisites	:	Basics of probability theory, conditional expectation

Learning Objectives

The reinforcement learning framework gathers several algorithms which allow an agent to take relevant sequential decisions when faced with an unknown environment providing a certain reward.

The aim is twofolds:

1) Introduce the framework of Markov Decision Processes which underlies many Reinforcement Learning techniques,
2) get hands dirty by applying the celebrated Q-learning method in a python notebook.

Main Subjects covered

- Markov decision processes and the Bellman equation
- Q-learning
- Deep Q-learning

Evaluation:

Labs

References

1. SUTTON R. and BARTO A., Reinforcement Learning: An Introduction, The MIT Press, Second edition, 2018

Some Recent Advances for Big Data Processing in the Cloud

Professor	:	Shadi IBRAHIM (INRIA – Rennes)
Estimated personal workload (beyond lecture and tutorial time)	:	3 to 5 hrs
Lectures and Tutorials	:	6 hrs (ENSAI) including 1 h of independent work
Software	:	Hadoop
Course materials	:	All course materials presentations, tutorials and hand-ons, libraries and codes will be available online on the course website in pdf and zip format.
Prerequisites	:	Attend the course: Big Data processing in Clouds: Hadoop

Learning Objectives

At the end of the lectures, the student will be able to identify the main performance bottlenecks when running Big data applications in Clouds and will know how the performance of Hadoop can be improved, accordingly.
During this conference, we will discuss several approaches and methods used to optimise the performance of Hadoop in the Cloud. We will also discuss the limitations of Hadoop and introduce state-of-the-art resource management systems and job schedulers for Big data applications including Mesos, Delay scheduler, ShuffleWatcher, and Tetrium.

Main Subjects covered

Approaches to optimize Hadoop in clouds (2.5 hrs)
Resource management and job scheduling for Big data applications: Mesos, Delay scheduler, ShuffleWatcher, Tetrium, etc (2.5 hrs)

Independent work (tentative): Students will be assigned to groups where each group will do a 15 -20 min presentation (1 hr)

Evaluation

During the session and/or a technical report to be submitted after the session

References

1. Apache Hadoop YARN: yet another resource negotiator. VAVILAPALLI Vinod Kumar, MURTHY Arun C., DOUGLAS Chris, AGARWAL Sharad, KONAR Mahadev, EVANS Robert, GRAVES Thomas, LOWE Jason, SHAH Hitesh, SETH Siddharth, SAHA Bikas, CURINO Carlo, O'MALLEY Owen, RADIA Sanjay, REED Benjamin, and BALDESCHWIELER Eric. In SOCC '13.
2. IBRAHIM Shadi, PHAN Tien-Dat, CARPEN-AMARIE Alexandra, CHIHOUB Houssem-Eddine, MOISE Diana, ANTONIU Gabriel. Governing energy consumption in hadoop through cpu frequency scaling: An analysis. In FGCS 2016.
3. PHAN Tien-Dat, IBRAHIM Shadi, ANTONIU Gabriel, BOUGE Luc. On Understanding the energy impact of speculative execution in Hadoop. In GreenCom2015.
4. YILDIZ Orcun, IBRAHIM Shadi, ANTONIU Gabriel. Enabling fast failure recovery in shared Hadoop clusters: Towards failure-aware scheduling.In FGCS 2016.

5. HINDMAN Benjamin, KONWINSKI Andy, ZAHARIA Matei, GHODSI Ali, JOSEPH Anthony D., KATZ Randy, SHENKER Scott, and STOICA Ion. Mesos: a platform for fine-grained resource sharing in the data center. In NSDI'11.
6. ZAHARIA Matei, BORTHAKUR Dhruba, SEN SARMA Joydeep, ELMELEEGY Khaled, SHENKER Scott, STOICA Ion. Delay scheduling: a simple technique for achieving locality and fairness in cluster scheduling. In EuroSys'10.
7. ZAHARIA Matei, KONWINSKI Andy, JOSEPH Anthony D., KATZ Randy, STOICA Ion. Improving MapReduce performance in heterogeneous environments. In OSDI'08.
8. AHMAD Faraz, CHAKRADHAR Srimat T., RAGHUNATHAN Anand, VIJAYKUMAR T. N. Shufflewatcher: Shuffle-aware scheduling in multi-tenant mapreduce clusters. In USENIX ATC 2014
9. HUNG Chien-Chun, ANANTHANARAYANAN Ganesh, GOLUBCHIK Leana, YU Minlan, and ZHANG Mingyang. 2018. Wide-area analytics with multiple resources. In EuroSys '18.

GANS

Professor	:	Ugo TANIELIAN (CRITEO)
Estimated personal workload (beyond lecture and tutorial time)	:	2 hrs
Lectures and Tutorials	:	6 hrs (ENSAI) including 1 h of independent work
Software	:	
Course materials	:	
Prerequisites	:	Knowledge of Machine Learning & Python programming, basic knowledge of pytorch

Learning Objectives

At the end of the lecture, the student will know:

- What are the challenges in generative modeling and mainly Generative Adversarial Networks (GANs).
- What are the pros and cons of GANs wrt to other generative techniques.
- What are the main variants of GANs and their characteristics.
- What were the main breakthroughs and the current difficulties in the GAN community.
- How to define, train, and evaluate GANs on both synthetic and real-world datasets.

Main Subjects covered

Generative Models has known a huge success in the past few years. In particular, Generative Adversarial Networks (GANs) have been proposed in 2014 as a new method efficiently producing realistic images. Since their original formulation, GANs have triggered a surge of empirical studies, and have been successfully applied to different domains of machine learning: video, sound generation, and image editing. This course will be an introduction to this powerful algorithm that is both extremely studied and but with many open questions remaining. A practical session will also be done where the students will implement, train, and evaluate different variants of GANs.

Evaluation:

- 60% In-class exercises,
- 30% Code quality and clarity,
- 10% Participation.

References

1. GOODFELLOW, I., POUGET-ABADIE, J., MIRZA, M., XU, B., WARDE-FARLEY, D., OZAIR, S., ... & BENGIO, Y. (2014). Generative adversarial nets. Advances in neural information processing systems, 27.
2. ARJOVSKY, M., CHINTALA, S., & BOTTOU, L. (2017, July). Wasserstein generative adversarial networks. In International conference on machine learning (pp. 214-223). PMLR.
3. KARRAS, T., LAINE, S., & AILA, T. (2019). A style-based generator architecture for generative adversarial networks. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 4401-4410).

Case Studies in Smart Data
MLOps: Machine Learning in a production environment

Professor	:	Thomas ZAMOJSKI (DATASTORM)
Estimated personal workload (beyond lecture and tutorial time)	:	1.5 – 2 hrs
Lectures and Tutorials	:	6 hrs (ENSAI) including 1 h of independent work
Software	:	Python, Docker
Course materials	:	
Prerequisites	:	Basic knowledge of Python Programming Language

Learning Objectives

At the end of the lecture, the student will know:
- What are the challenges in deploying and maintaining a machine learning model in operation.
- What are some best practices addressing these concerns.
- How to create a Docker image and run a container.
- How to serve a model as a service in python.
- Statistical methods for online and offline model monitoring.

Main Subjects covered

Machine Learning models are notoriously hard to put and maintain in production. But why is it so and what can we do about it?
In this course, we will explore the very latest trends in MLOps. We will learn about technologies such as Docker containers, FastAPI and MLFlow. We will also learn statistical methods to intelligently automate model monitoring and we will see how to put them in action via implementations in python packages such as scikit-multiflow and ruptures.

Evaluation

- 60% In-class exercises,
- 30% Code quality and clarity,
- 10% Participation.

References

1. TRUONG C., OUDRE L., VAYATIS N., Selective review of offline change point detection methods, Signal Processing, September 2019.
2. JAMES N.A., KEJARIWAL A., MATTESON D.S., Leveraging cloud data to mitigate user experience from ‘breaking bad’, 2016 IEEE International Conference on Big Data (Big Data).
3. WEB REFERENCES - 12 factors app: https://12factor.net

Second Semester

2nd Semester

TEACHING UNIT MSD-07 : INTERNSHIP

Supervisor	: François PORTIER (ENSAI)
ECTS Credits	: 30
Working time	: Full time internship, for a period between 4 and 6 months

Learning Objectives of the Teaching Unit

The internship is the main bridge between, on one hand, the scientific courses, tutorials and labs and, on the other hand, the world of work. It has two major objectives. First, consolidate students' ability to choose appropriate models, algorithms and computer resources to address real data applications and case studies, to realize proof of concepts and/or develop user solutions, and, finally, explain and provide appropriate arguments for the choices made. Second, place the students in total immersion in a professional environment, in autonomy, as part of a team, in interaction with specialists from the same or complementary fields.

Description

The MSc students are expected to work on topics defined in the internship agreement, under the supervision of a senior professional from the internship unit (private or public company, labs, research institutes…). Each MSc student will have an Ensai adviser who can be contacted for advice.

Acquired Skills

Become a highly skilled specialist in data science able to address complex tasks using up to date modeling tools and computer resources.

Pre-requisites

Complete the previous teaching unit from the Master's program.

UE-MSD07 - Internship – MSD07.1

End-of-Studies Internship

4-6 months from March to August

Objectives

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.

Students should be proactive and begin the search for an internship as early as possible to increase the chances of finding an interesting and relevant internship. Finding an internship is the exclusive responsibility of the student. ENSAI provides assistance in the search process.

This experience should allow for the student to apply the data-science and computer science theory and methods that they have learned during the 1st semester of coursework. Internship topics that are exclusively or almost exclusively oriented towards computer science tools will not be accepted.

The internship should allow students to meet at least two objectives:

- A technical objective: a task is given and, applying theoretical knowledge and skills, the student attempts to complete the task using to the best of his/her ability the resources at his/her disposal.
- A professional objective: the student is immersed in a professional context and must use the internship period to become more knowledgeable and at ease in such an environment, developing professional and personal skills to become a part of the team.

Evaluation

During the internship, students will write a master's thesis that will be examined by the jury and defended by the student in September.