Master of Science in Statistics for Smart Data

COURSE CATALOG
ACADEMIC YEAR 2017 / 2018

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 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 of Science “Statistics for Smart Data” at ENSAI.

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

The graduates of this MSc will be capable of processing and of 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 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.

Curriculum – Program Overview and Credits

	Semester Hours	ECTS Credits	Total in the block
UE-MSD01 – Statistical Models for Dependent Data			
Inhomogeneous Markov Models & Applications	30	2.5	
Graphical Models & Dynamic Networks	18	1.5	**5**
Dynamic Data Visualization	12	1	
UE–MSD02 – Machine Learning			
Machine Learning: Features Selection & Regularization Methods	18	1.5	
Deep Learning	30	2.5	**5**
Parallel Computing with R and Python	12	1	
UE–MSD03 – Smart Sensing			
Foundations of Smart Sensing	36	3	**5**
Advanced Topics in Smart Sensing	24	2	
UE–MSD04 – Models for Complex Data			
High-Dimensional Time Series	30	2.5	**5**
Functional Data Analysis	30	2.5	
UE-MSD05 – IT Tools			
IT Tools 1	30	2.5	**5**
IT Tools 2	30	2.5	
UE–MSD06 – Challenges for Smart Societies			
Energy Transitions: Quantitative Aspects	12	1	
Smart Data Project (8 weeks)	24	2	**5**
Topics & Case Studies in Data Science (conferences)	24	2	
(UE-MSD07 - French as a Foreign Language*)			
(French Summer Program [July-August at CIREFE]	(intensive)		**(8)**
(Courses for foreigners: Written/Oral French language S1 [at CIREFE])	(22)		
* for foreign students as needed			
TOTAL Semester 1	**360 H**	**30 credits**	
UE-MSD08- Internship	(4 to 6		**30**
End-of-Studies Internship	months)		
TOTAL Semester 2		**30 credits**	
TOTAL Academic Year	**360 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 list of these courses for September 2017 is the following.

Statistical Languages – R, Python	18 h
Multivariate Data Exploration	12 h
Markov Chain	12 h
Simulation Based Inference	6 h
Topics in Time Series	6 h

List of Professors and Lecturers

Code	Topic	Professor/Lecturer
Preliminary 1	Statistical Language: R	Matthieu MARBAC-LOURDELLE
Preliminary 2	Statistical Language: Python	Pierre NAVARO
Preliminary 3	Multivariate Data Exploration	Cesar SANCHEZ SELLERO
Preliminary 4	Markov Chains	Adrien SAUMARD
Preliminary 5	Simulation Based Inference	Myriam VIMOND
Preliminary 6	Topics in Time Series	Valentin PATILEA
MSD 01-1	Inhomogeneous Markov Models & Applications	Salima EL KOLEI Myriam VIMOND
MSD 01-2	Graphical Models & Dynamic Networks	Julien CHIQUET
MSD 01-3	Dynamic Data Visualization	Christophe BONTEMPS
MSD 02-1	Machine Learning: Features Selection & Regularization Methods	Fabien NAVARRO
MSD 02-2	Deep Learning	Badih GHATTAS Pavlo MOZHAROVSKYI
MSD 02-3	Parallel Computing with R and Python	Pavlo MOZHAROVSKYI
MSD 03-1	Foundations of Smart Sensing	Nancy BERTIN Cédric HERZET Aline ROUMY
MSD 03-2	Advanced Topics in Smart Sensing	Antoine CHATALIC Cédric HERZET Adrien SAUMARD
MSD 04-1	High-Dimensional Time Series	Valentin PATILEA Lionel TRUQUET
MSD 04-2	Functional Data Analysis	Jian Qing SHI
MSD 05-1	IT Tools 1 (GNU Linux & Shell Scripting, Hadoop & Cloud Computing)	François-Xavier BRU Shadi IBRAHIM
MSD 05-2	IT Tools 2 (NoSQL, Big Data Processing with Spark)	Pauline NERRIERE Hervé MIGNOT
MSD 06-1	Energy Transitions: Quantitative Aspects	TBA
MSD 06-2	Smart Data Project (8 weeks)	Industrial partners

Code	Topic	Professor/Lecturer
MSD 06-3	Topics & Case Studies in Data Science (conferences)	
	Bandit Theory	Romaric GAUDEL
	Is Data The New Currency of The Digital Economy?	Valeriu PETRULIAN
	New Trends in Cloud Computing	Shadi IBRAHIM
	Case Studies in Smart Data	Thomas ZAMOJSKI
MSD 07-1	French as a Foreign Language	CIREFE
MBD 08-1	End-of-Studies Internship	

Preliminary Courses

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

Statistical Language: R

Lectures and Tutorials: 6 hrs

Professor : Matthieu MARBAC-LOURDELLE (ENSAI)

Course Objectives

The purpose of this course is to help develop skills for R programing. R has the triple advantage of being free, comprehensive, and used extensively and increasingly. During this course, you will go from loading data to writing your own functions.

Course Description

This course is organized in three parts:

- The first part gives you the very basics on R. It describes the different objects (vector, matrix, array, list, factor), the elements of functions (input and output) and the ways to manipulate data (importation and exportation).
- The second part focuses on programing elements. It introduces the *if* and *else if* statements, the loops (while, for, repeat) and gives an introduction of vectorized coding (apply, lapply, sapply, sweep, replicate).
- The third part focuses on the use of R for data analysis. It presents the main functions for exploring and visualizing data. Finally, it presents the main functions to generate data from classical distributions.

Table of Contents

1) Fundamentals of R	2) Programs	3) Data analysis work-flow
(a) R objects	(a) If and If/Else statements	(a) Exploring data
(b) Functions	(b) Loops	(b) Visualizing data
(c) Manipulating data	(c) Vectorized code	(c) Distributions and modeling

Prerequisites

A laptop with R and Rstudio installed

References

- Winston Chang (2013), *R Graphics Cookbook*, O'Reilly.
- Pierre-André Cornillon *et al.* (2012), *R for Statistics*, Chapman & Hall.
- Richard Cotton (2013), *Learning R*, O'Reilly.
- Garrett Grolemund (2014), Hands-On Programming with R, O'Reilly.

Important message

The course will include many codes and programs on R. This software is available for download free of charge at http://cran.r-project.org. Examples and exercises will be done with Rstudio. Rstudio is a free and open-source integrated development environment for R, a programming language for statistical computing and graphics. This environment is available for download free of charge at https://www.rstudio.com/. **Students are asked to come in class with their laptop where R and Rstudio are installed.**

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

Statistical Language: Python

Lectures and Tutorials: 12 hrs

Professor : Pierre NAVARO (Université Rennes 1)

Course 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.

Course Description

- Setting up your Python environment
- Jupyter Notebook
- Arithmetic operations
- Creating strings.
- The "input" statement and combining strings
- Lists
- Using the "if" statement
- Loops
- **F**unctions
- NumPy
- Matplotlib
- Introduction to object orienting programming.

References

- Python documentation http://docs.python.org/
- Learning Python, Mark Lutz, David Ascher (O'Reilly)
- Python Scripting for Computational Science, Hans Petter Langtangen (Springer)
- How to Think Like a Computer Scientist: Learning with Python

http://interactivepython.org/runestone/static/thinkcspy/

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

Multivariate Data Exploration

Lectures and Tutorials: 12 hrs

Professor : Cesar SANCHEZ SELLERO (Universidad de Santiago de Compostela)

Course Objectives

This course provides an introduction to the main exploratory methods used to analyze multivariate data and will summarize their main characteristics. The concepts will be illustrated by applications using R packages. The contents are structured in the following chapters.

Course Description

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

- Everitt, B.S. (2005). An R and S-Plus companion to multivariate analysis. Springer.
- Everitt, B.S., Dunn, G. (2001). Applied multivariate data analysis. Hodder Education.
- Husson, F., Le, S., Pages, J. (2011). Exploratory multivariate analysis by example using R. CRC Press.
- Johnson, R.A., Wichern, D.W. (2007). Applied multivariate statistical analysis. Pearson Education.

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

Markov Chains

Lectures and Tutorials: 12 hrs

Professor : Adrien SAUMARD (ENSAI)

Course Objectives

Markov chains are a central family of random processes that naturally arises in various fields of application through modelisation. Markov chains allow also describing 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.

Course Description

- 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)
- Basic statistical inference.

References

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

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

Simulation Based Inference

Lectures and Tutorials: 6 hrs

Professor : Myriam VIMOND (ENSAI)

Course Objectives

The aim is to give a quick overview of the use of Monte Carlo experiments for statistical inference. The lecture covers Monte Carlo Integration, Monte Carlo for estimation and for hypothesis tests, the bootstrap.

Course Description

- Monte Carlo methods for statistical inference
- Resampling

Prerequisites

Probability theory, Statistical Inference, Bayesian Inference

Preliminary 6 – MSD - Before the start of the 1st Semester

Topics in Time Series

Lectures and Tutorials: 6 hrs

Professor : Valentin PATILEA (ENSAI)

Course Objectives

Many data structures are characterized by a time dependency of the observations. In this short course, first some basic topics on smoothing the series are presented. In the second part, the stationary time series are considered and the standard ARMA models for univariate stationary series are recalled. Some extensions of this standard setup are mentioned.

Course description

- Smoothing and forecasting (series decomposition, moving-averages, exponential smoothing)
- Stationary time series: definitions
- ARMA models and some extensions

Prerequisites

Basic knowledge in probability theory, statistical inference & R programming recommended

References

- Martin V., Hurn S. and D. Harris (2012). *Econometric Modelling with Time Series: Specification, Estimation and Testing*. Cambridge University Press.
- Shumway, R.H. and D.S. Stoffer (2012). *Time Series and Its Applications: With R examples*. Springer Texts in Statistics. New-York: Springer.
- Subba Rao, T., Subba Rao S. and C.R. Rao (eds) (2012). *Time Series Analysis: Methods and Applications*. Handbook of Statistics, Volume 30. Elsevier, North-Holland.

First Semester

UE-MSD01 – Statistical Models for Dependent Data – MSD 01.1 - 1st Semester

Inhomogeneous Markov Models & Applications

Lectures and Tutorials: 30 hrs

Professors : Salima EL KOLEI (ENSAI)
Myriam VIMOND (ENSAI)

Course Objectives

Homogeneous Markov chains are exploited in a broad range of applications. Nevertheless, in some situations homogeneous transition probabilities do not adequately model real processes. In these situations, Markov models with inhomogeneous rates, i.e., rates that are time-varying functions or that depend on covariates, could be more appropriate. In the first part of this course, the theory of these processes will be described and will be illustrated with applications in several areas (financial, aeronautics, meteorological…). The second part of this course will be devoted to Hidden Markov Models (HMM). After presenting the basic HMM, the framework is extended by considering nonhomogeneous hidden Markov models and Markov-switching models. Such models are used in finance, electricity prices, genomics... Bayesian methods, such that MCMC and particle filters, and the Expectation Maximization algorithm will be introduced and applied to infer in these models. Several real data applications will be considered to illustrate the methods.

Course Description

1. Homogeneous Markov Chain 1.1 Markov Chains 1.2 Continuous time Markov Chain Process 1.3 Higher Order Markov Chains 1.4 Multivariate Markov Chains 1.5 Applications	2. Inhomogeneous Markov Chain 2.1 Discrete Time 2.2 Continuous Time 2.3 Applications
3. Hidden Markov-Models 3.1 HMM architecture 3.2 Parameter Estimation 3.3 Applications	4. Non-Homogeneous Hidden Markov Model 4.1 NHMM architecture 4.2 Parameter estimation 4.3 Applications

Prerequisites

Probability theory, Markov Chains, Monte Carlo Methods, Statistical inference, Generalized Linear Models.

Course Evaluation

To be confirmed (written or oral exam or project)

References

[1] Ching, Huang, Ng and Siu. "Markov chains : models, algorithms and applications" (2nd ed.). New York : Springer; 2013.
[2] Iversen, Moller, Morales and Madsen (2017). "Inhomogeneous Markov models for describing driving patterns". IEEE Transactions on Smart Grid, Vol. 8 (2), p 581--588.
[3] Dymarski, Przemyslaw, ed. "Hidden Markov Models: Theory and Applications". InTech, 2011.
[4] Ailliot and Pene (2015). "Consistency of the maximum likelihood estimate for non-homogeneous Markov--switching models". ESAIM: Probability and Statistics, Vol. 19, p. 268--292.

UE-MSD01 – Statistical Models for Dependent Data – MSD 01.2 - 1st Semester

Graphical Models & Dynamic Networks

Lectures and tutorials: 18 hrs

Professor : Julien CHIQUET (INRA – Paris)

Course Objectives

Among the various statistical methods involving graphs, this course aims to give an introduction to two vast fields of Statistics extremely popular at the era of big data. First, random graphs, which are used when the data itself is available as a graph whose nodes are fixed and edges random. In this case, the objective is to adjust a model to unravel some particular organization of the data. Second, probabilistic graphical models, which give a 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. Large-scale of social or biological networks examples will illustrate the algorithms presented in this course.

Course Description

1. Basics on graphs
Tutorial : analysis of real networks with igraph

2. Randoms graphs analysis
- Spectral Clustering
- Stochastic Block Model (SBM)

Tutorial : Variational inference in the SBM

3. Graphical models
- Log-linear models
- Gaussian graphical models (GGM)

Tutorial : Sparse inference of high-dimensional GGM

Prerequisites

Basic knowledge in probability theory, mathematics & programming recommended

Course Evaluation

Small seminar thesis in groups of two; 15 minutes presentation in groups of two (10 minutes talk, 5 minutes questions) + tutorial reports

References

- Wainwright, Martin J., and Michael I. Jordan. "Graphical models, exponential families, and variational inference." *Foundations and Trends® in Machine Learning* 1.1–2 (2008): 1-305.
- Højsgaard, S., Edwards , D., Lauritzen, S. (2012). Graphical Models with R. Springer, New York.
- Bishop, C. (2000). Introduction to graphical modelling, 2nd edn. Springer, New York.
- Lauritzen, S.L. (1996). Graphical models. Clarendon Press, Oxford.
- Kolaczyk, Eric D., and Gábor Csárdi. *Statistical analysis of network data with R*. Vol. 65. New York: Springer, 2014.

UE-MSD01 – Statistical Models for Dependent Data – MSD 01.3 - 1st Semester

Dynamic Data Visualization

Lectures and tutorials: 12 hrs

Professor : Christophe BONTEMPS (Toulouse School of Economics - INRA)

Course Objectives

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. It is also 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'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.

Course Description

- Data visualization for data sciences
- Classics in Data visualization
- Visualizing in many dimensions
- Interactive and dynamic visualization

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: Density, histograms, random variable, correlations, statistical models.

Course Evaluation

The evaluation will be based on case studies based on data and chosen during the last course. Students will work in groups, write a report and defend this report orally.

References

- Bertin, Jacques. 1983. Semiology of Graphics, translation from Sémilogie graphique (1967).
- Cleveland, William S., & McGill, Robert. 1984. Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods. *J. Amer. Statist. Assoc.*, 79(387), 531-554.
- Tufte, Edward R. 2001. *The Visual Display of Quantitative Information*. 2 ed. Graphics Press.
- Few, Stephen. 2012. *Show me the numbers: Designing tables and graphs to enlighten*. 2 ed. Burlingame: Analytics Press.
- Munzner, Tamara. 2014. *Visualization Analysis and Design*. (AK Peters Visualizaon Series). A K Peters/CRC Press.
- Tukey, John W. 1977. *Exploratory data analysis*. Reading, Mass.
- Unwin, Antony, Theus, Martin, & Hofmann, Heike. 2006. *Graphics of large datasets: visualizing a million*. Springer Science & Business Media.

UE-MSD02 – Machine Learning – MSD 02.1 - 1st Semester

Machine Learning:
Features Selection & Regularization Methods

Lectures and tutorials: 18 hrs

Professor : Fabien NAVARRO (ENSAI)

Course Objectives

Starting from classical notions of shrinkage and sparsity, this course will cover regularization methods that are crucial to high-dimensional statistical learning. The syllabus includes feature selection and model selection, linear and nonlinear techniques for regression and for classification. The course will 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 (9h of lecture, 9h of computer lab sessions).

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.

Course Description

1. Subset Selection 1.1. Best Subset Selection 1.2. Stepwise Selection 1.3. Choosing the Optimal Model	**2. Shrinkage Methods** 2.1. Ridge Regression 2.2. The Lasso 2.3. Selecting the Tuning Parameter
3. Basis Expansions and Regularization 3.1. Smoothing Splines 3.2. Choosing the Smoothing Parameter	**4. Generalized Additive Models** 4.1. GAMs for Regression Problems 4.2. GAMs for Classification Problems

Prerequisites

Students are expected to have the following background: familiarity with linear algebra; a working knowledge of R or Python programming; familiarity with multiple linear regression.

Course Evaluation

Laptop exam or/and project

References

- Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference and Prediction. Springer. Free download.
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (Vol. 6). New York: springer. Free download.

UE-MSD02 – Machine Learning – MSD 02.2 - 1st Semester

Deep Learning

Lectures and tutorials: 30 hrs

Professors : Badih GHATTAS (Université de la Méditerranée)
Pavlo MOZHAROVSKYI (ENSAI)

Course Objectives

The course starts with a general introduction to machine learning and deep learning giving a brief overview of the problems which may be addressed using different kind of approaches. For the machine learning part we begin with a review of classification and regression trees and then focus on aggregation methods like bagging, random forests (RF), and boosting (AdaBoost algorithm and the gradient boosting optimisation). Support vector machine (SVM) are also discussed.
The second part of the 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. Further practical aspects will be addressed about the usage of Deep Learning to resolve typical problems like object recognition or tracking, and Image segmentation.
Presented material shall be motivated by the theoretical background together with real data illustrations. There will be specific labs for each topic majorly held in R, along with sparse Tensorflow and Python illustrations.

Course Description

- Introduction to machine learning and deep learning.
- Decision trees, bagging, and random forests.
- Boosting classifiers.
- Support vector machines.
- Stochastic gradient descent and the back-propagation algorithm.
- Neural networks for regression and classification.
- Convolutional neural networks, Restricted Boltzman Machines.
- Applications: Object detection, image segmentation.

Prerequisites

Regression analysis, gradient descent, (matrix) algebra, R, Python (basics).

Course Evaluation

Written Exam + Regular Labs

References

- Hastie, T., Tibshirani, R., Friedman, J. (2009). *The Elements of Statistical Learning*. Springer-Verlag
- Haykin, S.O. (2008). *Neural Networks and Learning Machines*. Pearson.
- Schapire, R.E., Freund, Y. (2012). *Boosting: Foundations and Algorithms*. The MIT Press.
- Vapnik, V.N. (1998). *Statistical Learning Theory*. Wiley-Blackwell.

UE-MSD02 – Machine Learning – MSD 02.3 - 1st Semester

Parallel Computing with R & Python

Lectures and tutorials: 12 hrs

Professor : Pavlo MOZHAROVSKYI (ENSAI)

Course Objectives

First, an overview of architectures for distributed computing is proposed: those based on central processing units (CPU) including multicore processor, multiprocessor node (single-node cluster), computing cluster; and the graphics processing unit (GPU). The content of the course is devoted to implementation of calculation and memory distribution for the mentioned architectures.

For the CPU parallel computation, the focus is put on two prevailing frameworks: Message Passing Interface (MPI) and the Map/Reduce model. First, parallelising using MPI shall be demonstrated in R with the Rmpi and extending it packages and ScientificPython library in Python. Second, realisation of the Map/Reduce model in Python and using the RHadoop packages are discussed. For the GPU-parallelisation, gpuR package shall be explored, a handy R-wrap to the OpenCL - an open computing language supporting GPU next to NVIDIA's CUDA.

(As the tools that are in the scope of the course evolve rapidly, the professor preserves the right to adapt the content to recent developments.)

Course Description

- Overview of distributed architectures.
- Message Passing Interface.
- Map/Reduce model.
- Graphics processing unit.

Prerequisites

Knowledge of R and Python

Course Evaluation

1h written exam + 1.5h lab

References

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

UE-MSD03 – Smart Sensing – MSD 03.1 - 1st Semester

Foundations of Smart Sensing

Lectures and tutorials: 36 hrs

Professors : Nancy BERTIN (INRIA, Centre Rennes-Bretagne Atlantique)
Cédric HERZET (INRIA, Centre Rennes-Bretagne Atlantique)
Aline ROUMY (INRIA, Centre Rennes-Bretagne Atlantique)

Course Objectives

Although most signals (audio, images,…) of interest belong to a very large dimensional ambient space, many of them possess some structure that contains the useful information and makes them intrinsically low dimensional or compressible. Such structures can be modeled and exploited to reduce the cost of their acquisition and their processing. Sparse representations are a powerful tool to express and represent such signals, typically by a small number of nonzero coefficients in an appropriate basis or dictionary. Combined with some appropriate design of the sensing device, they allow to acquire and to reconstruct signals with an extremely reduced number of measurements. This course will present theoretical and algorithmic frameworks and tools for low-dimensional representations of large-scale data and their compressed acquisition and recovery.

Course Description

Introduction:
Main definitions, NP-Hardness of the sparse recovery problem and the need for efficient algorithms, introduction to geometrical interpretations, brief history and overview of the field and its applications.

Sparse Representations :

- Sparsity, sparse approximations, sparsity measures
- Sparse recovery algorithms: greedy algorithms, thresholding algorithms and proximal operators, convex relaxation, message passing algorithms
- Theoretical guarantees of recovery (deterministic point of view) : Restricted Isometry Property, Null Space Property, Exact Recovery Conditions, mutual coherence Dictionaries: fixed dictionaries and sparsifying transforms, overcomplete dictionaries, dictionary learning

Applications: Sparse Component Analysis, image denoising and compression…

Compressive sensing

- Definition of compressive sensing and comparison with sparse representations
- Information theoretical interpretation
- Measurement matrix design (i.e. how many measurements and constraints) and major inequalities
- Random matrices and theoretical guarantees (probabilistic point of view) (Restricted Isometry property)
- Phase transition and the Donoho-Tanner curve
- Some applications will be studied in details among: compressed MRI acquisition and
- reconstruction, digital communications, image inpainting and interpolation, hyperspectral images, audio and acoustic applications.

Prerequisites

Basic linear algebra ; basic matrix algebra; basic Fourier analysis ; basic statistical foundations

Course Evaluation

Written exam + 2 projects

References

- E.J. Candes, M.B. Wakin (2008). An Introduction To Compressive Sampling. IEEE Signal Processing Magazine, March.
- G. Kutyniok (2013). Theory and Applications of Compressed Sensing, GAMM Mitteilungen 36, 79-101.

Main reference (will serve as a textbook):

- S. Foucart, H. Rauhut (2013). A mathematical introduction to compressive sensing. Birkhauser.

Additional references:

- M. Elad (2010). Sparse and Redundant Representations: From Theory to Applications in Signal and Image Processing. Springer.
- Y. Eldar and G. Kutyniok (2012). **Compressed Sensing: Theory and Applications**. Cambridge University Press.
- H. Boche, R. Calderbank, G. Kutyniok, J. Vybiral (editors) (2015). Compressed sensing and its applications. Birkhauser.

UE-MSD03 – Smart Sensing – MSD 03.2 - 1st Semester

Advanced Topics in Smart Sensing

Lectures and tutorials: 24h

Professor : Antoine CHATALIC
Cédric HERZET (INRIA, Centre Rennes-Bretagne Atlantique)
Adrien SAUMARD (ENSAI)

Course Objectives

This course collects some advanced topics in compressive sensing and related techniques. Although self-contained, a prerequisite to this course is the course of foundations of smart sensing, that focuses on the notion of vector sparsity.
In many applications, signals are structured in many ways that go beyond the basic notion of sparsity for vectors. We will thus introduce notions of sparsity for matrices, focusing on low rank matrices, and for groups.
We will also describe some emerging techniques of compressive learning, where data compression is oriented toward a subsequent learning task. In this context, data compression is also called sketching. Examples of sketched PCA and sketch k-means will be presented.
This course will present theoretical and algorithmic frameworks and discuss applications that are in the scope of the methods.

Course Description

1_Sparsity and compressive sensing with matrices.
2_Group sparsity.
3_Sketched learning.

Prerequisites

Basic linear algebra; basic matrix algebra; basic Fourier analysis; basic statistical foundations; foundations of smart sensing

Course evaluation

Written exam + 1 project

References

• S. Foucart, H. Rauhut (2013). A mathematical introduction to compressive sensing.
Birkhauser.
• M. Elad (2010). Sparse and Redundant Representations: From Theory to Applications in Signal and Image Processing.

UE-MSD04 – Models for Complex Data – MSD 04.1 - 1st Semester

High-Dimensional Time Series

Lectures and tutorials: 30 hrs

Professor : Valentin PATILEA (ENSAI)
Lionel TRUQUET (ENSAI)

Course Objectives

To model and forecast multivariate time series, practitioners often face a high dimensional problem due to the large number of parameters involved in the dynamics. Then the regularization techniques, originally introduced for linear regression models, could be particularly useful for fitting vector autoregressive models to the data. The first part of the course will present such approaches. Next, estimation of high dimensional correlation matrices will be considered. This problem also requires a special attention due to the lack of precision of the empirical covariance matrix, especially when the inverse of this matrix is the object of interest, as it is the case in some applications. An overview of some existing methods for getting more accurate estimates of these covariance matrices in a high dimensional setup will be presented. Finally, the increasing size of available databases has led to the development of factor models, especially in econometrics. Such models are a versatile approach to summarize information contained in large vectors of data. In the last part of the course, the fundamental factor models and the common inferences approaches will be presented and illustrated with real datasets, with a focus on dynamic factor models.

Course description

- Multivariate autocorrelation function. Vector autoregressive models. Stationarity and statistical inference of VAR models.
- Regularization methods for high-dimensional VAR processes.
- High-dimensional correlation matrices. Inference. Application to portfolio allocation.
- Factor models.

Prerequisites

Standard background in probability theory. Gaussian vectors. Variance-Covariance matrices. Linear projections in L^2. Basic notions of univariate time series: autocorrelation function, ARMA processes, least squares method.

Course evaluation

Written exam

References

- Lutkepohl, H. New introduction to multiple time series analysis. Springer. 2005.
- Tsay, R.S. Multivariate time series analysis: with R and financial applications. Wiley. 2014.
- Stock, James H., and Mark W. Watson. "Dynamic Factor Models." In The Oxford Handbook of Economic Forecasting. : Oxford University Press, 2011-07-08.

UE-MSD04 – Models for Complex Data – MSD 04.2 - 1st Semester

Functional Data Analysis

Lectures and tutorials: 30 hrs

Professor : Dr Jian Qing SHI (Newcastle University, School of Mathematics & Statistics)

Course Objectives

This course introduces ideas and methodology in functional data analysis (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.

Course Description

Chapter 1. Introduction (1h lecture)

Chapter 2. Representing functional data and exploratory data analysis
(3h lecture, 2h computer-based practicals) Including: basic expansions, smoothing, *fda* package

Chapter 3. Registration (2h lecture, 2h computer-based practicals)

Chapter 4. Functional principal component analysis (PCA) and analysis of functional time series (5 hour lecture, 2 hours computer-based practicals)
Including: Functional PCA, dynamical functional PCA, Analysis of functional time series, *fda* and *ftsa* packages

Chapter 4. Functional Linear regression models (5h lecture, 2h computer-based practicals)
Including: Functional linear regression models with scalar or functional response variable (function-on-scalar, scalar-on-function and function-on-function models)

Chapter 5. Bayesian nonparametric regression using Gaussian process (3 hour lecture, 2 hours computer-based practicals)
Including: Gaussian process regression analysis, *GPFDA* package

Chapter 6. Further problems (1 hour lecture)

Prerequisites

Statistical inference and methods, Multivariate statistical analysis

Course Evaluation

Project and report

References

- Ramsay, J. O. and Silverman, B. W. (2005). *Functional Data Analysis*. Springer.
- Ramsay, J. O., Hooker, G. and Graves, S. (2009). *Functional Data Analysis in R and Matlab*. Springer.
- Shi, J. Q. and Choi, T. (2011). *Gaussian Process Regression Analysis for Functional Data.* Chapman & Hall/CRC Press.
- Hormann, S. and Kidzinski, L. (2015). Dynamic functional principal componenets. arXiv 1210.7192v5
- Shang, H. L. (2013). ftsa: An R package for analysing functional time series. *The R journal*, 64-72.

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

IT Tools 1 (GNU Linux & Shell Scripting, Hadoop & Cloud Computing)

Lectures and tutorials: 30 hrs

Professors : François-Xavier BRU (Orange Cyberdefense)
Shadi IBRAHIM (INRIA - Rennes)

GNU Linux & Shell Scripting

(12 hrs – François Xavier BRU)

Course Objectives

This class teaches students the concepts that they should understand before they start working with GNU/Linux. During this course, students will install 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.

Course Description

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

Course Evaluation

Written evaluation or project

References

- C.PELISSIER, Unix, Editions Hermès
- B. FOX and C. RAMAY, Bash Reference Manual, Free Software Foundation

Hadoop & Cloud Computing

(18 hrs – Shadi IBRAHIM)

Course Objectives

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, Hadoop, and Spark. 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.

Course Description

Throughout the course we will cover the following topics:

- Cloud Computing: definitions, types, Challenges, enabling technologies, and examples (3 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 (3 hrs)
- Practical sessions on Hadoop (7.5 hrs)
  - How to use Virtual Machines and Microsoft Azure
  - Starting with Hadoop on Azure
  - Configuring HDFS
  - Configuring and Optimising Hadoop
  - Writing MapReduce applications

Prerequisites

- Familiar with Linux command-line
- Familiar with Java/Python

Course Evaluation

Written exam

References

[1] Cloud Types and Services. Hai Jin, Shadi Ibrahim, Tim Bell, Wei Gao, Dachuan Huang, Song Wu. Book Chapter in in the Handbook of Cloud Computing, Springer Press, 26 Sep 2010.
[2] Tools and technologies for building the Clouds. Hai Jin, Shadi Ibrahim, Tim Bell, Li Qi, Haijun Cao, Song Wu, Xuanhua Shi. Book Chapter in Cloud Computing: Principles Systems and Applications, Springer Press, 2 Aug 2010.

[3] A view of cloud computing. Michael Armbrust, Armando Fox, Rean Griffith, Anthony D. Joseph, Randy Katz, Andy Konwinski, Gunho Lee, David Patterson, Ariel Rabkin, Ion Stoica, and Matei Zaharia. 2010.. Commun. ACM 53, 4 (April 2010).
[4] The Google file system. Sanjay Ghemawat, Howard Gobioff, and Shun-Tak Leung. In SOSP '03.
[5] MapReduce: Simplified Data Processing on Large Clusters. Jeffrey Dean, Sanjay Ghemawat, OSDI, 2004.
[6] The MapReduce Programming Model and Implementations. Hai Jin, Shadi Ibrahim, Li Qi, Haijun Cao, Song Wu, Xuanhua Shi. Book Chapter in Cloud Computing: Principles and Paradigms.
[7] Apache Hadoop YARN: yet another resource negotiator. Vinod Kumar Vavilapalli, Arun C. Murthy, Chris Douglas, Sharad Agarwal, Mahadev Konar, Robert Evans, Thomas Graves, Jason Lowe, Hitesh Shah, Siddharth Seth, Bikas Saha, Carlo Curino, Owen O'Malley, Sanjay Radia, Benjamin Reed, and Eric Baldeschwieler. In SOCC '13.
[8] Spark: Cluster Computing with Working Sets. Matei Zaharia, Mosharaf Chowdhury, Michael J. Franklin, Scott Shenker, Ion Stoica, HotCloud 2010, June 2010.
[9] Discretized streams: fault-tolerant streaming computation at scale. Matei Zaharia, Tathagata Das, Haoyuan Li, Timothy Hunter, Scott Shenker, and Ion Stoica. In SOSP '13.

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

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

Lectures and tutorials: 30 hrs

Professors : Pauline NERRIERE (Equancy) - NoSQL
Hervé MIGNOT (Equancy) – Big Data Processing with Spark

Big Data Processing with Spark (18h – Hervé MIGNOT)

Course Objectives

Companies & organizations are collecting massive amounts of various data, making distributed storage & processing key technological challenges. Over the last decade, several cornerstone systems have been released to address these topics, such as Apache Hadoop and more recently Apache Spark. Promoted by a vivid open source world, distributed storage projects are blooming while Apache Spark is becoming a de facto standard for data processing. Beyond data processing and transformation, building data science applications using statistical and machine learning is now the new challenge, requiring both distributed learning and prediction engines. Also, dealing with streaming data for near real-time data processing is getting momentum as companies move to more event processing oriented architecture to cope with the data deluge they are facing.
The goal of this course 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 be set on practice & hands-on sessions.

Course Description

Course description (throughout the course we will cover the following topics)

- Distributed Storage & Computing: key concepts, origins, challenges, and examples. Introduction to Apache Spark, pySpark & SparkR (**3 hrs**)
- Apache Spark first hands-on session (**3 hrs**)
- Data sources, streaming and data manipulation (**3 hrs**)
- SparkSQL and Spark Streaming hands-on session (**3 hrs**)
- Distributed Machine Learning, pipelines, external libraries integration (**3 hrs**)
- SparkMLlib, pipelining, H20 hands-on session (**3 hrs**)
- Complete Project workshop session (**3 hrs + 3 hrs**)

Prerequisites

Familiar with Linux command-line; Familiar with Python; Familiar with R

Course Evaluation

Workshop evaluation and written exam

References

[1] **High Performamnce Spark.** Holden Karau, Rachel Warren. O'Reilly Media. June 2017.
[2] **Advanced Analytics with Spark. 2nd Edition.** Sandy Ryza, Uri Laserson,
Sean Owen & Josh Wills. O'Reilly Media. March 2017.
[3] Many articles on Databricks Blog: https://blog.databricks.com

NoSQL (12 hrs – Pauline NERRIERE)

Course Objectives

Understand fundamentals of NoSQL databases capabilities, features and the specific challenges NoSQL databases are addressing, compared to classic SQL databases.
Get some introduction to deploying and using NoSQL databases, such as MongoDB or CouchDB.

Course Description

NoSQL origins (history & players)
Key concepts with databases:
- CAP Theorem
- ACID transactions
- BASE capabilities

SQL / NoSQL high level comparison
NoSQL databases architecture
NoSQL on Hadoop
NoSQL databases overview & comparison (MongoDB, CouchDB, Cassandra, HBase, ElasticSearch…)
Cassandra introduction + lab
MongoDB introduction + lab
ElasticSearch introduction + lab

Prerequisites

Computer systems, architecture and database basic knowledge
SQL language practice

Course Evaluation

Questionnaire or Project

References

Many online resources are available

UE-MSD06-Challenges for Smart Societies – MBD 06.1 - 1st Semester

Energy Transitions: Quantitative Aspects

Lectures and tutorials: 12 hrs

Professor : TBA

Course Objectives

This lecture will provide a quantitative economic perspective on energy transitions. Using micro, macroeconomics and econometric tools, the lecture will present some proven and potential impacts on society of transforming energy in the past, present and future.

Course Description - TBC

Course Evaluation - TBC

References - TBC

UE-MSD06-Challenges for Smart Societies – MBD 06.2 - 1st Semester

Smart Data Project

Supervisors : Several industrial partners

Objectives

The main part of courses focuses on studying several facets of statistics, mathematics and computer sciences, according to the Big Data paradigm. One of the main objectives of this project is to apply all 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.

Description

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

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-Challenges for Smart Societies – MBD 06.3 - 1st Semester

Topics & Case Studies in Data Science

Conferences : 24 hrs

Professors	Romaric GAUDEL (ENSAI)
	Shadi IBRAHIM (INRIA - Rennes)
	Valeriu PETRULIAN
	Thomas ZAMOJSKI (DATASTORM)

Bandit Theory
(6 hrs – Romaric GAUDEL)

Objective

Nowadays, more and more decisions are made by computers. While some of these decisions arise from programs written by humans, some of them are the result of data analysis: the algorithm looks at the past to identify efficient decisions, and repeats these decisions in the present. In such a context, the decisions have a short term impact (they can be good or bad), but they also have a long term impact: the results of these decisions will be used to take future decisions.

During this course, we will focus on a "simple" setting (the Multi-Armed Bandit) and show that to be optimal in the long run, an algorithm as to act in a counter-intuitive way: from time to time, the algorithm as to take a decision which is sub-optimal given the past data (aka. "explore").

The course presents the setting, gives a sketch-proof for the need for exploration and contains a practical session to implement optimal algorithms.

Competences

- Identify settings needing for exploration
- Propose an alternative framework to A-B testing
- Develop simple Bandits algorithms

Prerequisites

- Not required, but recommended: basic knowledge about Machine Learning / Statistical Learning

Evaluation

During pratical session

References

- Sébastien Bubeck, Nicolò Cesa-Bianchi. Regret Analysis of Stochastic and Nonstochastic Multi-armed Bandit Problems. Foundations and Trends in Machine Learning 5, 1-122.

Is Data the New Currency of the Digital Economy?
(6 hrs – Valeriu PETRULIAN)

Objectives

Considered by some to be the "fuel" of our modern industrial systems, and by others as "the new currency" of the Digital Economy; in just few years, Data has become a central topic in our modern societies and economies and its importance is still growing.
This lecture aims to provide a business economics perspective on Data, as a modern trend in today's world. Using several fundamental economic, business and management concepts, the lecture will illustrate how Data irrigates all aspects of our lives and how it interrelates with other present economic and business trends.
Going further down into the exploration of the Digital Economy, the lecture will draw an overview of the behavior and organization of today's companies, thereby allowing prospective Data Specialist job applicants to assess Data from an individual perspective

Description

** Introduction
** Module 1: How did we get here?
** Module 2: The Digital (R)evolution.
** Module 3: Data in various industry settings.

Prerequisites

Items in bold must be read prior to the lecture:

- **Clayton CHRISTENSEN & al. Big Idea: What is disruptive innovation? Harvard Business Review, December 2015.**
- Jean TIROLE. Economie du bien commun (chapitre 14). Paris, PUF, 2016
- **Randy BEAN. How Companies Say They're Using Big Data. Harvard Business Review, April 28, 2017**
- Mary MEEKER. Internet Trends Reports. KPCB 2017.
- World Economic Forum. Digital Transformation of Industries. 2016
- Viktor Mayer-Schönberger, Kenneth Cukier. Big Data: A Revolution That Will Transform How We Live, Work, and Think. Houghton Mifflin, March 2013 (English). Ed. Robert Laffont, Paris 2014 (traduction française)
- Simon Chignard, Louis-David Benayer. Datanomics. FYP Editions, 2015

Internet resources:

- TED, TEDx conferences, Gartner', IDC

Evaluation

During pratical session

Some Recent Advances for Big Data Processing in the Cloud
(6 hrs - Shadi IBRAHIM)

Objectives
During this conference, we will discuss several approaches and methods used to optimise the performance of Hadoop in the Cloud. Finally, we will discuss the limitations of Hadoop and introduce Yarn and other resource management systems for Big data applications including Mesos.

Description
- Approaches to optimize Hadoop in clouds (3 hrs)
- Beyond Hadoop: Yarn, Mesos, etc (3 hrs)

Prerequisites
Attend the course: Big Data processing in Clouds: Hadoop

Evaluation
During the session

References
[1] Apache Hadoop YARN: yet another resource negotiator. Vinod Kumar Vavilapalli, Arun C. Murthy, Chris Douglas, Sharad Agarwal, Mahadev Konar, Robert Evans, Thomas Graves, Jason Lowe, Hitesh Shah, Siddharth Seth, Bikas Saha, Carlo Curino, Owen O'Malley, Sanjay Radia, Benjamin Reed, and Eric Baldeschwieler. In SOCC '13.

[2] Governing energy consumption in hadoop through cpu frequency scaling: An analysis. Shadi Ibrahim, Tien-Dat Phan, Alexandra Carpen-Amarie, Houssem-Eddine Chihoub, Diana Moise, Gabriel Antoniu. In FGCS 2016.

[3] On Understanding the energy impact of speculative execution in Hadoop. Tien-Dat Phan, Shadi Ibrahim, Gabriel Antoniu, Luc Bougé. In GreenCom2015.

[4] Enabling fast failure recovery in shared Hadoop clusters: Towards failure-aware scheduling. Orcun Yildiz, Shadi Ibrahim, Gabriel Antoniu. In FGCS 2016.

[5] Mesos: a platform for fine-grained resource sharing in the data center. Benjamin Hindman, Andy Konwinski, Matei Zaharia, Ali Ghodsi, Anthony D. Joseph, Randy Katz, Scott Shenker, and Ion Stoica. In NSDI'11.

[6] Delay scheduling: a simple technique for achieving locality and fairness in cluster scheduling. Matei Zaharia, Dhruba Borthakur, Joydeep Sen Sarma, Khaled Elmeleegy, Scott Shenker, Ion Stoica. In EuroSys'10.

[7] Improving MapReduce performance in heterogeneous environments. Matei Zaharia, Andy Konwinski, Anthony D. Joseph, Randy Katz, Ion Stoica. In OSDI'08.

Case Studies in Smart Data
(6 hrs – Thomas ZAMOJSKI)
Fault Detection For Gas Pipelines Using Tree Based Models.

Objectives
Get experience in modeling through a real case study of fault detection in a production environment.
Introduction to geolocalised information system (GIS), spatial queries cartographic visualization.
Review and application of tree based modeling.

Description
GRTGaz is the leading natural gas operator in Europe and operates the majority of the French network. One of its main task is to ensure security of gas transportation. In that effect, it performs excavations to go repair damaged pipelines due to corrosion, physical damage and others.
The localisation of faulty sections of the network today relies on a series of measurements and scientific based heuristics. The idea is to use all measurements available to propose a statistical model for the probabilities of finding a fault.
In this course, we will learn some basics of GIS and review tree based models. You will code classification models, validate them and propose a single model that will be tested on independent data.

Prerequisites
Some fluency with the R language and the dplyr or data.table package, or with Python3 and the pandas package.
Some notions of Decision Trees, Random Forests and Gradient Boosting will be helpful.
A computer with a minimum of 8GB of RAM with latest versions of R and/or python (R 3.4.x or Python 3.6.x).

Evaluation
60% Model performance.
30% Code quality and clarity.
10% Participation.

References
Article: Random Forests by BREIMAN Leo, published in Machine Learning, October 2001, Volume 45, Issue 1, pp 5-32
Book: Elements of Statistical Learning by FRIEDMAN J., TIBSHIRANI R., HASTIE T.
Article: Greedy Function Approximation: A Gradient Boosting Machine by FRIEDMAN Jerome H., published in Annals of Statistics, Volume 29, Number 5 (2001), 1189-1232.
WEB REFERENCES
Understanding GIS: http://www.ordnancesurvey.co.uk/support/understanding-gis/
Gradient Boosting: http://blog.kaggle.com/2017/01/23/a-kaggle-master-explains-gradient-boosting/

UE-MSD07 - French as a Foreign Language –MSD07.1 -1st Semester – For foreign students as needed

French: Language & Civilization

Professor : CIREFE (Rennes)

Course Objectives

French Language & Civilization courses allow students to develop and hone their knowledge of the language and culture of the country in which they are studying. These courses are focused on giving the students the linguistic skills they need for their daily life in France and for their integration into the ENSAI student body.

Course Description

Designed for foreign students who are following a full-time academic program in Rennes, these weekly evening courses give students practical written and/or oral French skills, necessary for practical life in France.

Course Evaluation

Quizz and Exams

References

Various textbooks and authentic audiovisual documents will be used.

UE-MSD08 - Internship – MSD08.1

End-of-Studies Internship

4-6 months from March to August

Objectives

This final phase of the MSc in Smart Data 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 statistical 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 an internship report that will be examined by the jury and defended by the student in September.