PROGRAM SYLLABUS

MASTER Data Science, IP Paris
2024-2025

PART 1

An Introduction to Machine Learning Theory

Instructor: Stephan Clémençon, Myrto Limnios

Credits: 3 ECTS

Grading :

Numerus clausus: 60

Language: French

Syllabus: Beaucoup d'applications modernes (génomique, finance, e-marketing, etc.) requièrent de manipuler et traiter des données de très grande dimension. La discipline qui développe et étudie des méthodes concrètes pour modéliser ce type de données s'appelle l'apprentissage statistique («statistical machine-learning»). Il s'agit, in fine, de produire des outils de prédiction et d'aide à la décision dédiés à une application spécifique. L'apparition d'algorithmes très performants pour la classification de données en grande dimension, tels que le boosting ou les Support Vector Machines dans le milieu des années 90, a progressivement transformé le champ occupé jusqu'alors par la statistique traditionnelle qui s'appuyait en grande partie sur le prétraitement réalisé par l'opérateur humain. En s'appuyant sur la théorie popularisée par Vapnik (The Nature of Statistical Learning, 1995), un nouveau courant de recherche est né: il se situe à l'interface entre les communautés mathématique et informatique et mobilise un nombre croissant de jeunes chercheurs tournés vers les applications liées à l'analyse de données massives. Dans ce module, on présentera le domaine, ses fondements théoriques, les problèmes qu'il permet d'aborder (apprentissage supervisé/non supervisé, batch/online, par renforcement, multi-tâche, asynchrone, etc.) et les approches algorithmiques les plus populaires.

«Nothing is more practical than a good theory» - V. Vapnik

Format: 6 sessions of lessons/practical lessons of 3h (1h30+1h30)

L’objectif du cours est de découvrir les enjeux et paradigmes du "machine learning", une discipline en plein essor à l’interface des mathématiques (probabilités/ statistiques, optimisation) et de l’informatique et qui joue aujourd’hui un rôle majeur en matière d’innovation technologique. Il s’agira ici d’en explorer quelques concepts et techniques essentiels, principalement autour du problème fondamental de la "classification supervisée" (i.e. "reconnaissance de formes"). Il se déroulera sur six séances de 3h incluant:

• une partie ‘cours magistral’ lors de laquelle seront formulés les problèmes et décrites certaines solutions de l’état de l’art ;
• une partie ‘travaux dirigés’ pour les séances d’exercices.

Séance 1 - 20/09:

• Introduction générale du cours : repères historiques, enjeux, applications, nomenclature des problèmes

• Le problème de la classification binaire (reconnaissance de formes) : Formalisme – Optimalité
Lectures conseillées: Chapitre 2 de (1), Chapitres 1 et 2 de (9), article (4)

Séance 2 - 27/09

• Théorie probabiliste de la classification - Minimisation empirique du risque
• Théorie de Vapnik-Chervonenkis – Complexité combinatoire - Moyennes de Rademacher
• Exercices Lectures conseillées: articles (3) et (4)

Séance 3 - 04/10

• Premières stratégies d'apprentissage supervisé, modélisation et moyennes locales: régression logistique - perceptron - arbres de classification – K-plus proches voisins - réseaux de neurones
Lectures conseillées: Chapitres 4 et 9 de (1)

Séance 4 - 11/10

• Evaluation de l'erreur et sélection de modèles : plan expérimental – bootstrap – validation croisée – minimisation structurelle du risque • Ensemble Learning: Bagging, Boosting et Forêts Aléatoires
Lectures conseillées: Chapitre 7 de (1)

Séance 5 - 18/10

• Les machines à vecteurs support (SVM) : linéaires/non linéaires
• «Kernel trick»: ACP, régression Lectures conseillées: (8) et (9)

Séance 6 - 25/10
• Au delà des problèmes d'apprentissage 'locaux' (classification, regression, estimation de densité): clustering, ranking, détection d'anomalies

References: Les «slides» du cours seront disponibles en version électronique. On se réfèrera en particulier aux documents suivants.

•Friedman, Hastie & Tibshirani (2009). The Elements of Statistical Learning. Third edition, Springer. Disponible en ligne.
•Bousquet, Boucheron & Lugosi (2004). Introduction to statistical learning theory. In O. Bousquet, U.V. Luxburg, G. Rätsch (editors), Advanced Lectures in Machine Learning, Springer, pp. 169-207, 2004. Disponible en ligne.
•Bousquet, Boucheron & Lugosi (2004). Concentration Inequalities. In Advanced Lectures in Machine Learning, Springer, pp. 208-240. Disponible en ligne.
•Kulkarni, G. Lugosi & S. Venkatesh (1998). Learning Pattern Classification. A Survey. 1948-1998 Special Commemorative Issue of IEEE Transactions on Infor mation Theory, vol.44, 2178-2206. Reprinted in S. Verdú, S.W. McLaughlin (editors.), Information Theory: 50 Years of Discovery, IEEE Press, New York, 1999. Disponible en ligne.

•Cesa-Bianchi & Lugosi (2006) Prediction, Learning, and Games. Cambridge. University Press.
•Devroye, Györfi & Lugosi (1996) A Probabilistic Theory of Pattern Recognition. Springer
•Györfi, Kohler, Krzyzak & Walk (2002) A Distribution-Free Theory of Nonparametric Regression. Springer
•Burgess. A Tutorial on SVM for Pattern Recognition. Kluwer. Disponible en ligne.
•Vapnik. The Statistical Nature of Learning Theory. Springer.

Deep Learning I

Instructors: Geoffroy Peeters, Loïc Le Folgoc (Télécom Paris, IP-Paris)

Credits: 3 ECTS

Grading : 30% labs/project + 70% written exam

Numerus clausus: NC

Language: English

Syllabus: Deep Learning (machine learning based on deep artificial neural networks) has become extremely popular over the last years due to the very good results it allows for tasks such as regression, classification or genera- tion. The objective of this course is to provide a theoretical understanding and a practical usage of the three main types of networks (Multi-Layer- Perceptron, Recurrent-Neural-Network and Convolutional Neural Network). The content of this course ranges from the perceptron to the generation of adversarial images. Each theoretical lecture is followed by a practical lab on the corresponding content where student learn to implement these networks using the currently three popular frameworks: pytorch, tensorflow and keras.

Lectures content:

- Multi-Layer-Perceptron (MLP): Perceptron, Logistic Regression, Chain rule, Back-propagation, Deep Neural Activation functions, Vanishing gradient, Initialization, Regularization (L1, L2, DropOut), Alternative Gradient Descent, Batch- normalization
- Recurrent Neural Network (RNN) : Simple RNN, Forward Propagation, Backward Propagation Through Time, Vanishing/ Exploding gradients, Gated Units (LSTM, GRU), Various architectures, Sequence-to-sequence, Attention model
- Convolutional Neural Network (CNN) : CNNs use sparse connectivity and weight sharing to reduce parameters and create more powerful networks , connections are organized in a convolution operation, CNNs now provide the state- of-the-art in a vast array of problems, we will see how CNNs work and we will implement them for classification problems

Labs content :

- text recognition, sentiment classification
- music generation
- image recognition
- image generation

Programming language

- Python (numpy, scikit-learn, matplotlib)
- DL frameworks: pytorch, tensorflow, keras
- Use Télécom computers, your own labtop or colab.research.google.com → needs a Google account → open one before the first Lab !

- a Graphics Processing Unit (GPU) will not be required, however if you have one, this will speed up the learning process

Kernel Machines: from shallow to deep learning

Instructor: Florence d'Alché

Credits: 3 ECTS

Grading: Practical session reports and oral (project)

Numerus clausus: 60

Language : English

In this course, kernel machines are introduced and deepened as a key tool to understand machine learning through a functional angle. Kernels are not only a means to handle nonlinear problems or complex data with a nonparametric approach, but they also shed light on the behavior of neural networks and support their statistical guarantees. We first introduce kernels, some elements of Reproducing Kernel Hilbert Space theory and the so-called kernel trick. Then we leverage representation theorems and show how in practice a linear method can be kernelized in regression, classification, dimension reduction and statistical testing. Kernel design, learning and approximation approaches are then presented and discussed. Kernel Machines are then extended to learning vector-valued functions, with applications to multitask, functional regression and structured output prediction. We then show how neural networks and kernel machines interfere with each other.

While kernel machines and their approximations are not necessarily shallow, deep neural network can be described as specific kernel machines (neural tangent kernel), or may benefit from kernel interpretation with deep kernel neural networks, transformers as kernel learning.

Agenda:

Session 1: Overview and warm-up. Functional angle to Machine Learning. Kernels and RKHS. Starting to learn within RKHSs (lecture)
Session 2: Learning within RKHSs (follow-up) (lecture / TP)
Session 3: Kernel design, learning and approximation: examples in regression, classification and statistical testing (lecture / TP)
Session 4: Learning vector-valued functions: novel representer theorem + multi-task, functional, structured outputs (lecture)
Session 5: TP (revisiting kernel machines with kernel approximation, structured output prediction)
Session 6 : Deep neural networks as kernel machines and vice-versa: neural tangent kernel, hybrid neural networks, double descent (lecture / TP)

Bibliography.

*Mohri, Mehryar, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2018.

*Belkin, M., Ma, S., & Mandal, S. (2018, July). To understand deep learning we need to understand kernel learning. In *International Conference on Machine Learning* (pp. 541-549). PMLR.

*Schölkopf, B., & Smola, A. J. (2002). *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press.

Monte Carlo methods and approximate inference

Instructor: Randal DOUC et Sholom SCHECHTMAN

Credits: 3 ECTS

Grading : Final exam + 3 QCMs during courses

Numerus clausus: 30

Language : English

Syllabus :
In this course, we will explore techniques for generating new observations from a given target distribution. Unlike generative models, where the target distribution is unknown and only a sample of observations is available, this course focuses on scenarios where the target distribution is specified. This context is particularly relevant in Bayesian statistics, which finds applications in various fields such as biostatistics, economics and finance, and environmental data analysis due to its capability to incorporate prior information and update beliefs with new data.

Prerequisite: Basic knowledge in probability and statistics.

Topics Covered:

We will concentrate on two main techniques:

1. **Variational Inference**: This technique, which underpins Variational Autoencoders, provides a framework for approximating complex distributions within a family of possible distributions .

2. **Markov Chain Monte Carlo (MCMC)**: A suite of algorithms that generate samples from a probability distribution based on constructing a Markov chain with the right asymptotic distribution.

Course Structure:

- **Introduction to Standard Procedures**: We will begin by introducing the fundamental methods of variational inference and MCMC.
- **Convergence Properties**: The course will delve into the convergence properties of these algorithms, ensuring a solid understanding of their theoretical foundations.
- **Practical Extensions**: We will extend our study to more complex scenarios and provide practical implementations, utilizing Python Notebooks for illustrating some algorithms.

The approach taken in this course will ensure that students not only understand the theoretical aspects but also are able to implement and adapt these techniques in diverse situations.

Practical Introduction to Machine Learning

Instructor: Rémi Flamary, Ekhine Irurozki

Credits: 3 ECTS

Grading : Practical session reports and oral

Numerus clausus: 60

Language : English

Syllabus : The objective of this course is to provide a practical introduction to the field of machine learning. We will discuss the different machine learning problems from unsupervised (dimensionality reduction, clustering and density estimation) to supervised (classification, regression, ranking). In this course we will introduce for each method the problem, provide its modeling as an optimization problem and discuss the algorithms that are used to solve the problem. The practical aspect of each method will also be discussed along with python code and existing implementations.

The course will be completed by practical sessions that will allow the students to implement the methods seen in the course on practical problems such as image classification and time series prediction (biomedical and climate data). The objective of the practical session will be not only to learn to use the methods but also to interpret their models and results with respect to the data and the theoretical models.

Course overview:

- Introduction
  - Machine learning problems
  - Knowing your data
  - Preprocessing
- Unsupervised learning
  - Dimensionality reduction and
  - Dictionary learning and collaborative filtering
  - Clustering and generative modeling
  - Generative modeling
- Supervised learning
  - Linear models and kernel methods for regression and classification

- Nearest neighbors and bayesian decision
- Trees and ensemble methods
- ML in practice
  - Find your problem
  - Model selection

PART 1-2

Advanced AI Methods for Graphs and NLP (ALTEGRAD)

Instructor: Michalis Vazirgiannis
Credits: 6 ECTS
Grading: 20% from lab assignments and 80% from a data challenge
Numerus clausus: NC
Language: English
Syllabus:
1. TEXT/NLP - Graph based Text Mining

- Graph-of-words GoWvis
- Keyword extraction (TFIDF, TextRank, ECIR'15, EMNLP'16)
- extractive summarization
- Sub-event detection in twitter streams
- graph based document classification: TW-IDF, TW-ICW, subgraphs abstractive summarization - ACL 2018 summarization

2. TEXT - NLP - Word & doc embeddings

- Word embeddings: word2vec-glove models, doc2vec, subword, Latent Semantic Indexing, context based embeddings
- doc similarity metrics: Word Mover's distance, shortest path kernels

3. Deep learning for NLP

- CNNs, RNNs LSTMs for NLP, text classification
- Meta-architectures
- Sequence to Sequence: Attention (HAN),
    - Domains: summarization
    - Translation, image captioning
- Unsupervised word sense detection/disambiguation
- French Linguistic resources: http://master2-bigdata.polytechnique.fr/FrenchLinguisticResources/

4. Graph kernels, community detection

Grakel python library: **https://github.com/ysig/GraKeL/**

5. Deep Learning for Graphs - node classification

- node embeddings (deepwalk & node2vec) for node classification and link prediction
- Supervised node embeddings (GCNN, ...)

6. Deep Learning for Graphs - Graph classification, GNNs

- graph CNNs
- message passing
- Graph - Auto-encoders

7. Sets embeddings - point clouds

8. Network Architecture Search - interpretability.

Course page: https://moodle.lix.polytechnique.fr/moodle/

Inscription to the course necessary:
https://docs.google.com/forms/d/11ZcoKzVenCEcVtgbxhuJ05CGBb2JrBWbPUoqQ7nOMHc/edit

Informative video for ALTEGRAD 2021:
https://www.dropbox.com/s/lcans4r4mlsryux/ALTEGRAD_VAZIRGIANNIS_2021-09-06.mov?dl=0

Prerequisites : Good level in Machine and Deep Learning, basic graph algorithms, python programming

Big Data Framework

Instructor : Duc Pham-Hi

Credits : 6 ECTS

Grading : The final mark of the module is a weighted average of 3 marks :

- Machine Learning (weight 2)
- Big Data with Hadoop (weight 5)
- Real Time Data and Analytics with Elastic Search (weight 3)

The 1st course is an ECE-exclusive online course with a final exam taken in situ at ECE. The 2nd course is evaluated by an exam (1 hour) and a continuous evaluation (labs and mini-projects). The 3rd course has a final exam (or a project, depending on students' orientations).

Numerus clausus : 30 students in Hadoop and ELK sessions, selected after the 1st exam.

Language : English.

Syllabus : The objectives of this course are the following :

- Get fundamentals in Machine learning techniques
- Discover the different components of a Big Data cluster and how they interact.
- Understand Big Data paradigms.
- Understand the benefits of open source solutions.
- Develop a Big Data project from scratch.
- Understand and implement distributed algorithms.
- Understand the advantages of SQL/NOSQL databases.

Description of the course:

The module Big Data Frameworks is composed of three courses :

- Machine Learning (approx. 6 hours of online self study with a final Test taken at ECE)
- Big Data with Hadoop (5 x 4 hours)
- Real Time Big Data Search and Analytics with Elastic Search (2 x 4 hours)

I. Machine Learning :

1. Introduction to Machine Learning
2. Octave
3. Dimensionality Reduction
4. Linear regression
5. Classification
6. Clustering

II. Big Data with Hadoop:

Apache Hadoop has been evolving as the Big Data platform on top of which multiple building blocks are being developed. This course presents the Hadoop ecosystem, Hadoop Distributed File System (HDFS) as well as many of the tools developed on it:

- MapReduce and YARN
- Hive and HBase
- Kafka, Flume, NiFi, Flink, Oozie, etc.

Students will also discover various subjects such as security, resource allocation and data governance in Hadoop.

III. Real time Elastic search and Analytics :

Powerful tool to analyze in real time tons of data available to provide insights to business.

Use cases (examples from last year, actual may differ) :

• Search (Wikipedia, Github), Logging Analytics (Blizzard), Security Analytics (Slack), Metrics

Analytics (NASA), Business Analytics (Tinder)

• High availability and Speed

• Elastic Stack:

• Collect (Beat)

• Ingest (Logstash)

• Store, search and analytics engine (Elasticsearch)

• Visualize (Kibana)

• The secret sauce of ES, Index, Shards (Primary, Replicas), Nodes, Cluster, Cross-Cluster

• Typical deployment in key use cases (search, logging, security)

Prerequisites :

Java, Python, Machine Learning and basic knowledge in Linux system administration and SQL

Convex Analysis and Optimization Theory

Instructor: Pascal Bianchi, Olivier Fercoq and Walid Hachem

Credits: 6 ETCS

Grading: Final exam

Numerus clausus: none

Language: English

Syllabus: The course is about taking a step back in order to understand the mathematical foundations for the construction of a large class of iterative methods. The first part of the course is about convex analysis. We shall review the properties of convex functions, Fenchel-Legendre transform, and introduce the duality theory in convex optimization. The second part of the course is about numerical algorithms. We shall see the conditions under which we can demonstrate the convergence of fixed point algorithms. This general approach makes it possible to obtain, as a corollary, the convergence of several emblematic algorithms such as the proximal gradient algorithm, or primal-dual algorithms. These algorithms are frequently used to solve optimization problems involving complex and structured regularizations, or optimization problems under constraints. They are frequently encountered in statistical learning, signal processing, and image processing. The objectives of the course are:

- Master the mathematical tools for the construction of optimization algorithms.
- Know how to demonstrate the convergence of iterates.
- Know how to numerically solve optimization problems involving non-differentiable terms and/or structured regularization.

Data Stream Processing

Instructor: Jérémie Sublime – Maurras Togbe – Mariam Barry

Credits: 6 ECTS

Grading:
- The practical sessions will make ⅔ of the mark
- The research paper presentation will make ⅓ of the mark

Numerus clausus: 70

Language: English

Syllabus: This course deals with the algorithms and softwares commonly used to process large data streams. It aims at understanding the main difficulties and specificities of this type of data, knowing what different types of streams exist, what are the theoretical models and practical algorithms to analyze them, and what are the right tools to process these streams.

After an introduction of what data streams are from a conceptual point of view, this class covers the question of data stream processing from two different angles:

1. A Machine Learning and Data Mining approach to cover the theoretical and algorithmic difficulties of learning from data streams: online learning vs incremental and batch learning, and sampling techniques.
2. A more practical approach with an introduction to the various systems and software that are used to handle these data.

In terms of organization, the course will consist of an alternance of lectures and practical sessions. Finally, during the last class the students will have to present a recent research article of their choice on the subject of data stream processing.

Objectives :
- Introduction to the concept of data stream processing
- Learning the basics on and how to use Data Stream Management Systems (DSMS)
- Understanding the main sampling techniques used for stream processing : sampling, sketching, etc.
- Understanding and using the main data stream processing algorithms

Prerequisites :
- Basics in SQL language
- Basics in Machine Learning (supervised and unsupervised)
- A knowledge of Java programming is recommended but not mandatory

Generalization Properties of Algorithms in ML

or Statistical Machine Learning and Convex Optimization 2023

Instructors: **Aymeric Dieuleveut and Hadrien Hendrikx**

Credits: 5ECTS (Orsay) - 6 ECTS (M2DS).

Numerus clausus: 40

Language: English.

Syllabus: The majority of learning problems are formulated as optimization problems. optimization problems, based on the observation of a data sample (training set). Optimization of an objective defined on the basis of this sample makes it possible to propose an estimator that has good performance on the training set. However, we are generally interested in the estimator's estimator, i.e. its performance on a new observation. observation. With the emergence of large quantities of data since the 2000s, the link between the algorithm used and the generalizability of the associated estimator has become a major issue. Today, the question of generalization is still a major research issue, both in its theoretical and practical aspects. both theoretical and practical aspects. In this course, we focus on all the theoretical and heuristic results both theoretical and heuristic, to tackle this problem. More specifically, we will firstly, we will study the various approaches that enable us to obtain theoretical guarantees theoretical guarantees for the generalization of algorithms, in particular approaches related to complexity, stability and early stopping methods (early stopping, stochastic approximation). In a second second part, we'll look at heuristic approaches and the differences (explained or observed) in the context of deep learning (non-convex and over-parametrized
Grading: Quizz + Présentation lecture/ projet.

Prerequisites : basic knowledge of convex optimisation and statistics. Having taken the Optimisation for Data Science course will give you a better understanding of the various algorithms involved.

Dates: Monday, afternoon (13/11, 20/11, 27/11, 04/12, 11/12, 15/01, 22/01, 29/01, 06/02, 13/02)
In short, this lecture aims at providing an overview of first order optimization with machine learning applications in mind. The course is composed of ten lectures.
1. Lecture 1: introduction to SML. Risk relaxation, Rademacher complexity.
Take away:
Understand the tradeoffs between Optimization and Statistics for ML algorithms.
2. Part I: Deterministic first order optimization
(a) Lecture 2: Convex Optimization (deterministic). (GD)
(b) Lecture 3&4: Acceleration, proximal, mirror descent.
Page 1
Syllabus - GPOMLA - SMLCO
Take away:

Diving into first order optimization methods, from high level insights to theoretical guarantees and advanced algorithms.
*Advanced - Performance estimation - a meta optimization point of view on optimization.
3. Part II: Stochastic optimization and generalization of ML algorithms
(a) Lecture 5: Stochastic Gradient descent 1. Non asymptotic convergence rates. Application to Machine learning: ERM of Risk minimization with Early stopping.
(b) Lecture 6: Stochastic Gradient descent 2. Asymptotic convergence rates, Stochastic approximation etc.
(c) Lecture 7: Case of Least-squares regression: Regularization (Tichonov, Early stopping), overparametrized setting, etc.
(d) Lecture 8: Stability (Bousquet-Elysseff)
Take away:
Understand stochastic optimization and SGD properties
Rethink generalization techniques and their links
*Insights on over-parametrized and Deep Learning regimes
4. Part IV - Adavanced topics
(a) Lecture 9. Performance estimation. A meta optimization point of view on worst case convergence guarantees.
(b) Lecture 10: Lower bound on stochastic convex optimization.

Lecture Evaluation
1. Quizz on moodle. (1/4 of the final mark)
Every Quizz is expected to take ∼1h. Quizz are meant to rely on the previous lectures' content and discover the next lectures.
2. Mid-term (1/4 of the final mark): 1 hour on optimization on 11/12. Lecture starts at 13h30 - 17h30 on 11/12.
3. Final Exam (1/2 of the final mark) - no documents. Exercises close to the ones given in the Exercises sheets will constitute about 2/3 of the exam.
4. No Homework but exercises sheets.
Remark: attending the lecture is mandatory. If you miss 2 lectures or more, your grade will be at best C (13/20), and you can't pass the exam if you have missed 4 lectures or more.

References :
- Rademacher and Gaussian Complexities: Risk Bounds and Structural Results, P. Bartlett, S. Mendelson
- The Tradeoffs of Large Scale Learning, L. Bottou, O. Bousquet - Stability and Generalization, O. Bousquet, A. Elisseef
- Train faster, generalize better: Stability of stochastic gradient descent, M. Hardt, B. Recht, Y. Singer
- Non-strongly-convex smooth stochastic approximation with convergence rate O(1/n), F. Bach, E. Moulines
- Understanding deep learning requires rethinking generalization, C. Zhang, S. Bengio, M. Hardt, B. Recht, O. Vinyals
- On early stopping in gradient descent learning, Y Yao, L. Rosasco, and A. Caponnetto
- Generalization properties of multiple passes stochastic gradient method, S. Villa
- Competing with the empirical risk minimizer in a single pass, R. Frostig, R. Ge, S. M. Kakade, A. Sidford

- Deep Learning and Generalization, O. Bousquet

Hidden Markov Models and Sequential Monte Carlo Methods

Instructor: **Chopin Nicolas**

Credits: 3 ECTS

Grading : Project-based (group of three students)

Numerus clausus: NC

Language: English

Syllabus: So-called hidden Markov chain (or state-space) models are time series models involving a "signal" (a Markov process (X_t) describing the state of a system) observed in an imperfect and noisy way in the form of data, e.g. $Y_t=f(X_t)+\epsilon_t$. These models are widely used in many disciplines:

- Finance: stochastic volatility (X_t is the unobserved volatility)...
- Engineering: target tracking (X_t is the position of a mobile whose trajectory we are trying to find; speech recognition (X_t is a phoneme).
- Biostatistics: Ecology (X_t=population size); Epidemiology (X_t=number of infected).

The aim of this course is to present modern methods of sequential analysis of such models, based on particle algorithms (Sequential Monte Carlo). The problems of filtering, smoothing, prediction, and parameter estimation will be discussed. At the end of the course, we will also briefly discuss the extension of such algorithms to non-sequential problems, notably in Bayesian Statistics.

At the end of the course, the student will be able to:
- state the main properties of HMM models
- implement a particle filter to filter and smooth a given HMM model
- estimate the parameters of such a model from different methods

Planning:

1. Introduction: definition of HMM (Hidden Markov models), main properties, notion of filtering, smoothing and prediction, forward-backward formulas.
2. Discrete HMMs, Baum-Petrie's algorithm
3. Gaussian linear HMM, Kalman algorithm
4. SMC algorithms for filtering an HMM model
5. Estimation in HMM models
6. Introduction to non-sequential applications of SMC algorithms

Prerequisites:
- 2A Simulation and Monte Carlo or similar course
- 3A courses of "Computational Statistics" and "Bayesian Statistics" are recommended but not mandatory.

References

Del Moral (2004). Feynman-Kac formulae, Springer; Chopin, N. and Papaspiliopoulos, O. (2020). An Introduction to Sequential Monte Carlo, Springer.

High Dimensional Statistics

Instructor: Evgenii Chzhen

Credits: 3 ECTS

Grading: Final exam

Numerus clausus: 15

Language: English

Syllabus: This course develops tools to analyze statistical problems in high-dimensional settings where the number of variables may be greater than the sample size. It is in contrast with the classical statistical theory that focuses on the behavior of estimators in the asymptotics as the sample increases while the number
of variables stays fixed. We will show that, in high-dimensional problems, powerful statistical methods can be constructed under such properties as sparsity or low-rankness. The emphasis will be on the non-asymptotic theory underlying these developments.

Topics covered:

- Sparsity and thresholding in the Gaussian sequence model.
- High-dimensional linear regression: Lasso, BIC, Dantzig selector, Square
- Root Lasso. Oracle inequalities and variable selection properties.
- Estimation of high-dimensional low rank matrices. Matrix completion.
- Inhomogeneous random graph model. Community detection and estimation in the stochastic block model.

Prerequisites: Solid knowledge of probability theory, mathematical statis-
tics, linear algebra. Notions of convex optimization.

Resources:
Alexandre Tsybakov. High-dimensional Statistics. Lecture Notes.

Master Class and Hi!ckathon

Instructor: Michael Jordan / Jérémy Harroch / Clément Farabet

Credits: 3 ECTS

Grading:

Numerus clausus : NC

Language: English

Syllabus:

Michael Jordan :

Jérémy Harroch :

Clément Farabet :

Hi!ckathon : du 18/11/2024 au 5/12/2024

Machine Learning for Climate and Energy

Instructor: Alexis Tantet, Victor Pellet

Credits: 3 ECTS

Grading:

Numerus clausus : NC

Language: English

Syllabus:

Natural Language Processing and Sentiment Analysis

Instructors: **Maria Boritchev, Chloé Clavel, Matthieu Labeau, Fabian Suchanek**

Credits: 6 ECTS

Grading: Practical session reports (60%) and Exam (40%).

Numerus clausus: 60

Language: English

Description: Natural Language Processing (NLP) is a very important field of research, which we can find applications of everywhere. During the last decade, deep Learning has profoundly transformed the way we do NLP, bringing huge performance gains in many traditional tasks involving natural language thanks to end-to-end neural models - relying mainly on the availability of textual data. The goal of this course is to provide students with (1) a theoretical understanding of modern NLP models through a review of the most important milestones in recent NLP (2) a practical understanding of the difficulties of performing machine learning tasks on natural language, and how to apply NLP techniques in novel circumstances.

Schedule:

- Lecture 1 - 13/09 - Introduction to Text Processing and Symbolic Text Representations
- Lab 1 – 20/09 - Introductory Lab on text processing
- Lecture 2 – 27/09 - Introduction to Language Modeling and Text Generation
- Lab 2 – 04/10 - Lab on Language Modeling
- Lecture 3 - 11 Word Embeddings, Algorithms and Applications
- Lab 3 – 18/10 - Topic modeling and classification
- Lecture 4 – 08/11 - Sequence models, Encoders and Decoders, Contextual Representations and Transfer Learning for Natural Language Processing Tasks
- Lab 4 – 15/11 - Machine Translation with Seq2seq models
- Lecture 5 - 22/11 - Structured Prediction in Natural Language Processing
- Lecture 6 – 29/11 - Large Language Models and Societal Impact
- Lecture 7 – 06/12 - Sentiment Analysis
- Lecture 8 - 13/12 - Conversational Systems
- Exam – 20/12
- **References**: Slides will be available. For more details, we will refer for lectures to relevant chapters of the following books (available online):

- *Natural Language Processing*, Jacob Eisenstein
- *Speech and Language Processing*, Dan Jurafsky and James H Martin
- *A primer on neural networks for NLP*, Yoav Goldberg

Non Parametric Estimation and Testing

Instructor: Vincent Divol

Credits: 3 ECTS

Grading: Final exam

Numerus clausus: NC

Language: English

Syllabus: Nonparametric statistics form a set of statistical methods for which no specific assumptions on the shape of the underlying distribution of the data (such as Gaussian) are made. Instead, model assumptions are typically phrased in terms of the regularity of a density or regression function. The latter are then approximated by flexible family of functions (based e.g. on wavelet decompositions or neural networks).

In this course, we will introduce key nonparametric models and methods which we will theoretically analyze, with a focus on nonasymptotic guarantees and rates of convergence.

Topics covered:

- estimation of the CDF
- density estimation
- nonparametric regression
- penalization
- minimax theory and adaptive estimation
- robustness

L. Devroye: A Course in Density Estimation. Birkhauser, Boston, 1987.

E. Giné, R. Nickl: Mathematical Foundations of Infinite-Dimensional Statistical Models, Cambridge University Press, 2015

A.Nemirovski: Topics in non-parametric statistics. Ecole d'Eté de Probabilités de Saint-Flour XXVIII – 1998. Lecture Notes in Mathematics, v.1738. Springer, 2000.

A.B. Tsybakov: Introduction to Nonparametric Estimation. Springer, New York, 2009.

L. Wasserman: All of Nonparametric Statistics. Springer, New York, 2006.

Optimization for Data Science

Instructors: **Alexandre Gramfort, Rémi Flamary**
Credits: 6 ECTS

Grading:
Labs. 2-3 Labs with Jupyter graded (30% of the final grade).
Project. Implementation of solvers for a machine learning model. 30% of final grade.
Exam. 3h Exam (40% of the final grade).

Numerus clausus: NC
Language: English

Syllabus : Modern machine learning heavily relies on optimization tools, typically to minimize the so called loss functions on training sets. The objective of this course is to cover the necessary theoretical results of convex optimization as well as the computational aspects. This course contains a fair amount of programming as all algorithms presented will be implemented and tested on real data. At the end of the course, students shall be able to decide what algorithm is the most adapted to the machine learning problem given the size of the data (number of samples, sparsity, dimension of each observation).

1. Introduction to optimization for data science
    - ML optimization problems and linear algebra recap
    - Optimization problems and their properties (Convexity, smoothness)
2. Smooth optimization : Gradient descent
    - First order algorithms, convergence for smooth and strongly convex functions
3. Smooth Optimization : Quadratic problems
    - Solvers for quadratic problems, conjugate gradient
    - Linesearch methods
4. Non-smooth Optimization : Proximal methods
    - Proximal operator and proximal algorithms
    - Lab 1: Lasso and group Lasso
5. Stochastic Gradient Descent
    - SGD and variance reduction techniques
    - Lab 2: SGD for Logistic regression
6. Standard formulation of constrained optimization problems
    - LP, QP and Mixed Integer Programming
7. Coordinate descent
    - Algorithms and Labs
8. Newton and quasi-newton methods
    - Second order methods and Labs
1. 9. Beyond convex optimization
    - Nonconvex reg., Frank-Wolfe, DC programming, autodiff

Planning:

12 weeks course+ labs and 13th week ewercises to prepare for exams.

References:

Book 1. Boyd & Vandenberghe: Convex Optimization. Chapters 2, 3 and 4 for a revision on convexity and chapter 9 for a revision on unconstrained optimization. Freely available here.
Book 2. Shalev-Shwartz & Ben-David: Understanding Machine Learning, from Theory to Algorithms. Chapters 1 and 2 for a frequentist introduction to Machine Learning. Freely available here.
Book 3. Bubeck: Convex Optimization: Algorithms and Complexity. Chapter 6 for additional proofs for stochastic gradient methods including SGD and SVRG. Freely available here.
Paper 1. Amir Beck and Marc Teboulle (2009), SIAM J. Imaging Sciences, A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems. Freely available here.
Paper 2. RMG et all (2019), Proceedings of Machine Learning Research, Volume 97, SGD: general analysis and improved rates freely available here.

Statistical Learning Theory

Instructor: Jaouad Mourtada (ENSAE)

Credits: 3 ECTS

Grading: Final exam. An exercise sheet (with exercises in the spirit of the exam) will be provided, and will be partly corrected in class.

Numerus Clausus: 50

Language: English

Syllabus: Supervised statistical learning refers to the following prediction problem: find a good rule to predict some output/label of interest (for instance, the object represented by an image), based on some associated input/features (for instance, the image itself), using a dataset of feature-label pairs. This is a core machine learning problem, with applications to computer vision, speech recognition and natural language processing, among many domains.
The aim of this course is to provide an introduction to the theory and principles underlying this general problem. The emphasis will be placed on mathematical and conceptual aspects, rather than on actual implementation: the key concepts and ideas will be illustrated through formal (often elementary) results, which will be proven in class.
Along the way, we will introduce some technical tools relevant in statistics and machine learning theory, including basic concentration inequalities and control of empirical processes, stochastic gradient methods, and approximation properties of some classes of functions.

Topics covered:

- Probabilistic setting: prediction, loss and risk;
- Universal consistency and No Free Lunch theorem;
- Empirical risk minimization and its analysis via uniform convergence. Approximation and estimation errors, overfitting and underfitting;
- Finite classes and Hoeffding's inequality; Rademacher complexity and Vapnik-Chervonenkis classes (upper and lower bounds);
- Model selection: complexity regularization and cross-validation;
- Nonparametric regression by histograms, bias-variance decomposition and convergence rates over Hölder classes; curse of dimension;
- Convex surrogate losses and their properties; Linear prediction, stochastic gradient methods; Links between regularization and optimization;
- Learning in high dimension and neural networks: approximation of 2-layer networks, interpolation and overfitting for linear regression.

Prerequisites: Knowledge of probability theory at an undergraduate level. Some notions in statistics (at the level of a first course in mathematical statistics) would be helpful, though this is not a strict requirement. Some general background in machine learning would also be a plus, though none is required to follow this course.

References:

Lecture notes will be provided before each class.

-- S. Boucheron, O. Bousquet, and G. Lugosi. Theory of classification: A survey of some recent advances. ESAIM: probability and statistics, 9:323–375, 2005. Link
-- L. Devroye, L. Györfi, and G. Lugosi. A Probabilistic Theory of Pattern Recognition, volume 31 of Applications of Mathematics. Springer-Verlag, 1996. Link
-- A. B. Tsybakov. Introduction to nonparametric estimation. Springer, 2009.
-- L. Györfi, M. Kohler, A. Krzyzak, and Harro Walk. A distribution-free theory of nonparametric regression. Springer Science & Business Media, 2002. Link
-- S. Shalev-Shwartz and S. Ben-David. Understanding machine learning: From theory to algorithms. Cambridge University Press, 2014. Link

PART 2

Data Camp

Instructor: **Thomas Moreau** **et** **Pedro Rodrigues**

Credits: 3 ECTS

Grading: Assignments, data challenge and project

Numerus clausus: NC

Language: English

Syllabus: a 5-day intensive course on data science in practice.

Description:

The aim of this course is to learn data science by doing. All aspects of completing a data science pipeline will be covered, from exploratory data analysis (EDA), feature engineering, parameter optimization to advanced learning algorithms. To put this in practice, you will **compete on a set of data challenges**, and you will also need to p**ropose your own challenge**! On GitHub you have some of the teaching materials. You **must** have a GitHub account to complete the course. Grade is a mix of your performance on the data challenge offered to the class as well as the challenge you will setup.

Each day you will have 50% of lectures and 50% of work on the competitive challenge using the RAMP website.

Day 1: Data wrangling

- Introduction to the workflow (VSCode, git, github, tests, ...)
- Advanced course on Pandas
- *Assignments:* numpy and pandas assignments using github.

Day 2: ML Pipelines and model evaluation

- Advanced scikit-learn: Column transformer and pipelines
- Parallel processing with joblib
- Generalization and Cross Validation
- *Assignment:* create your own scikit-learn estimator and cross-validation splitter.
- *Getting started on RAMP and introduction to the challenges.*

Day 3: Metrics and dealing with unbalanced data

- Presentation of the different ML metrics
- Problem of the metric with imbalanced data
- ML approaches to deal with imbalanced data
- *Working on data challenges*

<u>**Day 4: Feature engineering and model inspection**</u>

- Feature engineering and advanced encoding of categorical features
- Model inspection: Partial dependence plots, Feature importance
- *Working on data challenges*

<u>**Day 5: Ensemble methods and hyperparameter optimization**</u>

- From trees to gradient boosting
- Profiling with snakeviz
- Hyperparameter optimization
- *Working on data challenge*

Deep Learning for Computer Vision

Instructor: Stéphane Lathuilière, Jhony Giraldo

Credits: 3 ECTS

Grading:

Numerus clausus: NC

Language: English

Prerequisites: Basic knowledge of deep learning is required.

Course Description:
This course explores advanced topics in computer vision, focusing on deep learning techniques and recent developments in the field. Through a combination of lectures and practical sessions, students will develop a comprehensive understanding of key research challenges in computer vision and the corresponding methodologies.

Course Outline:
Week 1-2: Introduction to the main vision tasks
- Object detection and segmentation
- Depth estimation
- Human pose estimation
- Action recognition
- Practical sessions

Week 3-4: Self-Supervised Learning, Few-Shot Learning and Domain Adaptation
- Introduction to self-supervised learning
- Contrastive learning
- Few-shot learning: Problem formulation and techniques
- Domain adaptation: Theory and applications
- Recent advances in few-shot learning
- Practical session

Week 5: Recent Advances in Image and Video Generation
- Recap on Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs) and diffusion models
- Conditional image synthesis
- Text-to-image generation
- Video generation and editing

Week 6-7: Deep Architectures for Point Cloud Perception
- Introduction to point cloud data
- PointNet and PointNet++
- Point cloud segmentation

- Object recognition in point clouds
- Practical session
-

Please note that the specific topics and content may be adjusted based on the needs and preferences of the students and the availability of resources.

Operator Learning, Applications in Dynamical Systems and Uncertainty Quantification

Instructor: **Karim Lounici**

Credits: 3 ECTS

Grading: written final exam and homework

Numerus Clausus: 30

Language: English

Syllabus : Recent progress in the analysis of nonlinear dynamical systems was made possible by combining machine/deep learning techniques with the powerful Koopman framework for the estimation of transfer operators. This new approach was demonstrated to produce accurate forecasting of future states but also to provide pertinent interpretations on the dynamical system. In molecular dynamic applications, this new approach allows us to detect meta-stable states of a molecule and accurately simulate new trajectories at a reduced computational cost as compared to traditional computationally intensive methods. In this course, we will present a novel general framework for learning operators applicable to molecular dynamic to learn the transfer operator but more generally to any problem requiring conditional distribution learning. This approach combined with deep learning allows us to accurately learn complex nonlinear models and perform uncertainty quantification in a streamline manner. To this end, we will review several mathematical tools useful to develop statistical analysis methods and study their performances. Such tools include concentration inequalities, convex optimization, perturbation theory, minimax theory, some elements of kernel learning and deep learning.

Topics covered:

1. Principal Component Analysis and Canonical Correlation Analysis
2. RKHS framework for Operator Learning
3. Learning the transfer operator of a dynamical system: forecasting future state distribution and system interpretation
4. Neural conditional distribution learning: applications in uncertainty quantification

Suggested courses in first period:

- High Dimensional Statistics
- Kernel machine from shallow to deep learning

Introduction to Generative models

Instructor: Alain Durmus et Yazid Janati

Credits: 3 ECTS

Grading : Project-based on a research article

Numerus clausus: NC

Language : English

Syllabus :

This course introduces students to core concepts and recent developments in modern generative methods. In simple terms, generative modeling consists of learning a map capable of generating new data instances that resemble a given set of observations, starting from a simple prior distribution, most often a standard Gaussian distribution. This course aims to provide a mathematical and methodological introduction to generative models. These models have gained prominence for their ability to generate realistic data across diverse domains, making them a popular tool for researchers and practitioners in machine learning. Participants will learn about the methodological and theoretical foundations, as well as some practical applications associated with these models.

The first two lectures motivate the use of generative models, introduce their formalism, and present two simple yet relevant examples: models trained by minimum distance estimation and energy-based models. In the third and fourth lectures, we introduce normalizing flows, Variational Auto-encoders, and Generative Adversarial Networks.

Finally, we present score-based diffusion models and explain how they provide an algorithmic framework for the basic idea that sampling from the time-reversal of a diffusion process converts noise into new data instances. We shall do so following two different approaches: a first elementary one that only relies on discrete transition probabilities, and a second one based on stochastic calculus. After this introduction, we derive sharp theoretical guarantees of convergence for score-based diffusion models by assembling ideas from stochastic control, functional inequalities, and regularity theory for Hamilton-Jacobi-Bellman equations. If time permits, the course ends with an overview of some of the most recent and sophisticated algorithms such as flow matching and diffusion Schrödinger bridges (DSB), which bring an (entropic) optimal transport insight into generative modeling.

Lecture 1 : Introduction to generative models, minimum distance estimation and distance/divergence between distributions

Lecture 2 : Energy-based models

Lecture 3 : Normalizing flows and VAE Part I

Lecture 4 : VAE Part II and GAN

Lecture 5 : Diffusion models from a varitional perspective

Lecture 6 : Diffusion models from a the theory of reciprocal processes

Introduction to Operations Research

Instructor: Eric Soutil

Credits: 3 ECTS

Grading: final exam

Numerus clausus: NC

Language: French

Syllabus: La Recherche Opérationnelle (R.O.) est la discipline des méthodes scientifiques utilisables pour élaborer de meilleures décisions. Elle permet de rationaliser, de simuler et d'optimiser l'architecture et le fonctionnement des systèmes de production ou d'organisation. La R.O. apparaît comme une discipline-carrefour associant les mathématiques, l'économie et l'informatique.
Les apports de la R.O. sont visibles dans les domaines les plus divers : de l'organisation des lignes de production de véhicules à la planification des missions spatiales, de l'optimisation de portefeuilles bancaires à l'aide au séquençage de l'ADN ou à l'organisation de la couverture satellite des téléphones portables…
Tous ces problèmes sont de nature discrète ou combinatoire. Si l'existence d'une solution optimale est en général triviale, sa recherche de manière énumérative, même effectuée par les ordinateurs les plus puissants, pourrait demander plusieurs siècles de calcul.

Le but du cours est de présenter des problématiques d'optimisation combinatoire ainsi que des algorithmes efficaces pour leur résolution. Nous verrons ainsi des poblématiques d'optimisation dans les graphes (arbres couvrant de poids maximal, plus court chemin, voyageur de commerce et flot maximal) ainsi que la programmation linéaire continue et en nombres entiers. Des TP seront dédiés à l'implémentation de certains des algorithmes considérés.

Prequisites: None

Law and Ethics of Artificial Intelligence

Instructor: Winston Maxwell and Tiphaine Viard

Credits: 3 ECTS

Grading: 50% quiz, 50% final paper and presentation, with possible adjustment of final grade, up or down one point out of 20, based on class participation

Numerus clausus: NC

Language: English

Students will learn:

- why different families of AI (symbolic, machine learning, foundation models) present different risk profiles and require different approaches to regulation; discussion of systemic risks to democracy and national security

- AI ethics principles: what do they mean and do they serve a purpose?

- how risks of bias and discrimination manifest themselves in AI systems, and approaches to mitigate those risks; why achieving perfect fairness is impossible; introduction to human coginitive biases and why computers may be less (or more) discriminatory than humans.

- what purposes are served by AI explainability, why explainability prevents machine learning from being used in certain critical applications, and the main regulatory requirements for explainability

- what does "meaningful human control" of AI systems mean, and why is it important?

- what are the key elements of AI governance and accountability within corporations? What is an AI risk assessment?

- the legal framework around AI: the European General Data Protection Regulation (GDPR), Platform Regulation, the European AI Act, EU Charter of Fundamental Rights

Teaching will be structured around use cases, including autonomous lethal weapons, autonomous vehicles, facial recognition, social media use of algorithms, algorithms to detect terrorist risk; use of AI in criminal justice; use of AI in the workplace (recruiting, talent management)

Partially Observed Markov Chains in Signal and Image

Instructor: Wojciech Pieczynski

Credits: 3 ECTS

Grading: Quiz, project based on a research article

Numerus clausus: NC

Language: French

Syllabus:

Les modèles de Markov partiellement observes (MMPO) admettent de multiples applications dans des domaines les plus divers. Les modèles de Markov cachés (MMC), qui sont des MMPO de base, apparaissent comme les modèles parmi les plus simples permettant une recherche des réalisations des processus cachés à partir des processus observés dans les cas de grandes masses de données. Les calculs séquentiels – donc explicites et rapides - sont permis par la nature markovienne du couple (processus caché, processus observé).
Dans la première partie du cours on s'intéressera aux processus cachés discrets. On passera en revue les modèles classiques que sont les chaînes, les arbres, et les champs de Markov cachés. On précisera les traitements bayésiens correspondants, ainsi que l'estimation des paramètres permettant des traitements non supervisés. Par la suite, on exposera diverses extensions des Markov cachés classiques - semi-Markov cachés, Markov couples, Markov triplets, Markov évidentiels - dont certains récents. On proposera des illustrations en segmentation statistique d'images, qui est un problème important en traitement d'images.
Dans la deuxième partie on s'intéressera aux processus cachés continus. L'exposé des systèmes gaussiens classiques, rendant possible les filtrages optimaux de type Kalman, sera suivi par les descriptions des modélisations plus complexes de type Markov couple et Markov triplet. En particulier, on étudiera les systèmes à sauts permettant de faire un lien avec la première partie du cours. L'intérêt des notions traitées sera illustré par des applications en poursuite.
Une troisième partie, brève, contiendra des notions des machines de Boltzmann et des réseaux de connaissances profonds, qui peuvent être vues comme des modèles de Markov partiellement observés particuliers.

Reinforcement Learning

Instructor: Erwan Le Pennec

Credits: 3 ECTS

Grading: Project based on a research article.

Numerus clausus: NC
Language: English

Syllabus: This 20-hour course provides an introduction to reinforcement learning. It is based on the new edition of the book "Reinforcement Learning: An Introduction" by R. Sutton and A. Barto. Barto (available online at http://incompleteideas.net/book/the-book-2nd.html).

Outline:
1. Introduction to reinforcement learning and Markov decision processes
2. The bandit case
3. Tabular methods: prediction by dynamic programming, Monte Carlo method and TD Learning
4. Planning and learning for tabular methods
5. Approximate methods: prediction, planning and learning

PART 3

Audio and Music Information Retrieval

Instructor: Geoffroy Peeters, Gael Richard

Credits: 6 ECTS

Grading: 30% labs/project + 70% written exam

Numerus clausus: NC

Language: English

Syllabus: Audio and Music information retrieval is the interdisciplinary research field related to the extraction of semantic information from the audio signal; it allows the development of applications such as speech/music segmentation, recognition (environnemental sound classification, acoustic scene classification, musical instrument recognition), source separation, the estimation of specific music attributes (multi-pitch, tempo/beat, chord, structure), music identification by fingerprint (a la Shazam), cover detection or auto-tagging (into genre, mood)

This course presents the different facets of this field ranging from audio signal representation (Fourier, STFT, Constant-Q transform, audio features), music representation (pitch, chords, rhythm, structure), to pattern-matching and machine-learning models (DTW, HMM, generative/discriminative learning and deep learning).

Planning : 10 sessions of 4 hours + Exam

Big Data and Insurance Project

Instructor: Denis Oblin

Credits: 3 ECTS

Grading: business case and oral presentation by team

Numerus clausus: NC

Language: This course will be given in english or french with lecture material in french.

Syllabus: the objective of this course is to provide an insight of a specific industry and understand how data is leverage in this context. The course mix présentations (on insurrance industry, uses cases, project management insight) and testimony from insurance actors : established insurance company or start ups. You will have to identify a possible usecases and promote it to c-levers :

Course overview:
1) presentations:

- insurance industry insight
- how to find the question ? (not the easy part !)
- project management
- use cases & insurtech

2) 2-3 live testimony (to be updates), in french

Capstone Project

Instructors: Anna Korba & Charles-Albert Lehalle

Credits: 6 ECTS

Grading: Final report and presentation

Numerus clausus:

Language: English

Syllabus:

Period: 9 sessions, between mid-january and end of march.

Goal : The Data Science Capstone project provides an opportunity for students to carry out a defined piece of independent research or design, on a data science task proposed by an industrial partner. The skills it aims at developing include the capacity to define a research or design question, the ability to relate practical problems to existing knowledge and carrying out the research or design in a systematic manner, being given access to real data.

Students will work in groups. Group sizes (between 3 and 5) will depend on the project. The list of available projects will be provided to the students in October.

Mentors will provide weekly assistance to the students via 1h meetings, either remote or in person.

The results will be presented in a final project presentation (roughly 10 min. presentation + 5 min. questions) and a short report (5 pages max + appendices). The final grade of the course will include quality of presentation and report, of work done during the project, and satisfaction of the mentors.

Causal Inference

Instructor: **Marianne Clausel**

Credits: 3 ECTS

Numerus Clausus : 40

Language: English

Planning : 7 sessions of 4 hours + Exam

Grading: Exam (50%), Project (50%). It could be more practical with a data analysis or presentations based on recent research papers.

Syllabus : In machine learning, there has been great progress in obtaining powerful predictive models, but these models rely on correlations between variables and do not allow for an understanding of the underlying mechanisms or how to intervene on the system for achieve a certain goal. The concepts of causality are fundamental to have levers for action, to formulate recommendations and to answer the following questions: "what would happen if » we had acted differently?

The questions of causal inference arise in many areas (socio-economics, politics, psychology, medicine, etc.): depending on the context which drug to use to improve the patient's health? what marketing strategy for product placement should be used to influence consumer buying behavior, etc. The formalism of causal inference makes it possible to study these questions as a problem of classical statistical inference. The gold standard for estimating the effect of treatment is a randomized controlled trial (RCT) which is, for example, mandatory for the authorization of new drugs in pharmaceutical and medical research. However, RCTs are generally very expensive in terms of time and financial costs, and in some areas such as economics or political science, it is often not possible to implement an RCT, for example to assess the effectiveness of a given policy.

The aim of this course is to present the available methods to perform causal inference from observational data. We focus on both the theoretical framework and practical aspects (available software solution). In terms of application, the lecture will be illustrated with recent exemples mainly in the field of health : What is the effect of Hydrochloroquine on survival ? What would have happened if Italy's government had waited a week before imposing lockdown measures ? Ect.

Topics covered:
- The Neyman-Rubin potential outcome causal model for observational studies
- Matching, propensity scores
- Efficiency and double robustness, double machine learning
- Estimating treatment effect heterogeneity, causal forest
- Causal discovery : causal models, graphical models and markov conditions

References:

Hernan, Miguel A., and James M. Robins. Causal Inference. Chapman & Hall/CRC
Imbens, Guido W., and Donald B. Rubin. Causal Inference in Statistics, Social, and Biomedical Sciences
Jonas Peters, Dominik Janzing, Bernhard Schölkopf. Elements of Causal Inference: Foundations and Learning Algorithms
Articles : A Survey of Learning Causality with Data: Problems and Methods
Guo, Ruocheng and Cheng, Lu and Li, Jundong and Hahn, P. Richard and Liu, Huan.

Cloud Data Infrastructure

Instructor: Nicolas Travers, Jihane Mali

Credits: 3 ECTS

Grading: Project on real use cases, 5 deliveries

Numerus clausus: 40

Language: English

Syllabus: This course aims at describing how to model, distribute and optimize data in a distributed infrastructure dedicated to large-scale data management for an environmental point of view. The students will study techniques of denormalization, optimization of queries via modeling, sharding and indexing, information system constraints (consistency, persistence, distribution, etc.), and compare existing solutions and their specificities. A dedicated cost model is studied to study both financial and environmental impact of data infrastructures. The course will allow students to integrate the NoSQL eco-system for data management and to make relevant choices of adequate data infrastructure according to their own query needs (efficiency, environmental and financial).

Main themes :
- Learn data distribution/replication strategies & Compare replication algorithms (HDFS, DHT, GridFS),
- Distributed concurrency and consistency issues & algorithms (CAP, PACELC, 2PC, Paxos)
- Data modeling in a distributed context (NoSQL data families)
- Distributed cost models for IS optimization
- IS conception process over the Cloud

Recommended readings:
- T. Özsu, P. Valduriez, Principles of Distributed Database Systems, **4th Edition**, Springer (2020)
- S. Abiteboul, I. Manolescu, P. Rigaux, M.-C. Rousset, P. Senellart, « Web Data Management ». Cambridge University Press. http://webdam.inria.fr/Jorge/
- F. Abdelhedi, A.A. Brahim, F. Atigui, G. Zurfluh, “MDA-based Approach for NoSQL Databases Modelling”. In: DAWAK 2017. pp. 88–102 (2017)
- A. Chebotko, A. Kashlev, S. Lu, “A Big Data Modeling Methodology for Apache Cassandra”. In: IEEE International Congress on Big Data 2015. pp. 238–245. IEEE (2015)
- J. Mali, F. Atigui, A. Azough, N. Travers, “ModelDrivenGuide: An approach for implementing NoSQL schemas”. In: DEXA 2020: pp. 300-310
- Jihane Mali, Shohreh Ahvar, Faten Atigui, Ahmed Azough, Nicolas Travers: A Global Model-Driven Denormalization Approach for Schema Migration. RCIS 2022: 529-54
- Jihane Mali, Faten Atigui, Ahmed Azough, Nicolas Travers, Shohreh Ahvar: How to Optimize the Environmental Impact of Transformed NoSQL Schemas through a Multidimensional Cost Model? CoRR abs/2311.15406 (2023)
- Jihane Mali, Shohreh Ahvar, Faten Atigui, Ahmed Azough, Nicolas Travers: FACT-DM: A Framework for Automated Cost-Based Data Model Transformation. EDBT 2024: 822-825

- A. Fox, E.A. Brewer, “Harvest, Yield, and Scalable Tolerant Systems”. In: HOTOS 1999. pp. 174–178. IEEE (1999).
- F. Atigui, A. Mokrani, N. Travers, “DataGuide : une approche pour l'implantation de schémas NoSQL”. In: EGC 2020. pp. 407-408
- C. du Mouza, N. Travers, “Relevant Filtering in a Distributed Content-based Publish / Subscribe System”. NoSQL Data Models: Trends and Challenges vol. 1, pp. 203–244 (2018)
- R. Behmo, N. Travers, “Maîtriser les bases de données NoSQL”, https://openclassrooms.com/fr/courses/4462426-maitrisez-les-bases-de-donnees-nosql

Prerequisites: Database (E/R, SQL), basic knowledge in database optimization, UML class diagrams and notions in networks.

Cooperative Optimization for Data Science

Instructor: Andrea Simonetto

Credits: 3 ECTS

Grading: Project (python and literature) and written exam (2h)

Language: English

Numerus clausus: NC

Syllabus:

The course presents continuous optimization techniques that have been developed to deal with the increasing amount of data. In particular, we look at optimization problems that depend on large-scale datasets, spatially distributed data, as well as local private data.
We will focus on three different aspects: (1) the development of algorithms to decompose the problem into smaller problems that can be solved with some degree of coordination; (2) the trade-off of cooperation vs. local computation; (3) how to design algorithms that ensure privacy of sensitive data. This will lead to the study of Distributed optimization, Federated learning, and differential privacy.
The course is both theoretical and practical, requiring some degree of python programming skills as well as mathematical skills.

Indicative schedule:

#1 Class. Introduction: recap on convex models and algorithms. A model for a network of communicating and computing nodes. Parallel methods in optimization: Gauss method, Jacobi method, incremental methods. Distributed optimization (I): primal methods: gradient and gradient tracking.

#2 Class. Distributed optimization (II): dual methods: dual decomposition, ADMM; primal-dual methods and networked problems.

#3 Class. Intermezzo: Stochastic gradient descent, SAGA.
Federated optimization (I): the setting and the problem, its relation with distributed optimization and the main differences. Federated averaging and other momentum-based first-order algorithms.

#4 Class. Federated optimization (II): Robustness, Communication vs. computation trade-o , network scaling, SCAFFOLD, peer-to-peer FL and personalization.

#5 Class. Privacy issues in optimization and learning: the concept of privacy and how to enforce it.

Differential privacy

Deep Learning II

Instructor: Yohan Petetin

Credits: 3 ECTS

Grading: Session 1 : project - Session 2 : project 30% + writting exam 70%

Numerus clausus: NC

Language: English

Syllabus: Ce cours présente les architectures de type "réseaux de neurones profonds" qui ont permis, ces dernières années, de grandes avancées pour des problématiques de classification, de prédiction ou de détection. Il s'agit de la suite du cours 'Deep Learning I' de la première période et ce cours se focalisera avant tout sur les fondements et la compréhension théorique de modèles dits **génératifs** tels que les "Restricted Boltzmann Machines (RBM)", "Deep Belief Network (DBN)", "Generative Adversarial Network (GAN)" ou encore les "Graph Neural Networks" qui permettent le traitement de graphes de données. Les étudiants seront également amenés à implémenter ces modèles dans le langage de leur choix.

Planning

- Rappel des méthodes d'échantillonnage de type Monte Carlo (principe, échantillonneur de Gibbs);
- Restricted Boltzmann Machine (Définition, propriétés, théorème d'universalité, apprentissage par l'algorithme Contrastive Divergence);
- Deep Belief Network (Définition, apprentissage);
- Auto-encodeurs variationnels (l'approche variationnelle, les auto-encodeurs);
- Generative Adversarial Network (Principe, universalité, extensions);
- Graph Neural Networks (principe, applications)

Introduction to MLOps

Instructors: Gaëtan Brison, Awaïs Sani and Laurène David (IP-P, Hi! PARIS)

Credits: 3 ECTS
Grading:

- 20% Graded labs
- 20% Quiz
- 60% Final project

Numerus clausus: 30

Language: English

Syllabus: This course will teach the fundamentals of MLOps, which is defined as the set of tools and best practices for bringing Machine Learning into production. MLOps principles such as *data/model versioning*, *model serving and deployment*, *monitoring*, *continual learning* and *CI/CD automation* will be covered in depth during the duration of the course.

The course is split between theoretical sessions and practical labs. These labs will allow students to practice MLOps hands-on and will include examples using popular frameworks such as *MLFlow, Docker, GitHub, GitHub Actions, DVC, ML Flow, Streamlit* and *Hugging Face*.

The end goal of the course is to build a robust and fully automated machine learning pipeline from data cleaning to model deployment and monitoring using MLOps best practices.

Members of the DataForGood organization (https://dataforgood.fr/) will intervene at the end of the course to present the final project to students, which will center around a use case provided by the organization.

Planning:

- **Introduction to MLOps**: The course will start with an introduction to MLOps principals and show how they vary from traditional Machine Learning model development. A reminder on software development best practices such as version control, unit testing and CI/CD pipelines with Python will also be given.
- **Data pipelines/versioning**: This section will explain how data is ingested from different sources at the beginning of an MLOps pipeline. The section will also cover the importance of maintaining and tracking multiple versions of data as it changes over time to ensure reproducibility and to enable rollbacks to previous versions if needed.

- **Model development**: Students will learn how to create reproducible model training experiments and store important artifacts/metadata (model weights, parameters, metrics…) with model tracking and registry.
- **Model serving and deployment**: This section will cover how to deploy the trained model into a production environment to serve online predictions. Students will learn how to package a model and its dependencies, as well as how to create an API endpoint for it.
- **Monitoring and Continual Learning**: This section will focus on how to monitor deployed machine learning models, detect data distribution shifts and orchestrate automated retrainings using CI/CD.
- **Work on final project with DataForGood**: The course will end with free time for students to work on the DataForGood final project.

Prerequisites:

- Knowledge in Data Science (Data cleaning, Feature Engineering, Model training and evaluation)
- Knowledge in Python

References:

- C. Huyen. Designing Machine Learning Systems. O'Reilly Media Inc. 2022 **https://www.oreilly.com/library/view/designing-machine-learning/9781098107956/**
- Y. Haviv, N. Gift. Implementing MLOps in the Enterprise. O'Reilly Media Inc. 2023 **https://www.oreilly.com/library/view/implementing-mlops-in/9781098136574/**

Machine Learning research seminar

Instructor: Éric Moulines – EL Madhi El Mamdhi

Credits: 6 ECTS

Grading:

Numerus clausus: NC

Language: English

Syllabus: The field of machine learning has experienced exponential growth in recent years, revealing a wealth of promising and successful research opportunities. This seminar aims to address a selection of these innovative topics from current research. During the seminar, students will learn about the latest advances in machine learning, explore new areas and innovative ideas, and engage in the analysis and critique of recent scientific publications.

To achieve these goals, each student will be assigned 2-5 research papers to read and analyze thoroughly. Based on these primary sources, students will be required to engage with the surrounding literature and summarize their findings, critiques, and research proposals in a 4-page, double-spaced essay. This exercise will not only improve their understanding but also hone their analytical skills.

Peer reviews will play a crucial role in this seminar. Students will review and critique each other's work, providing constructive feedback to foster a collaborative learning environment. This process will help sharpen their critical thinking and evaluation skills, which are essential for their academic and professional growth.

At the end of the seminar, each group student will prepare and deliver a 25-minute presentation summarizing their research findings. These presentations will be given at the end of the semester in the form of a block seminar that will provide a comprehensive overview of the topics covered and allow for in-depth discussion among the participants.

We will cover the following topics:

- learning theory
- deep learning (architectures, deep learning theory, etc.)
- major applications of deep learning.
- optimal transport
- generative models
- reinforcement learning

- Optimization
- federated learning
- learning/economy interfaces
- Probabilistic ML

Machine Learning with Graphs

Instructor: Jhony H. Giraldo (Previously: Florence d'Alché-Buc)

Credits: 3 ECTS

Grading: Project defense

Numerus clausus : 50

Language: English

Syllabus: Graph data is ubiquitous. Any system with entities and relationships between them can be represented as a graph. Over the past decade, machine learning algorithms have made remarkable progress in fields such as natural language processing, computer vision, and speech recognition. This success is primarily due to deep neural network architectures' ability to extract high-level features from Euclidean-structured data like images, text, and audio. However, graph data has not received the same level of attention.

In this course, we will explore how to create machine learning models to extract high-level features from graph data, a process known as graph representation learning. The topics covered in this course include graph signal processing, graph neural networks (GNNs), such as graph convolutions and graph attention mechanisms, scalable GNNs for big data applications, self-supervised learning on graphs, spatiotemporal data analysis with GNNs, and graph transformers. This course also includes laboratory sessions to provide hands-on experience with these concepts.

Online Learning and Aggregation

Instructor: Solene Gaucher - Mathilde Tullii

Credits: 3 ECTS

Grading: Final exam

Numerus clausus: NC
Language: English

Syllabus: The aim of online learning is to provide efficient recursive algorithms of prediction when the data are arriving sequentially in a streaming way rather than as an array given once and for all. Whereas statistical learning is dealing with independent identically distributed data, the emphasis in online learning is on adversarial setting where the data are of arbitrary nature satisfying mild conditions. In this setting, one of the key ideas is to use, at each time instance, a suitable randomized choice from the given set of candidate predictors. Analogous techniques can be applied to solve the problem of aggregation, that is, to obtain procedures that predict almost as good as the best estimator in a given set. This course provides an introduction to online learning and aggregation focusing on theoretical aspects.

Topics covered:
-- Online classification in realizable case, halving.
-- Online gradient descent for convex and strongly convex loss. Online-to-batch conversion. Online linear regression.
-- Randomization by exponential weighting. Prediction with expert advice.
-- Adversarial multi-armed bandit problem.
-- Aggregation of estimators.
-- Gradient-free online learning. Continuum bandit problem.

References:
Shalev-Schwartz, S. (2011) Online learning and online convex optimisation. *Foundations and Trends in Machine Learning*, vol. 4, pages 107-194.

Tsybakov, A. (2020) Online learning and aggregation. Lecture Notes.

https://www.dropbox.com/s/lcans4r4mlsryux/ALTEGRAD_VAZIRGIANNIS_2021-09-06.mov?dl=0

Integer Optimization for Machine Learning

Instructor: Zacharie Ales

Credits: 3 ECTS

Grading: Exam and project

Numerus clausus: NC

Language: French (english course materials)

Syllabus: This course presents the main application of operation research to machine learning.

- Session 1 and 2 – Optimal decision trees
Decision trees are highly popular interpretable classifiers. They are generally trained using greedy algorithms such as CART which does not necessarily lead to an optimal solution. We will present several integer linear programming methods that can be used to obtain optimal decision trees.

- Session 3 and 4 - Linear Regression by Solving a Robust Optimization Problem
The linear regression problem is most commonly solved by introducing a bias term in the objective to prevent overfitting. In these sessions, we will see that it can be more effective to consider a robust modeling of the problem, in which we aim to minimize the error for the worst possible uncertainty in the data.

- Session 5 – From predictive to prescriptive analytics
In the context of prescription under uncertainty, classical operations research approach are usually not designed to efficiently leverage historical data. Conversely, machine learning approaches are efficient to predict values but going from good predictions to good prescriptions is not obvious. In this sesssion we introduce a method that combines operations research and machine learning to improve prescription under uncertainty.

- Session 6 - Exam

Prerequisites : Completion of an operation research course covering the following notions (e.g., the introductory course of the master): graph theory, simplexe algorithm, branch-and-bound.

Optimal Transport: Theory, Computations, Statistics, and ML applications

Instructor: Marco Cuturi

Credits: 3 ECTS

Grading: project on one of the projects proposed in class: python code + pdf writeup.

Numerus clausus: NC

Language: English

Syllabus: 8 séances de cours + 4 séances de TD = 18 heures

3 hours: Theory

- Monge and Kantorovich Problems, duality in OT.
- Gangbo-McCann and Brenier theorem.
- Closed forms: transport between Gaussians and transport in 1D
- Wasserstein space: JKO flow.

6 hours: OT solvers

- Algorithmic overview: network flow solvers in the discrete world, Benamou-Brenier formula in the PDE world.
- Statistical results and the curse of dimensionality
- Entropy regularized approaches to compute optimal transport.
- Differentiation of Optimal transport: unrolling / implicit.
- Unbalanced generalizations, quadratic / fused / low-rank extensions (GW).

3 hours TDs:

- 1D transport, transport between Gaussians
- Network flow solver type algorithms + SInkhorn algorithm

3 hours: Extensions

- Handling measures with the Wasserstein geometry: barycenters, clusters
- Computations wiht non-Euclidean costs, Entropic map estimator
- Dynamical formulations and links with flow matching / diffusion models

3 hours TDs:

- Differentiation of Sinkhorn's algorithm
- Transport between high-dimensional point clouds.

Recent Developments in Responsible AI

Instructor: Charlotte Laclau – Florence D’Alche-Buc

Credits: 3 ECTS

Grading: Students will be asked to write a blogpost article on a paper recently published at a highranking conference (eg. ICLR, NeurIPS) related to the topic of the course. A list of selected articles will be proposed by the lecturers. Blogposts will be published on the medium account of the master.

Numerus clausus: NC

Language: English

Syllabus: Artificial intelligence, as a transversal discipline, plays a central role in our modern society, driving vital advances and amplifying efficiency, well-informed decision-making and general practicality in our daily routines. This advanced master's course aims to provide students with a comprehensive understanding of the latest developments in Responsible AI. The course will explore various facets of Responsible AI, including interpretable AI, fairness in machine learning, robust machine learning, data privacy, and frugality. Students will delve into both theoretical foundations and practical implementations, equipping them with the skills to design and implement AI systems that are ethical, accountable, and aligned with societal values.

Description:

Representation Learning for Computer Vision and Medical Imaging

Instructor: Pietro Gori et Loïc Le Folgoc

Credits: 3 ECTS

Grading: Grading will be based on the practical session reports (50%) and written or oral exam (depending on the number of students) (50%)

Numerus clausus: NC

Language: English

Syllabus: 7 lectures divided into 1,5h of theory and 1,5h of practical session.

Description:

Good and expressive data representations can improve the accuracy of machine learning problems and ease interpretability and transfer. For vision tasks, handcrafting good data representations, a.k.a. feature engineering, was traditionally hard. Deep Learning has changed this paradigm by allowing us to automatically discover good representations from data. This is known as representation learning. The objective of this course is to introduce representation learning in computer vision and medical imaging applications.

We will cover the following topics:

- Introduction to Representation Learning for Vision

- Transfer Learning and Domain Adaptation

- Self-supervised and Contrastive Learning

- Knowledge Distillation

- Disentangled Representations

- Conditional Generative models

- Attention and Transformers

- Visualization and interpretability in Neural Networks

- Multimodal representation learning and Foundation models

Stochastic Approximation And Reinforcement Learning

Instructor: Pascal Bianchi – Walid Hachem

Credits: 3 ECTS

Grading: Written exam 2h.

Numerus clausus: NC

Language: English

Syllabus: This course provides a rigorous exploration of Iterative Stochastic Algorithms, also known as Stochastic Approximation. Covering a broad range of applications including optimization, game theory, Monte Carlo methods, and reinforcement learning, the course begins with a concise review of dynamical systems basics. Next, we introduce the Ordinary Differential Equation (ODE) method, a key tool to establish convergence. This method demonstrates how the discrete-time algorithm mirrors the behavior of an ODE, inheriting its convergence properties.

The second portion of the course concentrates on reinforcement learning, beginning with a comprehensive review of Markov decision processes. This foundational knowledge serves as a basis for understanding and establishing the convergence of key reinforcement learning algorithms, including Q-learning, SARSA, and gradient policy algorithms. Through a combination of theoretical grounding and practical exercises, participants will gain a clear, rigorous grasp of Iterative Stochastic Algorithms and their significance in various fields of application.

The course is mainly theoretical. It is a nice complement to E. Le Pennec's course on reinforcement learning.

Description:

Tail Events Analysis: Robustness, Outliers and Models for Extremes

Instructor: Pavlo Mozharovskyi

Credits: 3 ECTS

Grading: written exam

Numerus clausus: NC

Language: English

Syllabus: Analysis of events in the tail of the distribution constitutes an important topic in statistics. The aim of the course is threefold: construction of the estimators less sensible to contamination of the data, identification of the outlying observations, and modeling of extreme events. The course starts with the introduction to robust statistics and measures of robustness where the concepts of the influence function and of the breakdown value are given. Then, the simplest univariate robust estimators of location, scale, and skewness are regarded which behave consistently even if the data is contaminated. These are further generalised to robust estimators of multivariate location and scatter (Stahel-Donoho estimator, minimum covariance determinant estimator, S-estimators, MM-estimators). Robust regression as well as PCA estimators are also considered. An important notion in robust statistics—data depth—is then introduced for the multivariate and functional framework. Presentation of the concept of the data depth function is followed by studying most important depth notions such as Tukey or projection depth, and (multivariate) functional depths. The regarded above material is then applied to detection of outliers in multivariate and functional data, as well as the cell-wise outliers. Finally an introduction to extreme values analysis is given. Here, extremes are defined as the largest values of a considered dataset. Extreme value theory suggests natural models for block-maxima and excesses above large thresholds, which give rise to estimates of quantities of major interest for risk management, such as high quantiles, large return levels, or tail probabilities outside the range of observed data. This aims on providing guidelines to the students for applying extreme value models to answer such practical questions.

Format: 6 × 3.5 hours + exam

Planning:

Week 1: Introduction to robust statistics. Measures of robustness: influence function, breakdown value. Univariate robust estimators of location, scale, skewness. Multivariate location and scatter estimators. Robust regression and robust PCA. Reading:Rousseeuw and Leroy (1987), chapters 1–3; Huber and Ronchetti (2009), chapters 1, 3, 8, 11; Wilcox (2016), chapters 1–3, 10; Rousseeuw and Driessen (1999); Hubert et al. (2005).

Week 2: Lab session I. Univariate robust estimation. Robust multivariate estimation: projection pursuit, minimum covariance determinant estimator. Robust regression and ROBPCA.

Week 3: Multivariate data depth. Statistical data depth function: definition and properties, chosen notions. Identification of multivariate (row-wise and cell-wise) outliers. Reading: Becker et al. (2013), chapters 1, 2, 4; Wilcox (2016), chapter 6; Tukey (1975); Donoho and Gasko (1992); Zuo and Serfling (2000); Hubert et al. (2015); Rousseeuw and Bossche (2018).

Week 4: Data depth definition for non-Euclidean spaces. Data depth in infinite-dimensional setting. Identification of functional and curve outliers. Reading: Fraiman and Muniz (2001); López-Pintado and Romo (2009); Claeskens et al. (2014).

Week 5 : Extreme value statistics. The one dimensional case, distribution of maxima of large datasets, excesses above high thresholds. Inference methods and case studies. Multi-dimensional setting. Regular variation. Reading: Coles et al. (2001); Beirlant et al. (2006); Resnick (2013); De Haan and Ferreira (2007); Resnick (2007, 2013).

Week 6: Lab session II. Data depth: Tukey, projection, spatial depths, functional depths, applications to outlier detection. Extreme value analysis: return levels, probabilities of failure.

References:

Becker, C., Fried, R., and Kuhnt, S. E. (2013). Robustness and Complex Data Structures: Festschrift in Honour of Ursula Gather. Springer, Berlin–Heidelberg.

Beirlant, J., Goegebeur, Y., Segers, J., and Teugels, J. L. (2006). Statistics of extremes: theory and applications. John Wiley & Sons.

Coles, S., Bawa, J., Trenner, L., and Dorazio, P. (2001). An introduction to statistical modeling of extreme values, volume 208. Springer.

Cornillon, P., Guyader, A., Husson, F., Jegou, N., Josse, J., Kloareg, M., MatznerLober, E., and Rouviére, L. (2012). R for Statistics. Chapman and Hall/CRC, New York.

De Haan, L. and Ferreira, A. (2007). Extreme value theory: an introduction. Springer Science & Business Media.

Donoho, D. L. and Gasko, M. (1992). Breakdown properties of location estimates based on halfspace depth and projected outlyingness. The Annals of Statistics, 20(4):1803–1827.

Mosler, K. and Mozharovskyi, P. (2022). Choosing among notions of multivariate depth statistics. Statistical Science, 37(3):348–368.

Huber, P. J. and Ronchetti, E. M. (2009). Robust Statistics. Second Edition. John Wiley & Sons, Hoboken. Hubert, M., Rousseeuw, P. J., and Branden, K. V. (2005). Robpca: A new approach to robust principal component analysis. Technometrics, 47(1):64–79.