2nd-Level MASTER in Data Science and Statistical Learning (MD2SL)

SYLLABUS of the courses

Course	SDD	CFU	Hours
Block I – Bootcamp courses			
Mathematics and Statistics for Data Science		**8**	**64**
Optimization	MAT/09	2	16
Numerical calculus and linear algebra	MAT/08	2	16
Probability and stochastic processes	MAT/06	2	16
Statistical inference and modelling	SECS-S/01	2	16
Algorithmic Foundations and Programming Skills		**6**	**48**
Algorithms and programming in Python and R for data science	INF/01	3	24
Machine learning	ING-INF/05	2	16
Optimization for machine learning	MAT/09	1	8
Block II – Core courses			
Statistical Learning for Data Science		**6**	**48**
Statistical learning	SECS-S/01	2	16
Geo-spatial data analysis	SECS-S/01	2	16
Network data analysis	SECS-S/01	2	16
Supervised and Unsupervised Learning		**6**	**48**
Advanced machine learning	MAT/09	3	24
Deep learning, neural networks, and reinforcement learning	ING-INF/05	3	24
Complex Systems		**6**	**48**
Text mining and NLP	ING-INF/05	2	16
Network and media analysis	FIS/03	2	16
Complex system analysis	FIS/03	2	16
Decision Theory for Data Science		**7**	**56**
Bayesian inference and causal machine learning	SECS-S/01	3	24
Analytics in economics and business	SECS-P/06	3	24
Ethics and law for data science	IUS/01	1	8
Block III – Elective courses *Two tracks chosen from*			
1) Data Science for Economics		**4**	**32**
Experiments and real-world evidence in economics	SECS-P/02	2	16
Policy evaluation and impact analysis	SECS-P/06	2	16
2) Data Science for Business		**4**	**32**
Time series analysis	SECS-S/03	2	16
Optimization of financial portfolios	SECS-S/06	2	15

3) Data Science for Health		**4**	**32**
Health analytics and data-driven medicine	SECS-P/02	2	16
Environmental and genomic data analysis	MED/01	2	16

Hands-on labs	**SECS-S/01**	**3**	**24**

Total CFU/ hours of teaching	**50**	400

Seminars, real-case studies by colleagues and partners	**2**	16
Internship	**9**	225
Final project	**3**	

Total CFU/hours	**64**	

Block I – Bootcamp courses

Mathematics and Statistics for Data Science

Optimization

Basics

- Introduction to Optimization Problems and Models
- Local vs global optimization, convex and non convex problems
- Classification of Optimization problems
- Elementary examples

Introduction to Optimality conditions

- Descent and Feasible Directions
- First Order Conditions - Unconstrained Case
- First Order Conditions - Constrained Case
- Results for Convex Problems

Introduction to unconstrained local optimization methods

- Iterative Optimization Algorithms
- Gradient-Based Methods
- Line Searches
- Newton's Method

Stochastic gradient and variants

- Finite-Sum Problems
- The SGD algorithm
- Minibatch GD
- Momentum and Acceleration
- Adaptive Stepsizes

Basic constrained optimization methods

- Problems with convex constraints
- Gradient Projection Method
- Frank-Wolfe Method

Global optimization

- Exact global optimization methods
- Heuristic global optimization methods
- Bayesian optimization

Numerical calculus and Linear algebra

Topics:

1. Errors and machine arithmetic

2. Elements of matrix analysis

3. Numerical solution of linear systems: direct methods

4. Overdetermined systems, QR factorization and singular value decomposition (SVD)

5. Numerical solution of linear systems: iterative methods

Probability and stochastic processes

Probability

1. Discrete random variables: Probability distributions, probability mass functions, cumulative distribution functions, mean and variance. Discrete models.
2. Joint probability distribution, Marginal distributions, Conditional probability, conditional mean and variance. Discrete models.
3. Continuous random variables: Probability distributions, probability density functions, cumulative distribution functions, mean and variance. Conditional probability. Continuous models.
4. Convergence theorems and normal approximation. Poisson Process and applications.

Stochastic Processes

1. Introduction to Markov Chains and their transition matrix.
2. Classification of states, invariant distributions.
3. Simulated annealing and Metropolis algorithm.
4. Birth-and-death chains on finite state spaces.

Statistical inference and modelling

Learning outcomes:
This is a leveling course, revising the main concepts in Statistics, including essential ideas of data analysis, modelling (linear and generalized linear models, association structures) and inference (likelihood theory, confidence regions, significance testing, model building) with applications.

Topics:

1. *Dependence and independence*
   Marginalization and conditioning - association and regression - multivariate normal - types of independence
2. *Inference and linear models*
   Regression models - least-squares - multiple regression - extensions - tests
3. *Generalized linear models*
   Generalized linear models - binary data - likelihood inference - Poisson regression
4. *Model building*
   Purposes - prediction - explanation – case studies

Block I – Bootcamp courses

Algorithmic Foundations and Programming Skills

Algorithms and programming in Python and R for Data Science

Python:

- Introduction to Python and simple Data
- Python Modules and Functions
- Selections and Iterations
- Recursion and Strings
- Lists and Dictionary
- Classes and Objects, Files
- Analysis of Algorithms
- Sorting and Searching

R:

- Introduction to R: the R console, R packages, files .R
- Elementary objects of R: vectors, matrices, arrays, lists; different typologies of objects (numerical, characters, logical, factorial)
- Basic mathematical functions; personalization of functions
- The dataframe: definition and manipulation
- Data import and data export in R (.txt files, Excel files, Stata/SAS/SPSS files, .RData files)
- Manipulations of objects - 1: variable recoding, time variables, missing data, record linkage
- Manipulations of objects - 2: statistical descriptive analyses (tables, synthetic measures, basic graphical display)

Machine Learning

Part A

- Supervised versus unsupervised ML, essential probability theory, statistics, and distributions for ML, Bayesian versus frequentist interpretations for ML
- Linear models for supervised regression and classification
- The bias-variance decomposition, overfitting, underfitting, and model regularization
- Maximum Likelihood Estimation (MLE), the expectation-maximization (EM) algorithm, Maximum a Posteriori (MAP) versus Bayesian inference
- Connectionist models and introduction to artificial neural networks

Part B

- From neurons to artificial neural networks: training as a non-linear optimization problem
- Backpropagation and gradient-based methods
- Linear Support Vector Machines (SVMs)
- Non-linear SVMs and radial basis function networks
- Using the LIBSVM library

Block II – Core courses

Statistical Learning for Data Science

Statistical Learning

1. *Introduction to statistical learning*

- Statistical point of view of machine learning
- Data generating process
- Monte Carlo simulations

2. *Graphical models*

- Networks and concentration graph models
- DAG and Bayesian network

3. *Supervised statistical learning based on trees*

- CART algorithm
- Bagging and Random forest
- Boosted trees
- BART

4. *Interpretable statistical learning*

- Predicting vs explaining
- Interpretability, transparency, fairness

Geo-spatial data analysis

Topics:

- Introduction to spatial and geographical data
- Stochastic spatial processes and their properties
- Analysis of point process data
- Analysis of geodata random surface
- Analysis of areal data (lattice data)
- Spatial interaction data: gravity models
- Introduction to Geographical Information Systems

Network data analysis

Topics:

- Introduction to network data
- Network representation: types of relations, graph representation, matrix representation
- Hints on network visualization
- Descriptive analysis of network data: network statistics
- Descriptive analysis of network data: nodal statistics

- Modeling networks: introduction
- Latent Space Models
- Stochastic BlockModels
- Introduction to Exponential Random Graph Models

Block II – Core courses

Supervised and Unsupervised Learning

Advanced Machine learning
The course provides an introduction to basic concepts in machine learning.
Topics: *Learning theory* - bias/variance tradeoff, - Vapnik-Chervonenkis dimension and Rademacher complexity - cross-validation - regularization *Supervised learning* - linear regression, - logistic regression - support vector machines - neural networks *Unsupervised learning* - clustering - principal and independent component analysis *Semisupervised learning* - Laplacian support vector machines *Learning algorithms* - gradient descent - stochastic gradient descent - perceptron algorithm

Deep learning, neural networks, and reinforcement learning
Topics: - Deep Learning and Neural Network fundamentals - Optimization: backpropagation, stochastic gradient descent, overfitting - Hands on crash course on pytorch - Convolutional Neural Networks - Deep Learning Practicum, how to bootstrap a project, debug a model and track experiments - Sequence Learning and recurrent architectures

- Unsupervised Learning

Block II – Core courses

Complex Systems

Complex system analysis

Topics:

1. *Introduction to complexity in nonlinear dynamics*

- Dynamical systems in 1D, 2D and 3D
- Fixed points and stability
- Bifurcation theory
- Discrete maps
- Chaos
- Examples and applications
- Fractal geometry: box counting dimension
- Strange attractors
- Multifractals, generalized fractal dimensions

2. *Nonlinear time-series analysis*

- Diagnostic of chaos
- Embedding and Takens' theorem
- Correlation dimension and Grassberger-Procaccia method

Text Mining and NLP

Topics**:**

Text Mining and Information Retrieval

- Introduction to text-mining
- IR models (boolean and vectorial) term-document matrix; tf-idf
- Inverted Index
- linguistic preprocessing steps: tagging, stop-word removal, lemmatization, stemming
- query with wildcard (n-gram); spelling correction; edit distance
- performance evaluation (Precision, Recall)

Natural Language Processing:

- Probabilistic language models (n-gram),
- Text classification (sentiment analysis)
- Word meaning, vector semantics word embeddings
- POS (Part of Speech) tagging
- NE (Named Entity) recognition

Modern NLP:

- NN-based approaches (sentiment analysis and language modelling)
- Word and sentence embeddings: training and use of the models

Network and media analysis

Topics**:**

The birth of a theory

- Konigsberg bridge problem

From graphs to networks: stylized facts

- scale invariance of the degree distribution
- small-world phenomenon
- modularity

Network representations

- adjacency matrix of monopartite, binary/weighted, undirected/directed networks
- adjacency matrix of bipartite, binary networks
- multilayer networks, hypergraphs

Network properties

- degree(s), strength(s)
- degree(s) and strength(s) correlations
- clustering coefficient(s)
- motifs
- reciprocity

Paths, walks, distances, centralities

- trails, circuits, paths, cycles
- degree centrality
- closeness centrality
- betweenness centrality
- eigenvector centrality
- Katz centrality
- Page Rank centrality

A primer on static models: the Erdos-Renyi model

- local, meso and macro-scale properties of the Erdos-Renyi model

A primer on static models: the MaxEnt model

- local, meso and macro-scale properties of the MaxEnt model

Early network reconstruction models

- perturbed MaxEnt
- Minimum Density algorithm
- Iterative Proportional Fitting algorithm

Statistical mechanics of networks

- ensembles of networks
- Exponential Random Graphs formalism
- Erdos-Renyi model...again
- Configuration Model

Network reconstruction

- from the Configuration Model to the fitness model
- applications of the fitness model to economic and financial networks

Network mesoscale structures

- communities
- core-periphery structure
- bow-tie structure
- modelling mesoscale structures: the Stochastic Block Model

Block II – Core courses

Decision Theory for Data Science

Analytics in economics and business
The aim of this course is to teach students how to apply advanced machine learning techniques in economics and management using hands-on empirical tools for different data structures. We will bridge the gap between applications of methods in published papers and practical lessons for producing your own research. After introductions to up-to-date illustrative contributions to literature, students will be asked to perform their own analyses and comment results after applications to microdata provided during the course
Topics: - New Tricks for Econometrics and Artificial Intelligence - Statistical Learning with Sparsity: The Lasso and Generalizations - Classification and Regression Trees - Using Big Data for Measurement and Research - Using Big Data for Measurement and Research

Bayesian Inference and Causal Machine Learning

Topics:
 	
Introduction to the potential outcome approach
- Potential Outcomes
- Definition of Causal Effects
- Learning about Causal Effects: Multiple Units
- The Stable Unit Treatment Value Assumption
- The Assignment Mechanism
  • The role of the assignment mechanism in causal inference
  • A Classification of Assignment Mechanisms
  • Classical Randomized Experiments
 	
An introduction to Bayesian inference
- Reflections on the 'classical approach' for statistical inference
- The likelihood approach: a precursor of Bayesian inference
- A first encounter of the Bayesian approach
 	
Model-based Bayesian inference in completely randomized studies
- Brief review of the Fisher's exact p−value approach and Neyman's repeated sampling approach to completely randomized experiments
- Bayesian model-based imputation in the absence of covariates
- Simulation methods in the model-Based approach
- Dependence between Potential Outcomes
- Model-Based Imputation with Covariates
- Super-Population Average Treatment Effects
- Model-Based Estimates of the Effect of the NSW Program in R

Observational studies and Bayesian inference
Regression in observational studies

- Estimands
- Unconfoundedness
- Identify causal effects under unconfoundedness
- Overlap
- Outcome-regression-based estimation
- Strategies to reduce model sensitivity

The role of the propensity score in the design and analysis of observational studies

- Propensity score
- Balancing property of propensity score
- Propensity score and unconfoundedness
- Estimation of the propensity score
- Propensity score stratification
- Propensity score matching
- Propensity score regression
- Estimated versus true propensity score

Ethics and Law for Data Science

- *Governo dei dati e tutela dei diritti fondamentali*
- *Data Localization e sovranità dei dati*
- *La protezione dei dati: profili introduttivi*
- *Big data e automazione*
- *Big data e IOT*
- *Responsabilità dell'ISP*

Hands-on labs

Module A

1.a) Introduction to R and STATA

- Overview
- The basics (objects, manipulation, basic statements, missing data)
- Reading data from files
- Probability distributions
- Basic statistical models
- Graphical procedures
- Packages overview

2.a) Data Modeling for policy evaluation:

- Regression analysis
- Matching and subclassification
- Inverse Probability Weighting
- IPW with Regression Adjustment
- Entropy Balancing

3.a) Machine Learning (ML) tools for Econometrics

- Predictive analysis
- ML to select control variables and/or instruments
- ML to build counterfactuals (when no control group is available)
- Heterogeneity of treatment effects

Block III – Elective courses

Data Science for Economics

Experiments and real-world evidence in economics
This course aims at introducing students to the empirical study of behaviour and the methodological issues related to the inference of causation by means of experimental and related techniques. Standard statistical analysis of experimental data will also be covered, together with a number of examples from real-world phenomena which have been studied with these techniques. The course is almost self-contained and does not require a specific background apart from basic statistics.
Topics: 1. From theory to data (and the way back). Introduction to behavioural and experimental economics. 2. Learning from the data. Correlation is not causation. In search for practicable ways to go beyond correlations in social and economic phenomena. - The controlled solution: Experiments (online, in the laboratory, in the field). - The less controlled solution: Natural and Quasi-experiments. 3. Statistical analysis of experimental data. Mediator variables, modulator variables, specific statistical tests, multiple testing of hypotheses. 4. Case studies. - Examples of controlled experiments and their analysis (e.g., risky behaviors, addiction, - strategic behaviors, moral dilemmas, marketing, persuasion, nudging). - Examples of natural experiments and their analysis (e.g., Italian clemency bill and - criminal behaviors). - Examples of quasi-experiments and their analysis (e.g., evaluating educational - programs in primary schools).

Policy evaluation and impact analysis
The course will introduce most popular approaches for causal inference in econometrics with an emphasis on applications to different economic topics in industrial organization, economics of innovation, health economics, and corporate finance. A variety of methods are illustrated with a handson-tool approach, combining theory and practice. The objective is to develop a critical understanding of the iterative research process leading from real economic issues to the choice of the best tools available from the analyst kit.
Topics: *Introduction to microeconometrics*: - Structure, Endogeneity, and Identification Problems - Least-squares, Probit, and Logit Estimators - Static panel data - Dynamic panel data Applications: Firms, Productivity, and Technical Change (Industrial Organization, Economics of Innovation) *The Evaluation Problem*

- Randomization and Matching Models
- The Difference-in-difference Estimators
- Instrumental Variables
- Regression Discontinuity Design

Applications: Mergers & Acquisitions (Corporate Finance); Competition and Market Power (Industrial Organization)

Causality and Non-linear Models

- Quantile Regressions
- Multinomial Models
- Models for Count Data
- Survival/Duration Analyses
- Models with Control Functions
-

Applications: Location Choices (Economic Geography); Firms' Bankruptcy (Corporate Finance); Pharmaceutical Markets (Health Economics).

Block III – Elective courses

Data Science for Business

Time Series Analysis
The teaching provides an introduction to modeling and forecasting methods for time series data, which are illustrated through real-world economic and financial problems using R for Statistical Computing. Familiarity with basic probability calculus and matrix algebra are required.
Topics: - Stationary processes - Linear models for stationary processes: AR, MA, and ARMA - Trend stationary processes - Integrated processes - Linear models for integrated processes: ARIMA - Model building and model diagnostics - Model-based forecasting and assessment of forecast error - Seasonal ARIMA models

Optimization of Financial Portfolios
Topics: - Risk aversion and expected utility theory - Mean-variance approach and portfolio theory - The Capital Asset Pricing Model (CAPM): theory and evidence - Stylized facts of financial markets, with MATLAB applications (part 1); covariance matrix shrinkage for portfolio optimization (part 2).

Block III – Elective courses

Data Science for Health

Health analytics and data-driven medicine

Part A

This part of the course focuses on the application of machine learning to healthcare. The lecture slides presented during this part cover the following subset of topics of the MIT graduate course "Machine learning for healthcare" taught by Professors Peter Szolovits and David Sontag:

1) What makes healthcare unique?
2) Overview of clinical care
3) An example of analysis of clinical data
4) Risk stratification
5) Learning with noisy labels
6) Machine learning for cardiac imaging
7) Disease progression modelling
8) Reinforcement learning for healthcare

Part B

The student will be introduced to topics and tools of health econometrics. The aim is to provide the student with the ability to critically evaluate pros and cons of different empirical strategies to perform their own investigations in the context of new case studies.

1) Why econometrics for healthcare?
2) Experiments and observational analyses
3) Censored and truncated samples
4) Survival/duration models
5) Count data regression models
6) Functional forms and the generalized linear model
7) Modelling choices: multinomial regression models
8) Quantile regression models

Environmental and genomic data analysis

Through real data examples, the student will learn how to implement and compare different machine learning (ML) models and to use complex algorithms in clinical and epidemiological contexts. The course will focus on two applications. After brief introductions to the specific problems, guided practicals in R software will be proposed.