Block I – Data Science Bootcamp

Mathematics and Statistics for Data Science

- Optimization
  - Introduction to Optimality conditions
  - Introduction to unconstrained local optimization methods
  - Stochastic gradient and variants
  - Basic constrained optimization methods
  - Global optimization
  - Exact global optimization methods
  - Heuristic global optimization methods
  - Bayesian optimization
- Numerical Calculus and Linear Algebra
  - Errori e aritmetica finita
  - Elementi di analisi delle matrici
  - Risoluzione numerica di sistemi lineari: metodi diretti
  - Sistemi sovradeterminati (fattorizzazione QR e decomposizione ai valori singolari)
  - Risoluzione numerica di sistemi lineari: metodi iterativi stazionari
  - Precondizionatori
  - Generalità sui Metodi di Krylov
  - GMRES
- Probability and stochastic processes
  - Probability:
    - Discrete random variables: Probability distributions, probability mass functions, cumulative distribution functions, mean and variance. Discrete models.
    - Joint probability distribution, Marginal distributions, Conditional probability, conditional mean and variance. Discrete models.
    - Continuous random variables: Probability distributions, probability density functions, cumulative distribution functions, mean and variance. Conditional probability. Continuous models.
    - Convergence theorems and normal approximation. Poisson Process and applications.
  - Stochastic Processes:
    - Introduction to Markov Chains and their transition matrix.
    - Classification of states, invariant distributions.
    - Simulated annealing and Metropolis algorithm.
    - Birth-and-death chains on finite state spaces.
- Statistical inference and modelling
  - Inference and linear models:
    - Statistical thinking
    - Frequentist (classical) inference
    - Exploring associations
    - Significance tests
    - Prediction
  - Generalized linear models:
    - Non-normal responses
    - Regression with a binary response
    - Binary data
    - The general linear logistic model
    - Inference and prediction
    - Generalized linear models
    - Contingency tables and Poisson models
    - Log-linear models
    - The Ising model in 3 binary variables

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, .R Data 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)
- Introduction to Machine Learning
  - 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
  - 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
  - Introduction to statistical learning:
    - Statistical point of view of machine learning
    - Data generating process
    - Monte Carlo simulations
  - Graphical models:
    - Networks and concentration graph models
    - DAG and Bayesian network
  - Supervised statistical learning based on trees:
    - CART algorithm
    - Bagging and Random forest
    - Boosted trees

- BART
- Interpretable statistical learning:
  - Predicting vs explaining
  - Interpretability, transparency, fairness

Geo-spatial and network data modelling

- Network data modelling:
  - 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
  - Exponential Random Graph models
  - Stochastic blockmodels
  - Latent space models
- Geo-spatial data modelling:
  - 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

Supervised and Unsupervised Learning

Machine learning

- Introduction to supervised learning and regression.
- Classification problems.
- Online learning: the perceptron learning algorithm.
- Gradient descent and stochastic gradient descent: analysis, MATLAB implementation, backpropagation.
- Unsupervised learning. MATLAB implementation of principal component analysis and spectral clustering.
- Introduction to statistical learning theory.
- Structural risk minimization and support vector machines.
- Trade-off between sample size and precision of supervision.
- A comparison of approximation error bounds for neural networks and linear approximators.
- Application of neural networks to optimal control problems.
- Radial basis function interpolating networks and their application to surrogate modeling and optimization.
- Connection between supervised learning and reinforcement learning.

Deep learning, Neural Networks, and Reinforcement learning

- Sequence learning and recurrent networks
- Attention mechanisms
- Graph learning
- Explainable machine learning
- Explainable deep learning

Complex Systems

Text Mining and NLP

Network and media analysis

- Introduction to complex networks:
  - networks definition;
  - network representation;
  - degree and ANND.
- Introduction to Twitter data:
  - data structure
- User features and power law distributions:
  - information per user: tweets, retweets, followers and friends
  - power law distributions: scale free networks
  - verified users
- Gonzalez-Bailon user classification
- The retweet network:
  - building retweet network from data;
  - visualize the network;
  - assign attributes to nodes.
- Centrality measures:
  - Hub and Authorities;
  - Page Rank;
  - Node betweenness
- Community detection algorithms:
  - Girvan-Newman and the definition of Modularity;
  - The importance of null-model;
  - Louvain community detection

Complex system analysis

- Dynamical systems in 1D, 2D and 3D
- Fixed points and stability
- Bifurcation theory
- Discrete maps
- Chaos
- Turing instability in reaction diffusion models
- Examples and applications

Decision Theory for Data Science

Bayesian Inference and Causal Machine Learning

Analytics in economics and business

- General introduction
- New Tricks for Econometrics and Artificial Intelligence
- Statistical Learning with Sparsity: The Lasso and Generalizations
- Classification and Regression Trees
- Matrix Completion and Networks
- Using Big Data for Measurement and Research
- Neural Networks
- Mining Text and Images

Ethics and law for Data Science

- Governo dei dati e tutela dei diritti fondamentali
- Data Localization e sovranità dei dati
- Riconoscimento facciale
- Data Governance: considerazioni introduttive
- Big data e automazione
- Big data e IOT

Responsabilità dell'ISP

Block III – Elective Courses

Data Science for Economics

- Experiments and real-world evidence in economics
  - From theory to data (and the way back). Introduction to behavioral and experimental economics.
  - 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.
  - Statistical analysis of experimental data. Mediator variables, modulator variables, specific statistical tests, multiple testing of hypotheses.
  - 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
  - Introduction to microeconometrics:
    - Structure, Endogeneity, and Identification Problems
    - Least-squares, Probit, and Logit Estimators
    - Static panel data
    - Dynamic panel data
  - The Evaluation Problem:
    - Randomization and Matching Models
    - The Difference-in-difference Estimators
    - Instrumental Variables
    - Regression Discontinuity Design
  - Causality and Non-linear Models:
    - Quantile Regressions
    - Multinomial Models
    - Models for Count Data
    - Survival/Duration Analyses
    - Models with Control Functions

Data Science for Business

- Business analytics
- Optimization of Financial Portfolios
  - Financial assets, returns, statistical features of returns
  - Portfolio choice criteria: expected utility vs. Markowitz mean-variance
  - Mean-variance portfolio selection in action
  - Further topics: dealing with high-dimensional portfolios; constraints on concentration and turnover; the Black-Litterman model; sensitivity w.r.t. inputs ("estimation risk"); mean-VaR and mean-CVaR portfolio selection
  - Portfolio optimization in Matlab: 'quadprog' function and 'portfolio' object via Financial Toolbox

Data Science for Health

- Health analytics and data-driven medicine
  - Causal inference in healthcare with MEPS data
  - Predictive healthcare and patient outcome (digital records, diagnostic procedure and intervention)
  - Clinical trials and prescription behavior: market analysis and regulation
  - Epidemiology and COVID-19
- Environmental and genomic data analysis

Hands-on Labs

- Hands on R and STATA for Data Science:
  - Introduction to R and STATA
  - Data Modeling for policy evaluation:
  - Data Modeling: inference and predictive analysis
  - Data Modeling: causal machine learning
- Hands on Python for Data Science:
  - Introduction to Python for data science:
  - Unsupervised learning
  - Dimensionality reduction
  - Neural networks and deep learning
  - Support Vector Machine