3 ECTS

19D031

Foundations of Data Science

Prerequisites to Enroll

Although not mandatory, some knowledge of Python, Jupyter notebooks, and algebra is recommended.

Students must have their own laptop or desktop computer. Last version of Anaconda with Python and R should be installed, following instructions to be sent. Check hardware requirements to run Anaconda here.

Notebooks might be run as well using Google Collaboratory.

Further instructions will be sent prior to the start of the course.

Overview and Objectives

This is an intensive 20-hour course based on a hands-on approach using Jupyter notebooks, all material is motivated by specific information retrieval and data analysis questions and each thematic unit concludes with a small project. The course provides basic training in data analysis and machine learning with Python and R..

The course will be delivered by Joan Verdú, Head of Consulting and Knowledge Transfer of the BGSE Data Science Center, in collaboration with Data Scientists affiliated to the Data Science Center.

Course Outline

Online classes last 4 days, during the mornings. In the afternoons, students are supposed to work on the materials and assignment, with TA assistance.

The course provides training in data analysis and machine learning with Python and R and evolves along the thematic following units:

1. Programming with Python. Main topics:
   - Intro to Jupyter Notebooks
   - Loops, control flow

19D031

Foundations of Data Science

- Lists, maps, reductions
- Functions and classes
- Inputs and Outputs

Keywords: data types, functions, objects

2. Data analysis with Python
   - Inputs and Outputs
   - Series and Dataframe
   - Group, apply, combine
   - Merge and concat
   - Non-rectangular data

Keywords: pandas, database management

3. Data visualization
   - Elements of data visualization
   - Scatter plots
   - Line plots
   - Exploration plots: barplots, boxplots
   - Advanced plots: correlation, regression, biplots
   - Special plots
   - Reporting using visualization

Keywords: seaborn, plotly

4. Data preparation
   - Handling missing data: imputation methods
   - Feature transformation and engineering: normalization, dimensionality reduction, category encoding

   Keywords: sklearn

5. Supervised learning
   1. Linear models for regression
      - Linear models and non-linear feature maps
      - Model evaluation
      - Convexity
      - Bias-Variance tradeoff
      - Penalized likelihood and lasso
      - Cross validation and model selection
   2. Linear models for classification
      - Logistic regression

- Misclassification, ROC, AUC
- Class imbalance
- Generative vs. discriminative models

3. Non-linear models: decision trees
   - Decision trees
   - Variable selection
   - Forests
   - Bagging and boosting

Keywords: sklearn, linear models, cross validation, regularization, lasso, trees, ensembles, boosting

6. Unsupervised learning
   - Continuous latent variables
   - PCA and SVD
   - Probabilistic PCA
   - Factor analysis
   - ICA
   - Matrix Factorization
   - Multidimensional Scaling

Keywords: clustering, factors, independent component analysis and matrix factorization

7. Intro to data science in R
   1. R basic programming
      - Loops, control flow
      - Lists, vectors, matrices, dataframes
      - Functions
   2. Data exploration and visualization in R
      - Scatter plots
      - Line plots
      - Exploration plots
      - Advanced plots
   3. Sample data preparation, supervised and unsupervised learning with R
      - Data preparation: missing data imputation, feature transformation
      - Supervised learning: linear, non-linear
      - Unsupervised learning: PCA, clustering

Keywords: tidyr, data.table, caret, ggplot2, plotly

3 ECTS

19D031

Foundations of Data Science

Required Activities

(Online) Class participation is compulsory. During classes there will be little projects that students are expected to perform while in class.

Students will have access to videos corresponding to most of the topics, where the notebooks and materials are explained in advance. They are supposed to watch them, so that we can speed up during online sessions and focus on Q&A and solve the short exercises that are embedded in the notebooks.

At the end of each module (4 of them, one per day) there will be a project assignment, to be done:

1) Individually, during the same day, with some support from TAs.
2) As an extension, 1 more month will be given to complete the same project, this time in groups of 3 students, with some support from TAs.

Evaluation

Attendance at classes, and submission of projects.

- Projects will be given at the last class of each module (one per day)
- The individual project should be delivered the same day. Expect an effort of 2-4 h each one.
- The home projects (continuation of the individual ones) will be in groups of 3.
- Students will be given 4 weeks to submit the extended version of projects.

The grade will be the average of these projects (40% individual projects, 60% group projects).
There might be some quizzes to recap materials by the end of each module, not graded.

Students wanting to register to other data science courses that have Foundations of DS as pre-requisite must pass the individual projects with an average grade >=6.5 (out of 10), and a minimum grade of 5 in each of the individual projects.

Competences

☒ Construct a global vision of the situation of the problem based on knowledge of the synergies between advanced statistical methods, computing and business analysis to generate added value.

☒ Modeling and predicting high-dimensional data with advanced statistical methods in the field of data science in order to improve strategic decision making.

☒ Apply the knowledge of programming languages, computer programs and advanced services in the Cloud to solve the problems that are presented to the data scientist.

3 ECTS

19D031

Foundations of Data Science

☒ Solve the real problems that arise in the fields of study through the accurate analysis of the data.

☐ Visualize and interact with high-dimensional data in order to contextualize the information and facilitate subsequent decision-making.

☐ Communicate with conviction in English the results and implications of the required analytical study using a language related to the receiver.

☐ Work in a heterogeneous team of researchers in the field of the economic analyst using specific group techniques.

☐ Own and understand knowledge that provides a basis or opportunity to be original in the development and / or application of ideas, often in a research context.

☐ That students know how to apply the acquired knowledge and their ability to solve problems in new or unfamiliar environments within broader (or multidisciplinary) contexts related to their area of study.

☐ That the students be able to integrate knowledge and face the complexity of making judgments based on information that, being incomplete or limited, include reflections on the social and ethical responsibilities linked to the application of their knowledge and judgments.

☒ That the students know to communicate their conclusions and the knowledge and last reasons that sustain them to specialized and non-specialized publics in a clear and unambiguous way.

☐ That students have the learning skills that allow them to continue studying in a way that will be largely self-directed or autonomous.

Learning Outcomes

☒ Elaborate and estimate probabilistic prediction models based on certain data.

☐ Predict random processes.

☒ Apply supervised and semi-supervised learning algorithms.

☐ Apply search algorithms and estimation methodologies in networks through observation of data.

☐ Apply mathematical and computational analysis of social, business and economic networks knowing the theory and optimization algorithms.

3 ECTS

19D031

Foundations of Data Science

☒ Predicting information needs based on decisions that must be made.

☐ Apply mathematical theory and statistics on data sets from disparate disciplines.

Materials