Practical Data Science for Finance

Professor: Alberto Santini
E-mail: alberto.santini@upf.edu
Office hours: by appointment
Course Type: Compulsory
Credits: 4
Term: First

Course Description

This course is an introduction to Data Analytics and Machine Learning, with applications to finance. We first focus on problems that are independent of specific context or area of application; for example, how to determine which model works best among a set of candidates, how to correctly estimate the accuracy of a model, how to make our models generalise well to new data.

We present a minimal statistical framework used to understand learning, and then dive deeper into two main problems in supervised learning: regression (predicting numbers) and classification (predicting non-numeric labels).

Objectives

To attain proficiency in the following areas:

- General competences:
  - Analyse and manipulate data: data reading, cleaning, visualisation and exploration.
  - Learning from a mathematical and statistical perspective; supervised, unsupervised and reinforcement learning problems.
  - Modelling: selecting, evaluating and comparing models. Bias and variance of an estimator.
  - Regression and classification problems from the Machine Learning point of view.
  - End-to-end Machine Learning pipelines using Python.
- Specific methods and concepts:
  - Estimating a relationship: models, hypothesis and loss functions, bias, and variance, over- and under-fitting.
  - Training a model: gradient descent and stochastic gradient descent.

  - Performing model selection via hold-out, cross-validation, and the bootstrap. Hyperparameter tuning.
  - Regularisation: Ridge- and LASSO-regularised regression models.
  - Classification: possible loss functions; linear, non-linear and tree-based models.
- Tools:
  - Using Python and Jupyter notebooks.
  - Basic familiarity with data libraries, including pandas, pyplot, seaborn.
  - Basic familiarity with ML libraries, including sklearn.

Methodology

Classes are both frontal lectures and practical laboratories, in a roughly 50-50% proportion. The lab classes complement the theory ones, by getting the students familiarised with the Python programming language and its main data science libraries.

The competences, the learning outcomes, the assessment elements and the quality of the learning process included in this Teaching Plan will not be affected if during the academic trimester the teaching model has to switch either to a hybrid model (combination of face-to-face and on-line sessions) or to a complete on-line model.

Evaluation criteria

Three elements concur in the final mark:

- Participation. 20% of the mark.
- Project work. 30% of the mark. Students will apply their knowledge to a real-life problem. They are expected to use the computer tools they learnt to use during the lab classes.
- Final exam. 50% of the mark. Contains questions about theory only.

Only the final exam and the participation marks are carried over to an eventual re-take. There is no retake for the project work.

Students are required to attend 80% of classes. Failing to do so without justified reason will imply a Zero grade in the participation/attendance evaluation item and may lead to suspension from the program.

Students who fail the course during the regular evaluation are allowed ONE re-take of the evaluation, in the conditions specified above. If the course is again failed after the retake, the student will have to register again for the course the following year.

In case of a justified no-show to an exam, the student must inform the corresponding faculty member and the director(s) of the program so that they study the possibility of rescheduling the

exam (one possibility being during the "Retake" period). In the meantime, the student will get an "incomplete", which will be replaced by the actual grade after the final exam is taken. The "incomplete" will not be reflected on the student's Academic Transcript.

Plagiarism is to use someone else's work and present it as one's own without acknowledging the sources in the correct way. All essays, reports or projects handed in by a student must be original work completed by the student. By enrolling at any UPF-BSM Master of Science and signing the "Honor Code," students acknowledge that they understand the schools' policy on plagiarism and certify that all course assignments will be their own work, except where indicated by correct referencing. Failing to do so may result in automatic expulsion from the program."

Calendar and Contents

Class	Topics
1	Theory: introduction to data analytics and machine learning.
2	Theory: estimators, hypothesis and loss function, reducible and irreducible error, training and testing a model. Lab: introduction to Python.
3	Theory: bias and variance, overfitting and underfitting, examples. Lab: introduction to pandas to read and manipulate data, data imputation.
4	Theory: training as an optimisation problem, training algorithms, gradient descent. Lab: introduction to data visualisation, pyplot and seaborn.
5	Theory: training in big-data settings, stochastic and mini-batch gradient descent, acceleration techniques. Lab: python exercises.
6-7	Theory: model selection and the workflow of the machine learning scientist, the holdout validation method, variance of the holdout validation method. Lab: a complete example using the holdout validation method, information leakage and how to avoid it during model selection.
8-9	Theory: leave-one-out cross-validation, k-fold cross-validation, revising the workflow of the machine learning scientist. Lab: a complete example using cross-validation.
10-11	Theory: the bootstrap, its advantages and drawbacks; the 0.632 and the 0.632+ methods. Lab: python exercises.
12-13	Theory: accuracy and interpretability of a model, feature selection, introduction to regularisation. Lab: implementing stepwise feature selection.
14	Theory: Ridge and LASSO regularisation, similarities and differences. Lab: visualising the impact of regularisation on the model.
15	Theory: classification problems and measures of accuracy of classification models; unbalanced datasets and how to deal with them. Lab: unbalanced datasets and SMOTE.
16-17	Theory: linear classifiers; the maximal margin classifier, the support vector classifier. Lab: complete example using linear classifiers.
18	Theory: non-linear classifiers, support vector machines; tree-based classifiers, trees and forests. Lab: complete example for classification.

Reading Materials/ Bibliography/Resources

A good reading is the book “An introduction to Statistical Learning”, freely available on-line. In the book you will find most of the theory topics covered by this course, and many more.

The interested student could then progress to the more advanced “Elements of Statistical Learning” (also freely available on-line).

A good book for the practical part is “Introduction to computation and programming using Python”. The “Scipy lecture notes” can also prove very valuable.

Other good books are “Python for Data Analysis”, by McKinney, and “Building Machine Learning Systems with Python”, by Richert and Coelho.

In general, a student attending all classes will not need any book to pass this course. The professor will provide the students with the Jupyter Notebooks used as interactive lecture notes.

Bio of Professor

Alberto Santini is a Marie Curie Post-Doctoral fellow at ESSEC Business School, in Paris. He is on leave from the Department of Economics and Business of UPF, which he joined in 2017, holding a tenure-track assistant professor position. Before that, he was a Postdoctoral Researcher at RWTH Aachen. He obtained his PhD from the University of Bologna. His main research interests are in the field of Operational Research and Machine Learning. https://santini.in/