Practical Data Science for Operations Management

Professor: Alberto Santini
E-mail: alberto.santini@upf.edu

Objectives

Have a good understanding of basic concepts in Machine Learning (ML):

- What is data, information, and knowledge from the point of view of ML, what is data analytics and what is the typical workflow of a data analyst.
- What are the most common supervised and unsupervised ML tasks.
- Assessing the quality of predictions, the bias-variance trade-off, overfitting and underfitting.
- General setup of a learning problem, and how prediction methods you already know (e.g., linear regression) fit into this setup.
- What algorithms we use to train Machine Learning models, what are their pros and cons.

Have a good understanding of the two main tasks in supervised learning, regression and classification. Have a working knowledge of what are the challenges when training models for these tasks, and how we can overcome them. Have a working knowledge of the main methods to solve regression and clustering problems efficiently and with high accuracy.

Be able to use a Python3 Jupyter Notebook to solve practical problems: data input and cleaning, data visualisation, modelling, model tuning and training, making estimations.

Description

We give an overview of the main problems in Machine Learning and Data Analysis. We discuss the most popular models used to solve these problems, the algorithms to train the models, and ways to assess their effectiveness. Finally, we discuss how they can be used in an Operations Management context.

Methodology

The course will have a hands-on approach. In the first lectures we will introduce ML and the main theoretical ideas (frontal classes). We will then setup our computers to use Python, Jupyter and the main Data Science packages, and we will start the practical work (laboratory classes).

Evaluation criteria

Three elements concur in the final mark:

- **Class attendance**. Since most of our classes are lab classes, attendance is extremely important. It will count for 20% of the mark (2% for each session).
- **Project work**. As this is a practical course, you will be mostly evaluated based on your ability to develop a Machine Learning project. The project can be carried on in groups of 4-5 people, and there will be a broad offering of diverse projects to choose from. This item counts for 40% of the final mark.
- **Final exam**. The final exam is used to assess the individual level of knowledge and understanding of each student. It will include questions covering topics from all the classes. This item counts for 40% of the final mark.

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

As with all courses taught at the UPF BSM, students who fail the course during regular evaluation will be allowed one re-take of the examination/evaluation. Students that pass any retake exam should get a **5 by default as a final grade for the course.** If the course is again failed after the retake, students 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 of the programme 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 another's work and to 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 "Honour 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 programme."

Prerequisites

We will assume familiarity with basic concepts of real analysis, linear algebra, probability and statistics. It is recommended, although not compulsory, to have some familiarity with a programming language.

Calendar and Contents

Session	Topics
1-2	Introduction to ML and Data Science. The workflow of a data scientist. Supervised and unsupervised learning. Training and assessing the quality of a predictor in the supervised case. The bias-variance trade-off. Over- and underfitting.
2-3	Supervised learning as an optimisation problem. Main algorithms to solve the problem of training a predictor: gradient descent, stochastic gradient descent, acceleration techniques.
4	Linear and polynomial regression. Adding regularisation terms: Ridge, LASSO, and ElasticNet. Regularising via reducing variables.
5	Lab class: setting up our jupyter environment. Reading input data with pandas. Data cleaning and data visualisation with pyplot + seaborn.
6	Lab class: regression. Selecting a model, training and tuning it with sklearn. Assessing the quality of our models.
7	Classification problems. Linear classifiers. Maximal margin classifiers and support vector classifiers. Beyond the linear case: support vector machines and kernels.
8	Lab class: classification. Selecting a model. Evaluating with ROC/AUC. The case with unbalanced data and the SMOTE method.
9	Classification problems. Decision trees. Ensemble methods based on decision trees: bagging and boosting. Robust training of the model: bootstrapping and cross-validation.
10	Lab class: ensemble methods. Bagging classifier and regressor with sklearn. AdaBoost with sklearn. Doing k-folds and bootstrapping.

Reading Materials/ Bibliography/Resources

A great reading is the book "An introduction to Statistical Learning" (http://www-bcf.usc.edu/~gareth/ISL/). In the book you will find the theory topics covered by this course, and many more. The interested student could then progress to the more advanced "Elements of Statistical Learning" (https://web.stanford.edu/~hastie/ElemStatLearn/). A good book for the practical part is "Introduction to computation and programming using Python" https://mitpress.mit.edu/books/introduction-computation-and-programming-using-python-0). The scipy lecture notes (https://www.scipy-lectures.org/) 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. Online resources will probably prove more useful than a book, should the student hit a roadblock with the practical project implementation.

Bio of Professor

Alberto Santini has joined the Department of Economics of UPF in September 2017. Before, 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 (mostly combinatorial optimisation) and Machine Learning (mostly in how it can interact with Operational Research). Find his full cv at https://santini.in/