Proposal for a MAP-I UC 2022/2023

Foundations and applications of Machine Learning
Fundamentos e Aplicações de Aprendizagem Automática

A. Programmatic component

1. Context

Big data is overwhelming nearly every field of knowledge, from life sciences, Internet, finances and banking or social networks. Daily, 2.5 quintillion bytes of data are being generated. It is becoming more and more important to be able to make sense of and to communicate all the knowledge that it represents. It is often the case that generated datasets have high dimensionality, complexity, heterogeneity, which associated to their large volumes signify a hard analysis problem. Modern Machine Learning (ML) techniques represent a powerful approach to the analysis of such large-scale datasets, by deriving novel representations that augment domain knowledge and supporting informed decision-making processes. ML is considered by many to be the driver of the next wave of innovation and Artificial Intelligence. There are currently worldwide companies such as Facebook, Google or NVidia that are making significant breakthroughs by developing disruptive ML solutions and incorporating them into new products. Many other companies are following similar paths making record investments in ML.

The application of ML technologies is certainly not only business oriented. Life sciences disciplines such as physics and biology are also steadily progressing with the applications of ML to the analysis of Terabytes to Petabytes of data generated by modern devices and technologies. One exciting example is the field of biomedicine. Biomedical data, including those generated from large-scale genomic projects, electronic health records or clinical exams is growing at an unprecedented scale. ML is increasingly a critical tool to extract value from these data on the different domains.

The high demand for ML specialists to work in problems, such as self-driving cars, DNA genome analysis or cancer prediction, climate change and many other fields prompts the need to train the next generation of computer scientists with the theoretical and practical knowledge of Machine Learning that allows them to develop projects that use the latest technologies following the best implementation practices. See for instance the McKinsey Global Institute report: "*A significant constraint on realizing value from big data will be a shortage of talent, particularly of people with deep expertise in statistics and machine learning.*" - From: Big data: The next frontier for innovation, competition, and productivity, 2011, McKinsey Global Institute.

In order to tackle this shortage and high demand of professionals with a solid background in ML and big data analytics, we propose a curricular unit that teaches Machine Learning using the state-of-the-art technologies, including the most recent software libraries and platforms.

2. Pedagogical objectives and learning skills

The main objectives are to supply the students with adequate knowledge and skills in the core principles and techniques of Machine Learning. Thus, students completing this unit should:

- Learn the fundamentals of ML – regression, classification, clustering, deep learning.
- Understand the connection between learning and optimization. Build optimal data representation models.
- Learn how to implement and apply predictive, classification, clustering, information retrieval and deep learning algorithms to real datasets.
- Develop a critical view and be able to choose, apply and evaluate the most adequate problem solving techniques in ML;
- Be able to design, specify, implement and validate advanced software tools for specific data analysis problems; assess the quality of the models using the relevant error metrics;
- Be able to interact with professionals from different knowledge domains in the process of software development and generate adequate reporting.

Other more transversal skills are also approached in this unit, such as:

- To use modern software tools to implement reproducibility and portability during development, testing, and deployment.
- To conduct a short research project, being able to formulate a research problem, to review significant literature, to evaluate existing solutions and propose alternative approaches implementing those in new software tools;
- To write reports and scientific publications explaining the work developed;
- To be able to communicate with other researchers within multidisciplinary teams, in many cases international;
- To present orally the work developed.

3. Program

The overall aim is to provide an introduction to modern ML approaches for several problems including regression, classification, clustering, and deep learning, among others.

Brief program:

- Fundamentals of ML. What is ML and what are the challenges?
- Supervised versus Unsupervised ML.
- Data Pre-processing and exploration
    - Detection of outliers; Standardization; Transformation; Dimensionality reduction (Principal Component Analysis, Multi-Dimensional Scaling); Split data into training and test sets
- Model evaluation and model selection methods
    - Cost (loss) function. Cost function convergence. Iterative gradient descent algorithm. Learning curves.

    - Performance measures (error, confusion matrix, sensitivity/specificity, ROC curves); Train and test paradigm; Cross validation; Bias and Variance; Overfitting. Regularization.
- Classification methods
    - Decision Trees; K-NN; Linear and Nonlinear (Kernel); Support Vector Machines; Neural Networks; Logistic Regression, Ensemble approaches: Bagging, Boosting; Random Forests; Gradient Boosted Decision Trees
- Regression
    - Univariate and Multivariable linear regression; Performance measures: RMSE, R-Squared; Results (Coefficients, residuals) interpretation. Batch/mini batch/stochastic gradient descent.
- Unsupervised learning - fundamentals
    - Distance (similarity) measures
    - K-means clustering; Hierarchical clustering (different measures and methods); t-SNE.
    - Data compression
- Deep Learning fundamentals
    - Shallow and Deep Neural networks; Network architectures (feed-forward, convolutional, recurrent, auto-encoders); Hyper-parameter optimization; Input data transformation; Applications to unsupervised, classification and regression problems; Introduction to deep generative models
- Reinforcement learning and deep reinforcement learning: an introduction
- Information visualization
    - Scatter plots; Boxplots; Heatmaps; Trees and Dendrograms
- Introduction to *scikit-learn*, *numpy*, *matplotlib*, *pandas* and *tensorflow/ keras* Python packages. Implementation and testing of ML pipelines. Create notebooks and *docker* containers for portability and predictability during development, testing, and deployment.

Learning Outcomes:

- Identify potential applications of ML in practice.
- Describe the core differences in analyses enabled by regression, classification, and clustering.
- Select the appropriate ML task for a potential application.
- Apply regression, classification, clustering, retrieval, recommender systems, and deep learning.
- Represent your data as features to serve as input to machine learning models.
- Assess the model quality in terms of relevant error metrics for each task.
- Build an end-to-end application that uses ML at its core.
- Implement ML algorithms and data analysis pipelines in Python, in particular taking advantage of *scikit-learn, scipy, tensorflow/ keras* and *pandas* packages.

4. Pedagogical strategies

This unit will consist of theoretical-practical (TP) lectures to introduce and examine ML methods and algorithms. Lecture material will consist of slides prepared by the lecturers and references to textbooks and scientific articles. Practical programming exercises and small projects will be discussed. Here, the appropriate programming libraries will be introduced. A data-driven approach to teaching will be followed. The course will also contain a semester project where students will tackle real-life problems and datasets. Students may propose a dataset under their thesis project.

Introduction to research

This unit will require that students develop skills that are essential for scientific research, namely:

- Characterize the state-of-the-art. This will be achieved by reading recommended textbooks, scientific papers and webpages.
- Concise presentation of results. The report for the final project should be written as an extended article. At the end of the semester, a presentation should also be done to present and defend the work.
- Team spirit. Students should work in small teams with division of tasks.
- Reproducible research and pipelines. Students should apply recent methodologies to allow their results to be easily reproduced by others in different settings and platforms.
- Advancing state-of-the-art. By using competitive platforms students will get a sense on how their proposed solutions are advancing current best performing solutions.

Tentative calendar

- Class 1 [Lecturer: Pedro Ferreira, FCUP] - Fundamentals of ML, Basic supervised ML pipeline, Data Pre-processing and exploration.
- Class 2 [Lecturer: Pedro Ferreira, FCUP] - Unsupervised learning (Clustering, PCA, MDS)
- Class 3 [Lecturer: Rita Ribeiro, FCUP] – Regression
- Class 4 [Lecturer: Petia Georgieva, UA] – Classification Algorithms, Training process
- Class 5 [Lecturer: Pétia Georgieva, UA] - Model selection and performance evaluation, Ensemble methods.
- Class 6 [Lecturer: Miguel Rocha, UMinho] - Neural Networks; Deep Learning (Introduction)
- Class 7 [Lecturer: Miguel Rocha, UMinho] - Deep Learning (Recurrent and convolutional neural networks; transfer learning; applications of DL)
- Class 8 [Lecturer: Miguel Rocha, UMinho] - Deep Learning - generative models; introduction to deep reinforcement learning
- Final Project Presentation

Evaluation criteria

The evaluation will consist of a Project applying Machine and/or Deep Learning approaches to a selected problem approved by the teaching staff. The evaluation of the project will consist on a written report - 70% - and a presentation- 30%)

5. References

- *Machine Learning: The Art and Science of Algorithms that Make Sense of Data*, 1st Edition. Peter Flach.
- *Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. Aurélien Géron. O'Reilly
- *Python Machine Learning*. Sebastian Raschka. Packt Publishing
- *An Introduction to Statistical Learning: With Applications in R*. Daniela Witten, Gareth James, Robert Tibshirani, and Trevor Hastie. Springer Texts in Statistics Book
- *Machine learning in genomic medicine: a review of computational problems and data sets*. Michael K. K. Leung; Andrew Delong ; Babak Alipanahi ; Brendan J. Frey. Proceedings of the IEEE (Volume: 104 , Issue: 1 , Jan. 2016)
- *Deep Learning in Python.* F. Chollet., Manning, 2018.
- *Deep Learning*. I. Goodfellow, Y. Bengio, A. Courville. MIT Press, 2016.
- *Machine Learning*, T. Mitchell, McGraw-Hill, 1997

B. Instructors Team

1. Presentation of the team

The team includes four docents from the three universities involved in the MAPI program: Miguel Rocha (U. Minho; coordinator), Pedro Ferreira and Rita Ribeiro (U. Porto) and Petia Georgieva (U. Aveiro). Altogether, the team has a long-standing experience in teaching and researching in the area of this proposal: Machine Learning and its applications. The team has proposed and successfully run this curricular unit in MAPI for the last 3 editions.

2. Coordinator

Miguel Rocha is an Associate Professor with Habilitation at the Department of Informatics and Senior Researcher of the Centre of Biological Engineering, at the University of Minho. He is currently the Director of the Master in Bioinformatics and elected member of the Scientific Council of the School of Engineering.
He has a background in computer science, a PhD thesis in machine learning, and a vast curriculum in Machine Learning and Bioinformatics/ Systems Biology, including over 200 publications in peer-reviewed journals/conferences (ORCID: 0000-0001-8439-8172), 8 projects as a PI, and a patent application. He supervised 15 PhD students and over 60 master students. He is the responsible docent of several curricular units related to Bioinformatics, Machine Learning/ Data Mining and programming, in both first degree and master courses.
More details are available in the URL: http://ceb.uminho.pt/People/Profile/mrocha

3. Team members

Pétia Georgieva (*http://wiki.ieeta.pt/wiki/index.php/Petia Georgieva*) is an Associate Professor with the Dep. of Electronics Telecommunications and Informatics of the University of Aveiro, a Researcher with the Institute of Electronics Engineering and Telematics of Aveiro (IEETA) and a collaborator member of Institute of Telecommunications, Aveiro. Previously she had academic positions with the University of Porto (2001-2003) and research visiting positions with the Rowan University, USA, Carnegie Mellon University, USA, University of Lancaster, UK and Bulgarian Academy of Sciences, Bulgaria.
Her interests are in the area of machine learning, deep learning, data mining, signal processing and control. Her work involves the development of novel techniques for high dimensional problems (including neuro-computing, brain computer interfaces, image processing) and data science approach in bio-chemical processes and more recently in optical communications. She supervised 9 PhD students and over 40 master students.
Dr. Georgieva is a Senior member of IEEE and a Senior Member of International Neural Network Society (INNS).

Her research is funded by sponsors such as EU, Portuguese Foundation for Science and Technology (FCT) and industry.
Publications: http://orcid.org/0000-0002-6424-6590

Pedro G. Ferreira is an Assistant Professor at the Department of Computer Science, Faculty of Sciences of University of Porto and an affiliated researcher at i3s/ipatimup and at LIAAD-INESCTEC, the Artificial Intelligence and Decision Support Lab of the University of Porto. He graduated in Systems and Informatics Engineering from the University of Minho in 2002 and completed his PhD in Artificial Intelligence from the same University in 2007. From 2008 to 2012, he was a Postdoctoral Fellow at the Center for Genomic Regulation, Barcelona and from 2012 to 2014, at the Functional Population Genomics and Genetics of Complex Traits group, University of Geneva. He was involved in several major international consortia including ICGC-CLL, ENCODE, GEUVADIS and he is an active member of the GTEx consortium. In 2015, he was awarded an FCT Investigator Starting grant and he joined Ipatimup/i3s. He has experience in the genomics start-up environment where he developed information systems for personal genomics data interpretation. His main research focuses on the development of methods for a diverse set of problems in genomic data science. In particular, he is interested in unraveling the role of genomics on human health and disease. In order to achieve this goal he applies and develops data-analytical models using machine learning and probabilistic methods to analyze and interpret diverse, complex and large-scale genomic datasets.
Personal webpage: pgferreira.net
Publications: https://scholar.google.com/citations?user=X097-20AAAAJ

Rita P. Ribeiro (https://www.dcc.fc.up.pt/~rpribeiro) is an Assistant Professor at the Department of Computer Science of the Faculty of Sciences of the University of Porto (FCUP) and a Researcher at the Artificial Intelligence and Decision Support Lab (LIAAD) of the Institute for Systems and Computer Engineering, Technology and Science (INESCTEC). She holds a PhD in Computer Science from the University of Porto. Her main

research topics are imbalanced domain learning, outlier detection, evaluation issues on learning tasks and problems related to social good. She has been involved in several research projects concerning environmental problems, fraud detection, and predictive maintenance applications. She is a member of the program committee of several conferences, serving as a reviewer of several journals and has been involved in the organization of some scientific events. Currently, she is also the director of the Masters in Data Science at FCUP.
Publications:
https://www.authenticus.pt/en/profileOfResearchers/publicationsList/15959
http://orcid.org/0000-0002-6852-8077