Postgraduate Course
Machine Learning Lab (MSc)

Instructor Information

Luis A. Hernández Gómez
E-mail: luisalfonso.hernandez@upm.es
Work Phone: +34 91 549 57 00 Ext: 4082

Course Information

Course Description

In this course students will address practical problems on the application of the variety of Machine Learning methods presented in the Predictive and Descriptive Learning course. Experimental activities will cover both predictive or supervised learning (from linear and logistic regression to random forest and SVM) and descriptive or unsupervised prevised learning (principal component analysis and cluster analysis). Several realistic and practical scenarios and use cases will be addressed (as those proposed in Kaggle competition, www.kaggle.com). Students will practice using "scientifically-oriented" languages and environments, mainly working with R language. They will also approach the principles of parallel computing for large-scale machine learning experiencing with cluster computing frameworks for big data processing. In particular students will learn how to develop large-scale machine learning solutions using the MLlib of Spark Open Data Processing PLatform using Python language. A practical introduction to *streaming learning* will be also provided. Through all lab activities students will have to gain practice on model accuracy using cross-validation and on how to draw precise conclusions and valuable interpretations from machine learning results and models.

Prerequisites

Attending to Predictive and Descriptive Learning course

Previous exposure to a programming language, such as MATLAB, R or Python.

Course Goal

In this laboratory students will be learn how to apply the variety of Machine Learning methods presented in the Predictive and Descriptive Learning course to practical scenarios. Students will experience with both "scientifically-oriented" processing environments and cluster computing frameworks for big data processing.

Summary of intended course outcomes

The students will acquire the skill to apply the variety of Machine Learning methods on to practical scenarios. Main course outcome will be to consolidate the theoretical study of machine learning techniques along this Master's Programme. Through hands-on experience case studies students will learn how to select and accurately assess the performance evaluation of machine learning methods. They will also acquire solid criteria on what could be best model for a given data and task as well to be able to draw precise conclusions and interpretations from experimental results. By the end of the course, students should be able to:

- Understand how to apply the most used models and techniques for predictive and descriptive learning to different real scenarios.
- Design a proper experimental methodology for accurately assessing and gaining knowledge from the use of each one of the particular machine learning technique.
- Work with both "scientifically-oriented" processing environments and cluster computing frameworks for big data processing that can be used in a wide range of applications in science and industry.

Syllabus

Introduction to Machine Learning Lab
Introduction to "large-scale processing" and Spark
Designing a Machine Learning System
Programming with Resilient Distributed Datasets RDD's in Python
Spark MLlib Introduction
Introduction to R

Linear Regression
Linear Regression in R
Principles of Parallel Computing through Linear Regression parameters estimation
Linear Regression with Spark

Classification
Classification methods with R
Dimension Reduction and High-Dimensional Data in Spark
Classification methods with Spark

Resampling methods
Cross-Validation and Bootstrap with R
Model Selection and Regularization with Spark

Tree-Based Methods
Decision tress, Bagging, Random Forests and Boosting with R
Decision tress, Bagging, Random Forests with Spark

Support Vector Machines
Kernels and Support Vector Machines with R
Support Vector Machines with Spark

Descriptive Learning
Principal Components Analysis, K-means and Hierarchical Clustering with R
Principal Components Analysis and K-means with Spark
OnLine Learning
Introduction to real-time machine learning with Spark Streaming

Textbooks:

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An introduction to statistical learning*. Vol. 112. New York: springer, 2013.

Karau, Holden, Andy Konwinski, Patrick Wendell, and Matei Zaharia. *Learning Spark: Lightning-Fast Big Data Analysis*. " O'Reilly Media, Inc.", 2015.

Pentreath, Nick. *Machine Learning with Spark*. Packt Publishing Ltd, 2015.

Bibliography:

- Ryza, Sandy, Uri Laserson, Sean Owen, and Josh Wills. *Advanced Analytics with Spark: Patterns for Learning from Data at Scale*. O'Reilly Media, Inc., 2015.
- Nandi, Amit. *Spark for Python Developers*. Packt Publishing Ltd, 2015.
- McKinney, Wes. *Python for data analysis: Data wrangling with Pandas, NumPy, and IPython*. " O'Reilly Media, Inc.", 2012.

Student Assessment Criteria

Practical activities will be assessed through personal and group projects through short reports and public sessions for presentation and discussion of experimental planning, methodological design and evaluation and critical interpretation of results.

Personal projects	20%
Group project	80%